Identifying Fine-grained Forms of Populism in Political Discourse: A Case Study on Donald Trump's Presidential Campaigns

Ilias Chalkidis[α] **Stephanie Brandl**[β] **Paris Aslanidis**[γ]
[α] Department of Computer Science, University of Copenhagen, Denmark
[β] Copenhagen Center for Social Data Science, University of Copenhagen, Denmark
[γ] National Centre for Social Research (EKKE), Greece
ilias.chalkidis[at]di.ku.dk stephanie.brandl[at]sodas.ku.dk paslanidis[at]ekke.gr

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of instruction-following tasks, yet their grasp of nuanced social science concepts remains underexplored. This paper examines whether LLMs can identify and classify fine-grained forms of *populism*, a complex and contested concept in both academic and media debates. To this end, we curate and release novel datasets specifically designed to capture populist discourse. We evaluate a range of pre-trained (large) language models, both open-weight and proprietary, across multiple prompting paradigms. Our analysis reveals notable variation in performance, highlighting the limitations of LLMs in detecting populist discourse. We find that a fine-tuned RoBERTa classifier vastly outperforms all new-era instruction-tuned LLMs, unless fine-tuned. Additionally, we apply our best-performing model to analyze campaign speeches by Donald Trump, extracting valuable insights into his strategic use of populist rhetoric. Finally, we assess the generalizability of these models by benchmarking them on campaign speeches by European politicians, offering a lens into cross-context transferability in political discourse analysis. In this setting, we find that instruction-tuned LLMs exhibit greater robustness on out-of-domain data.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Llama Team, 2024; Gemini Team, 2025) exhibit unprecedented capabilities, excelling in a plethora of instruction-following tasks such as open-domain question answering, solving arithmetic and coding problems, engaging in creative writing, and more (Bommasani et al., 2023; Chiang et al., 2024).

However, we know little about how well LLMs perform on tasks involving social science concepts that researchers use to investigate sociopolitical phenomena. In this paper, we explore whether

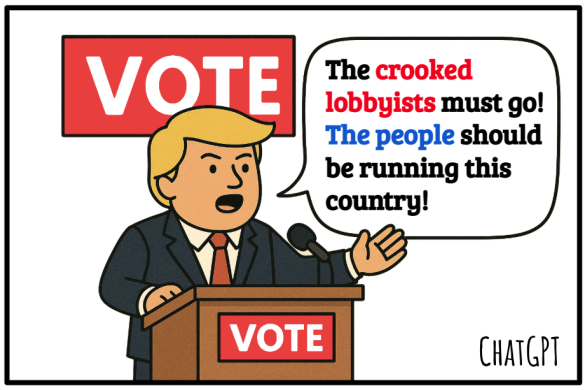

Figure 1: Depiction of populist political discourse under our working definition, with **negative** invocations towards *elites*, and **positive** ones towards the *people*.

LLMs can adequately understand the concept of *populism* (Figure 1) and detect its presence in political discourse. Our goal is to enable the analysis of large-scale corpora and extract valuable insights into how politicians employ populist language.

Working Definition – What is populism? In our work, we follow *discursive approaches* (Laclau, 2005; Stavrakakis, 2024), where populism leverages the principle of popular sovereignty to emphasize the idea of the common "people" and its opposition to various (political, economical, etc.) elite groups that game the system to serve their own ends (Aslanidis, 2024).[1] Hence, there are two core elements in identifying populism in political discourse: (i) people-centrism, i.e., positive invocations of the "people", and (ii) anti-elitism, i.e., negative invocations of "elites" (Figures 1-2).

Scope of our Work The scope of our work is to address the challenge of detecting nuanced manifestations of populism by leveraging contemporary Natural Language Processing (NLP) techniques to analyze large-scale textual corpora. By minimizing the limitations of manual annotation and

[1] We present background on the concept of populism, and alternative approaches on populism in Appendix D.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 919–936

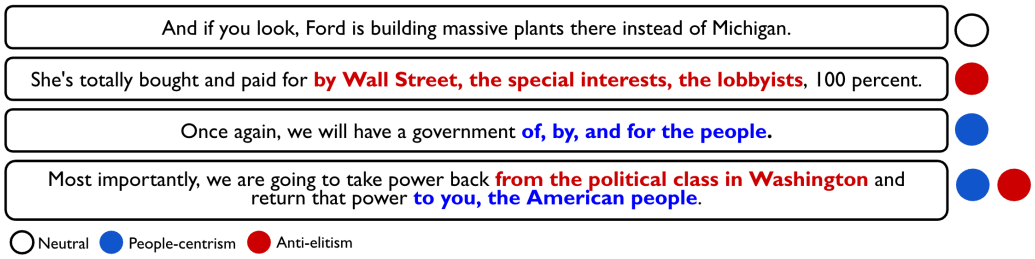

Figure 2: Color-coded examples from the TRUMP-2016 dataset (Section 2.2) with an example per class (neutral, people-centrism, anti-elitism, both). Relevant parts of the sentence are highlighted in red or blue.

observational bias, our approach enables systematic identification of discursive patterns associated with populism. As a case study in contemporary politics, we examine the populist content of Donald Trump's political rhetoric from 2016 to 2025, with particular attention to his campaigns in the 2016, 2020, and 2024 U.S. presidential elections.

Contributions We curate and release 3 new datasets (Section 2.2) that we use to train and benchmark a variety of pre-trained (large) language models, including both open-weight and proprietary models in various settings (Section 3-4). We utilize our best-performing model to automatically classify and analyze hundreds of Donald Trump's speeches over a decade, gaining valuable insights into how he instrumentalizes populism as part of his discourse (Section 5). Lastly, we benchmark the best-performing models on speeches from European politicians' campaigns to examine the potential of transfer learning on out-of-domain data and potential performance discrepancies (Section 6).

2 Task & Datasets

2.1 Task Definition

The goal of this task is to develop a method to identify fine-grained forms of populist discourse. As mentioned earlier, under our working definition, populism has two main components: *people-centrism* and *anti-elitism* that need to operate in tandem before we can label a discursive body, e.g., a political speech or a social media post, as a populist one. To enable a fine-grained analysis of political content and take advantage of higher-frequency data for examining variations in populist discourse, we move beyond full speeches or paragraphs as units of analysis and instead focus on a lower-level discursive structure: the sentence. We frame the task as a three-way multi-label classification, where each sentence can be labeled as neutral, anti-elitist, or people-centric. The co-occurrence of the latter two labels indicates a *fully populist* sentence, see the last sentence in Figure 2 for an example.

Dataset Name	#S	#I	Labeled
TRUMP-2016	70	∼15K	✔
TRUMP-CHRONOS	713	∼656K	✘
EU-OOD	5	∼1.8K	✔

Table 1: Datasets released as part of our study. #S refers to number of speeches, and #I to numbers of instances.

2.2 Datasets

As part of this study, we curate and release three datasets for the identification of fine-grained forms of populist discourse as described in Table 1. We release our datasets and code to support the replication of our results and to facilitate further experimentation with alternative methods or corpora.[2][3]

Donald Trump (2016) We curate and release a dataset of 70 presidential campaign speeches by Donald Trump during the Republican Party primaries and the 2016 presidential campaign, dubbed TRUMP-2016. The speech transcripts, collected from UC Santa Barbara's American Presidency Project,[4] range from June 2016 to January 2017. All transcripts were manually curated to correct transcription errors, remove audience interventions from these rallies, and eliminate remarks by guest speakers or third parties. We also include the *date* and *location* as relevant metadata.

The speeches in this set were split by sentences and annotated based on the 3-class labeling defined in Section 2.1 by 4 undergraduate Political Science students, under the supervision of our domain expert.[5] We use this dataset for training and evaluation, splitting it chronologically into training (56 speeches) and test (14 speeches) subsets.

[2] https://huggingface.co/collections/coastalcph/identifying-populist-rhetoric

[3] https://github.com/coastalcph/populism-llms

[4] https://www.presidency.ucsb.edu/

[5] See Appendix C for details on the annotation process.

Category		#Instances	
Neutral	(N)	13,910	(92.6%)
Anti-elitism	(AE)	826	(5.5%)
People-centrism	(PC)	517	(3.4%)

Table 2: Distribution of the 3 categories (classes) in the TRUMP-2016 dataset. Some sentences are labeled with both positive categories (AE, PC).

Table 2 presents the label distribution. Under our strict definition and when annotating at the sentence level, populist invocations appear relatively sparse in political discourse, even in the case of a characteristically populist politician. However, a qualitative examination can consider them as providing the core sentiment (tone) to a whole speech.

Donald Trump (2015-2025) We also curate and release an additional, much larger collection of 713 non-annotated speeches by Donald Trump, ranging over a decade from June 2015 to April 2025, dubbed TRUMP-CHRONOS, for a total of 656,136 sentences. This additional corpus is used to derive insights on the historical trends (2015-2025) of populist discourse in Donald Trump's campaign speeches. After consulting online sources to produce a complete list of Donald Trump rallies in this period, we sourced the transcripts of a subset of the speeches from UC Santa Barbara's American Presidency Project collection and used the publicly available version of OpenAI's Whisper (Radford et al., 2023) to auto-transcribe the remainder by extracting audio from the videos of the speeches that we found on YouTube, C-SPAN and other online platforms. All transcripts were curated manually, similarly to the TRUMP-2016 dataset, and they reflect the actual speech addressed to the audience, rather than the prepared remarks that campaign managers routinely hand over to the press before or after a campaign event. We also include the *date*, *location*, and *campaign period* as metadata.

European Campaign Speeches In addition, we curate and annotate a collection of six speeches by five selected European leaders who recently ran for office and are widely considered to have relied on populist discourse during their campaigns. These are: Marine Le Pen (Rassemblement National, France), Alice Weidel (AfD, Germany), Herbert Kickl (FPÖ, Austria), Éric Zemmour (Reconquête, France), and Alexis Tsipras (SYRIZA, Greece). The first four politicians are identified as right-wing populists, while the last one is identified as a left-wing populist. We aim to use these speeches as Out-of-Domain (OOD) data to evaluate the robustness of the examined models-dubbed EU-OOD. Since the speeches are in the speaker's native language (French, German, and Greek), we auto-translate them to English using DeepL API.[6]

3 Experiments

3.1 Methods

Baseline We first consider a linear SVM (Cortes and Vapnik, 1995) classifier using TF-IDF scores as features for 10K n-grams, where $n \in [1, 2, 3]$.[7]

PLMs We evaluate Pre-trained Language Models (PLMs), also known as BERT-like (Devlin et al., 2019) encoder-only models. We use the large (approx. 350-400M parameters) versions of RoBERTa (Liu et al., 2019) and DeBERTa-V3 (He et al., 2023). RoBERTa has initially been pre-trained with the Masked Language Modeling (MLM) objective, while DeBERTa-V3 uses an ELECTRA-like, Replaced Token Detection (RTD) objective, on large web corpora. We fine-tune both PLMs as 3-way multi-label classifiers using a shallow MLP on top of the final sentence representation.[7]

Open-weight LLMs We also consider new-era LLMs. We use RLHF'd instruction-tuned versions of LLMs that exhibit strong performance on diverse NLP tasks in a zero-shot setting. We benchmark a variety of publicly released open-weight models, specifically: Meta's Llama 3.1 (8/70B) (Llama Team, 2024), Google's Gemma 3 (12/27B) (Gemma Team, 2025), and Alibaba's Qwen 3 (8/14/32B) (Qwen Team, 2025).[8] We evaluate all models in a zero-shot setting, while we also present results fine-tuning Llama 3.1 (8B) and Qwen (8/14B) using a LoRA (Hu et al., 2022; Dettmers et al., 2023) adapter. The base prompt we use is available in Table 11, dubbed as *Base*.[7]

Proprietary LLMs Finally, we consider two recent top-tier closed-source proprietary LLMs. We use OpenAI's GPT-4.1 -version 04/2025- (OpenAI Team, 2025) and Google's Gemini 2.5 Flash -version 05/2025- (Gemini Team, 2025) as a service (inference via their API).

[6] https://www.deepl.com/en/pro-api

[7] We present additional details in Appendix A.

[8] We use Qwen 3 models in non-thinking mode.

Group	Model Name	Size	FT	Neutral	Anti-Elitism	People-centrism	Macro F1
BASELINE	TFIDF-SVM	30K	✔	.974 ± -	.531 ± -	.385 ± -	.630 ± -
PRE-TRAINED LMS	RoBERTa	355M	✔	**.976** ± -	.661 ± -	**.602** ± -	**.746** ± -
	DeBERTa	418M	✔	.971 ± -	.639 ± -	.583 ± -	.731 ± -
OPEN-WEIGHT LLMS	Llama 3.1	8B	✘	.814 ± .011	.320 ± .005	.245 ± .009	.460 ± .008
			✔	.965 ± -	.646 ± -	.497 ± -	.703 ± -
		70B	✘	.850 ± .010	.439 ± .009	.231 ± .006	.507 ± .008
	Gemma 3	12B	✘	.813 ± .008	.318 ± .002	.220 ± .009	.437 ± .000
		27B	✘	.819 ± .009	.365 ± .007	.232 ± .006	.472 ± .007
	Qwen 3	8B	✘	.864 ± .004	.368 ± .002	.286 ± .005	.506 ± .004
			✔	.967 ± -	.621 ± -	.553 ± -	.714 ± -
		14B	✘	.916 ± .001	.491 ± .000	.340 ± .005	.582 ± .002
			✔	.972 ± -	**.674** ± -	.575 ± -	.740 ± -
		32B	✘	.875 ± .001	.458 ± .000	.292 ± .001	.542 ± .000
PROPRIETARY LLMS	GPT 4.1	n/a	✘	.850 ± -	.461 ± -	.201 ± -	.503 ± -
	Gemini 2.5 Flash	n/a	✘	.903 ± -	.513 ± -	.306 ± -	.574 ± -

Table 3: Classification results across all models and classes. We present F1-scores (mean ± std),[9] best-performing performance per class in bold, and the runner-up underlined. The second column indicates the size of the model counted in total parameters. The third column (FT) indicates task-related fine-tuning with the TRUMP-2016 dataset.

3.2 Main Results

In Table 3, we present the main results of our experiments using F1-score on the test subset of Donald Trump's speeches (TRUMP-2016). For each model, we report the mean and standard deviation over multiple runs.[9] The main observation is that the examined task proves to be very challenging for all examined methods, with the best-performing method, the fine-tuned RoBERTa classifier, demonstrating a macro-F1 score of 0.746. This finding aligns with existing work on coding populist discourse (Bonikowski et al., 2022) that has adequately documented the complex nature of any task that involves consistently identifying populist tropes in political discourse. We further observe that identifying people-centrism is harder compared to anti-elitism; with a consistent performance discrepancy between the two populist categories across all models.

The fine-tuned RoBERTa and DeBERTa classifiers perform almost on par, substantially outperforming LLMs in a zero-shot setting (approx. +20-30 p.p.). We present additional details and experiments in different settings, i.e., data augmentation, and models with context, in Appendices A-B.

Open-weight LLMs struggle to perform in the zero-shot setting, which include a brief concept definition and task description in the prompt.[9][10] The best-performing open-weight model, Qwen (14B), achieves a macro-F1 of 0.582. Overall, we observe that the Qwen 3 models substantially outperform and are also more robust (low±std), compared to Llama 3.1 and Gemma 3 models. Similarly, top-notch proprietary LLMs demonstrate poor performance. The best-performing model, Gemini 2.5 Flash, achieves a macro-F1 of 0.574, scoring just below the best-performing open-weight LLM.

To render the comparison more comprehensive, we also fine-tune Llama 3.1 (8B) and Qwen 3 (8B, 14B) using LoRA adapters. Fine-tuning leads to substantial performance improvement, by approximately 20-25 p.p., although still substantially below the fine-tuned PLMs, except for Qwen 3 (14B).

4 Analyses

Given the poor zero-shot performance of LLMs, we experiment with more elaborate prompting strategies, i.e., *prompt tuning*, for the best-performing open-weight and proprietary LLMs (Section 4.1). We also present a *meta analysis* where we reevaluate models based on a re-annotation of a part of the dataset by our domain expert (Section 4.2). We further examine *word-specific relevance* scores to detect which (type of) words drive the models or lead to common errors (Section 4.3).

4.1 Prompt Tuning

Our prompt tuning strategy includes augmenting the prompt and exploring the following settings, as demonstrated in Table 11: (a) **Baseline**, where

[9] For each open-weight zero-shot model, we report the standard deviation (± std) over multiple runs to account for prompt instability. We present additional details in Appendix A.

[10] See Table 11 in the Appendix for details on prompts

Prompt Setting		N	AE	PC	Avg
Baseline		.916	.491	.340	.582
Context-Aware		.827	.336	.252	.472
Distribution-Aware		.932	.517	.358	.602
K-Shot	K=8	**.943**	**.554**	**.386**	**.628**
	K=16	.932	.520	.392	.615
	K=32	.920	.456	.374	.583
	K=64	.929	.487	.375	.597
RAG-Shot	K=8	**.943**	.528	.377	.616
	K=16	.938	.494	.396	.609
	K=32	.929	.455	.364	.582
	K=64	.926	.438	.356	.573

Table 4: Results of Qwen 3 (14B) performance across different prompt settings. *Baseline* as in Table 3.

we provide the model with the working definition of populism (Section 1) and a brief task description (original prompt, as used to build Table 3), (b) **Context-Aware**, where we present up to five preceding sentences from the same speech after the *Baseline* prompt to augment the context, (c) **Distribution-Aware**, where we include the label distribution after the *Baseline* prompt (Table 2), (d) **K-shot**, where we explore few-shot prompting, and present K randomly-selected training examples, K/4 per category (incl. examples with both people-centrism and anti-elitism), appended after the *Baseline* prompt, (e) **RAG-Shot**, where the K demonstrated training examples are retrieved based on similarity to the target (examined) sentence.

Qwen 3 – Alternative Prompting In Table 4, we present the results for Qwen 3 (14B).[11] We observe that presenting preceding sentences as context (*Context-Aware*) harms performance. We hypothesize that the model becomes heavily biased by preceding sentences, rather than using them as assistive context, e.g., to resolve coreferences, as suggested by the prompt, leading to misclassifications. In contrast, there is a considerable performance improvement (approx. +2-3 p.p.) when we present the expected label distribution (*Dist-Aware*), as the model is informed and subsequently biased towards the majority neutral category.

Finally, for few-shot prompting, we observe that presenting a small number of examples (*K-Shot*, $K \in [8, 16]$) leads to an improvement (approx. +3-4 p.p.), while more, $K \in [32, 64]$, have a lesser effect. We have comparable results when presenting similar examples (*RAG-Shot*).

[11] We present results with Llama 3.1 (8B) in Appendix B.

Prompt Setting		Classes			
K	Mode	N	AE	PC	Avg
0	Base	.903	.513	.306	.574
	Think	.868	.429	.272	.523
	Dist	.931	.557	.372	.620
32	Base	.912	.558	.329	.600
	Think	.874	.416	.309	.533
	Dist	.921	.612	.337	.623
128	Base	.925	.597	.351	.624
	Think	.860	.448	.261	.523
	Dist	**.928**	**.609**	**.346**	**.628**

Table 5: Results of Gemini 2.5 Flash performance across different prompt settings. K refers to the number of presented examples for *K-shot*.

Gemini 2.5 – Alternative Prompting Given the previous results, we focus our analysis on *Baseline* (here *Base*), and further enhancement via *K-Shot* (for $K \in [0, 32, 128]$) and *Distribution-Aware* prompting to Gemini 2.5 Flash, the best-performing proprietary LLM. We further evaluate Gemini's *Thinking* mode. This enables the model to generate up to 1000 tokens of "thinking" (reasoning), before classifying the sentence in question.

The results are presented in Table 5. Comparing the performance of the base prompt across different Ks (0,32,128), we observe that *K-shot* demonstrations improve the results by 3-5 p.p., while providing the label distribution leads to similar improvements; similar to our findings for Qwen 3 (14B).

Surprisingly, we observe that enabling "thinking" (reasoning) leads to much lower performance across the board, something we can coin as "*overthinking*". Specifically, overthinking nudges the model toward labeling an additional 9% of total sentences as populist (i.e., anti-elitist and/or people-centric) compared to the non-thinking mode. In contrast, only 3% of sentences are flipped in the opposite direction, i.e., are classified as neutral. In other words, enabling "thinking" mode leads to a substantial labeling bias towards populist classes.

On the other hand, revealing the label distribution leads to the opposite phenomenon, which we can coin as "*hypervigilance*", a bias towards the neutral/dominant class. In other words, when exposed to the label distribution, the model becomes noticeably more cautious on the task; an additional 5% of sentences are labeled as neutral.

Method	FT	N	AE	PC	Avg	Prev.
Students	-	.975	.658	.577	.737	-
RoBERTa	✔	**.983**	.724	**.695**	**.800**	**.746**
Qwen 3 (8B)	✔	.972	.653	.565	.730	.714
Qwen 3 (14B)	✘	.951	.612	.374	.646	.628
	✔	.977	**.730**	.598	.768	.740
Gemini 2.5	✘	.931	.677	.290	.633	.628

Table 6: Results of the meta evaluation. We consider the best-performing settings for zero-shot models: Qwen 3 (14B ✘,*K-Shot*, K=8), and Gemini 2.5 (*K-Shot*, K=128 with distribution). The last column shows the original results comparing to students' annotations.

4.2 Meta Evaluation

Deciding whether a sentence is populist is not a trivial task, neither for models nor for humans. A certain level of disagreement between annotators is to be expected, as the task is based on *interpreting* an abstract social science concept. Hence, to put the models' performance into perspective, we let the domain expert on this project re-annotate the test set of TRUMP-2016, acting as an *Oracle*. For this meta-evaluation, we re-evaluate the best-performing models in their best-performing setting on the re-annotated test set, see Table 6 for results.

We treat the re-annotation as the gold standard and evaluate student annotations accordingly. The students achieve a macro-F1 of 0.737. However, they are outperformed by the RoBERTa model (m-F1: 0.8), and the fine-tuned Qwen 3 (14B) (m-F1: 0.768), and perform on par with the fine-tuned Qwen 3 (8B). Notably, the performance of the fine-tuned RoBERTa is assessed approx. 0.05 higher, while the remaining models assessed comparably to before. From a practical perspective, these results suggest that fine-tuned models, especially the RoBERTa classifier, achieve greater consistency than student annotators, effectively overcoming the "noise" introduced by non-expert labeling, thus exhibiting higher agreement with the domain expert.

4.3 Word-specific Relevance

To better understand which words are relevant for the classifiers to detect populism, we take a closer look at the TF-IDF weights as well as relevance scores for the fine-tuned RoBERTa model.

As a first step, we analyze the feature coefficients of the TFIDF-SVM model. We observe that the five most indicative expressions (n-grams) for classifying a sentence as *anti-elitism* are 'donors', 'establishment', 'insiders', 'rigged', and 'special interests'. Respectively, the five most indicative expressions for *people-centrism* are 'the American people', 'for you', 'the people', and 'forgotten'.

We further apply Layer-wise Relevance Propagation (LRP) (Ali et al., 2022) to the fine-tuned RoBERTa model, a feature attribution method that assigns a positive or negative continuous relevance score to each token in the input. This relevance score indicates its influence on the predicted class label. We extract the 5 "most relevant" tokens that were assigned positive scores from each correctly classified sentence and group them based on their part-of-speech tag, see Table 13.

Similar to the TF-IDF coefficients, we find that keywords like 'establishment', 'rigged', and 'special' are indicative of *anti-elitism* as well as 'corrupt', 'Washington', 'Hillary', and 'system'. There is also an overlap between the most relevant words for *people-centrism* like 'American'/'people', but also 'government'/'power'. Most relevant words that indicate neutral sentences are more general, which is not surprising since the model classifies them as such in the absence of populism.

When we conduct the same analysis for the misclassified sentences, we see that similar keywords like 'American', 'people', 'dishonest', 'crooked' and 'Hillary' mislead the model into classifying sentences as populist. A manual inspection of the misclassified sentences shows that RoBERTa tends to misclassify sentences as *anti-elitism* that include references to political corruption or the media, and references to the American people as *people-centrism*. In both cases, the sentences are mostly ambiguous, i.e., the context would need to be resolved with the adjacent sentences. The model however, without access to preceding sentences, heavily relies on keywords based on their high frequency among populist training sentences.

5 Analysis of Donald Trump's Populism

To demonstrate the potential application of fine-tuned models in political science research, we utilize our best-performing model, RoBERTa, on the extended dataset TRUMP-CHRONOS, which spans the decade 2015-2025. While our primary unit of analysis for NLP purposes is the sentence, we introduce a scoring scheme that assigns a discrete value to each complete speech (S), capturing the overall intensity of its populist content. Based on prior work (Aslanidis, 2018), we refer to this per-speech metric as the *Populism Discourse Index* (PDI).[12]

[12]See Appendix E for more details on how the *Populism Discourse Index* (PDI) is computed.

Campaign Name		#S	$\overline{\text{PDI}}(\pm)$
2016 Primaries	(06/15-07/16)	220	1.67 ± 2.37
2016 Election	(07/16-11/16)	133	6.30 ± 4.92
2020 Election	(06/19-11/20)	98	2.44 ± 2.01
2024 Election	(11/22-11/24)	141	1.50 ± 1.03

Table 7: Statistics across Trump's campaign periods. #S refers to the number of speeches. $\overline{\text{PDI}}$ refers to mean PDI across campaign speeches.

As shown later on, collecting PDI scores enables the statistical treatment required to address the research questions posed in our case study. This coding scheme leverages sentence-level granularity to compute a PDI for each of the 713 speeches in our dataset, alongside its associated metadata such as date and location, as the unit of analysis in downstream statistical modeling.

5.1 Research Questions & Analysis

To analyze Donald Trump's populist discourse, we answer three critical research questions (RQ):

RQ1: *"Does Donald Trump's populism substantially fluctuate over time?"*

Theoretical Rationale The dominant ideational approach in populism studies conceptualizes populism as a "thin-centered ideology" (Mudde, 2004, 2007; Mudde and Kaltwasser, 2017). Under this view, populism constitutes the core worldview and authentic political identity of populist politicians. Accordingly, the intensity of their populist discourse should remain relatively stable across contexts (null hypothesis), which in our case can be examined across electoral campaigns (Table 7).

Results The results of a one-way ANOVA (Ross and Willson, 2017) analysis decisively reject the null hypothesis, indicating a significant effect of campaign period on PDI (populist intensity) scores. The data reveal a clear strategic pattern: Trump's populist discourse intensity peaks dramatically during the 2016 general election ($\overline{\text{PDI}} = 6.30$) compared to the primaries ($\overline{\text{PDI}} = 1.67$). This surge is followed by statistically significant ($p < 0.001$) declines in subsequent campaigns in 2020 ($\overline{\text{PDI}} = 2.44$), and 2024 ($\overline{\text{PDI}} = 1.50$). The consistent differences across all campaign periods thus point to highly systematic rather than random variation.

Our findings challenge binary conceptions of populism that dominate academic and public debates. Rather than a fixed ideological trait, populist discourse emerges here as a variable mode of political meaning-making, modulated

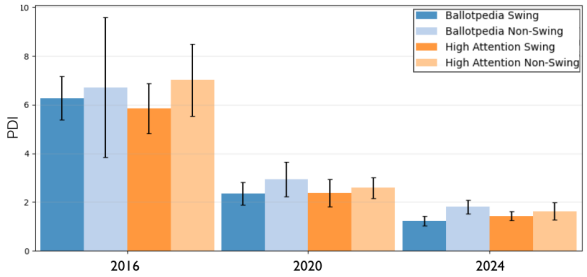

Figure 3: PDI scores split by campaign and state type.

across time and contexts. This supports the idea that scholars should shift to dynamic, *gradational* approaches in populism research (Aslanidis, 2016).

RQ2: *"Does Trump's populist discourse intensity vary between swing states and non-swing states?"*

Theoretical Rationale In U.S. elections, due to the winner-take-all system at the state level, candidates focus disproportionately on key battlegrounds, known as *swing states* (Shaw et al., 2024). As populist discourse is considered an effective tool for activating the voter base through emotional appeals and in-group/out-group distinctions (Mudde and Kaltwasser, 2017; Norris and Inglehart, 2019), we expect Trump to deploy higher populist intensity in swing states where mobilization efforts are most critical. Lacking unanimity on swing state status, we test two different clustering schemes (see Table 15). To test for statistical significance, we conduct a series of independent-samples t-tests comparing mean PDI scores between swing and non-swing states per campaign and state clustering.

Results Our analysis provides no evidence for the swing state targeting hypothesis. Independent t-tests comparing populist discourse intensity between swing and non-swing states yielded significant results in only 1 of 6 tests in the 2024 campaign under the Ballotpedia definition (Table 12). Contrary to theoretical expectations, these significant results reveal that Trump deployed *less* populist rhetoric in swing states than in non-swing states. For detailed results, see Tables 16-12.

Our findings suggest that Trump's populist discourse was not geographically targeted in general. The only significant differences run counter to expectations, with *lower* levels of populist rhetoric in swing states during 2024. That may reflect a strategic shift toward moderation in competitive contexts, where aggressive populist messaging risks alienating the critical pool of undecided voters (Dai and Kustov, 2022).

RQ3: *"How does Donald Trump's populism spread out within a speech?"*

Theoretical Rationale Our final research question examines the strategic placement of populist invocations within speeches. According to our definition, populism is a discursive operation that lacks programmatic content, so we expect it to primarily appear in bursts that aim to frame in-group/out-group identities. Therefore, we expect Trump to set the overall tone of his framing by clustering populist discourse at the beginning and end of his speeches, with sparse populist content in between.

Results We analyze sentence-level positioning, as a ratio of populist sentences in different positions (bins) among all populist sentences, employing a 3-bin scheme (Opening/Body/Closing 20/60/20%), and conduct pairwise paired t-tests between bins. We find significant differences in people-centrism and overall populism between all bins. As shown in Figure 5, populist discourse is patterned, with approx. 27% of all populist sentences appearing in the closing of speeches. *People-centrism* is heavily back-loaded, with nearly 40% of people-centric sentences appearing in the closing of speeches compared to less than 20% in the opening. We do not find significant differences for *anti-elitism*.

6 OOD Exploration – The case of Europe

Given that our TRUMP-2016 dataset is limited to Donald Trump and the specific historico-political context of the 2016 U.S. presidential election, we test the models' performance in Out-of-Domain (OOD) campaign speeches. We evaluate the best-performing models (RoBERTa, Qwen 3 (14B), and Gemini 2.5 Flash) on a small set of speeches by European politicians (EU-OOD). For Qwen 3 (14B) and Gemini 2.5 Flash, we run zero-shot experiments adding the label distribution from the TRUMP-2016 dataset to the prompt (*Distribution-Aware* from Tables 4-5), as a minimal intervention informing on the general data skewness.

In Figure 4, we present the results using the sentences in the speakers' native languages (EU-OOD) and translated into English (-EN). We observe that both LLMs, Qwen 3 and Gemini 2.5 Flash, perform substantially better (approx. +10 p.p.) compared to the fine-tuned RoBERTa classifier. The performance of the fine-tuned RoBERTa classifier declines by approx. 24 p.p., while the performance of both LLMs is comparable across datasets. This is a strong indication that our classifier has substantially overfitted Donald Trump's style of discourse and manifestation of populism.

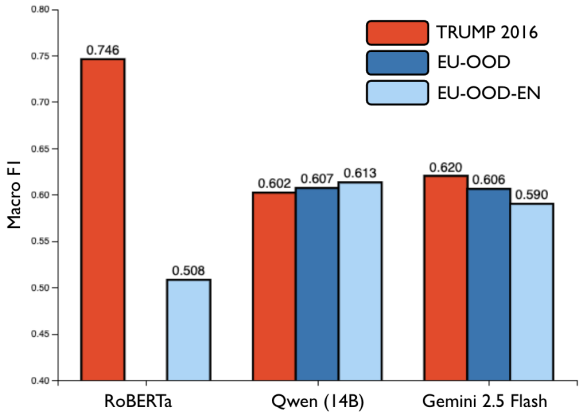

Figure 4: Performance across models on the original EU-OOD dataset and translated (-EN).

7 Related Work

Bonikowski et al. (2022) examine the social scientific concepts of populism, nationalism, and authoritarianism in U.S. presidential campaign speeches from 1952-2020. They deploy a RoBERTa classifier, which they fine-tune in an active learning fashion. To do so, they manually annotate 2,224 paragraphs in total across six non-mutually exclusive frames, including populism, which they code in binary terms, e.g., populist or not. In their analysis, they further binarize speeches by labeling them as populist if at least one paragraph contains populist appeals. Paragraphs were selected as the unit of analysis because sentences were deemed too granular, while full speeches would have too low a signal-to-noise ratio for effective classification. They find that the fine-tuned RoBERTa classifiers substantially outperform Random-Forest classifiers.

Klamm et al. (2023) develop MoPE (Mentions of the People and the Elite), a cross-lingual dataset for entity-level detection of populist references in parliamentary debates. They create a hierarchical annotation scheme to identify mentions of "the People" and "the Elite" in text, regardless of specific terminology used. They annotate 9,297 mentions across German Bundestag speeches (2017-2021) and 1,423 mentions in English EuroParl data, training transformer-based models to automatically detect these entity references. While complementary to our work, their approach focuses on identifying specific entity mentions (e.g., "taxpayers", "government") rather than classifying populist discourse

patterns at the sentence level.

Zhang and Schroeder (2024) detect populist discourse in Chinese social media by analyzing over 100K posts as their unit of analysis. They employ manual annotation for binary classification (populist vs. not) to train a Chinese RoBERTa model. Their methodology adapts populism to the Chinese context as an intersection of populism and ultra-nationalism, targeting foreigners, Westerners, pro-Western intellectuals, and political elites. Using topic modeling and correspondence analysis, they identify distinct populist themes including "corruption", "betrayal", "super-national treatment", and "immoral West". Their work appears to primarily capture nationalist sentiment rather than populist discourse, as evidenced by their focus on cultural invasion and foreign enemies.

Erhard et al. (2025) develop PopBERT, a transformer-based multi-label classifier for detecting populist discourse in German parliamentary speeches. Their approach extends beyond core populist dimensions to identify "host ideologies" (left-wing socialism vs. right-wing nativism) that attach to populist rhetoric, according to their working definition. They fine-tune a German BERT model using active learning for data sampling and employ a highly permissive annotation aggregation strategy where any sentence labeled populist by at least one of five coders is considered positive, leading to low inter-annotator agreement.

Halterman (2025) introduces a methodology for using LLMs to generate synthetic training data to later train supervised text classifiers rather than employing LLMs directly for annotation tasks. In a populism-focused validation, among other use cases, the author uses GPT-3.5 to generate 5,357 synthetic populist manifesto statements across 27 European countries in 22 languages. The synthetic sentences, along with 36,509 negative examples generated from policy position descriptions, are used to train a sentence-level binary classifier using the SetFit few-shot framework without any gold-standard hand-labeled training data.

Tao et al. (2025) examine the use of online populist discourse in China. Their methodology relies on Semantic Role Labeling (SRL) techniques to extract semantic triplets from text and a rule-based method to classify triplets based on a curated dictionary. They use the developed tool to examine the operationalization of people-centrism and anti-elitism, considering three trending events that occurred in China between 2019 and 2021.

The aforementioned work, although highly relevant to our work, does not examine (use and evaluate) new-era LLMs. In many cases, the authors do not use gold-standard annotations for training and evaluating models. The conceptualization of populism is binary (populist or not) and is usually conflated with other concepts, such as nativism. There is no elaborate analysis on how Donald Trump instrumentalizes populism as part of his discourse.

8 Conclusion

To the best of our knowledge, this is the first work on populism detection employing LLMs. As discussed throughout our work, the task is challenging for humans, as well as models, due to subjectivity and contextualization. Our work focuses on identifying populism in Donald Trump's campaign speeches. We find that a fine-tuned RoBERTa classifier vastly outperforms both open-weight and top-notch proprietary LLMs in zero-shot fine-grained populism identification. Only after fine-tuning, LLMs reach comparable performance to the much smaller RoBERTa model. Providing information about the label distribution, i.e., making the LLMs aware that labels are heavily skewed towards non-populist sentences, improves zero-shot performance as well as few-shot prompting. Including "reasoning", on the other hand, seems to have the opposite effect and decreases the performance.

An analysis of a large unlabeled collection of Trump's campaign speeches spanning 2015-2025, utilizing our best-performing model, RoBERTa, shows that his populist rhetoric decreased over time. We do not find a significant discrepancy in populist rhetoric between swing and non-swing states. Furthermore, we observe that populism, and specifically people-centrism, is significantly back-loaded within speeches. We finally conduct a small-scale, explorative experiment on European speeches. Results show a clear drop in performance for RoBERTa and comparable results for LLMs, outperforming RoBERTa by 10%.

In conclusion, although LLMs struggle with the ambiguity and contextualization of populism detection, the best-performing models perform on a comparable level to human annotators. Future research should explore further prompt tuning techniques and apply these models to larger, more diverse datasets with high-quality annotations, or explore other research questions (hypotheses).

Acknowledgments

In line with the suggestion of the "Generative AI Disclosure Statement" introduced by ACM FAccT 2026, we report the use of generative AI tools that were used in this work. We primarily used generative AI tools for spell-checking, grammar, and style editing, i.e., minor recommendation edits on sentence phrasing. We reviewed the suggested edits, and accepted only those that we considered as having a positive impact on the writing quality of our work. We also used generative AI tools to prototype snippets of code, e.g., for generating plots, that we rigorously checked for their validity.

Limitations

Annotated Trump Dataset Our newly released annotated dataset, TRUMP-2016 (Section 2.2), is relatively small (70 speeches, ∼15K sentences) and focuses exclusively on a single individual, Donald Trump, within the specific context of the 2016 U.S. presidential election. As a result, models trained on this dataset, such as RoBERTa, adopt a narrow understanding of populism and struggle to generalize across contexts (see Section 6). Accordingly, our analysis (Section 5) is confined to Trump, and our findings should not be generalized to other political figures. In future work, we plan to expand our datasets and analyses to include a more diverse set of political actors across different ideological orientations, geographic regions, and languages, where different colloquial expressions are used to denote the elite (e.g. 'regenten' in Dutch, 'la casta' in Spanish and Italian, etc.).

Augmenting Context In our study, we primarily classify sentences in isolation, i.e., out-of-context, which limits our ability to incorporate relevant contextual cues, such as surrounding sentences. In many cases, populist discourse is context-dependent, and key interpretive clues are referentially underspecified, often taking the form of pronouns or deictic expressions. For example, in the sentence *"They are acting against you!"*, the referents of 'they' and 'you' may carry strong political significance but cannot be identified (disambiguated) without additional context.

As shown in Section 4.1, providing as context up to five preceding sentences to Qwen 3 (14B), and also Llama 3.1 (8B) in Appendix B.3, hurts the classification performance. In preliminary experiments, we also considered a variant of the fine-tuned RoBERTa classifier that is enhanced with additional context. Specifically, we augmented the input with up to 7 preceding sentences separated by the special <sep> token (Appendix B.2), but this yielded no performance gain. Similarly, we tested Gemini 2.5 Flash with the full speech as context and observed negative results.

These negative results are a strong indicator that models cannot effectively utilize contextual information and are rather negatively biased (misguided). Going forward, we plan to explore alternative forms of contextual augmentation or consider using substantially larger models that may leverage extended context effectively and faithfully.

Subjectivity The concept of populism is heavily debated in both academia and the media (see Section 1). Identifying and classifying populist discourse is inherently subjective and often context-sensitive, much like many other social science concepts. Therefore, annotating populist discourse is a labor-intensive process that demands substantial training resources to achieve acceptable inter-annotator agreement. In our experience, two issues posed particular challenges for human annotators: (a) negative references to the media, and (b) references to corruption. Both can qualify as anti-elitist, depending on framing and context. For instance, if the media are framed as part of an establishment pitted against the people, such references should be coded as anti-elitist. However, this framing is not always present. Similarly, anti-corruption discourse qualifies as populist, provided that the allegedly corrupt actors are discursively framed as members of the elite. Yet in many instances, corruption references lack such framing. These challenges again highlight the importance of contextual interpretation when identifying populist discourse. Overall, a comprehensive study should complement computational results with qualitative insights to draw robust conclusions about the populist character of a given political phenomenon.

Ethical Considerations

The developed datasets include speeches of Donald Trump and other European, mostly right-wing, political leaders from electoral campaign rallies, already publicly online; either transcribed by the UC Santa Barbara's American Presidency Project or as videos on the public domain (YouTube, etc.). There are mentions to individuals, i.e., praising political allies or attacking political opponents, e.g., 'Hillary

Clinton', as expected in political rhetoric. The data may contain derogatory language against protected groups based on race, gender, or other characteristics. The authors do not endorse, but rather condemn, the use of such language. The use of data -in their original form- is critical to identifying populist cues. The intended use of collecting these data is to accommodate scientific research and the rigorous analysis of populism. Furthermore, the data can be used (re-annotated) to examine other social phenomena, as nativism, nationalism, or the use of derogatory language. The intended use of our data is limited to the scope of research, and by no means used to create caricature chatbots.

References

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. Xai for transformers: Better explanations through conservative propagation. In *International conference on machine learning*, pages 435–451. PMLR.

Paris Aslanidis. 2016. Is Populism an Ideology? A Refutation and a New Perspective. *Political Studies*, 64(1_suppl):88–104.

Paris Aslanidis. 2018. Measuring populist discourse with semantic text analysis: an application on grassroots populist mobilization. *Quality & Quantity*, 52(3):1241–1263.

Paris Aslanidis. 2024. *Populist Mobilization*. Oxford University Press.

Rishi Bommasani, Percy Liang, and Tony Lee. 2023. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146.

Bart Bonikowski, Yuchen Luo, and Oscar Stuhler. 2022. Politics as usual? measuring populism, nationalism, and authoritarianism in u.s. presidential campaigns (1952–2020) with neural language models. *Sociological Methods & Research*, 51(4):1721–1787.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, and 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*.

Corina Cortes and Vladimir Vapnik. 1995. Support Vector Networks. *Machine Learning*, 20:273–297.

Yaoyao Dai and Alexander Kustov. 2022. When do politicians use populist rhetoric? populism as a campaign gamble. *Political Communication*, 39(3):383–404.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, Minneapolis, Minnesota.

Lukas Erhard, Sara Hanke, Uwe Remer, Agnieszka Falenska, and Raphael Heiko Heiberger. 2025. Popbert. detecting populism and its host ideologies in the german bundestag. *Political Analysis*, 33(1):1–17.

Google Gemini Team. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. Technical report.

Google DeepMind Gemma Team. 2025. Gemma 3 - Technical Report. Technical report.

Andrew Halterman. 2025. Synthetically generated text for supervised text analysis. *Political Analysis*, 33(3):181–194.

Kirk A Hawkins, Ryan E Carlin, Levente Littvay, and Cristóbal Rovira Kaltwasser. 2018. *The ideational approach to populism: Concept, theory, and analysis*. London: Routledge.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Christopher Klamm, Ines Rehbein, and Simone Paolo Ponzetto. 2023. Our kind of people? detecting populist references in political debates. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1227–1243, Dubrovnik, Croatia. Association for Computational Linguistics.

Ernesto Laclau. 2005. *On Populist Reason*. Verso.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Cas Mudde. 2004. The Populist Zeitgeist. *Government and Opposition*, 39(4):541–563.

Cas Mudde. 2007. *Populist Radical Right Parties in Europe*. Cambridge University Press.

Cas Mudde and Cristóbal Rovira Kaltwasser. 2017. *Populism: A very short introduction*. Oxford University Press.

Pippa Norris and Ronald Inglehart. 2019. *Cultural Backlash: Trump, Brexit, and Authoritarian Populism*. Cambridge University Press.

OpenAI. 2023. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

OpenAI OpenAI Team. 2025. Introducing GPT-4.1 in the API. Technical report.

Qwen Qwen Team. 2025. Qwen3 - Technical Report. Technical report.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Andrew Ross and Victoria L. Willson. 2017. *One-Way ANOVA*, pages 51–62. SensePublishers, Rotterdam.

Daron R Shaw, Scott L Althaus, and Costas Panagopoulos. 2024. *Battleground: Electoral College Strategies, Execution, and Impact in the Modern Era*. Oxford University Press.

Yannis Stavrakakis. 2024. *Populist Discourse - Recasting Populism Research*. Routledge.

Yannis Stavrakakis and Giorgos Katsambekis. 2024. *Research Handbook on Populism*. Edward Elgar Publishing.

Yuzhou Tao, Zhiqin Zhan, Han Zhou, Jingshi Kang, and Shaojing Sun. 2025. Measuring chinese online populist discourse: an automated semantic text analysis method. *Chinese Journal of Communication*, 18(2):121–141.

Kurt Weyland. 2001. Clarifying a Contested Concept: Populism in the Study of Latin American Politics. *Comparative Politics*, 34(1):1–22.

Kurt Weyland. 2017. Populism: A Political-Strategic Approach. In Cristóbal Rovira Kaltwasser, Paul Taggart, Paulina Ochoa Espejo, and Pierre Ostiguy, editors, *The Oxford Handbook of Populism*, pages 48–73. Oxford University Press, Oxford.

Yuan Zhang and Ralph Schroeder. 2024. "it's all about us vs them!": Comparing chinese populist discourses on weibo and twitter. *Social Media + Society*, 10(1):20563051241229659.

A Additional Experimental Details

A.1 Baseline

We train a linear SVM-based classifier using the Scikit-learn library.[13] We use the `TfidfVectorizer` to vectorize the inputs with min_df=20, max_df=0.5, for 10K n-grams, where $n \in [1, 2, 3]$ with a lower-cased vocabulary. We use the `svm.SVC` with a linear kernel as a multi-output classifier.

A.2 PLMs

Fine-tuning SetUp We fine-tune both PLMs (RoBERTa and DeBERTa-V3) as 3-way multi-label classifiers using a shallow MLP on top of the final `<s>`, also known as `[CLS]`, representation. We train both models for a total of 5 epochs (iterations over the training examples), using cosine learning rate scheduling with a maximum learning rate of $1e{-}5$, a warmup ratio of 0.1, and weight decay of 0.01.[14] We use mini-batches of 32 examples, the maximum we can use, to maximize the potential of having populist (labeled as PC and/or AE) examples in every mini-batch, since positive examples are very sparse (7.4% of the training set), and even sparser considering individual categories (AE, PC). We also truncate all examples to 64 tokens, which is much larger compared to the average sentence length (approx. 15 words), to speed up training.

Upsampling To further tackle the intense class-imbalance, we use a trivial up-sampling (5×) of the positive examples (ending up with approx. 37% positive examples) for fine-tuning, which leads to improved performance compared to using the original dataset (no augmentation, see Table 8). In Appendix B.1, we present results using more elaborate data augmentation strategies with negative results (Appendix B.1).

A.3 LLMs

Prompts and SetUp In Table 11, we present the exact (verbatim) wording of all examined prompts. As mentioned above, the vast majority of our experiments are with LLMs out-of-the-box. We use 4-bit quantized versions. We run 6 runs for all open-weight models using 3 different seeds ($s \in [21, 42, 84]$) for all random engines, e.g., numpy, torch, etc. At the same time, we also present the options in two different orders: (i) where a-d are neutral, anti-elitism, people-centrism in order, both, and (ii) where a-d are both, anti-elitism, people-centrism, and neutral in order. We report the mean and standard deviation per class across all 6 runs.

Augmentation	N	PC	AE	Avg
None	.975	.625	.566	.722
Up-sampling (5×)	.975	.661	.602	.746
Synthetic	.969	.589	.574	.710
Curated Synthetic	.975	.690	.584	.750
Paraphrased	.972	.629	.586	.729

Table 8: Results of RoBERTa using different augmentation techniques (settings).

Finetuned LLMs We additionally present results with fine-tuned versions of Llama 3.1 (8B), and Qwen (8B) using LoRA adapter ($\alpha = 32$, $r = 16$, and a dropout rate of 0.05) targeting the projection layers, following best practices. We fine-tune these models for 1 epoch, since they rapidly overfit, as the main point of fine-tuning is to optimize the selection among 4 predefined options (a-d). During training, we mask the whole input (instruction) up to the expected generated when computing the loss. We do not run different option orders for fine-tuned models, since they have been fine-tuned to follow a pre-defined order.

B Additional Experiments

B.1 Data Augmentation

We consider alternative data augmentation strategies to improve PLMs' performance. All strategies aim to augment the training set with positive (labeled as PC and/or AE) examples. The strategies are: (a) **Up-sampling (5×)**: Up-sampling the ratio of examples labeled with positive categories (Anti-elitism and/or People-centrism) by a factor of 5, leading to 5.6K positive examples in total, (b) **Synthetic**: Augmentation with 1.2K synthetic positive (Anti-elitism and/or People-centrism) examples generated by Llama 3.1 (70B) and ChatGPT. (c) **Curated Synthetic)**: Augmentation with a curated selection of 500 out of the 1.2K synthetic examples. (d) **Paraphrased**: Augmentation with 3.5K examples labeled with positive categories (Anti-elitism and/or People-centrism) paraphrased by Llama 3.1 (8B).

[13] https://scikit-learn.org/stable/

[14] We select the core hyper-parameters after a brief grid-search for learning rate ($lr \in [1e-5, 3e-5, 5e-5]$ and epochs ($e \in [3, 5, 10]$) using a subset of the training set as a development set, tracking the macro-F1.

Setting	N	PC	AE	Avg
Baseline	.975	.661	.602	.746
Context-Aware	.975	.649	.575	.733

Table 9: Results of RoBERTa with and without context.

In Table 8, we observe that naive up-sampling (5×) leads to a considerable performance improvement, compared to the baseline (*None*) using the original dataset. Contrary, the vanilla generation of synthetic data (*Synthetic*) has a negative impact. Manually curating (sub-sampling) the synthetic data (*Curated Synthetic*) leads to a similar performance compared to naive up-sampling. Lastly, augmentation via paraphrasing (*Paraphrased*) has a trivial impact.

B.2 Context-Aware RoBERTa

We also considered a variant of RoBERTa that is enhanced with additional context. Specifically, we use an input, the examined sentence, as before, and up to 7 preceding sentences separated by the special <sep> token. The hypothesis is that since populist invocations are heavily context-dependent, presenting context alongside the sentence will lead to performance improvement. As we observe in the results of Table 9, the performance deteriorates. We hypothesize that RoBERTa cannot effectively use the context as additional information and is negatively biased by the context, i.e., if a populist sentence is preceding then the model classifies the examined sentence as such, similar to Qwen 3 (14B) prompted in a similar setting (Table 4 in Section 4.1). The context-aware model is also substantially slower encoding 512 (8×64) tokens in comparison to 64 tokens.

B.3 Llama 3.1 - Prompting Tuning

In Table 10, we present the results for Llama 3.1 (8B) for alternative prompts, similar to our analysis with Qwen 3 (14B) in Section 4. Similarly to Qwen 3 (14B), we observe that presenting preceding sentences as context harms performance, while in contrast, there is a considerable performance improvement (approx. -3%) when we present the expected label distribution (Dist-Aware) compared to our base prompt (Baseline). Lastly, we observe that presenting labeled examples (K=32) has a substantially negative impact (approx. -10%) in the case of Llama 3.1, which may indicate that smaller LLMs may not have the potential of effectively utilizing few-shot demonstrations. In Section 4, we find that both Qwen 3 (14B) and Gemini Flash 2.5 benefit from few-shot demonstrations (Tables 4-5).

Setting	N	AE	PC	Avg
Baseline	.814	.320	.245	.460
Context-Aware	.507	.212	.116	.278
Dist-Aware	**.879**	**.368**	**.269**	**.505**
K-Shot (32)	.704	.275	.186	.388

Table 10: Results of Llama 3.1 (8B) performance across different prompt settings.

C Annotation Process

Dataset Curation Our datasets excludes town hall events where Donald Trump exclusively responds to open-ended questions from the host or audience members. However, we include town halls that begin with a speech by Trump, omitting the subsequent Q&A portion from the transcript. We also exclude formal events (inauguration speeches, joint addresses to Congress, State of the Union addresses) and speeches where Trump is not addressing the public at large, (e.g., invitation-only events at think tanks, lobby groups, trade unions, galas, fund-raisers, or other civil society associations) since these are expected to focus on specific policy issues. However, we include speeches at GOP conventions (regional and national), and annual/semi-annual CPAC events because of their rally-like nature. The excluded speeches comprise a very small fraction of the overall roster of public events that Trump headlined.

Student Training The Political Science students were recruited to help build the annotated dataset of Trump speeches. The students resided in the US, and received compensation from an internal university grant (∼$13 an hour). They received extensive training from our domain expert, using a detailed coding scheme (see "Coding Manual" below) designed to identify key markers of populist discourse and to distinguish populism from related political science concepts such as nationalism, nativism, socialism, authoritarianism, etc. The training phase began with a pilot phase in which the students annotated out-of-domain sample speeches and received detailed feedback. This was followed by a second stage in which the students independently annotated a subset of the Trump speeches drawn from the target dataset, without further feedback. On this in-domain material, they achieved an inter-annotator agreement of Krippendorff's α =

Setting	Prompt Text
Base	You are a helpful AI assistant with expertise in identifying populism in public discourse. Populism can be defined as an anti-elite discourse in the name of the "people". In other words, populism emphasizes the idea of the common "people" and often positions this group in opposition to a perceived elite group. There are two core elements in identifying populism: (i) anti-elitism, i.e., negative invocations of "elites", and (ii) people-centrism, i.e., positive invocations of the "people". You must classify each sentence in one of the following categories: (a) No populism. (b) Anti-elitism, i.e., negative invocations of "elites". (c) People-centrism, i.e., positive invocations of the "People". (d) Both people-centrism and anti-elitism populism.
Content-Aware	*[Base]* + Here are the preceding sentences for context: [Up to 5 preceding sentences] When classifying a sentence, focus primarily on the content of that specific sentence. Use the context of preceding sentences only to resolve coreferences (e.g., identifying who "they" or "you" refer to) or to disambiguate when the sentence is ambiguous on its own.
Speech-Aware	*[Concept Desc.]* + The following sentence is part of a presidential campaign rally speech by [Speaker name] on [Date] at [Location].
Distribution-Aware	*[Base]* + The label distribution is (a) No populism (92%), (b) Anti-elitism (4%), (c) People-centrism (2%), (d) Both people-centrism and anti-elitism (2%).
K-shot	*[Base]* + The following sentences are in category (a) No populism: [K/4 examples] The following sentences are in category (b) Anti-elitism populism: [K/4 examples] The following sentences are in category (c) People-centrism populism: [K/4 examples] The following sentences are in category (d) Both people-centrism and anti-elitism populism: [K/4 examples]
RAG-Shot	*[Base]* + Here are the most similar [K] sentences from the training set, accompanied by their label: [K most-similar examples with labels] When classifying a sentence, focus primarily on the content of that specific sentence.

Table 11: Prompt Settings explored in Section 4.1 - "Qwen 3 Exploration" and "Gemini 2.5 Exploration". All presented prompt settings are followed by the question: *"Which is the most relevant category for the sentence: S?"*

0.751. After establishing sufficient agreement, the remaining speeches were divided among the students for independent annotation. This final phase produced a dataset of 15,025 sentences, with an average sentence length of approximately 15 words.

Overall, the same procedure was used for coding the European annotated dataset. Students were recruited from Panteion University of Social and Political Sciences (Athens, Greece) to code the French and Greek speeches. Speeches by Kickl and Weidel were annotated by one of the co-authors who is fluent in German, under the supervision of our domain expert. The students resided in Greece and were compensated with ∼25€ per speech.

Coding Manual The coding manual outlines a structured, step-by-step process for annotators to identify and classify populist discourse in political texts.[15] Annotators are instructed to begin with a holistic reading of the entire speech to grasp its overall context and to mentally map out the in-group ("we") and out-group ("they") actors. Coding is done at the sentence level, with each sentence assessed for the presence of social actors (the "actor set"). The first step is to check whether the sentence contains any actors; if not, it is skipped. If actors are present, the coder must determine whether the sentence reflects people-centrism, i.e. references to a moral and sovereign "people", and/or anti-elitism,

[15]The full manual in PDF format is here.

i.e. references to a corrupt elite usurping political power. These two codes can appear independently or together, depending on the content of the sentence. Coders are trained to assess both explicit and, where justified, implicit references, and to use context to resolve ambiguities. A flowchart is provided to assist in decision-making, but consistency and neutrality in interpretation are emphasized.

D Background on Populism

Populism has attracted growing scholarly attention in recent decades, yet it remains a contested concept, with multiple theoretical traditions offering distinct analytical perspectives (Stavrakakis and Katsambekis, 2024). Populism encompasses a broad range of political expressions, spanning both left- and right-wing movements and including authoritarian and democratizing projects.

On the political right, figures such as Donald Trump, Jair Bolsonaro, Viktor Orbán, and Marine Le Pen are routinely characterized as populists; on the left, the label has been applied to politicians like Bernie Sanders, Alexandria Ocasio-Cortez, Jeremy Corbyn, and Jean-Luc Mélenchon. Owing to its prominence in political debates, *populism* has become a highly contested label, often equated with *extremism* and *anti-democratic* views, or reduced to simple *demagogy*.

The academic literature on the topic broadly divides into three major schools of thought:

(a) *Top-down manipulation theories* conceptualize populism as a power-seeking strategy employed by non-ideocratic, opportunistic leaders who bypass institutional mediation to establish direct, quasi-personal relationships with unorganized mass constituencies (Weyland, 2001, 2017);

(b) *Ideational approaches* define populism as a thin-centered ideology that attaches itself to core ideologies (e.g. socialism, conservatism) and frequently undermines democratic institutions through its anti-establishment stance and rejection of standard liberal democratic norms (Mudde, 2004, 2007; Mudde and Kaltwasser, 2017; Hawkins et al., 2018);

(c) *Discursive approaches* define populism as a form of political discourse that constructs a fundamental antagonism between "the people" and "the elites" in order to challenge the status quo (Laclau, 2005; Stavrakakis, 2024; Aslanidis, 2024).

The discursive school of thought aligns with understanding populism as a collective action frame, i.e. a strategic interpretive schema employed by political entrepreneurs to unify diverse societal grievances into a single narrative that champions popular sovereignty over elite domination (Aslanidis, 2016, 2024). This constructivist lens emphasizes the role of meaning-making in sociopolitical mobilization and highlights how populist actors weave together seemingly disparate demands under an "us" versus "them" (specifically, "people" versus "elites") rhetoric. This approach is particularly well-suited to the computational detection of populism, given its reliance on specific discursive patterns and lexical choices.

E Populism Discourse Index (PDI)

The scoring scheme per sentence (s) is defined as follows:

$$\text{Score}(s) = \begin{cases} 0 & \text{if } p(s) \text{ is N} \\ 1 & \text{if } p(s) \text{ is AE} \vee \text{PC} \\ 3 & \text{if } p(s) \text{ is AE} \wedge \text{PC} \end{cases} \quad (1)$$

where p is the prediction of the model for a given sentence (s), N is neutral, AE is anti-elitism, and PC is people-centrism. The *exponential boost* assigned to *fully populist* sentences (a constant set to 3 in our implementation) is intended to capture the loaded cognitive and affective impact of characteristically populist declarations such as *"The political elite is the enemy of the American people"* or *"The system must work for all Americans, not just those at the top"*.

In addition, we apply an *adjacency multiplier* to pairs of sequential (adjacent) sentences (s_k, s_{k+1}) coded as anti-elitist (AE) and people-centric (PC), or vice versa. This multiplier (set to 1.5) reflects the discursive effect of consecutive but distinct populist claims. For instance, consider the following pair of adjacent sentences:

	s	$p(s)$
s_k	*"The system is rigged."*	AE
s_{k+1}	*"The people must rise up."*	PC

The score for this sentence pair (s_k, s_{k+1}) is $1+1=2$ points, under simple additive scoring, but given the *adjacency multiplier* applied, the score increases to $(1 \times 1.5) + (1 \times 1.5) = 3$ points. The rationale is that such a pair effectively equals the rhetorical force of a *fully populist* sentence (AE $\wedge$ PC) from the perspective of the audience.

In the computation of PDI, we exclude sentences that have fewer than 3 words, since the vast

Campaign	Clustering	PDI Diff.	PDI p-value	PDI Cohen's d
2016	Ballotpedia	-0.445	.785	-0.090
	High Attention	-1.170	.182	-0.238
2020	Ballotpedia	-0.584	.290	-0.291
	High Attention	-0.214	.629	-0.106
2024	Ballotpedia	-0.577	.001***	-0.579
	High Attention	-0.200	.270	-0.193

Table 12: Statistical Test Results – Swing (S) vs Non-Swing ($\neg S$) State Populist Discourse. Difference is defined as: $\overline{\text{PDI}_S} - \overline{\text{PDI}_{\neg S}}$. Negative values (-) indicate lower populist rhetoric in swing states. *** p < 0.001.

majority are interjections, e.g., *"Wow!"*, *"Incredible!"*, rhetorical questions, e.g., *"Right?"*, *"Guess what?"* or trivial text in general.[16] For similar reasons, we also filter out all sentences that begin with the string *"Thank_"*. The resulting dataset comprises 604,391 sentences (excluded sentences account for 7.9% of the original dataset). Based on the scoring from Equation 1, the speech-wise score (PDI_S) is defined as:

$$\text{PDI}_S = \frac{1}{N} \sum_{s=i}^{N} \text{AdjScore}(s_i) \qquad (2)$$

where AdjScore is the adjusted scoring function, w.r.t. adjacent populist sentences, and N is the total number of sentences (s_i) in a single speech (S).

F XAI Analysis

In Table 13, we present the top-5 relevant tokens per class across the most common Part-Of-Speech (POS) part of our XAI word-relevance analysis in Section 4.3.

G Analysis of Donald Trump's Populism - Supplementary Results

In Table 14, we present detailed results of the pairwise t-tests comparing PDI scores across campaign periods (RQ3 - Section 5). In Table 15, we present the clustering of swing states vs non-swing states by campaign period. In Table 12, we present the statistical test results for the swing states vs non-swing states analysis (RQ2 - Section 5). In Figure 5, we present the distribution of populist content across speech related to RQ3 in Section 5.

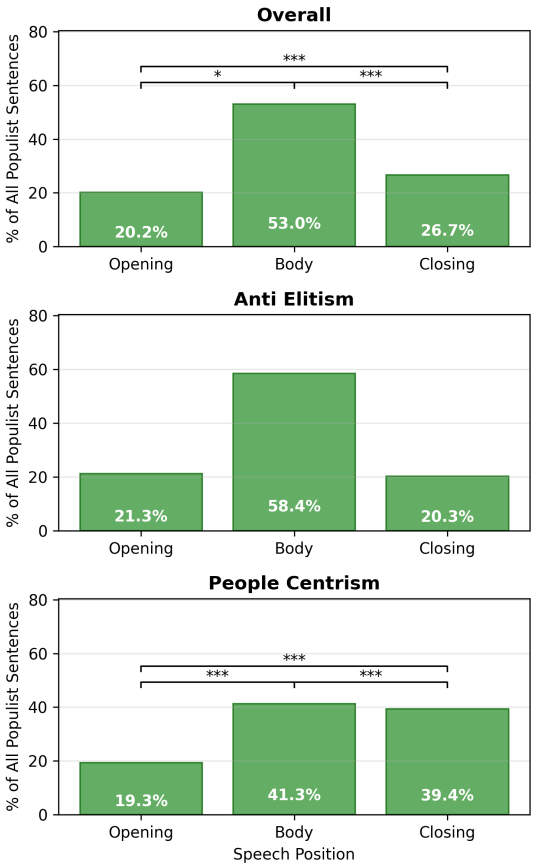

Figure 5: Distribution of populist content across speech. Bars depict % of populist sentences in a given bin among all populist sentences) for the 3-bin scheme using a balanced 20-60-20 split. Significance levels of comparisons between the opening/closing bin and all other bins are depicted as *** (p<0.001), * (p<0.05). No connections means that the comparison is not significant.

[16] We observe that in our annotated dataset (TRUMP-2016), only 3 out of ∼15K have less than 3 words and are labeled as populist (Anti-Elitism and/or People-centrism).

POS	Neutral	Anti-elitism	People-centrism
ADJ	great, good, right, bad, ok	corrupt, entire, american, political, special	american, corrupt, political, special, new
NOUN	way, people, percent, numbers, job	system, establishment, people, government, drain	people, government, power, interests, system
PRON	you, i, a, that, they	the, our, you, ourselves, us	every, you, ourselves, the, our
PROPN	african, trump, fantastic, hampshire, mr.	washington, hillary, voter, d.c., november	voter, washington, americans, justice, nation
VERB	know, look, doing, do, see	rigged, serve, hear, win, fighting	serve, fighting, fight, believe, believes

Table 13: Top-5 relevant tokens per class across the most common Part-Of-Speech (POS) tags, applied to correctly classified sentences for the fine-tuned RoBERTa model. Relevance scores were computed with LRP.

Comparison	t-value	p-value	Mean Difference (Δ)	Cohen's d
2016 Primaries vs 2016 Campaign	-11.867	0.000***	-4.63	-1.303
2016 Primaries vs 2020 Campaign	-2.816	0.005**	-0.78	-0.342
2016 Primaries vs 2024 Campaign	0.777	0.438 (n.s.)	+0.16	0.084
2016 Campaign vs 2020 Campaign	7.315	0.000***	+3.86	0.974
2016 Campaign vs 2024 Campaign	11.307	0.000***	+4.80	1.367
2020 Campaign vs 2024 Campaign	4.728	0.000***	+0.94	0.622

Table 14: Pairwise t-tests comparing PDI scores across campaign periods, with Cohen's d effect sizes.

Campaign	Ballotpedia Swing States		Speeches	High-Attention States (75th %ile)		Speeches
2016	AZ, CO, FL, IA, MI, NV, NH, NC, OH, PA, VA, WI	12	123 vs 10	FL, NC, OH, PA, CO	5	81 vs 52
2020	AZ, FL, GA, IA, MI, MN, NV, NH, NC, OH, PA, TX, WI	13	82 vs 16	PA, NC, FL, MI, WI, AZ, MN, OH	8	68 vs 30
2024	AZ, GA, MI, NV, NC, PA, WI	7	73 vs 68	IA, PA, NC, NH, MI, WI	6	89 vs 52

Table 15: Swing State Clustering Scheme by Campaign: Speech counts shown as swing vs. non-swing states.

Metric	Clustering Scheme	2016 Campaign		2020 Campaign		2024 Campaign	
		Swing	Non-Swing	Swing	Non-Swing	Swing	Non-Swing
PDI	Ballotpedia	6.269 (n=123)	6.714 (n=10)	2.350 (n=82)	2.933 (n=16)	1.226 (n=73)	1.804 (n=68)
	High Attention	5.845 (n=81)	7.015 (n=52)	2.379 (n=68)	2.594 (n=30)	1.431 (n=89)	1.631 (n=52)

Table 16: Descriptive Statistics: Mean PDI scores by Campaign, State Type, and Clustering Scheme. Sample sizes (n) in parentheses.