AI Explainability in Classifying Political Speeches and Interviews

Institute of Computer Science, Brandenburgische Technische Universität Cottbus-Senftenberg

Juan-Francisco Reyes

pacoreyes@protonmail.com

This study applies explainable AI techniques to understand the linguistic features of classifying *speeches* and *interviews* in political discourse. Using logistic regression, statistical tests, and *Shapley values*, *SHAP*, we create a more understandable version of the predictions made by *BERT* models in this *NLP* classification task. This study delves into the role that recognizable linguistic features play in speech and interviews in both feature-based and neural models. Specifically, it examines the extent to which BERT models depend on linguistic structures for their predictions, using *anonymization*. Built on findings from classic and modern linguistic literature, in addition to improving the interpretability of neural models, it highlights the means to identify important (*global*) "*political discourse features*" that distinguish speeches and interviews: nominalization frequency, discourse marker frequency, personal pronoun usage, and interjection frequency.

Keywords: Natural Language Processing, explainable AI, political discourse, BERT, SHAP

1. Introduction

This study investigates the classification of political discourses into two primary modes of communication: *speeches* and *interviews*. Speeches, characterized by their unidirectional nature, allow 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., 2018) 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 (Gupta et al., 2020; Szczepański et al., 2021).

One motivation of our research is the precise automatic 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 maximum automation, assuming integration with other downstream *natural language processing* (*NLP*) tasks that operate autonomously, such as information extraction, information retrieval, text mining, or other text classification task to create a cohesive and efficient analytical pipeline.

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: monologic discourse (speeches and similars) and dialogic discourse (interviews and similars). 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*, operationalized computationally using *spaCy* (Honnibal & Montani, 2017) and various lexicons. The study employs a logistic regression 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 under context variability or *anonymization* conditions.

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 compels the model to prioritize linguistic structures over contextual cues, addressing concerns that BERT might otherwise classify based on recurrent themes rather than linguistic patterns. Utilizing *SHAP* (*Shapley Additive Explanations*) (Nohara et al., 2019) 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:

1. *RQ1*: How can the BERT model's classification decisions be explained, particularly regarding the significance of linguistic features that play a role in classifying speeches and interviews?
2. *RQ2*: How does training 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 trained on non-anonymized data?

The contribution of this work is three-fold: (1) we advance the field of *XAI* (*Explainable Artificial Intelligence*) by innovating methodologies that clarify the decision-making processes of AI in the nuanced domain of political discourse, showcasing a significant step forward in NLP; (2) we foster interdisciplinary research, melding computational techniques with political science and linguistics, thereby opening new avenues for research across these fields; and (3), by exploring AI transparency and accountability, alongside providing novel

insights into political language, this study not only contributes to the development of more accurate political analysis tools but also enriches linguistic research with a deeper understanding of political communication.

2. Related Work

Due to the scarse of specific literature on the linguistic structures of speeches and interviews, we also reviewed the research work on 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. Several previous research studies have the linguistics of discourse text based on the number of speakers, suggesting that the number of speakers in a discourse can influence the features of the text (Kashiha, 2021; Koplenig et al., 2019; Mauranen, 2023; Mendhakar, 2022; Wells, 2006; Zare & Tavakoli, 2016). Gumperz (1982) explored the use of language in different social and cultural contexts, concluding that speakers use different discourse strategies in monological (one speaker) and dialogical (two speakers or more) contexts.

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 and Carter (1995) pointed out that conversational language usually employs shorter words. Biber and Finegan (1994) elucidated that dialogic discourse favors shorter sentences, aligning with the conversational, interactive, and real-time processing demands of spoken interactions. Larsson and 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 of 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 (Halliday, 1994; Billing, 2008; Yao, 2009) 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 (Fowler et al., 1979; Van Dijk, 1998; Fairclough, 2001; Chilton, 2004) 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 notable contrast, Holmes (1990), Tottie (1991), Dingemanse (2021), and Burenko & Fedorova (2020) 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, 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 (Yong, 2002).

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

Table 1

Comparative of linguistic features of Speeches and Interviews discourse according to literature review.

Feature	Speeches	Interviews
Sentence Length	Long	Short
Word Length	Long	Short
Sentence Complexity	High	Low

Feature	Speeches	Interviews
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

2.1 Explainable NLP

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, making them unfeasible for handling large volumes of data—a standard at the time.

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

The explainability of linguistic features in NLP models is crucial to understanding how linguistic features, such as syntax, semantics, and morphology, impact decision-making is crucial (Jurafsky & Martin, 2023). 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 and 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 Yin and Neubig (2022) and Li (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., 2020; 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 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 (Dwork, 2006; Sweeney, 2002).

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

3.1 Datasets

Through a manual selection process, we gathered audio/transcribed political discourses representing speeches 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 limits on the maximum number.

We parsed public discourses using an ad-hoc web-scraping tool, from targeted websites, predominantly from presidents and vice-presidents (78%), and a smaller fraction from other political figures and government officials (22%). We annotated the dataset through a three-round of annotation with seven annotators classifying samples as "speech" or "interview" independently. The *Inter-Annotator Agreement* (*IAA*) analysis achieved a Cohen's Kappa score of 0.964. In the final round, we introduced a blind review, where two annotators were unaware of initial classifications, further reducing bias. An additional curator established the gold standard in case of disagreements.

A comprehensive 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. The data labeling was partially self-reported but primarily determined through human annotation, ensuring a methodical approach to data categorization and analysis. In our process of dataset preparation, we implemented comprehensive measures to ensure 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. 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 using the placeholder "ENTITY". 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.

In the case of interviews, we preserved the dialogue exchanges between the interviewer(s) and interviewee(s) entirety. This approach aligns with the anticipated context of the classification models' deployment in automated systems, which aims to streamline text processing. Minimizing text preprocessing, such as selective removal of text segments (such as the interviewer's), is crucial to maintain the dialogue's integrity. By treating the interactions of both parties as a cohesive unit, we ensure that the model processes the conversational flow naturally, which is representative of how such systems would operate in real-world scenarios.

In total, we created three datasets, as described in Table 2: *FeatDataset*, *TextDataset*, and *TextDatasetAnonym*, with NER Anonymization. Both text datasets, TextDataset and TextDatasetAnonym, for fine-tuning the BERT model were built 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". The main goal in this procedure was to minimize the contextual information, such that the BERT models focus on linguistic structures and not on the context named entities give.

Figure 1

Examples in TextDataset and TextDatasetAnonym.

TextDataset	***TextDatasetAnonym***
Our eventual goal is a total withdrawal of all outside forces. But as long as North Vietnam continues to hold a single American prisoner, we shall have forces in South Vietnam. The American prisoners of war will not be forgotten by their Government. I am keeping my pledge to end America's involvement in this war. But the main point I want to discuss with you today and the main theme of my report to the Congress is the future, not the past. (…)	Our eventual goal is a total withdrawal of all outside forces. But as long as GPE continues to hold a single NORP prisoner, we shall have forces in GPE. The NORP prisoners of war will not be forgotten by their Government. I am keeping my pledge to end GPE's involvement in this war. But the main point I want to discuss with you DATE and the main theme of my report to the ORG is the future, not the past. (…)

The datasets comprise a collection of political discourse texts representing a wide range of speeches and interviews covering most of the political issues in the U.S.A. FeatDataset containing 1,089 discourse texts was 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), N*ational 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). After random shuffling, the TextDataset and TextDatasetAnonym were built from 867 discourse texts coming from: *The American Presidency Project* (816), *Rev.com* (34), *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 speakers' gender distribution was highly skewed, featuring 98.5% males and only 1.5% females. Sampling was initially random but was later adjusted for convenience to achieve class balance. As seen in Table 2, TextDataset and TextDatasetAnonym were equally balanced, ending with a stratified number of examples per class to enhance the evaluation of BERT models' performance and their explainability.

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* / 441 *i* • Test: 108 *s* / 111 *i*	• Train: 1868 *s* and 1868 *i* • Validation: 233 *s* ad 233 *i* • Test: 234 *s* and 234 *i*	• Train: 1868 *s* and 1868 *i* • Validation: 233 *s* ad 233 *i* • Test: 234 *s* and 234 *i*
Metadata	• title (document) • source_url • politician_name • gender (speaker) • publication_date	• title (document) • source_url • politician_name • gender (speaker) • publication_date	• title (document) • source_url • politician_name • gender (speaker) • publication_date
Data Period	1939-2023	1939-2023	1939-2023

Note: (*) *s* = speech, *i* = interview. (1) The segments were sliced using the sliding window approach. (2) TextDataset and TextDatasetAnonym were equally balanced. (3) The three datasets freely available in Hugging Face: Speech-vs-Interview-Dataset, DOI: 10.57967/hf/2651; Speech-vs-Interview-Dataset-Anonym, DOI: 10.57967/hf/2650; and GitHub: https://github.com/pacoreyes/speech_interview_classification.

3.2 Experimental Set-up

Based on the existing literature on the linguistic features of speeches and interviews, we developed a set of rules to extract the measurement of the frequency of ten specific features. These rules relied on the advanced linguistic capabilities provided by spaCy, including *sentence segmentation*, *part-of-speech* tagging, statistical and rule-based *morphology*, *lemmatization*, and *dependency parsing*.

During the feature engineering process, we implemented 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:

1. *Sentence Length*: The count of words (tokens) in a sentence, excluding punctuation.
2. *Word Length*: The count of the number of characters in words (tokens), excluding those punctuation.
3. *Sentence Complexity*: The count of the number of adverbial clauses per sentence.
4. *Passive Voice Frequency*: The count of the occurrences of passive voice constructions within sentences.
5. *Lexical Word Frequency*: The count of lexical words (nouns, verbs, adjectives, adverbs) within sentences.
6. *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.
7. *Personal Pronoun Frequency*: The count of the instances of personal pronouns in each sentence.
8. *Interjection Frequency*: The count of how often interjections appear within sentences.
9. *Modal Verb Frequency*: The count of modal verb occurrences in sentences.
10. *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 to convey relationships and connections, which usually signal transitions, emphasize information, indicate contrast, introduce examples, express cause and effect, and more.

Although promising because of its role in interviews, *question frequency* was disregarded as a feature of interest because spaCy does not offer a native approach to detect it, and a simple account of question marks does not reflect the turn-taking dynamic in discourse. On the other hand, the ambiguity and complexity of questions in political discourse may make detecting question frequency difficult.

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 a logistic regression model 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 both text datasets, 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. 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*, a BERT model trained with TextDataset, 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*, another BERT model trained with TextDatasetAnonym, 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).

Overall, the developed methodological framework aimed to assess if BERT models could be modified to emphasize linguistic structure rather than thematic context, which involved an experimental design that examined the influence of linguistic features on BERT's decision-making through SHAP analysis for class-wide interpretation. Initially, a chi-square test identified terms with significant semantic biases toward any political discourse class, replacing the first 50 biased terms with neutral terms in each class of the datasets to reduce thematic content classification bias. We split the test dataset from both the TextDataset and TextDatasetAnonym in the two classes, with 234 examples per class; then, we used SHAP to identify which features (words or tokens) are more important in the model's predictions, class by class.

Features were categorized into six groups using previously used spaCy rules for linguistic feature detection, aggregating mean absolute SHAP values across classes and dataset versions. Features 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.

4. Results

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

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

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). The average word length in speeches was 4.39, slightly longer than the 4.25 observed in interviews. Sentence complexity scores differed, with speeches showing a mean of 1.36 and interviews a mean of 1.63. Passive voice frequency was low in both formats, with interviews displaying a slightly higher mean. 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. Lexical word frequency measures indicated a broad range in both speech and interview formats, with maximum values slightly higher in speeches. The mean values for nominalization were 0.80 in speeches and 0.74 in interviews. Personal pronoun frequency was more frequent in interviews, with a mean of 1.90, compared to 1.76 in speeches. 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. Modal verb frequency was relatively low in both formats, with mean values of 0.30 for speeches and 0.33 for interviews. 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 the presence of outliers or a long-tail distribution.

Figure 2

Histogram and Box Plot for Interjection Frequency in Speeches and Interviews.

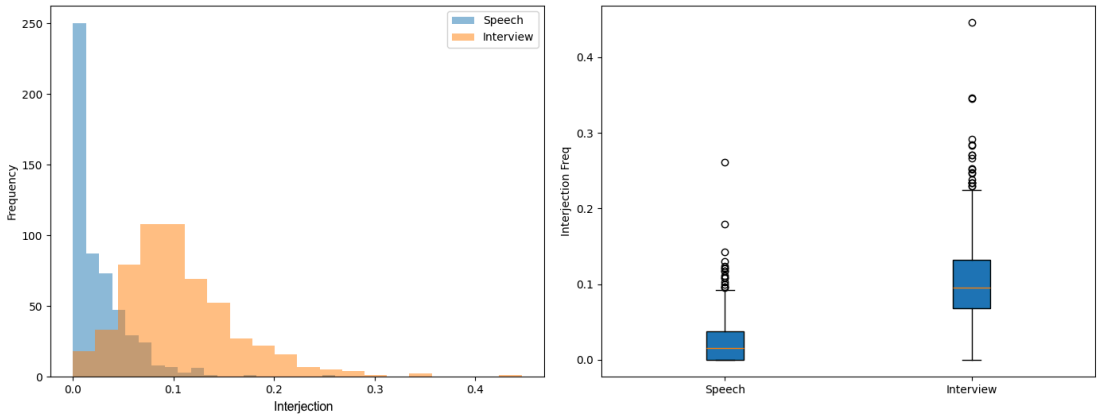
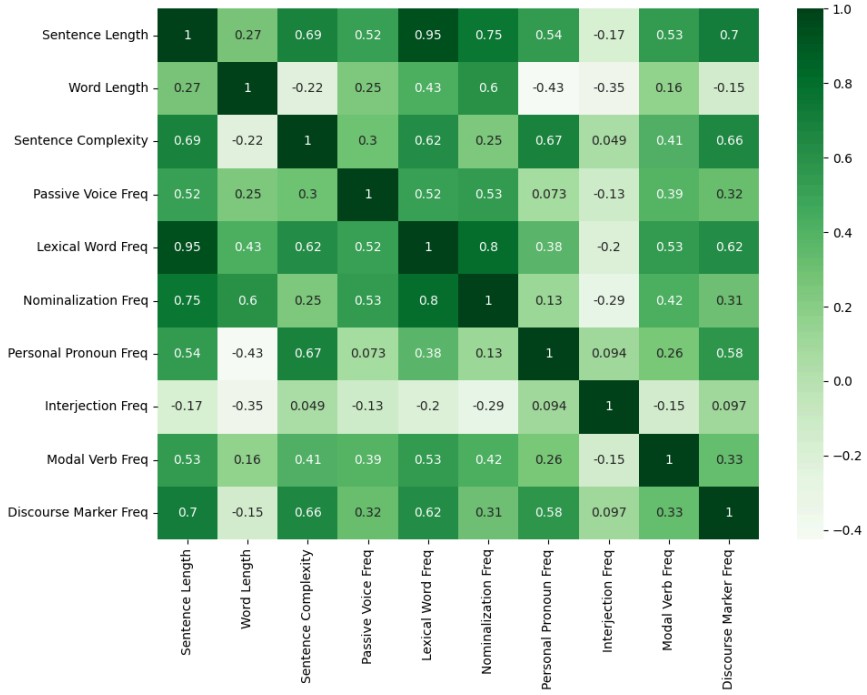


Figure 2 presents histograms illustrating the distributions of interjection frequencies within each class. The interview histogram displays a distribution that differs notably from speeches. Specifically, the interview distribution shows a broader spread of interjection frequencies, whereas a sharp peak and a pronounced long tail characterize the speech distribution. The speech class distribution's kurtosis is higher than the interview class's.

Correlation analyses were conducted to explore relationships among various linguistic features within the speech and interview classes. Figure 3 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 3

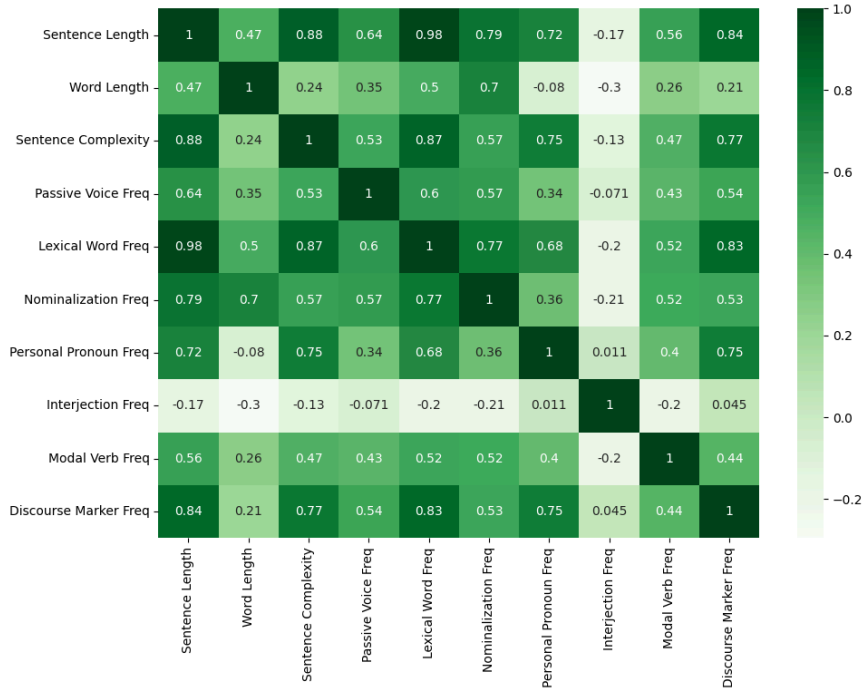
Correlation analysis for the Speech class.



In the interview class, as depicted in Figure 4, 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.

Figure 4

Correlation analysis for the Interview class.



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 that primarily serve grammatical purposes, being those usually shorter. Likewise, a sentence with more nominalizations will have more lexical words because nominalizations—nouns derived from verbs or adjectives—add to the count of nouns, thereby increasing the sentence's overall lexical content. Therefore, we pruned sentence length, sentence complexity, and lexical word frequency. Given that SHAP explanations focus on the word or *subword* level, we also pruned passive voice due to their impracticality in evaluating them at the token level. Finally, since we know that BERT models evaluate word length for their predictions, we also pruned this feature in the logistic regression analysis. We will investigate later if token length from SHAP explanations aligns with the claim found in the literature that longer words are more prevalent in speeches.

4.1 Feature Importance

Several significant findings were observed in the logistic regression analysis, 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 (*SEβ*) 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 *SEβ* of 0.932, a *z*-value of -4.426, and a *p*-value of less than .001, with *e^β* of 1.61 × 10^-02 further highlights the tendency for a lower frequency of modal verbs in speeches compared to interviews.

Table 4

Logistic Regression Results for the Classification of Speeches and Interviews.

Feature	*β*	*SEβ*	*z*	*p*	*e^β*
const (intercept)	2.936	0.457	6.426	< .001	18.8
Interjection Frequency	**-22.330**	1.575	-14.176	< .001	2.00e-10
Modal Verb Frequency	**-4.126**	0.932	-4.426	< .001	0.0161
Discourse Marker Frequency	**0.518**	1.247	0.415	**.678**	1.68
Personal Pronoun Frequency	**1.079**	1.095	0.986	**.324**	2.94
Nominalization Frequency	**2.569**	1.159	2.217	.027	13.0

Discourse marker frequency, however, did not show significantly effect the classification outcome, with a β of 0.518, an *SEβ* 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 $SE\beta$ 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 usage 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 $SE\beta$ 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.

4.2 Models Performance

The logistic regression model, 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 an 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 BERT1, compared BERT2, 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.

Table 5

Summary of performance metrics of the Logistic Regression, BERT1 and BERT2 for classifying Speeches and Interviews.

Metric	Logistic Regression	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**
Confusion Matrix (*)	*s* *i* *s* 96 16 *i* 8 100	*s* *i* *s* 231 3 *i* 6 228	*s* *i* *s* 227 7 *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			

Metric	Logistic Regression	BERT1	BERT2
Precision	0.862	0.987	0.970
Recall	0.926	0.974	0.979
F1-score	0.893	0.981	0.974

Note: (*) *s* = speech, *i* = interview. (1) Across-class metrics are macro and class-wise metrics are not averaged. (2) The BERT models are freely available in Hugging Face Hub: Speech-vs-Interview-Classification-BERT, DOI: 10.57967/hf/2649, and Speech-vs-Interview-Classification-BERT-Anonym, DOI: 10.57967/hf/2648.

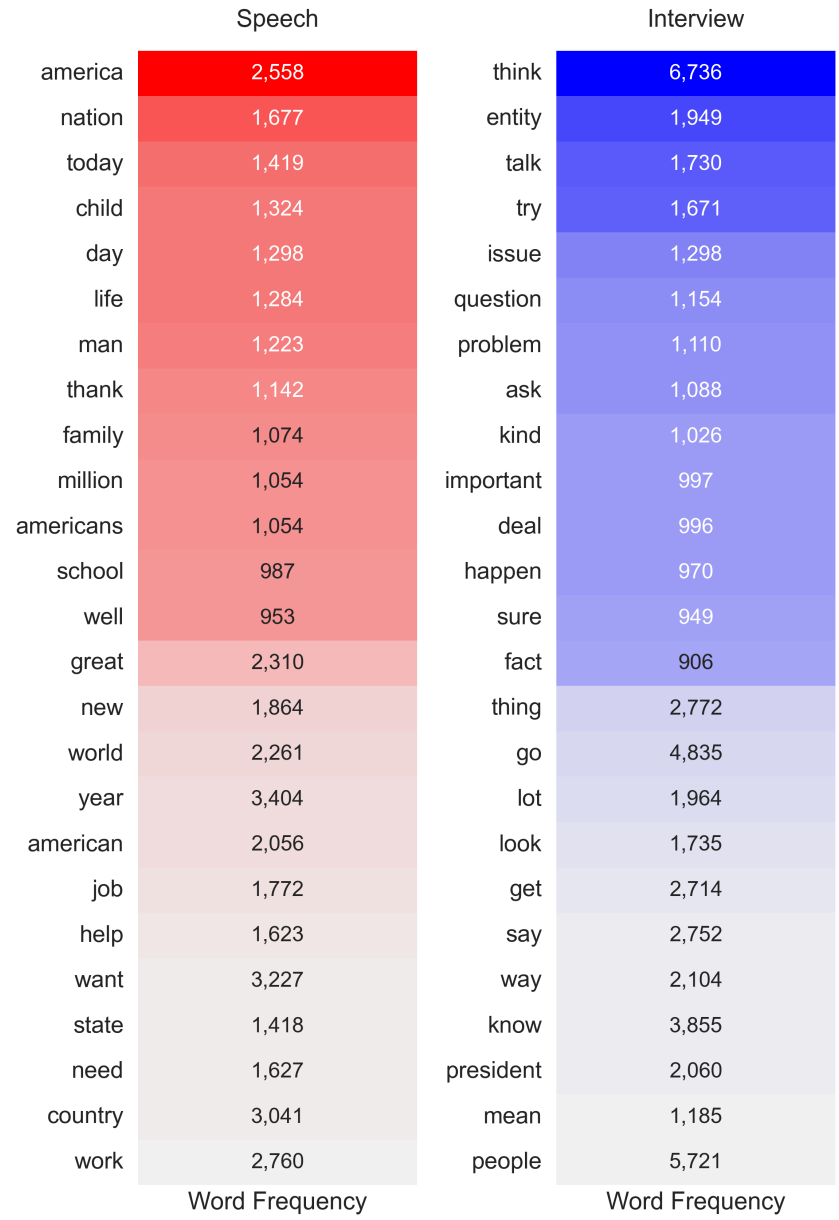
The performance of both BERT models reinforces their capabilities to understand and analyze political communication, offering insights into their capabilities and limitations in handling complex linguistic patterns. Specifically, the slight performance dip in the BERT2 and its higher AUC-ROC score (0.995) provide empirical evidence that anonymization pushes the BERT model to rely more on linguistic cues rather than thematic content, thereby improving generalization capabilities.

4.3 Bias Mitigation

The chi-square analysis to identify terms with statistically significant biases towards either class in the three splits of the Text Dataset 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. As shown in Figure 5, 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 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*".

Figure 5

Heatmap of Word Frequency Bias in Political Speeches and Interviews from Chi-Square Analysis.



Note: First 50 (lowered and lemmatized) terms per class ordered by chi-square coefficients showing number of term occurrences.

4.4 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.

Table 6

Summary of t-tests for Speech and Interview Classes based on SHAP values.

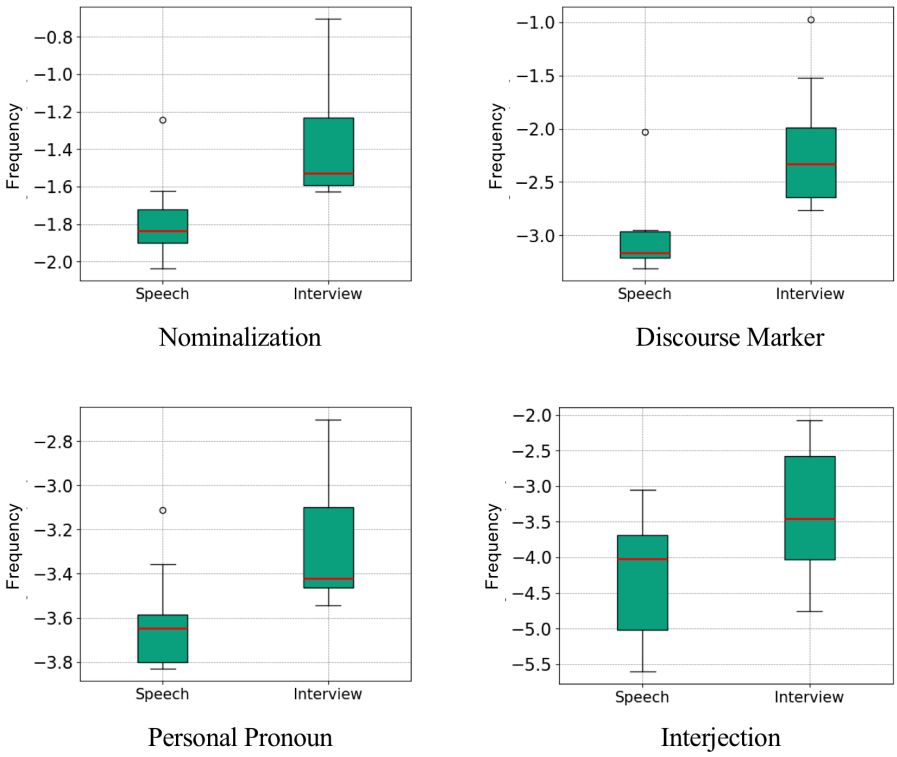
Model	Features	stats	p-value	Speech		Interview	
				M	SD	M	SD
BERT1	Nominalization Frequency	-4.2	**< .001**	-1.79	.20	-1.38	.30
	Discourse Marker Frequency	-4.9	**< .001**	-3.04	.33	-2.23	.53
	Personal Pronoun Frequency	-3.9	**.001**	-3.64	.21	-3.29	.26
	Interjection Frequency	-2.9	**.008**	-4.26	.78	-3.34	.91

Model	Features	stats	p-value	Speech		Interview	
				M	SD	M	SD
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

T-test results on SHAP values from BERT1 revealed that the four features serve as critical discriminators in the model's classification process. 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 6 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.

Figure 6

Comparison of features between Speech and Interview classes from SHAP values analized in BERT1.



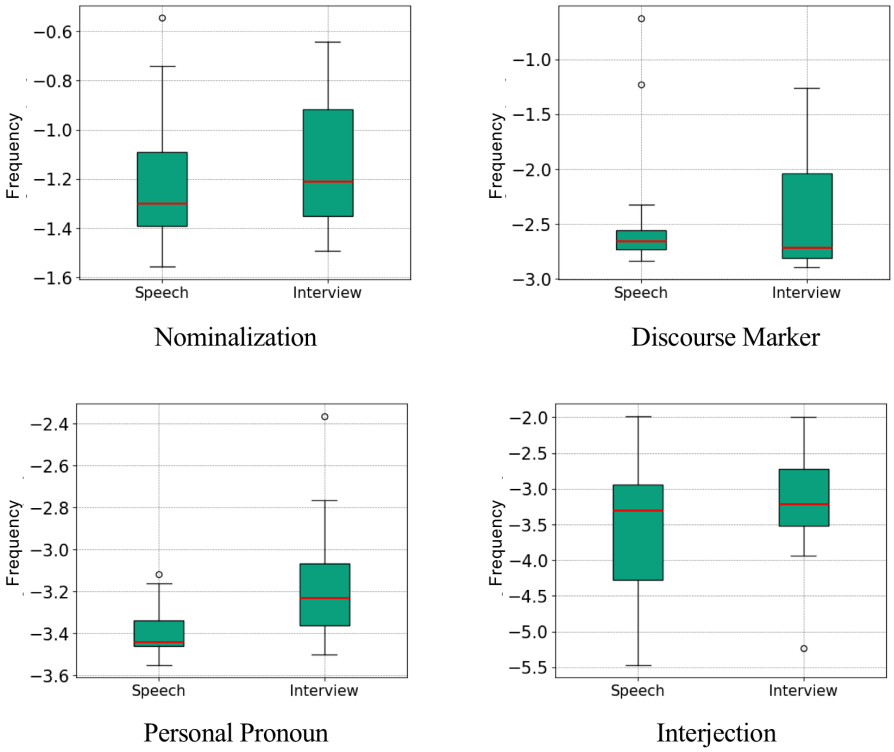
However, the results from BERT2 show a different pattern, particularly for nominalization frequency and interjection frequency, where the p-values indicate a lack of statistical significance. This suggests that with anonymization, BERT2's ability to rely on these specific linguistic features diminishes, 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 suggests that BERT2 has a reduced sensitivity to the specific linguistic characteristics, and it may leverage other aspects of the data or rely on a more generalized understanding of the text to make its classifications.

Plots in Figure 7 show that anonymizing the dataset leads to a more uniform distribution of SHAP values across features, indicating that BERT2's reliance on these linguistic structures is less pronounced when semantic and contextual identifiers are removed.

Figure 7

Comparison of features between Speech and Interview classes from SHAP values analized in BERT2.



5. Discussion

The following analysis should be understood under the important assumption that the classification between speeches and interviews is performed automatically, allowing minimal text preprocessing. Thus, interviews retain both interviewer(s) and interviewee(s) contributions, influencing the metrics of the interview class. Therefore, the conclusions drawn here should be interpreted with this context in mind, which may pose a limitation or function as a benefit, contingent upon the further downstream NLP task in the pipeline.

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 analysis 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 must 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 ($\varkappa = 21.27$) compared to interviews ($\varkappa = 16.21$), which signals that passive voice is used either very frequently or very infrequently. This discrepancy maybe 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.

Table 7

Alignment of Literature Review and Descriptive Statistic Analysis.

Feature	Speech		Interview		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

Note: Comparison made by mean values.

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 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 2 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, logistic regression analysis 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 are not statistically significant or 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 expectations. 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 the non-anonymized dataset 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) suggest that BERT models rely on these linguistic cues to differentiate between the two types of discourse. This relationship indicates that BERT's classification decisions can be partially explained by its sensitivity to these linguistic features, answering RQ1 positively.

Unlike BERT1, BERT2 relies less on the lexical and structural features of the dataset (Table 6), with the notable exception of Personal Pronoun Frequency. This observation is crucial for answering RQ2, which suggests that anonymization could make the BERT model's decisions

more explainable by decreasing dependence on thematic context and prompting the model to base its decisions more on linguistic structures. However, 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.

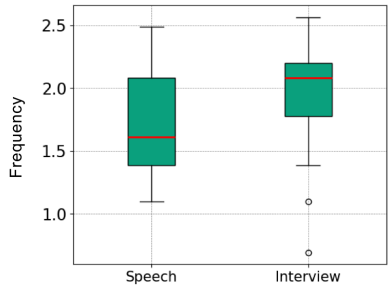
Therefore, RQ2 is partially answered since the performance metrics of BERT2 prove that the model relies less on thematic data but not on the studied linguistic features, as we expected; all that without detriment of its performance. The following are explanations that may contribute to our understanding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

1. 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.
2. 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.
3. 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.

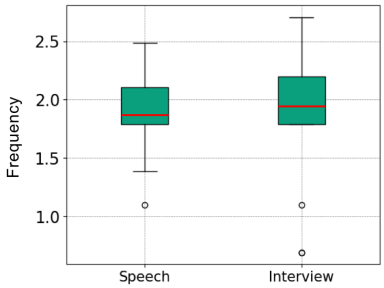
One possible explanation for the previous ideas is the preference of BERT models in subwords after the limitation that NER anonymization brings to the model. Analyzing the word length (token length) of the SHAP's attributed features (tokens) before and after the NER anonymization, we observed a shift in the model's focus from relying on potentially identifying longer tokens ("*whole*" words) to shorter tokens (subwords). As the model no longer has access to certain features that could have been crucial for its predictions, it appears to compensate by placing greater emphasis on the remaining subwords. Therefore, despite the statistical insignificance resulted from a t-test ($p = 0.325$) of word length, it is possible to observe in Figure 8 a tendency towards shorter words after the NER anonymization treatment. This shift reinforces the idea that subword tokenization is a key component in the model's predictive capability in the context of anonymized inputs.

Figure 8

Comparison of Token (Word) Length of SHAP attributed features between BERT models with and with no Anonymization for Speech and Interview Classification.



Word Length in BERT



Word Length in BERT with NER Anonymization

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 underscores how NER anonymization can influence the linguistic features SHAP identifies as significant, notably by reducing the average word length within both classes.

6. Conclusions

Our investigation shed lights 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 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 at identifying linguistic features in the text by capturing the complexities of political language; nevertheless, 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 machine learning models contributed enormously to the explainability of the NLP models and the comprehension of linguistic phenomena.

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 the logistic regression results (Table 4), five linguistic features—measured and extracted automatically—were enough to achieve high performance in distinguishing speeches and interviews.

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 scarifying 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.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. In *KDD*. https://doi.org/10.48550/arXiv.1907.10902.
2. Akpatsa, S. K., Li, X., & Lei, H. (2021). A Survey and Future Perspectives of Hybrid Deep Learning Models for Text Classification. In Sun, X., Zhang, X., Xia, Z., & Bertino, E. (Eds.). *Artificial Intelligence and Security* (pp. 358-369). Springer. https://doi.org/10.1007/978-3-030-78609-0_31.
3. 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). https://doi.org/10.30743/ll.v4i1.2519.
4. Austin, J. L. (1962). *How to Do Things with Words*. Clarendon Press.
5. Bakhtin, M. M. (1981). *The Dialogic Imagination: Four Essays* (C. Emerson & M. Holquist, Trans). University of Texas Press.
6. Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024.
7. Biber, D. (1992). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, 15(2), 133-163. https://doi.org/10.1080/01638539209544806.

8. Biber, D., & Finegan, E. (1994). *Sociolinguistic Perspectives on Register*. Oxford University Press.
9. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Pearson Education.
10. 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.
11. Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.
12. Burenko, T., & Fedorova, D. (2020). Interjections Functioning in Modern American Discourse. *Lviv Philological Journal*. https://doi.org/10.32447/2663-340x-2020-8.4.
13. Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, 538(7623), 20-23. https://doi.org/10.1038/538020a.
14. Chilton, P. (2004). *Analyzing Political Discourse: Theory and Practice*. Routledge.
15. Clark, K., Khandelwal, U., Levy, O., & Manning, C. (2019). What Does BERT Look at? An Analysis of BERT's Attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 276-286. https://doi.org/10.18653/v1/W19-4828.
16. Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viegas, F., & Wattenberg, M. (2019). Visualizing and Measuring the Geometry of BERT. *arXiv*. https://doi.org/10.48550/arXiv.1906.02715.
17. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 1, 4171-4186. https://doi.org/10.18653/v1/N19-1423.
18. Dwork, C. (2006). Differential Privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming - Volume Part II (ICALP 2006)*, Springer-Verlag, Berlin, Heidelberg, 1-12. https://doi.org/10.1007/11787006_1.
19. Dingemanse, M. (2021). Interjections. *Oxford Handbook of Word Classes*. https://doi.org/10.31234/osf.io/ngcrs.
20. Fairclough, N. (2001). *Language and power* (2nd ed.). London: Longman.
21. Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979). *Language and Control*. Routledge & Kegan Paul.
22. Grzybek, P., Stadlober, E., & Kelih, E. (2006). The Relationship of Word Length and Sentence Length: The Inter-Textual Perspective. In R. Decker & H-J Lenz (Eds.). *Advances in Data Analysis* (pp. 611-618). Springer, Heidelberg. https://doi.org/10.1007/978-3-540-70981-7_70.
23. Gumperz, J. J. (1982). *Discourse Strategies*. Cambridge University Press.
24. Gupta, S., Bolden, S. E., Kachhadia, J., Korsunska, A., & Stromer-Galley, J. (2020). PoliBERT: Classifying political social media messages with BERT. In *Proceedings*

Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) Conference. Washington, DC.

25. Halliday, M. A. K. (1994). *An Introduction to Functional Grammar*. Edward Arnold.
26. Heeman, P., Byron, D., & Allen, J. (1998). Identifying Discourse Markers in Spoken Dialog. *arXiv*. https://doi.org/10.48550/arXiv.cmp-lg/9801002.
27. Hirst, G. (2001). Book Reviews: Longman Grammar of Spoken and Written English. *Computational Linguistics*, 27(1), 132-139. https://doi.org/10.1162/089120101300346831.
28. Holmes, J. (1990). Hedges and Boosters in Women's and Men's Speech. *Language & Communication*, 10(3), 185-205. https://doi.org/10.1016/0271-5309(90)90002-S.
29. Honnibal, M., & Montani, I. (2017). *spaCy: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*.
30. 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.
31. Htut, P., Phang, J., Bordia, S., & Bowman, S. (2019). Do Attention Heads in BERT Track Syntactic Dependencies?. *ArXiv*. https://doi.org/10.48550/arXiv.1911.12246.
32. Hyland, K. (2005). Metadiscourse: Exploring interaction in writing. *Continuum*.
33. Jawahar, G., Sagot, B., & Seddah, D. (2019). What Does BERT Learn about the Structure of Language?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3651-3657. https://doi.org/10.18653/v1/P19-1356.
34. Jumelet, J., & Zuidema, W. (2023). Feature Interactions Reveal Linguistic Structure in Language Models. *arXiv*, 8697-8712. https://doi.org/10.48550/arXiv.2306.12181.
35. Jurafsky, D., & Martin, J. H. (2023). *Speech and Language Processing* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/.
36. 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.
37. 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.
38. 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.
39. 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.
40. 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. In *Journal of Low Power Electronics and Applications*, 11(2), 18. https://doi.org/10.20944/PREPRINTS202101.0621.V1.
41. Li, B. (2022). Integrating Linguistic Theory and Neural Language Models. *arXiv*. https://doi.org/10.48550/arXiv.2207.09643.

42. Li, Z., Zhou, Q., Li, C., Xu, K., & Cao, Y. (2020). Improving BERT with Syntax-aware Local Attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 645-653. https://doi.org/10.18653/v1/2021.findings-acl.57.
43. 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.
44. Liu, N. F., Gardner, M., Belinkov, Y., Peters, M. E., & Smith, N. A. (2019). Linguistic knowledge and transferability of contextual representations. *arXiv*. https://doi.org/10.48550/arXiv.1903.08855.
45. Mauranen, A. (2023). *Reflexively speaking: Metadiscourse in English as a lingua Franca*. De Gruyter. https://doi.org/10.1515/9783110295498.
46. McCarthy, M., & Carter, R. (1995). Spoken Grammar: What Is It and How Can We Teach It? *ELT Journal*, 49(3), 207-2018.
47. 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.
48. Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one? *arXiv*. https://doi.org/10.48550/arXiv.1905.10650.
49. Mosca, E., Demirtürk, D., Mülln, L., Raffagnato, F., & Groh, G. (2022). GrammarSHAP: An Efficient Model-Agnostic and Structure-Aware NLP Explainer. In *Proceedings of the First Workshop on Learning with Natural Language Supervision*, 10-16. https://doi.org/10.18653/v1/2022.lnls-1.2.
50. Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019). Explanation of Machine Learning Models Using Improved Shapley Additive Explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. https://doi.org/10.1145/3307339.3343255.
51. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*.
52. 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.
53. Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.
54. Raskin, A., & Harris, T. (2023). The A.I. Dilemma. The Singju Post. Retrieved from https://singjupost.com/discussion-the-a-i-dilemma-march-9-2023-transcript/.
55. Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. In *Transactions of the Association for Computational Linguistics*, 8, 842-866. https://doi.org/10.1162/tacl_a_00349.

56. 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.
57. Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.
58. Sweeney, L. (2002). k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5), 557-570. https://doi.org/10.1142/S0218488502001648.
59. Szczepański, M., Pawlicki, M., Kozik, R., & Choraś, M. (2021). New explainability method for BERT-based model in fake news detection. *Scientific Reports*, 1, 2375. https://doi.org/10.1038/s41598-021-03100-6.
60. Tannen, D. (1981). *Conversational Style: Analyzing Talk Among Friends*. Ablex Pub. Corp.
61. Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4593-4601. https://doi.org/10.18653/v1/P19-1452.
62. The pandas development team (2020). pandas-dev/pandas: Pandas. Zenodo. https://doi.org/10.5281/zenodo.3509134.
63. Tottie, G. (1991). Conversational Style in British and American English: The Case of Interjections. *Journal of Pragmatics*, 15(1), 13-28.
64. Van Dijk, T. A. (1998). *Ideology: A Multidisciplinary Approach*. Sage Publications.
65. Vanni, L., Corneli, M., Mayaffre, D., & Precioso, F. (2020). From text saliency to linguistic objects: learning linguistic interpretable markers with a multi-channels convolutional architecture. *arXiv*. https://doi.org/10.48550/arXiv.2004.03254.
66. Waskom, M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open Source Software*, 6(60), 3021. https://doi.org/10.21105/joss.03021.
67. 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.
68. 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.
69. Yao, J. (2009). A Critical Analysis of Nominalization in News Discourse. *Journal of Xinzhou Teachers University*..
70. Yin, K., & Neubig, G. (2022). Interpreting Language Models with Contrastive Explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 184-198. https://doi.org/10.18653/v1/2022.emnlp-main.14.
71. Yong, C. (2002). Functions and Features of Personal Pronouns in Discourse. *Journal of Changde Teachers University*.
72. 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.

73. Zafar, M., Donini, M., Slack, D., Archambeau, C., Das, S., & Kenthapadi, K. (2021). On the Lack of Robust Interpretability of Neural Text Classifiers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 3730-3740. https://doi.org/10.18653/v1/2021.findings-acl.327.

74. 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. *arXiv*. https://doi.org/10.48550/arXiv.2112.12444.

75. 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.

76. Zhao, W., Joshi, T., Nair, V., & Sudjianto, A. (2020). SHAP values for Explaining CNN-based Text Classification Models. *arXiv*. https://doi.org/10.48550/arXiv.2008.11825.