information

MDPI

Article

The Power of Words from the 2024 United States Presidential Debates: A Natural Language Processing Approach

Ana Lorena Jiménez-Preciado [1], José Álvarez-García [2,*], Salvador Cruz-Aké [1] and Francisco Venegas-Martínez [1]

[1] Escuela Superior de Economía, Instituto Politécnico Nacional, Av. Plan de Agua Prieta 66, Miguel Hidalgo, Mexico City 11350, Mexico; ajimenezp@ipn.mx (A.L.J.-P.); scruza@ipn.mx (S.C.-A.); fvenegas1111@yahoo.com.mx (F.V.-M.)
[2] Departamento de Economía Financiera y Contabilidad, Instituto Universitario de Investigación para el Desarrollo Territorial Sostenible (INTERRA), Facultad de Empresa Finanzas y Turismo, Universidad de Extremadura, Avda. de la Universidad, n° 47, 10071 Cáceres, Spain
* Correspondence: pepealvarez@unex.es

Abstract: This study analyzes the linguistic patterns and rhetorical strategies employed in the 2024 U.S. presidential debates from the exchanges between Donald Trump, Joe Biden, and Kamala Harris. This paper examines debate transcripts to find underlying themes and communication styles using Natural Language Processing (NLP) advanced techniques, including an n-gram analysis, sentiment analysis, and lexical diversity measurements. The methodology combines a quantitative text analysis with qualitative interpretation through the Jaccard similarity coefficient, the Type–Token Ratio, and the Measure of Textual Lexical Diversity. The empirical results reveal distinct linguistic profiles for each candidate: Trump consistently employed emotionally charged language with high sentiment volatility, while Biden and Harris demonstrated more measured approaches with higher lexical diversity. Finally, this research contributes to the understanding of political discourse in high-stakes debates through NLP and can offer information on the evolution of the communication strategies of the presidential candidates of any country with this regime.

Keywords: presidential debates; natural language processing; sentiment analysis; Jaccard similarity; type–token ratio; measure of textual lexical diversity

MSC: 68T50; 76M55; 68T10

Academic Editor: Arkaitz Zubiaga

Received: 20 November 2024
Revised: 11 December 2024
Accepted: 18 December 2024
Published: 25 December 2024

Citation: Jiménez-Preciado, A.L.; Álvarez-García, J.; Cruz-Aké, S.; Venegas-Martínez, F. The Power of Words from the 2024 United States Presidential Debates: A Natural Language Processing Approach. *Information* **2025**, *16*, 2. https://doi.org/10.3390/info16010002

1. Introduction

Presidential debates have always been an essential part of the democracy of any country with this regime. They provide a crucial platform for candidates to present their vision, policies, and leadership qualities directly to the voters. Presidential debates often influence voter decisions, potentially swaying the outcome of elections. Debates offer an opportunity to observe candidates under pressure, revealing aspects of their character and knowledge, as well as their ability to think independently. The language used in debates reflects the broader political discourse and can provide an understanding of the prevailing concerns and values of the electorate. In an era of increasing political polarization and media fragmentation, debates remain among the few occasions where candidates must engage directly with each other and address a wide range of issues before a national audience.

The study of presidential debates has a rich scholarly tradition. Seminal work by [1] laid the foundation for understanding these events' rhetorical structure and impact. Furthermore, Ref. [2] has developed several frameworks for analyzing debate content and its

effects on voter perceptions. Studies from [3,4] have specifically examined the linguistic features of political discourse in debates, providing features about the nature of debate analysis, encompassing rhetoric, political communication, and linguistic studies.

In recent years, the field of Natural Language Processing (NLP) has revolutionized political discourse analysis, including presidential debates. NLP techniques allow for the systematic examination of large volumes of text data, revealing patterns and insights that might be missed by traditional qualitative analyses. Methods such as sentiment analysis, as described by [5], enable researchers to quantify the emotional tone of candidates' statements. The n-gram analysis and lexical diversity measurements, discussed by [6], provide tools for examining the complexity and variety of language used in debates. When combined with qualitative interpretation, these computational approaches offer a powerful means of dissecting the nuances of political communication in debate settings.

The 2024 U.S. presidential debates were particularly noteworthy, as they featured the unexpected withdrawal of the incumbent president, Joe Biden, and the subsequent face-off between the former president, Donald Trump, and Vice President Kamala Harris. These debates played a significant role in shaping public opinion. This study focuses on the candidates' linguistic patterns, sentiments, and rhetorical strategies. NLP techniques, including an n-gram analysis, sentiment analysis, and linguistic diversity measurements, seek to discover topics, emotional tones, and communication styles that characterize these crucial political events. The methodology combines a quantitative text analysis with qualitative interpretation, drawing on transcripts from the CNN-hosted debate between Donald Trump and Joe Biden on 27 June 2024, and the subsequent ABC News-hosted debate between Donald Trump and Kamala Harris on 10 September 2024. The proposed analysis is based on the debates and shows how the candidates articulated their positions, responded to challenges, and attempted to connect with voters during these pivotal moments in the 2024 presidential campaign.

This paper is organized as follows: Section 2 briefly reviews the literature, highlighting the findings from the analysis of texts under different contexts, with a specific emphasis on presidential speeches; Section 3 provides the methodology, starting with text cleaning and generating unigrams, bigrams, and trigrams; Section 4 focuses on the sentiment analysis of the speeches and explores the relationship between the first and second debates; Section 5 presents the results from the Jaccard similarity coefficient and the Type–Token Ratio to analyze the diversity of the lexicon used in the speeches; Section 6 gives conclusions and provide recommendations.

2. Brief Literature Review

Analyzing political communication has always been crucial for understanding democratic processes and election dynamics. Amongst the diverse aspects of this field, presidential debates have become a key area of research, providing valuable insights into candidate behavior, public perception, and electoral results. Political communication is a diverse discipline covering media effects, campaign strategies, and public opinion formation. In this sense, Ref. [7] offers an overview of the field, emphasizing its interdisciplinary nature and crucial role in molding democratic discourse. Likewise, Ref. [8] explores the changing landscape of political communication in the digital age, highlighting the growing significance of social media and personalized messaging in political campaigns.

The investigation of presidential debates falls at the intersection of political communication and rhetorical analysis. The influential work in [1] establishes the foundation for comprehending the format and influence of these crucial political occurrences. The authors maintain that debates function as a method of sharing information and as ritualistic performances that uphold democratic values. Moreover, Ref. [9] develops this line of

inquiry by proposing a functional theory of political campaign discourse. The author provides a framework for analyzing debate content and categorizing statements into acclaims, attacks, and defenses. Subsequent research has widely adopted this approach, offering a systematic method for comparing candidates' rhetorical strategies across different debates and election cycles.

The linguistic aspects of political debates have been receiving more attention. In this sense, Ref. [3] led a comparative analysis of the language used by candidates in U.S. presidential debates. This study revealed distinct lexical patterns associated with political ideologies and personal styles and showed the potential of computational linguistics in uncovering subtle aspects of political communication that might be missed by traditional qualitative analyses.

Some studies have explored Donald Trump's distinctive communication style, particularly in the context of the recent U.S. presidential debates. In this sense, Ref. [4] analyzed Trump's language during the 2016 presidential debates and identified key features such as grandiosity, informality, and dynamism. The authors suggest that Trump's unconventional rhetorical approach influenced his electoral success by setting him apart from traditional politicians. Building on this, Ref. [10] exhibits a critical discourse analysis of the 2016 debates, highlighting Trump's use of simple language, repetition, and emotionally charged expressions. The work by [11] introduces a novel framework, the LLM-POTUS Score, which analyzes candidates' "Policies, Persona, and Perspective" (3P) and how they resonate with the "Interests, Ideologies, and Identity" (3I) of key audience groups. It examines the effectiveness of different debating strategies and their impact on various audience segments.

Recent studies have further expanded the application of NLP in political sentiment analyses. For instance, Ref. [12] examined the politicization of immigration in Spanish parliamentary debates by applying NLP techniques. Their study highlighted how ideological stances and political positions (government versus opposition) influenced the framing of immigrants, either as a "threat" or as "victims. Another significant contribution can be found in [13]. The study utilized word embeddings to analyze political language in Austrian parliamentary speeches. By mapping the semantic relationships among words, they could trace shifts in the political discourse over time and across different parties. The authors exposed the evolution of language in presidential debates, revealing subtle changes in candidates' rhetorical strategies and policy positions.

Integrating NLP techniques with traditional content analysis methods has also yielded valuable knowledge. In this sense, Ref. [14] combined topic modeling with a sentiment analysis to study public opinion on Twitter during the 2016 U.S. presidential debates. Their approach captured the sentiment of public reactions and identified key debate topics.

Recent advances in computational linguistics and NLP have revolutionized political discourse analyses. In the same way as the present study, but focusing on different aspects of the U.S. 2024 debates, Ref. [15] analyzes the language patterns in Trump and Harris's 2024 presidential debate, spotting framing values, emotional appeals, and ideological markers. Their research showed thematic differences, with Harris often framing issues around recovery and empowerment, while Trump focused on crisis and decline narratives. Based on the above, Ref. [16] proposes a context-based disambiguation model for sentiment concepts using the bag-of-concepts technique, demonstrating how semantic augmentation through commonsense knowledge can improve the accuracy of sentiment analyses in political discourse.

Furthermore, recent studies have begun incorporating multimodal analysis approaches. For instance, Ref. [17] shows how social media logic influences political communication during elections, emphasizing the need to consider multiple communication

channels. For instance, Ref. [18] further highlights the importance of considering visual elements alongside textual analysis in political communication, particularly in debates where television images interact with social media discourse.

The approach presented in this document combines a traditional debate analysis with NLP techniques and aims to contribute to understanding political communication in the context of high-stakes presidential debates. As NLP methodologies evolve, they promise to reveal patterns in political language, enhancing the understanding of how debates shape public opinion and influence electoral outcomes.

3. Methodology: Preprocessing and Stopwords Removal

The analysis of the presidential debates began with the transcript of the Trump–Biden discussion held on 27 June 2024, in Atlanta, Georgia, presented by CNN. Jake Tapper and Dana Bash moderated the debate. Subsequently, the transcript of the second debate between Donald Trump and Vice President Kamala Harris, on 22 July 2024, was taken after President Joe Biden unexpectedly withdrew from the 2024 presidential race and endorsed Vice President Kamala Harris as his replacement. This decision came after weeks of mounting pressure from democrats, citing concerns about Biden's age (81) and health, as well as his perceived inability to defeat Donald Trump in the November election. The Harris–Trump debate took place on 10 September 2024, in Philadelphia, and was moderated by David Muir and Linsey Davis, both of whom are ABC News anchors. The transcripts from both sources were retrieved from [19,20].

The first step was to isolate the comments made by the presidential candidates from all the debate transcripts, focusing only on the content provided by the candidates while excluding interjections from moderators and audience members. Along these lines, Ref. [21] pointed out that this targeted extraction allows a more accurate representation of each candidate's discourse and rhetorical strategies. After extracting the candidate comments, a thorough text-cleaning process was implemented, including

1. The removal of special characters and punctuation;
2. The conversion of all text to lowercase;
3. The elimination of numerical digits;
4. The removal of extra whitespace.

The following steps are NLP tasks that help to normalize the text data, reduce noise, and improve the consistency of the corpus; see [6]. A required step in the preprocessing pipeline was the removal of stopwords, which are common words (e.g., "the", "is", "and") that typically do not contribute significant meaning to the analysis of text content. Their removal is important since it improves the focus on content-bearing words by eliminating high-frequency functional words; the analysis can concentrate on words that carry more semantic weight; see [22]. Likewise, removing stopwords helps reduce the dimensionality of the data, which can significantly improve the computational efficiency in later analyses, as shown in [23]. This efficiency not only saves time but also enhances the effectiveness of text mining techniques, as they can reveal underlying patterns in the text, according to [24].

The next step involves tokenizing each word. This process breaks down the continuous text into discrete units; see [25]. In addition, the study generated n-grams (specifically bigrams and trigrams) to capture multi-word expressions and phrases that are particularly significant in political discourse, allowing the identification of common collocations and recurring phrases used by the candidates, as in [26]. In the Biden–Trump debate, Joe Biden delivered a total of 8383 words, with 2932 words remaining after excluding stopwords. In comparison, Donald Trump spoke 9959 words, of which 3357 words remain after removing the stopwords. In the second debate, Donald Trump uttered 9675 words (3442 after

excluding the stopwords), while Kamala Harris used 6702 words. Harris's text contained 2747 words, excluding the stopwords.

3.1. First Presidential Debate's Bag of Words

Figure 1 represents the first presidential debate's unigram, bigram, and trigram analyses. The frequency distribution of the unigrams (single words) provides information on each candidate's main themes and rhetorical strategies. Let $f_{Trump}(w)$ and $f_{Biden}(w)$ be the frequency of word w in the speeches of Trump and Biden, respectively.

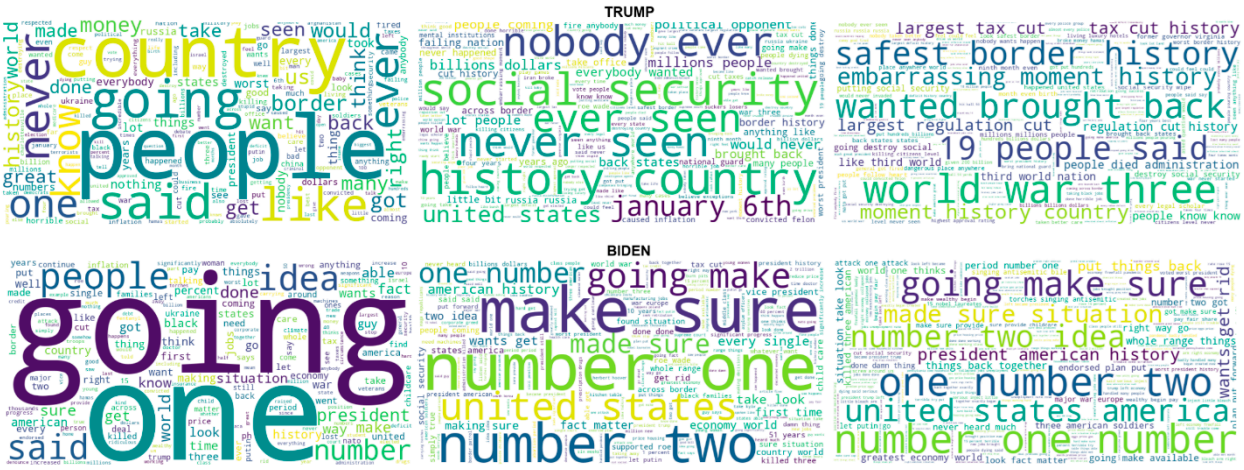

Figure 1. Trump–Biden presidential debate's unigram, bigram, and trigram word cloud, without stopwords. Source: Authors' own elaboration.

For Trump, the most frequent unigrams were

1. f_{Trump}("people") = 71
2. f_{Trump}("country") = 46
3. f_{Trump}("going") = 45

In contrast, Biden's most frequent unigrams were

1. f_{Biden}("going") = 42
2. f_{Biden}("one") = 42
3. f_{Biden}("people") = 38

The high frequency of "people" in both candidates' speeches (ranking 1st for Trump and 3rd for Biden) suggests a populist approach, attempting to connect with the electorate. However, Trump's more frequent use of this term f_{Trump}("people") > f_{Biden}("people") may indicate a stronger emphasis on populist rhetoric. Trump's frequent use of f_{Trump}("country") = 46 suggests a focus on national issues and patriotic themes. In contrast, Biden's repeated use of "one" f_{Biden}("one") = 42 could indicate an attempt to present unified or singular solutions.

Nevertheless, the bigram analysis shows clearer patterns in the candidates' speeches. Let $f_{Trump}(w_1, w_2)$ and $f_{Biden}(w_1, w_2)$ represent the frequency of bigram (w_1, w_2) in Trump's and Biden's speeches, respectively. Trump's most frequent bigrams were

1. f_{Trump}("history", "country") = 9
2. f_{Trump}("social", "security") = 9
3. f_{Trump}("never", "seen") = 8

Biden's most frequent bigrams were

1. f_{Biden}("number", "one") = 14
2. f_{Biden}("make", "sure") = 13
3. f_{Biden}("number", "two") = 9

The bigram "social security" in Trump's speech f_{Trump}("social", "security") = 9 suggests a focus on welfare and retirement issues, potentially appealing to older voters. In contrast, Biden's frequent use of "make sure" f_{Biden}("make", "sure") = 13 could be interpreted as an attempt to project confidence and certainty in his proposed policies.

Finally, trigrams provide even more context for the candidates' messaging. Let $f_{Trump}(w_1, w_2, w_3)$ and $f_{Biden}(w_1, w_2, w_3)$ be the trigram frequency in Trump's and Biden's speeches. Trump's most frequent trigram was f_{Trump}("world", "war", "three") = 4, centering on international relations and potential global conflicts, explicitly referring to the Ukraine–Russia conflict. In contrast, Biden's most frequent trigram was f_{Biden}("number", "one", "number") = 7. This repetitive pattern could indicate a structured approach to presenting ideas. However, Biden's performance was highly criticized due to his appearance of freezing up repeatedly and fumbling even in prepared lines, making communicating his stance on issues complex. In that sense, Biden is characterized as an "anti-charismatic" leader; see [27]. Table 1 displays the 20 total top-frequency words identified by the unigram, bigram, and trigram analyses of the first debate.

Table 1. Trump–Biden top-frequency words from the unigram, bigram, and trigram analyses.

Trump					
1-g	**Count**	**2-g**	**Count**	**3-g**	**Count**
people	71	history country	9	world war three	4
country	46	social security	9	wanted brought back	3
going	45	never seen	8	safest border history	3
said	39	nobody ever	8	19 people said	3
like	36	ever seen	7	embarrassing moment history	3
never	32	January 6th	7	moment history country	3
ever	29	united states	6	largest tax cut	2
know	26	millions people	6	tax cut history	2
one	26	would never	6	largest regulation cut	2
border	24	billions dollars	6	regulation cut history	2
us	23	lot people	5	people know know	2
money	23	political opponent	5	people died administration	2
history	23	people coming	5	like third world	2
got	23	everybody wanted	5	third world nation	2
right	22	back states	5	putting social security:	2
think	22	brought back	5	going destroy social	2
many	21	border history	5	destroy social security	2
would	21	failing nation	5	millions millions people	2
back	20	cut history	4	social security wipe	2
get	20	things done	4	happened united states	2
Biden					
1-g	**count**	**2-g**	**count**	**3-g**	**count**
going	42	number one	14	number one number	7
one	42	make sure	13	one number two	7
people	38	number two	9	going make sure	5
idea	35	united states	9	number two idea	4
said	34	going make	7	united states America	4
number	32	one number	7	made sure situation	3
president	30	made sure	6	president American history	3
done	26	American history	6	wants get rid	2
world	25	take look	5	put things back	2
sure	23	every single	5	things back together	2
make	22	wants get	6	whole range things	2
get	21	economy world	4	situation take look	2
fact	21	fact matter	4	going make available	2
able	20	making sure	4	greatest economy world	2
know	20	first time	4	world one thinks	2
way	20	two idea	4	period number one	2
situation	19	vice president	4	number two got	2
look	18	get rid	4	right way go	2
time	16	social security	4	killed three American	2
got	16	states America	4	three American soldiers	2

In the formation presented in Table 1, several thematic analyses emerge. For instance, the first theme appears to focus on economics, as Trump frequently used. f_{Trump}("billions", "dollars") = 6, and "largest tax cut", f_{Trump}("largest", "tax", "cut") = 2. Likewise, Trump makes frequent references to history, f_{Trump}("history") = 23, f_{Trump}("history", "country") = 9, often in the context of superlatives (e.g., "largest tax cut history", "largest regulation cut history"). This result suggests a strong emphasis on economic issues, particularly tax policy. While also touching on economic themes, Biden's language focuses less on specific fiscal measures.

Another topic is related to border security; Trump's language shows a significant focus on border concerns, f_{Trump}("border") = 24, f_{Trump}("safest", "border", "history") = 3, reflecting his emphasis on immigration policy. Biden's speech shows a different focus on this topic. From the standpoint of structured language, Biden's frequent use of ordered phrases, like "number one", f_{Biden}("number", "one") = 14, and "number two", f_{Biden}("number", "two") = 9, can be interpreted as an attempt to appear organized and methodical in his arguments. Both candidates use action verbs frequently, but in different contexts. Biden often uses "make sure", f_{Biden}("make", "sure") = 13, and Trump uses phrases like "never seen", f_{Trump}("never", "seen") = 8, and "ever seen", f_{Trump}("ever", "seen") = 7, emphasizing unique actions or situations.

3.2. Second Presidential Debate's Bag of Words

Figure 2 presents the data displaying unigrams, bigrams, and trigrams from the second presidential debate. Like the first presidential analysis, each candidate exhibits a frequency distribution of unigrams, bigrams, and trigrams. Let $f_{Trump}(w)$ and $f_{Harris}(w)$ characterize the frequency of the words, w, in Trump's and Harris' speeches, respectively.

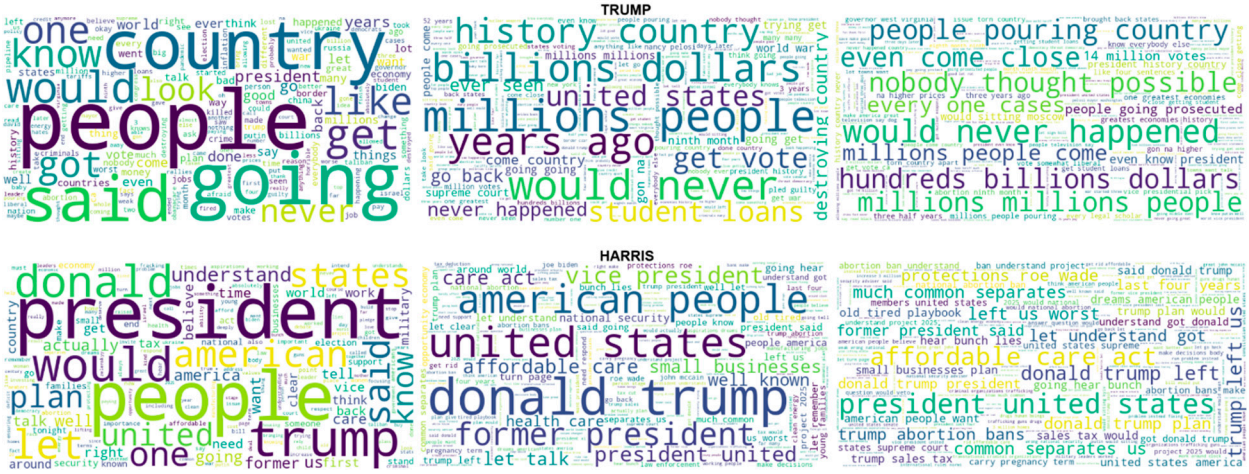

Figure 2. Trump-Harris presidential debate's unigram, bigram, and trigram word cloud, without stopwords. Source: Authors' own elaboration.

For Trump's second debate, the most frequent unigrams were

1. f_{Trump}("people") = 80
2. f_{Trump}("country") = 55
3. f_{Trump}("going") = 50

In contrast, Harris's most frequent unigrams were

1. f_{Harris}("president") = 55
2. f_{Harris}("people") = 49
3. f_{Harris}("trump") = 38

The high frequency of "people" in both candidates' speeches (ranking 1st for Trump and 2nd for Harris) suggests a continued populist approach, attempting to connect with the electorate. However, Trump's more frequent use of this term f_{Trump}("people") > f_{Harris}("people") may indicate a stronger emphasis on populist rhetoric. Trump's frequent use of f_{Trump} ("country") = 55 suggests focusing on national issues and patriotic themes. In contrast, Harris's repeated use of "president" f_{Harris}("president") = 55 could indicate an attempt to emphasize her experience and readiness for the role or to critique the current administration.

The bigram analysis reveals clearer patterns in the candidates' speeches. Trump's most frequent bigrams were

f_{Trump}("millions", "people") = 9
f_{Trump}("would", "never") = 8
f_{Trump}("years", "ago") = 7

Harris's most frequent bigrams were

f_{Harris}("donald", "trump") = 32
f_{Harris}("united", "states") = 20
f_{Harris}("american", "people") = 18

The bigram "millions people" in Trump's speech f_{Trump}("millions", "people") = 9 suggests a focus on large-scale issues or achievements, potentially appealing to a sense of magnitude. Harris's frequent use of "donald trump" f_{Harris}("donald", "trump") = 32 indicates a strategy of directly addressing or criticizing her opponent.

Trigrams provide even more context for the candidates' messaging. Trump's most frequent trigram was f_{Trump}("would", "never", "happened") = 4, potentially indicating a focus on hypothetical scenarios or criticisms of alternative policies. Harris's most frequent trigram was f_{Harris}("president", "united", "states") = 7, emphasizing the role and responsibilities of the presidency.

From the point of view of the n-gram analysis, Trump's language tends to focus on broad, populist themes and hypothetical scenarios. At the same time, Harris's discourse centers more on specific critiques of her opponent and emphasizes the presidential role. These patterns reflect the different positions of the candidates: Trump as the challenger seeking to differentiate himself and criticize the Biden administration and Kamala Harris as the incumbent democratic nominee to continue with Joe Biden's project. Table 2 displays the 20 total top-frequency words identified in the unigram, bigram, and trigram analyses of the second debate.

Table 2 shows the top-20-frequency words that can be used to detect the main topics that can be identified. On economic issues, both candidates placed a significant emphasis on economic matters, but with different approaches. Trump's repeated use of "billions" and "millions" suggests an attempt to emphasize the magnitude of his economic policies and their impact. The phrase "greatest economies history" focuses on past economic achievements. On the other hand, Harris emphasized "small businesses" and "affordable care" more specifically, suggesting an attempt to connect with middle-class voters and address healthcare concerns. The "sales tax" could indicate a focus on tax policy differences.

The divergence in the economic rhetoric aligns with the findings from [28], which argue that incumbents tend to focus on macro-level economic achievements, while challengers often emphasize the specific economic challenges voters face. Likewise, the candidates' language revealed different approaches to discussing leadership and governance. Harris's frequent references to "Donald Trump" and "former president" indicate a strategy of directly criticizing her opponent. The phrase "president united states" emphasizes the responsibilities and expectations of the role. Discussing leadership aligns with [2], the

functional theory of political campaign discourse, since, as mentioned before, challengers tend to attack the incumbent's record, while incumbents focus on their achievements.

Table 2. Trump–Harris top-frequency words from the unigram, bigram, and trigram analyses.

Trump					
1-g	**Count**	**2-g**	**Count**	**3-g**	**Count**
people	80	millions people	9	would never happened	4
going	50	would never	8	people pouring country	3
country	55	billions dollars	7	nobody thought possible	3
said	46	years ago	7	hundreds billions dollars	3
would	42	history country	7	millions millions people	3
get	40	united states	6	even come close	3
know	32	get vote	6	millions people come	3
like	32	student loans	6	every one cases	3
one	32	never happened	6	largest regulation cut	3
got	29	ever seen	6	14 million votes	3
never	27	destroying country	5	even know president	3
look	27	go back	5	would sitting Moscow	3
president	26	come country	5	president history country	3
years	25	millions millions	4	millions people pouring	2
go	25	ninth month	4	one greatest economies	2
done	22	trying get	4	greatest economies history	2
back	21	supreme court	4	know everybody else	2
world	21	going get	4	like four sentences	2
good	20	going going	4	going to higher	2
come	20	going to	4	going higher prices	2
Harris					
1-g	**count**	**2-g**	**count**	**3-g**	**count**
president	55	Donald trump	32	president united states	7
people	49	united states	20	affordable care act	7
trump	38	American people	18	Donald trump left	5
would	32	former president	11	trump left us	4
Donald	32	vice president	10	left us wors	4
let	26	president united	8	much common separates	3
American	26	affordable care	7	common separates us	3
said	25	care act	7	Donald trump plan	2
one	22	small businesses	6	let understand got	2
states	22	let talk	6	protections roe wade	2
united	21	well known	6	trump abortion bans	2
plan	20	health care	5	former president said	2
know	20	national security	5	Donald trump president	2
understand	19	people America	4	last four years	2
actually	17	trump left	4	dreams American people	2
America	17	left us	4	small businesses plan	2
us	17	going hear	4	trump sales tax	2
well	17	let understand	4	sales tax would	2
going	17	turn page	4	old tired playbook	2
former	16	let remember	4	American people want	2

The debate also touched on countless social problems. Trump's language suggested a focus on immigration and education: f_{Trump}("people", "pouring", "country") = 3 and f_{Trump}("student", "loans") = 6. The phrase "people pouring country" likely refers to immigration issues, while "student loans" indicates a discussion of education costs. On the other hand, Harris emphasized healthcare and women's rights: f_{Harris}("affordable", "care", "act") = 7, f_{Harris}("protections", "roe", "wade") = 2, and f_{Harris}("trump", "abortion", "bans") = 2. Harris's focus on the Affordable Care Act and abortion rights suggests an attempt to mobilize the democratic base and highlight differences with her opponent on these

issues. The divergence in the social issue focus reflects the broader ideological differences between the two parties, as the authors of [29] noted in their analysis of asymmetric politics in the United States.

Finally, both candidates employed distinct rhetorical strategies. Trump used language suggesting hypothetical scenarios and comparisons: f_{Trump}("would", "never") = 8, f_{Trump}("would", "never", "happened") = 4, and f_{Trump}("even", "come", "close") = 3. These phrases indicate a strategy of presenting counterfactuals and emphasizing the uniqueness of his presidency. In contrast, Harris employed language aimed at creating a connection with voters and turning the page: f_{Harris}("let", "understand") = 4, f_{Harris}("turn", "page") = 4, and f_{Harris}("American", "people", "want") = 2. These phrases suggest an attempt to empathize with voters and position herself as a change candidate.

The contrasting rhetorical strategies align with the analysis of presidential debates in [30], which highlights how candidates use language to construct their persona and connect with voters. The candidates' language differences can be interpreted through political communication theory. As the authors of [31] note, incumbents often focus on their achievements and broad national themes, while challengers tend to critique the current administration and emphasize their readiness for office. This pattern is evident in the word choices of Trump and Harris.

Furthermore, the frequent use of opponent language by Harris aligns with what [2] terms "functional theory" in political campaign communication, in which candidates often engage in acclaim (self-praise), attack (the criticism of opponents), and defense strategies. Harris's repeated references to "Donald Trump" suggest a strong emphasis on the attack function, while Trump's focus on "people" and "country" may represent more of an acclaimed strategy. These linguistic patterns provide understanding into the candidates' campaign strategies and the dynamics of the 2024 presidential race.

4. Sentiment Analysis of Debates and Semantic Similarity

The sentiment analysis implementation combines (Robustly Optimized BERT Pretraining Approach) RoBERTa-based deep learning with contextual embeddings from Bidirectional Encoder Representations from Transformers (BERT) to provide semantic insights through contextual embeddings of the political discourse. Following [5], the sentiment score for each speech segment was calculated using

$$S(T) = \alpha \sum_{i=1}^{n} R(w_i) + \beta \sum_{j=1}^{m} C(p_j) \quad (1)$$

where $R(w_i)$ represents the RoBERTa sentiment score for word w_i, $C(p_j)$ represents the contextual modifier for phrase p_j, and α and β are weighting parameters optimized for political discourse. To capture semantic relationships between the speakers' statements, BERT embeddings, with a cosine similarity analysis, were implemented. For each speech segment, T, they generated a contextual embedding vector, $E(T)$, using BERT:

$$E(T) = \text{BERT}(T) \in R^d \quad (2)$$

where $d = 768$ is the dimension of BERT's hidden states. The semantic similarity between any two segments, T_i and T_j, was then computed using cosine similarity:

$$sim(T_i, T_j) = \frac{E(T_i) \cdot E(T_j)}{||E(T_i)|| \cdot ||E(T_j)||} \quad (3)$$

This approach revealed significant patterns in the rhetorical strategy. Trump maintained a high semantic consistency across the debates (average self-similarity: 0.82), while showing a lower similarity with opponents (Trump–Biden: 0.43; Trump–Harris: 0.39). These findings align with previous research on political discourse consistency, as in [32].

Figure 3 displays the temporal evolution of the sentiment scores during the first presidential debate between Trump and Biden. The analysis reveals distinct patterns in the rhetorical strategy, with Trump exhibiting a greater sentiment volatility (range: 0.50–0.95) compared to Biden's more measured approach (range: 0.50–0.85). Trump's discourse shows frequent peaks and troughs, indicating rapid shifts between positive and negative sentiments, while Biden maintains a more consistent emotional tenor throughout the debate.

Figure 3. Trump and Biden sentiment speech over time. Source: Authors' own elaboration.

Figure 4 presents the sentiment analysis from the second debate between Trump and Harris. Trump's pattern of high sentiment volatility persists (range: 0.45–1.0), showing even greater amplitude than in the first debate. In contrast, Harris demonstrates a more controlled emotional range (range: 0.50–0.90), with fewer extreme fluctuations. This pattern aligns with previous research on differences in the political discourse, as in [15], while also reflecting the candidates' distinct rhetorical strategies.

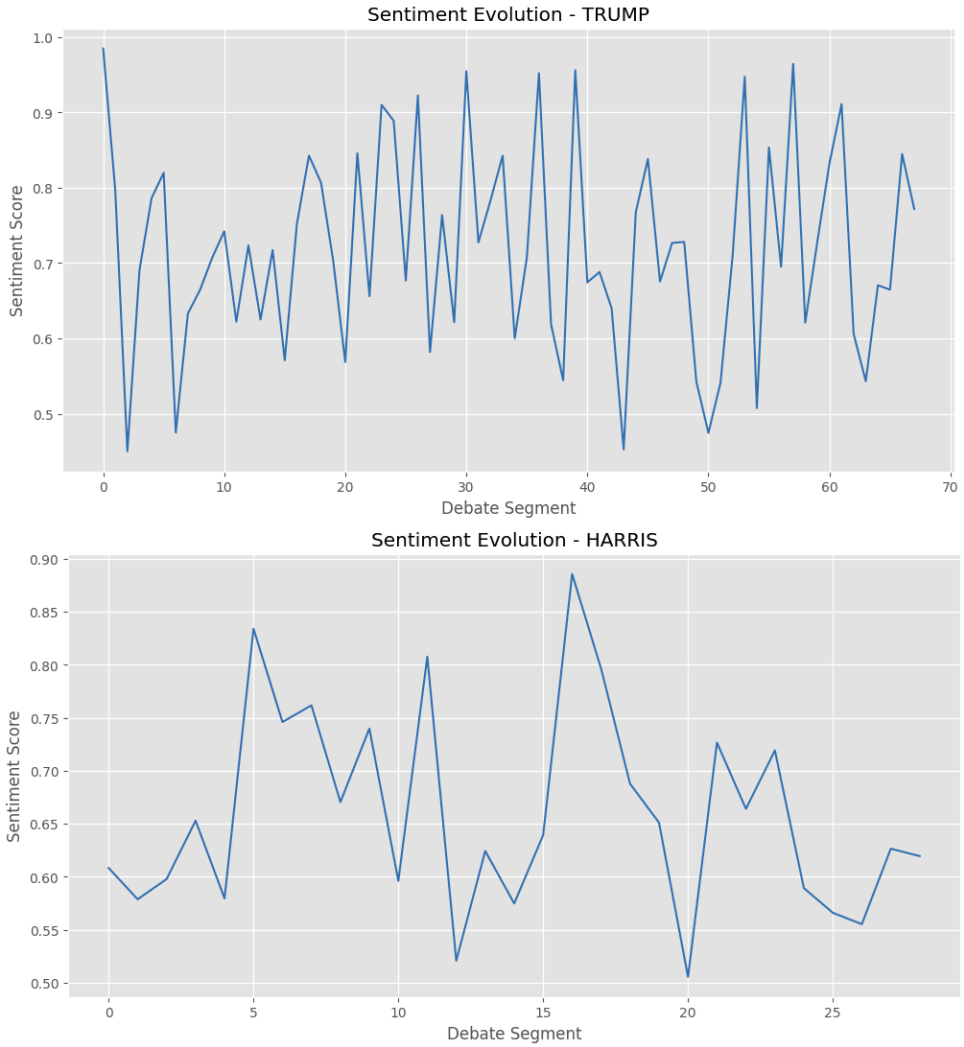

Figure 4. Trump and Harris sentiment speech over time. Source: Authors' own elaboration.

The comparative analysis of Figures 3 and 4 reveals several key findings:

a. Sentiment stability: Trump's mean sentiment scores remained relatively stable across both debates (0.749 to 0.719), despite increased volatility in the second debate.
b. Opponent adaptation: both of the democratic candidates maintained lower sentiment volatility than Trump (Biden: σ = 0.08; Harris: σ = 0.09; Trump: σ = 0.14), suggesting a deliberate strategy to project stability and measured leadership.
c. Temporal patterns: all of the candidates showed distinct temporal patterns in their sentiment evolution, with key inflection points often corresponding to significant debate topics such as economic policy, healthcare, and foreign relations.

The sentiment analysis reveals distinct patterns (Figures 3 and 4) that were not captured by using the bag of words. Trump's sentiment volatility (σT = 0.14) exceeded both Biden's (σB = 0.08) and Harris's (σH = 0.09), suggesting a more dynamic rhetorical strategy. The normalized sentiment range (R) for each candidate can be expressed as

$$R_c = \frac{\max(S_c) - \min(S_c)}{\text{mean}(S_c)} \quad (4)$$

where c represents the candidate. This leads to

- Trump: R_{Trump} = 0.67 (first debate), 0.71 (second debate).
- Biden: R_{Biden} = 0.41.

- Harris: R_{Harris} = 0.48.

The implementation of RoBERTa revealed significant changes in the rhetorical strategy among the debates. Trump's average sentiment remained relatively stable (0.749 to 0.719), while his segment count increased substantially (42 to 68), indicating a more aggressive debate strategy. The semantic coherence between the debates, measured using BERT embeddings, can be expressed as

$$C(D_1, D_2) = \frac{1}{n}\sum_{i=1}^{n} \max_j sim\left(T_i^1, T_j^2\right) \tag{5}$$

where T_i^1 represents segments from the first debate and T_j^2 from the second, revealing a stronger thematic consistency in Trump's rhetoric (C = 0.76) compared to the democratic candidates (C = 0.58). The emotional range of each candidate, calculated using RoBERTa's fine-grained sentiment classification, showed distinct patterns:

$$E_c = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(S_i - \overline{S}\right)^2} \tag{6}$$

where E_c represents the emotional range for candidate c, S_i represents individual sentiment scores, and $\overline{S}$ is the mean sentiment. In this sense, the analysis reveals the following

- Trump: E_{Trump} = 0.31 (higher volatility).
- Biden: E_{Biden} = 0.19 (more measured).
- Harris: R_{Harris} = 0.23 (moderate volatility).

These findings demonstrate the effectiveness of combining advanced NLP techniques in analyzing political discourse, providing quantitative support for qualitative observations about candidates' rhetorical styles.

5. Empirical Results and Discussion

5.1. Jaccard Similarity Coefficient

A Jaccard similarity analysis was implemented to understand the linguistic similarities and differences between the two debates. The Jaccard similarity coefficient, represented as $J(A, B)$ is a statistical measure used to assess the similarity and dissimilarity of sample sets. For two sets, A and B, it is calculated as the size of the intersection divided by the size of the union of the sample sets.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

The Jaccard similarity coefficients for Donald Trump's speeches across the two debates were as follows:

- Unigram Jaccard similarity: 0.3430.
- Bigram Jaccard similarity: 0.0494.
- Trigram Jaccard similarity: 0.0101.

The Jaccard similarity coefficients quantitatively measure the overlap in Trump's vocabulary and phraseology between the two debates. A higher coefficient indicates more similarity, with 1.0 representing identical sets and 0.0 representing no overlap. Likewise, a unigram similarity of (0.3430) suggests moderate consistency in individual word choice across the debates, showing that while Trump's core vocabulary is stable, there is also significant variation, possibly reflecting differences in the debate topics or the strategies employed against different opponents.

The bigram (0.0494) and trigram (0.0101) similarities are notably lower, indicating substantial differences in phrase usage between the two debates. This sharp decrease in similarity, along with the transition from unigrams to bigrams and trigrams, is consistent with findings in other political discourse studies, where longer n-grams tend to show more significant variation due to their sensitivity to context and topic shifts [21]. These results suggest that while Trump maintained some consistency in his overall vocabulary, the specific combinations of words and phrases varied significantly between debates. Some of the reasons are

- Different debate opponents (Biden vs. Harris) necessitate different rhetorical strategies.
- Shifts in debate topics or focus areas between the two events.
- Changes in the political landscape or campaign strategy in the time between debates.

5.2. Type–Token Ratio

In addition, the Type–Token Ratio (TTR), which is a quantitative measure of lexical diversity in a text, showing the richness and variety of vocabulary used by speakers. The TTR is calculated as

$$TTR = \frac{V}{N} \tag{8}$$

where V is the number of unique words (types) in the text, and N is the total number of words (tokens). For this ratio, the full speech must be included, meaning that stopwords must be reincorporated. Finally, Ref. [33] states that a higher TTR indicates greater lexical diversity, suggesting a more varied vocabulary use. The TTR result is presented in Table 3.

Table 3. TTR lexical diversity.

Frist Debate	**Trump**	**Biden**	**Second Debate**	**Trump**	**Harris**
Total words	9957	8379	Total words	9670	6701
TTR	0.1168	0.1421	TTR	0.1234	0.1876

In both debates, Donald Trump consistently used more words than his opponents due to his dominant speaking presence, which could be attributed to various factors such as debate strategy, speaking style, or potential interruptions and overtalk. In addition, Trump's TTR increased slightly from the first debate (0.1168) to the second (0.1234), indicating a marginal improvement in linguistic diversity. This slight increase suggests relatively consistent vocabulary usage across both debates, with only a minor expansion in the variety of words used.

Both of Trump's opponents demonstrated higher TTR values than Trump in their respective debates. Biden's TTR (0.1421) and Harris' TTR (0.1876) were higher than Trump's in each discussion, suggesting that both Biden and Harris employed a more diverse vocabulary in their responses, which is an unexpected result considering Biden's criticism in the first debate.

For the second debate, Kamala Harris exhibited the highest TTR (0.1876) among all the participants, despite using the fewest total words (6701). Harris employed the most diverse vocabulary relative to the length of her speech and a more concise, yet varied, communication style. The difference in the TTR between Trump and his opponents was more pronounced in the second debate (0.0642) than in the first (0.0253). This more significant gap could be attributed to Harris's high lexical diversity or a shift in debate topics that allowed more varied vocabulary use.

5.3. Measure of Textual Lexical Diversity

In contrast to the TTR, the Measure of Textual Lexical Diversity (MTLD) analysis provides a more robust assessment of lexical sophistication, addressing the length sensitivity limitations of the traditional TTR; the MTLD score is calculated by evaluating the TTR sequentially through the text until it reaches a standard factor size. Following [34], the total of the factors can be expressed as

$$Factor = \sum_{i=1}^{n} F_i + \frac{1 - TTR_{partial}}{1 - threshold} \tag{9}$$

where F_i represents the complete factors (segments where the TTR drops to the threshold), and $TTR_{partial}$ is the TTR of the incomplete segment *threshold* (the standard TTR cutoff 0.72). The implementation involves a bidirectional calculation—forward and backward through the text—with the final MTLD score being the mean of both directions:

$$MTLD_{final} = \frac{MTLD_{forward} + MTLD_{backward}}{2} \tag{10}$$

This bidirectional approach, as noted in [34], helps mitigate any potential sequence effects in the text. The empirical application of this methodology to the debate corpus reveals distinct patterns in lexical deployment. For instance, the higher MTLD scores achieved by Biden (28.62) and Harris (37.86) compared to Trump's scores (20.87 and 26.68, respectively) indicate not just a greater lexical diversity, but specifically a more sustained deployment of varied vocabulary throughout their discourse.

This computational approach provides several advantages over traditional TTR measures by maintaining sensitivity to text internal variation while controlling length effects. Likewise, MTLD captures the length and structure of candidates' speeches. For instance, the democratic candidates consistently demonstrated higher sustained vocabulary variation, despite using fewer total words.

5.4. Discussion

The empirical findings from our sentiment analysis, Jaccard similarity coefficient, and the TTR and MTLD approaches contribute to the broader theoretical understanding of the political discourse in several ways. First, the results obtained align with the findings of [15] regarding the distinct rhetorical patterns between Trump and Harris, though the present paper provides additional quantitative support through lexical diversity measurements.

The sentiment volatility observed in Trump's discourse (σ_{Trump} = 0.14) compared to his opponents (σ_{Biden} = 0.08, σ_{Harris} = 0.09) suggests a deliberate rhetorical strategy rather than merely stylistic differences. This pattern aligns with the concept provided by [16] of contextual disambiguation, where sentiment shifts serve specific communicative purposes in political discourse. The higher TTR values for the democratic candidates (Biden: 0.1421; Harris: 0.1876) than Trump (0.1168, 0.1234) challenge traditional assumptions about the relationship between vocabulary diversity and perceived debate effectiveness.

These findings suggest several theoretical implications:

(a) Rhetorical adaptation: the variation in the sentiment patterns between the debates indicates that candidates adapt their rhetorical strategies based on their opponents, supporting the dynamic nature of political discourse.

(b) Lexical sophistication: the unexpected higher lexical diversity among the democratic candidates suggests that the public perception of debate performance may be more influenced by delivery and timing than vocabulary range.

(c) Strategic consistency: the Jaccard similarity analysis reveals how candidates maintain core messaging while adapting to different debate contexts, supporting theories of strategic political communication.

As the authors of Ref. [17] suggest, future research directions could explore the integration of multimodal analyses incorporating non-verbal cues and social media interactions, as well as the application of advanced topic modeling techniques like Latent Dirichlet Allocation (LDA); however, as expressed in the Conclusions Section, this will be the subject of a subsequent study.

6. Conclusions

This study analyzed the linguistic patterns and rhetorical strategies used in the 2024 U.S. presidential debates, explicitly focusing on the interactions between Donald Trump, Joe Biden, and Kamala Harris. By utilizing NLP techniques, such as an n-gram analysis, sentiment analysis, and measurements of lexical diversity, this research has provided features for the candidates' communication styles and strategies, aligning with and extending the recent work in computational political discourse analysis in [15].

The analysis revealed distinct linguistic profiles for each candidate. Donald Trump consistently employed more emotionally charged language with a high sentiment volatility, spanning the full range of the sentiment scale (-1 to $+1$) in both debates. It is worth mentioning that this aligns with previous research on Trump's rhetorical style, which is characterized by extreme language and stark contrasts [4]. In contrast, Joe Biden and Kamala Harris demonstrated more measured approaches, with Biden avoiding the extremes of the sentiment scale and Harris showing a slightly more expansive, but still more moderate, range than Trump. These findings align with the observations in [16] about the importance of context in sentiment disambiguation within political discourse.

The empirical findings also highlighted significant differences in the lexical diversity among the candidates. Contrary to expectations, particularly given criticisms of Biden's debate performance, both Biden and Harris demonstrated a higher TTR and MTLD than Trump, indicating a diverse vocabulary usage. This result advises that perceived debate performance may only sometimes correlate with lexical diversity and points to the complex nature of effective political communication, as noted by [17] in their analysis of political communication dynamics.

The Jaccard similarity analysis of Trump's language across the two debates revealed moderate consistency in individual word choice but substantial differences in phrase usage. This shows the adaptability of political rhetoric to different opponents and changing campaign contexts while suggesting some stability in core messaging. This finding contributes to our understanding of rhetorical consistency in political discourse, particularly in high-stakes debates.

Several limitations and opportunities for future research emerged from this study. While our focus on textual analyses provided valuable insights, it did not capture non-verbal aspects of communication, which the authors of [18] demonstrate, that can be vital in debate performance. Future research could address these limitations through the integration of advanced topic modeling techniques like LDA to expose more subtle thematic patterns. Another future work could focus on the development of political discourse–specific sentiment analysis models, incorporating domain-adapted transformers, or the implementation of multimodal analysis frameworks, incorporating non-verbal cues and social media interactions.

This study contributes to the growing body of research applying computational methods to political discourse analyses. By providing quantitative visions into the linguistic and rhetorical dimensions of presidential debates, offering a valuable complement to traditional

qualitative analyses of political communication. Moreover, the proposed methodological framework demonstrated how NLP techniques can be effectively applied to analyze political discourse, opening new avenues for research in this field. As political landscapes continue to evolve and new computational tools emerge, such data-driven approaches will likely play an increasingly important role in understanding and interpreting the dynamics of political discourse.

This research establishes a foundation for future studies that employs more sophisticated analytical techniques while maintaining the focus on understanding how language shapes political communication and public perception in presidential debates. Finally, it is important to highlight that debates serve as platforms for candidates to present their policies and positions, influencing public sentiment and perceptions of leadership. In this sense, the present study has relevance in understanding economic and social expectations. But, as suggested by contemporary researchers in the field, the integration of additional computational methods and multimodal analysis approaches could further improve our understanding of these crucial political events.

Author Contributions: Conceptualization, data gathering, simulations, numerical tests, methodology, formal analysis, investigation, writing—original draft preparation, and writing—review and editing, A.L.J.-P., S.C.-A., J.Á.-G. and F.V.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The debate transcriptions are public, and the test results are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Jamieson, K.H.; Birdsell, D.S. *Presidential Debates: The Challenge of Creating an Informed Electorate*; Oxford University Press: Oxford, UK, 1988.
2. Benoit, W. *Communication in Political Campaigns*; Peter Lang: New York, NY, USA, 2007.
3. Savoy, J. Trump's and Clinton's Style and Rhetoric during the 2016 Presidential Election. *J. Quant. Linguist.* **2018**, *25*, 168–189. [CrossRef]
4. Ahmadian, S.; Azarshahi, S.; Paulhus, D.L. Explaining Donald Trump via communication style: Grandiosity, informality, and dynamism. *Personal. Individ. Differ.* **2017**, *107*, 49–53. [CrossRef]
5. Liu, B. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*; Cambridge University Press: Cambridge, UK, 2015.
6. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: Cambridge, UK, 2008.
7. Graber, D.A.; Smith, J.M. Political Communication Faces the 21st Century. *J. Commun.* **2005**, *55*, 479–507. [CrossRef]
8. McNair, B. *An Introduction to Political Communication*; Routledge: New York, NY, USA, 2017.
9. Benoit, W.L. *Political Election Debates: Informing Voters About Policy and Character*; Lexington Books: Lanham, MD, USA, 2013.
10. Kreis, R. The "Tweet Politics" of President Trump. *J. Lang. Politics* **2017**, *16*, 607–618. [CrossRef]
11. Liu, Z.; Li, Y.; Zolotarevych, O.; Yang, R.; Liu, T. LLM-POTUS Score: A Framework of Analyzing Presidential Debates with Large Language Models. *arXiv* **2024**. [CrossRef]
12. Chulvi, B.; Molpeceres, M.; Rodrigo, M.F.; Toselli, A.H.; Rosso, P. Politicization of Immigration and Language Use in Political Elites: A Study of Spanish Parliamentary Speeches. *J. Lang. Soc. Psychol.* **2024**, *43*, 164–194. [CrossRef]
13. Rudkowsky, E.; Haselmayer, M.; Wastian, M.; Jenny, M.; Emrich, Š.; Sedlmair, M. More than Bags of Words: Sentiment Analysis with Word Embeddings. *Commun. Methods Meas.* **2018**, *12*, 140–157. [CrossRef]
14. Albalawi, R.; Yeap, T.H.; Benyoucef, M. Using Topic Modeling Methods for Short-Text Data: A Comparative Analysis. *Front. Artif. Intell.* **2020**, *3*, 42. [CrossRef] [PubMed]
15. Wicke, P.; Bolognesi, M.M. Red and Blue Language: Word Choices in the Trump and Harris 2024 Presidential Debate. *arXiv* **2024**, arXiv:2410.13654. [CrossRef]

16. Rajabi, Z.; Valavi, M.R.; Hourali, M. A Context-Based Disambiguation Model for Sentiment Concepts Using a Bag-of-Concepts Approach. *Cogn. Comput.* **2020**, *12*, 1299–1312. [CrossRef]
17. Verdegem, P.; D'heer, E. Social Media Logic and Its Impact on Political Communication During Election Times. In *Managing Democracy in the Digital Age*; Schwanholz, J., Graham, T., Stoll, P.T., Eds.; Springer: Cham, Switzerland, 2018; pp. 119–135. [CrossRef]
18. Shah, D.V.; Hanna, A.; Bucy, E.P.; Wells, C.; Quevedo, V. The Power of Television Images in a Social Media Age: Linking Biobehavioral and Computational Approaches via the Second Screen. *Ann. Am. Acad. Polit. Soc. Sci.* **2015**, *659*, 225–245. [CrossRef]
19. CNN. CNN Presidential Debate. READ: Biden-Trump Debate Transcript. 2024. Available online: https://edition.cnn.com/2024/06/27/politics/read-biden-trump-debate-rush-transcript/index.html (accessed on 28 June 2024).
20. ABC News. Your Voice Your Vote 2024. READ: Harris-Trump Presidential Debate Transcript. 2024. Available online: https://abcnews.go.com/Politics/harris-trump-presidential-debate-transcript/story?id=113560542 (accessed on 30th September 2024).
21. Savoy, J. Lexical Analysis of US Political Speeches. *J. Quant. Linguist.* **2010**, *17*, 123–141. [CrossRef]
22. Rajaraman, A.; Ullman, J.D. *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2011.
23. Berry, M.W. *Text Mining: Applications and Theory*; John Wiley & Sons: Chichester, UK, 2010.
24. Silge, J.; Robinson, D. *Text Mining with R: A Tidy Approach*; O'Reilly Media: Sebastopol, CA, USA, 2017.
25. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 2009.
26. Tan, C.M.; Wang, Y.F.; Lee, C.D. The Use of Bigrams to Enhance Text Categorization. *Inf. Process. Manag.* **2020**, *30*, 529–546. [CrossRef]
27. Wagner-Pacifici, R. Anticharismatic Authority: Joe Biden's Approximation of the Ideal Type. *Politics Soc.* **2024**, *52*, 241–267. [CrossRef]
28. Vavreck, L. *The Message Matters: The Economy and Presidential Campaigns*; Princeton University Press: Princeton, NJ, USA, 2009.
29. Grossmann, M.; Hopkins, D.A. *Asymmetric Politics: Ideological Republicans and Group Interest Democrats*; Oxford University Press: Oxford, UK, 2016.
30. *Jamieson, K.H. Eloquence in an Electronic Age: The Transformation of Political Speechmaking*; Oxford University Press: Oxford, UK, 1988.
31. Kenski, K.; Jamieson, K.H. *The Oxford Handbook of Political Communication*; Oxford University Press: Oxford, UK, 2017.
32. Sclafani, J. *Talking Donald Trump: A Sociolinguistic Study of Style, Metadiscourse, and Political Identity*; Routledge: New York, NY, USA, 2017. [CrossRef]
33. Johansson, V. *Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective*; Working Papers; Lund University, Department of Linguistics and Phonetics: Lund, Sweden, 2008; Volume 53, pp. 61–79.
34. McCarthy, P.M.; Jarvis, S. MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behav. Res. Methods* **2010**, *42*, 381–392. [CrossRef] [PubMed]