POPOLARE: A Populism and Polarization Classification Framework for Italian Texts

Anonymized for double blind review[1] [0000−1111−2222−3333]

Anonymized
anon@anon.org

Abstract. The influence of political discourse in forming public opinion has intensified the need for tools to capture complex ideological patterns. This requires innovative, data-driven approaches to analyze and interpret political language with both precision and transparency. This paper presents POPOLARE, a two-fold populism and political polarization framework for Italian political speeches based on Natural Language Processing and Machine Learning. Individual transcripts are transformed into textual document representations and then aggregated to derive representations for each speaker. Based on these representations, populism and polarization classification tasks are performed. A key novelty lies in the use of Generative AI for data annotation, and explainability techniques for model interpretation. Results show that simple models combined with lexical representations perform best, and that interpretable features enhance both accuracy and transparency. POPOLARE provides a replicable approach for ideological analysis, with future directions including multilingual extension and deeper use of explainable AI.

Keywords: Populism · Polarization · Natural Language Processing · Explainability.

1 Introduction

The intersection of political science and Artificial Intelligence (AI) presents a dynamic frontier for computational social science. As political communication moves increasingly online, vast textual corpora–from parliamentary transcripts to tweets–become available for automated analysis. At the same time, Natural Language Processing (NLP) and Machine Learning (ML) technologies have reached high levels of efficiency that allow for accurate text classification, and meaningful linguistic, semantic, and emotional analysis.

Among the most critical challenges facing democracies today are populism and political polarization. Populism typically employs a binary narrative, e.g., "the pure people" versus "the corrupt elite"; while polarization refers to the intensification of ideological divides that erode common ground and democratic dialogue. Understanding how these phenomena are embedded in language, and how they vary across individuals, institutions, and time, is essential to both political theory and practice.

Studies in this area are limited, both in scope and number, and face two main challenges. First, the lack of ground truth. Classification approaches are most often supervised, and as nuanced and relatively recent tasks, populism and polarization detection is lacking in large and established benchmark datasets on which to train such models. Second, the opacity and complexity of models. Current NLP approaches are largely based on Large Language Models (LLM), whose behavior is still largely uninterpretable. This limitation is particularly detrimental due to the subtelty of the task, which may often require human supervision and intervention. Relying on "black-box" models who act as oracles without providing proper explanations of their classification, and thus information for recourse, is in direct contrast with the proper use of models in a healthy online discourse.

This paper proposes POPOLARE, (POpulism and POLARization Extraction) a computational approach to identifying and quantifying populist and polarized language in political texts focusing on *a*) how populism and political polarization can be automatically detected and measured in political texts, and *b*) what linguistic, semantic, and emotional features are most strongly associated with populism and political polarization. POPOLARE exploits an NLP pipeline for processing and classifying political texts in terms of their populist and polarized content, and enriches this analysis by unveiling features characterizing populist and polarized text, both at a token-level and at a level.

2 State of the Art

This section outlines foundational research at the intersection of political science and AI, a dynamic, interdisciplinary field. Growing interest in political populism and polarization aligns with recent advances in NLP, particularly with LLMs, enabling large-scale text analysis.

Populism and Political Polarization. In [19], a review of 154 NLP studies on political polarization is presented, identifying ideological scaling and supervised classification as primary methods. Challenges in generalization, the prevalence of supervised learning with diverse labeling and features (Bag-of-Words, n-grams, embeddings), and frequent application of topic modeling and sentiment analysis were highlighted. The study emphasized the importance of interpretable models for transparency and trustworthiness in politically sensitive contexts. The authors in [6] reviewed 61 studies on political polarization and sentiment analysis in parliamentary debates, focusing on ideology detection, polarization measurement, and position scaling. The review noted variable transcription granularity and identified five main methods: dictionary-based, statistical machine learning, rule-based, similarity comparison, and word frequency analysis. Most studies utilized Bag-of-Words, some employed embeddings, often with metadata. In [18], populism is defined, based on [3], as a dichotomy between "the pure people" and "the corrupt elite". The authors identified key linguistic features of populist discourse. The employed methodology involves analyzing Facebook posts using text mining, topic modeling (lemmatization, Bag-of-Words, LDA), and sentiment

analysis to detect populist communication and characterize its emotional tone. The authors in [13] analyzed emotional dynamics in Italian populist communication, finding populist rhetoric to be more emotionally charged than non-populist discourse, with distinct emotional strategies across right-wing, left-wing, and hybrid populist parties. This work highlights the importance of emotion analysis in modeling populism and polarization. Methodologically, it used manually labeled data and a Random Forest classifier, leveraging the expert-coded PopuList dataset [22] for ground-truth labels in a supervised learning framework. The authors in [8] employed the PopuList alongside CHES [9] and POPPA [16, 17] expert surveys to model populism as a continuous variable using low-cost machine learning methods. These expert-coded datasets provide essential ground-truth for supervised learning, as also seen in [1]. Finally, [23] compared BERT and GPT for political text classification, finding BERT to maintain a slight advantage in usability and computational efficiency despite GPT's in-context learning capabilities.

Explainability. Explainability has found ample use in NLP models [12], often in the form of token importance, which assigns a weight to each input token estimating its contribution to the model prediction [14]. Algorithms extracting such explanations often enjoy two properties: they are extracted *post-hoc* from a trained model, which allows for a plug-and-play use; and are often *model-agnostic*, thus can be applied to any model regardless of its architecture. LIME [21], which proposes a model distillation algorithm, and SHAP [11], which instead relies on sensitivity analysis, are the most prominent such algorithms. LIME generates synthetic data, labels it with the model of interest, and then learns an interpretable linear model. The linear model, which effectively acts as a surrogate model for the model of interest, offers feature importance by design, thus allowing explanation of the original model by proxy. SHAP instead estimates feature importance by quantifying the rate of change of the model's predictions on a large set of data perturbations: the larger the change in model predictions when a feature is perturbed, the higher the importance of the feature.

As model-agnostic algorithms, LIME and SHAP operate *observationally* on the data, thus treating the model as a black box to probe. This approach is severely limited, as it relies on input perturbation, which is prone to out of distribution sampling [7], and only provides a surface-level understanding of the model. Probing [2] provides a deeper-level understanding by addressing the latent features learned by a model. Rather than create surrogate models on the data itself like in LIME, probing instead learns a surrogate model, named *probe*, on the *latent* representation, thus aiming to understand the latent instead of the surface representation. Labelling is another key difference in probing: probes are trained to predict high-level features of the text, rather than the model label, thus they are trained as detectors of such features. A probe with high performance on a high-level feature is a strong indicator that the model has indeed learned such feature. Abstract concepts [10], semantic roles [4], basic logical circuits [15], or sentiment [5], which are not explicit in the text, can instead be estimated in the latent representations. This approach is particularly effective for language

models, which have state of the art performance on a plethora of diverse tasks, each requiring different understandings of the text.

3 Methodology

This section describes POPOLARE, a two-level framework for populism and political polarization classification at both document- and speaker-level, incorporating XAI methods to enhance result interpretability. Formally, let $D = \{d_1, \ldots, d_n\}$ be a set of n political speeches, where $d_{i,j} \in D_i$ is the j^{th} comment of speaker i, where D_i is the set of their speeches. Let $S = \{s_1, \ldots, s_m\}$ be a set of m speakers belonging to a set P of political parties. The set of features $\mathcal{F}(d_{i,j})$ is extracted from each document $d_{i,j} \in D_i$. These features are then aggregated using an aggregation function $\mathcal{A}$ (i.e., mean) to represent the speakers in terms of their speeches. While widely discussed in the literature, populism and political polarization lack universally accepted definitions. Given the diversity of interpretations, this work adopts the definitions provided in Definitions 1 and 2.

Definition 1. *[Definition of Populism] Populism is the set of ideas and attitudes centered on the belief that society is fundamentally divided into two antagonistic groups: the ordinary people and a corrupt or detached elite. It emphasizes the primacy of the popular will, often expressing discontent with established institutions and political leadership. Populism can manifest across the political spectrum, appearing in both left-wing and right-wing movements.*

Definition 2. *[Definition of (Political) Polarization] Polarization is defined as the set of processes and attitudes that contribute to an increasing divide between opposing political camps, particularly along the left-right spectrum. It involves the growing extremization and distancing of opinions, beliefs, and identities, often diminishing opportunities for compromise and intensifying political conflict.*

Following Definitions 1 and 2, populism and polarization are treated as conceptually distinct and independent phenomena. Accordingly, POPOLARE addresses them separately through the Document- and Speaker-level Populism Evaluation (Definition 3) and Polarization Evaluation (Definition 4) problems.

Definition 3 (Document and Speaker-level Populism Evaluation). *The Document-level Populism Evaluation problem consists of learning a classification model f that, given an unlabeled document $d_i \in D$, predicts its populism labels $f(d_i) \in \{0, 1\}$ with 0 for non-populist, and 1 for populist. The Speaker-level Populism Evaluation problem consists of learning both a classification f_c and a regression f_r model that – based on the aggregated features of speakers' speeches – given an unlabeled speaker $s_i \in S$, predicts its populism levels $f_c(s_i) \in \{0, 1\}$ and $f_r(s_i) \in [0, 4]$.*

Definition 4 (Document and Speaker-level Polarization Evaluation). *The Document-level Polarization Evaluation problem consists of learning a classification model f that, given an unlabeled document $d_i \in D$, predicts its (political) polarization labels $f(d_i) \in \{Left, Center, Right\}$. The Speaker-level Polarization Evaluation problem consists of learning both a classification f_c and a*

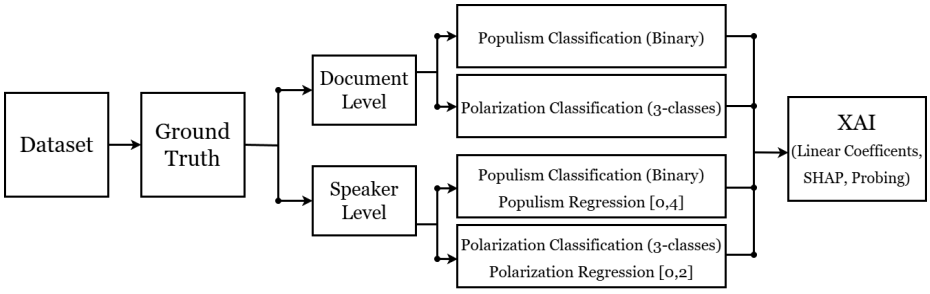


Fig. 1: General Methodology Schema.

regression f_r. model that - based on the aggregated features of speakers' speeches – given an unlabeled speaker $s_i \in S$ – predicts its (political) polarization level $f_c(s_i) \in \{Left, Center, Right\}$ and $f_r(s_i) \in [0, 2]$.

Based on Definitions 3 and 4 we implement the POPOLARE framework following the pipeline shown in Figure 1.

Ground Truth Annotation via Generative AI. Ground truth labels are fundamental for implementing supervised learning methods. Due to this, POPOLARE requires that each text is labeled according to its degree of populism and polarization for the training phase. Since text-level political and populist labels often lack, POPOLARE proposes a two-level annotation strategy. First, a randomly selected, party-representative sample of texts is manually annotated using a numerical left-right scale to measure polarization, along with a set of binary-labeled features that are later aggregated to assess populism. Then, the annotated texts are used to prompt a Large Language Model (LLM), such as Gemini, to automatically annotate the rest of the dataset. Manual annotations provides the ground truth for evaluating model performance. If accuracy is adequate, e.g., $>= 0.80$, the LLM is used to label the full dataset.

Textual Representations. While raw text can be directly processed by transformer models, traditional ML methods require structured input. To support both, POPOLARE allows for multiple textual representations that encode different linguistic or semantic dimensions. This enables both comparative evaluation of feature relevance for modeling populism and polarization, and systematic optimization across model–representation configurations. Textual representations — such as TF-IDF (Term Frequency- Inverse Document Frequency) vectorization, numerical features describing grammatical and structural traits, and document embeddings — can be used independently or combined to exploit a broader view of text. Furthermore, these can be enriched with additional information such as affective dimensions, capturing emotional or evaluative content.

Modeling. POPOLARE provides a two-level modeling pipeline to analyze populism and polarization allowing to test different combinations of algorithms and text representations, but differ in analytical granularity. At the document-level, POPOLARE allows classifying texts based on their degree of populism and

political polarization. Furthermore, it supports higher-level classification via aggregation of text-level predictions across political actors, i.e, speakers, and organizations, such as parties, coalitions.

Document-Level. At the document level, populism classification is framed as a binary task, i.e., populist vs. non-populist, while political polarization is modeled as a three-class problem, i.e., left, center, right.

Speaker/Organization-Level. At a higher level, document-level predictions are aggregated to infer the populist and polarized orientation of entities such as parties, politicians, or social media profiles. Aggregation methods include averaging and standard deviation of features and predicted scores. For transformer-based models, concatenating all texts into a single input is possible, though limited by the context window of the model. At the speaker/organization level, populism is modeled both as a binary classification task (populist vs. non-populist) and as a regression task with continuous outputs in the range $[0, 4]$. Similarly, polarization is modeled as a three-class classification task (left, center, right) and as a regression task with continuous outputs in the range $[0, 2]$.

Explainability. Model explainability is a central component of POPOLARE as the significance of modeling populism and polarization lies not only in accurate prediction, but also in understanding the underlying drivers of these phenomena. Due to the multifaceted nature of polarization and populism, POPOLARE provides a plethora of explanations, each addressing a different facet of the data.

- *Global explanations.* Global explanations provide a global view of the model, allowing POPOLARE to peer into the general behavior of the model. Linear models provide out-of-the-box global explanations in the form of feature relevance: the weights associated to each feature encode, by design, the contribution that said feature has on the model's prediction. This is particularly effective when the model leverages interpretable features with a clear semantic meaning, as it is the case of TF-IDF, grammatical, structural, and sentiment features.
- *Local explanations.* Unlike global explanations, local explanations provide a fine-grained understanding of the model, allowing their user to observe the behavior of the model in a simpler setting. POPOLARE offers post-hoc model-agnostic feature relevance with SHAP.
- *Latent features induction.* The previous families only tackle observed features, e.g., the importance of a given token in a text, while ignoring features encoded in the latent representation of the model. POPOLARE leverages probing to induce hidden features learned by the model.

Together, these techniques form a flexible interpretability toolkit. They differ in assumptions, implementations, and objective, making it possible to tailor the interpretability process to the nature of the model and the representation used. This layered and flexible approach enables both high-level and fine-grained insights into how models detect and classify populist and polarizing discourse, thereby advancing both academic inquiry and the potential for applied impact.

Table 1: Gemini evaluation for populism and political polarization features.

Feature	Precision	Recall	F1	r	MSE
Manichean	**0.91**	**0.89**	**0.90**	0.80	**0.10**
People-centrism	0.80	0.80	0.80	0.60	0.20
Anti-elitism	**0.91**	**0.89**	**0.90**	0.80	**0.10**
Emotional appeal	0.88	0.85	0.86	0.73	0.12
Polarization	0.86	0.83	0.83	**0.81**	0.18

4 Case Study

We present a case study on Italian political speeches. Most works focus on the English language, thus in this case study we focus on an understudied language. Similarly, the literature focuses on short, informal texts, e.g., social media posts, while our case study focuses on long speeches given by politicians in a more formal context.

4.1 Dataset

Our case study is on the Italian subset of ParlaMint 4.1[1], a multilingual corpus of European parliamentary debates. The Italian corpus consists of transcribed Senate plenary sessions from 2013 to 2022. Starting from 172,296 speeches, several filtering steps were applied to exclude irrelevant content—such as procedural remarks and interventions from marginal parties. To ensure analytical robustness, only parties with at least 2,000 speeches were retained. After additional filtering based on text length, the final dataset (`ParlaMint-IT`) includes 10,840 speeches. After preprocessing, the dataset's length distribution—initially skewed toward short texts—was balanced around 7,500 characters. Speech-length variability across parties was reduced, and party representation normalized by merging affiliated groups and resolving naming inconsistencies. Despite residual imbalance, the cleaned dataset improves cross-party comparability by ensuring sufficient and consistent representation across the political orientations.

4.2 Data Annotation with Gemini

Following the literature[2], POPOLARE extracts a set of populism-related features. These capture key dimensions of populist rhetoric, and are implemented as indicator variables indicating the nature of each text:

- *Manicheism*: frames politics as a moral conflict between good and evil, with clear enemies.
- *People-Centrism*: emphasizes the sovereignty of ordinary citizens.

[1] `https://github.com/clarin-eric/ParlaMint`.
[2] Populism and Political Parties Expert Survey: `https://poppa-data.eu/`

– *Anti-Elitism*: depicts elites as corrupt or disconnected from "the people".
– *Emotional Appeal*: relys on emotion-driven language over rational argument.

Furthermore, political polarization is annotated along the Left–Right axis using three classes, i.e., Left, Center, Right. Due to the unavailability of such labels in ParliaMint-IT, POPOLARE labels them with Gemini [20], a LLM well-suited for few-shot in-context learning. The model receives three labeled examples as prompts and classifies new inputs accordingly. A validation set of 53 manually labeled texts spanning multiple parties and time periods was used, with 3 for prompting and 50 for evaluation. Results (Table 1) — in terms of Precision, Recall, F1-score, Mean Squared Error (MSE), and Spearman's rank correlation (r) — show that Gemini performs reliably with binary labels for populism and three-class labels for polarization, achieving strong performance on the four populism features ($F1 > .8$, $r \in [.6, .8]$). The obtained labels provide the ground truth for training models to classify both political polarization and populism.

4.3 Textual Representations & Feature Extraction

To accommodate models such as SVM and LightGBM, which are not designed for raw text, we preprocess the raw speech transcriptions, and encode them in four types of textual representations, each capturing different linguistic or semantic features:

– TF-IDF: Each document is encoded using corpus-level TF-IDF values. Given the high dimensionality of the resulting feature vectors, dimensionality reduction is applied via Singular Value Decomposition (SVD), compressing each vector to 300 components.
– Document Embeddings: We tested two variants of document embeddings using pretrained Word2Vec embeddings derived from the itWaC corpus[3], which comprises about one billion Italian web-domain words:
  • Standard Doc Embedding (*Doc*): Formed by averaging the embeddings of all lemmatized words in a document.
  • POS-Filtered Doc Embedding (*PosDoc*): Created by averaging only embeddings corresponding to semantically rich parts of speech, i.e., nouns, verbs, and adjectives.
– Linguistic Profiles (*LGT*): Each text is represented by a 144-dimensional vector generated through Profiling-UD[4], which extracts interpretable linguistic features related to raw text characteristics, lexical diversity, morphosyntactic patterns, and syntactic structures. Unlike previous representations, here features retain semantic, and thus interpretable, value. Each feature captures some fine-grained linguistic information not explicitly encoded in standard embedding spaces.

[3] `https://www.lilec.it/lisa/itwac/`.
[4] `http://www.italianlp.it/demo/profiling-ud/`.

Table 2: Polarization task: document-level performance of models on different extended text representations. Dataset with even class distribution.

Representation	SVM	RF	LightGBM
$TF-IDF^{+}$	**0.71**	0.59	0.65
Doc^{+}	0.57	0.57	**0.61**
$PosDoc^{+}$	0.58	0.57	**0.60**
LGT^{+}	0.41	0.52	**0.54**

In addition to textual representation, we extract sentiment and emotional information, and lexical indicators. Specifically, *i)* a $1-5$ sentiment strength score, computed with a pre-trained multilingual BERT model; *ii)* 28 indicators of the emotional richness and variability of the text, computed with a pre-trained RoBERTa model; and *iii)* ranked TF-IDF scores within each political alignment, i.e., Left, Center, Right, to obtain salient words which are one-hot encoded binary vectors indicating the presence or absence of these ideological markers. Models are trained either on *base* representations, which do not include these additional features, or on *enriched* representations, which instead do. We indicate with $TF{-}IDF^{+}$ (respectively, Doc^{+}, $PosDoc^{+}$, LGT^{+}) the enriched representations.

4.4 Document-Level Modeling

To model populism and political polarization at the document-level we tested three machine learning algorithms, i.e., linear Support Vector Machines (SVM), Light Gradient Boosting Machine (LightGBM), and Random Forest (RF).[5] Model selection was performed with 5-fold cross-validation strategy. In addition to traditional ML models, we also employ a fine-tuned transformer-based model[6] trained on a development set and then tested on held-out data, comprising 25% of the original dataset.

Political polarization is modeled as a single-label multi-class ordinal classification task with three classes, corresponding to increasing levels of ideological extremity. On the other hand, for modeling populism each document is analyzed for the presence of the four binary populist traits, i.e., Manichean, People-centrism, Anti-elitism, and Emotional appeal, individually. Additionaly, another label is assigned to each document: 1 for documents with at least 2 populist traits, and 0 for others. We name this class *Threshold index*.

Political Polarization. To mitigate class imbalance[7], we apply random undersampling to downsample each class to $1,651$ instances, and then train models using extended representations as inputs. As shown in Table 2, a linear SVM on a TF-IDF representation yields the best performance (0.71), suggesting that linear models can leverage sparse, high-dimensional representations effectively.

[5] Tree models trained on maximum depth $\in [3, 15]$, SVM on a slack weight of 1.

[6] `dbmdz/bert-base-italian-xxl-cased`

[7] Class 0 (Left): $1,651$ instances, Class 1 (Center): $4,957$), and Class 2 (Right): $4,232$.

Table 3: Opaque vs Interpretable models on polarization. Performance of a linear SVM, trained on a TF-IDF representation, and a fine-tuned BERT model. Reported are Precision (P), Recall (R), F1 score, and Accuracy (A) of all classes, and their macro average

	SVM				**BERT**			
Class	**P**	**R**	**F1**	**A**	**P**	**R**	**F1**	**A**
Left	0.72	0.71	0.72		0.64	0.62	0.62	
Center	0.66	0.63	0.65		0.76	0.74	0.75	
Right	0.73	0.77	0.75		0.75	0.79	0.77	
macro avg	0.71	0.71	0.71	0.71	0.72	0.71	0.72	0.74

Table 4: Document-level classification of populism using the threshold index on different representations (base, enriched, and raw text). Performances measured as macro average of the F1-scores for each class.

Representation	*Base*			*Enriched*			*Raw*
Model	**SVM**	**RF**	**LightGBM**	**SVM**	**RF**	**LightGBM**	**BERT**
$TF-IDF$	**0.81**	0.71	0.79	**0.80**	0.74	0.79	
Doc	0.72	0.77	**0.78**	0.74	0.77	**0.79**	
$PosDoc$	0.75	**0.77**	**0.77**	0.76	0.77	**0.78**	
LGT	0.61	0.72	**0.74**	0.65	0.74	**0.76**	
Raw							0.78
Average	0.74 ± 0.05			0.75 ± 0.03			0.78 ± 0.0

Doc^+ and $PosDoc^+$ show lower performance, likely due to their higher semantic abstraction but lower local detail compared to TF-IDF. LGT^+ results in lower performance, indicating that such features are either less discriminative for the task and not fully exploitable by the models. Moreover, LightGBM and Random Forest tend to perform better on denser data or with less sparse feature sets, while SVMs works better with traditional text representations like TF-IDF.

Focusing on the SVM model (Table 3), it shows balanced performances across classes, with F1-scores of 0.72 (Left), 0.65 (Center), and 0.75 (Right). Interestingly, this is on par with the much more complex fine-tuned BERT model, which achieved a slightly higher macro F1-score of 0.72, driven mainly by better precision and recall for the Center class (0.76 and 0.74, respectively). The performance gap between the two models is marginal and class-dependent. While BERT outperforms the SVM on the Center class, the SVM achieves better scores on the Left class. These results suggest that, with appropriate feature engineering and data balancing, traditional models like SVMs can rival transformer-based approaches even in such a nuanced tasks.

Populism. For populism, we leverage the threshold index, and classify texts as populist if they contain at least 2 populist traits. The results in Table 4 sum-

Table 5: Polarization and populism performances on speaker level, measured as macro average of the F1 scores for each class. Comparison of different representations.

	Polarization			Populism		
Model	**SVM**	**RF**	**LightGBM**	**SVM**	**RF**	**LightGBM**
$TF-IDF^{+}$	**0.74**	0.64	0.73	**0.83**	0.81	**0.83**
Doc^{+}	0.62	0.62	**0.63**	0.81	0.82	**0.83**
$PosDoc^{+}$	0.61	0.62	**0.63**	0.81	**0.82**	0.81
LGT^{+}	0.43	0.59	**0.60**	0.65	0.79	**0.81**

marize results of models using different text representations. Performances are comparable across representations, above 0.65(with one exception) and reaching up to 0.81. Variance across representation is also low, at around 10^{-3} for both the base and enriched representation. The SVM model trained on a TF-IDF representation achieved the highest F1-score of 0.81. It's also worth noting that LightGBM consistently delivered strong results across various representations, often in the high 0.70s. In the same table, we compare these interpretable models with an opaque BERT model fine-tuned on the raw text repreentation. BERT's performance remains comparable to that of the traditional classifiers, reaffirming the earlier observations made in the context of polarization modeling. Specifically, BERT achieved an overall macro averaged F1-score of 0.78. The overall performance aligns well with the best results from traditional models, especially when considering the range of F1-scores around $0.70 - 0.80$ observed previously.

4.5 Speaker-Level Modeling

POPOLARE provides a higher-level text analysis by characterizing speakers' overall rhetorical style and ideological positioning, based on the aggregation of their speeches. Speaker-level features are derived by average and standard deviation of the document-level representations of all their respective speeches. This allows POPOLARE to capture both central tendencies and intra-speaker variance. The populism label, aggregated to a real value $\in [0, 4]$ is discretized for populism-classification tasks (average values of ≤ 1 given class 0, the remaining given class 1), and used as-is for regression tasks. The polarization label, aggregated to a real value $\in [0, 2]$, is instead categorized into three classes at increasing thresholds $[0.75, 1.45]$. The per-speaker aggregation also allows POPOLARE to infer a speaker's political party and ideological orientation.

Table 5 presents the best models for both polarization and populism, on all representations. While LightGBM performed best in three out of four text representation cases, SVM with $TF-IDF^{+}$ representation achieved the highest overall performance for both tasks. The performance difference between these models is minimal, particularly for populism, making SVM the preferable choice due to its simplicity and interpretability. Notably, all models achieved their best

Table 6: Party Affiliation and Political Orientation by Speaker.

(a) Party

Party	Precision	Recall	F1
FI-BP	0.72	0.70	0.72
Lega	0.88	0.90	0.89
M5S	**0.91**	**0.92**	**0.91**
PD	0.85	0.91	0.88
PdL	0.69	0.46	0.55
macro avg	0.81	0.78	0.79

(b) Political orientation

Orientation	Precision	Recall	F1
Cdx	0.92	0.81	0.86
Csx	0.88	0.92	0.90
Dx to Xdx	**0.93**	**0.95**	**0.94**
Pigliatutto	**0.93**	0.93	0.93
macro avg	0.91	0.90	0.91

results with the $TF - IDF^{+}$ representation, underscoring its effectiveness for this task.

4.6 Party Affiliation and Political Orientation

Our analyses consistently show the effectiveness of the TF-IDF representation. While its results are occasionally comparable to other models, its efficiency and interpretability make it the preferred choice for practical applications. Due to this, following experiments focusing on Party affiliation and Political orientation classification employs only the TF-IDF representation and the linear SVM. To address the limited number of examples in certain classes, we excluded parties with fewer than 47 instances. As shown in Table 6a, all parties achieve a F1-score up to 0.70, except PdL, that also shown lowest Recall value (0.46). In contrast, all the other parties achieve both higher precision and recall values, with M5S obtaining the best performance (0.91), suggesting some re-occuring unique speech patterns.

If lines between party affiliations appear to be silghtly blurred, political orientation shows easier identification – see Table 6b. The orientation with lowest F1 score still has a F1-score of 0.86, suggesting that while some speech patterns are somewhat unique among parties, they are even more so among political orientations.

4.7 Explainability

The explainability framework adopted in this study is based on three complementary approaches: coefficient analysis of linear models, SHAP explanations, and probing classifiers, all tackling the polarization task. The first two analyzed a linear SVM model, while the third a BERT model. Each of these methods contributes to a deeper understanding of the decision-making processes of the models on a document-level and a speaker-level.

Feature	Importance
presidente_avg	0.66
stato_avg	0.39
stars_std	0.22
articolo_avg	-0.29
legge_avg	-0.39
quando_avg	-0.58

Table 7: Left

Feature	Importance
quando_avg	0.65
decreto_avg	0.33
legge_avg	0.31
politica_avg	-0.33
fatto_avg	-0.52
signor_std	-0.55

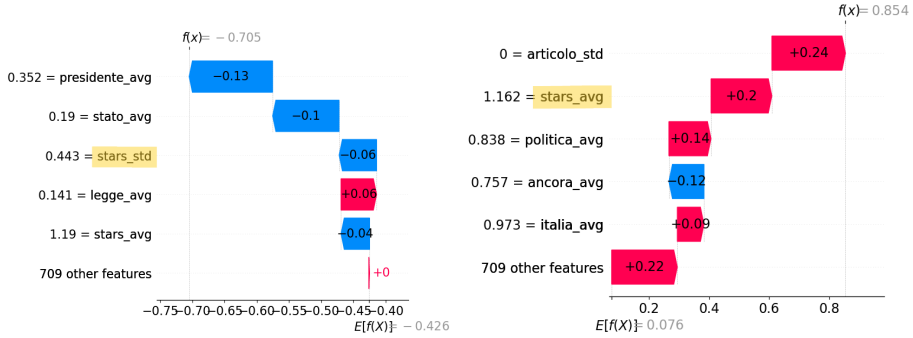
Table 8: Center

Feature	Importance
quando_avg	0.54
fatto_avg	0.51
molto_avg	0.35
decreto_avg	-0.36
stato_avg	-0.38
articolo_std	-0.93

Table 9: Right

Table 10: Features with higher and lower linear coefficients for the SVM model in the speaker level for polarization classification.

Fig. 2: Local feature importance for a correctly (left) and incorrectly (right) classified instance on a polarization task.



Coefficient Analysis. Coefficient analysis was run on a linear SVM model classifying political affiliation at a speaker level – Table 10 reports the top-3 most influential features, both in positive and negative direction. The model shows a bizzare behavior: expectedly inconsequential adverbs ("quando", "fatto") show high importance, but are accompanied by more characteristic nouns, e.g., "stato" (state), "presidente" (president), and "legge" (legal clause).

SHAP Explanations. To analyze individual predictions, SHAP values were computed for both correct and incorrect classifications – see Figure 2.

In the case of a positive classification, the model has put high negative importance on specific tokens: "presidente" (president) and "stato" (state), which indicate that in this instance, these two expressions tend to lower the polarization score. The model has, in this case, learned that speeches often mentioning high super-partes governmental authorities are indicators of low polarization. This further confirms that the model has learned to recognize traits such as people-centrism, proper of populistic and polarization traits, and leverage them. Other features show, in proportion, a negligible or null importance.

In the opposite case of a misclassification, instead the model shows higher confusion: many features show moderate to high importance, and the most im-

Table 11: Probe performance on the BERT model fine-tuned for populism classification at the document level.

Populistic Feature	Precision	Recall	F1
Manicheism	**0.83**	**0.83**	**0.83**
People-centrism	0.80	0.81	0.80
Anti-elitism	0.82	0.82	0.82
Emotional appeal	0.79	0.79	0.79

portant are generic, e.g., "ancora" (yet), "articolo" (clause), "Italia" (Italy), etc. While lexical features and sentiment-based metrics are often influential, interpretation remains challenging due to the nuanced and context-dependent nature of their contributions.

Probing Classifiers. Probing methods interrogate the internal representations of a fine-tuned document-level BERT model for polarization classification. The first probe is trained to detect the presence of *populistic* speech patterns, namely the relative frequency of tokens characterizing populistic speeches. The probe achieves a moderate performance of 0.74 F1-score, suggesting that while the model may have learned populistic patterns, they are not central to polarization. Moving to document-level populism, BERT shows instead to have learned to recognize the 4 high-level traits characterizing populistic discourse, with probes on manicheism, people-centrism, anti-elitism, and emotional appeal showing high performance, as shown in Table 11.

In conclusion, the multi-faceted explainability strategy adopted in this study provides converging evidence that interpretable, lexicon-based features are more reliable and robust at the speaker level, while latent features derived from dimensionality reduction dominate at the document level. Transformer-based models, though less transparent, encode rich latent representations that align with higher-level discourse characteristics and can be effectively probed for fine-grained traits.

5 Conclusions

The increasing availability of political discourse data and advancements in NLP offer unprecedented opportunities to investigate political phenomena like populism and polarization. This paper introduced the POPOLARE methodology, designed to automatically detect and measure populism and political polarization in texts, and to identify their influential linguistic, semantic, and emotional features. POPOLARE features an innovative annotation strategy using Generative AI for efficient ground truth generation, and employs distinct modeling approaches for document and speaker levels. At the document level, populism is modeled via binary classification (populist/non-populist), while polarization

uses a three-class classifier (Left/Center/Right). At the speaker level, aggregated representations enable binary classification for populist speakers and regression for continuous populism (0-4) and polarization (0-2) scores.

POPOLARE natively includes several explainability techniques (local and global feature importance, probing) to provide transparent insights into model predictions. A key methodological contribution lies in integrating diverse feature types and emphasizing model interpretability, aligning with the interpretive needs of political science research.

Experiments revealed several important findings. TF-IDF representations paired with simple models like linear SVMs frequently outperformed more complex models on both populism and polarization. LightGBM also showed consistent strong performance. Interpretable features (linguistic, emotional) maintained predictive power post-aggregation, while high-dimensional semantic embeddings tended to lose importance, likely due to noise reduction. Contrary to expectations, transformer-based models (e.g., BERT) did not consistently outperform classical algorithms. This highlights that for domain-specific contexts with limited data, traditional models with robust preprocessing and engineered features can match or exceed complex architectures.

In conclusion, the POPOLARE framework provides a robust method to effectively model populism and polarization in complex political corpora, balancing performance, interpretability, and representation richness. Speaker-level aggregation, notably, enhances model stability, pointing towards a promising direction for future research focused on political actors.

References

1. A., C., M., R., I., R.C.: Will the real populists please stand up? a machine learning index of party populism. European Journal of Political Economy **82** (2024). https://doi.org/https://doi.org/10.1016/j.ejpoleco.2024.102529
2. Alain, G., Bengio, Y.: Understanding intermediate layers using linear classifier probes. In: ICLR (Workshop). OpenReview.net (2017)
3. C., M.: The populist zeitgeist. Government and Opposition **39**(4) (2004)
4. Conia, S., Navigli, R.: Probing for predicate argument structures in pretrained language models. In: ACL (2022)
5. Di Palma, D., De Bellis, A., Servedio, G., Anelli, V.W., Narducci, F., Di Noia, T.: Llamas have feelings too: Unveiling sentiment and emotion representations in llama models through probing. arXiv preprint arXiv:2505.16491 (2025)
6. G., A., R., B.: Sentiment and position-taking analysis of parliamentary debates: a systematic literature review. Journal of Computational Social Science **3**, 245–270 (2020). https://doi.org/https://doi.org/10.1007/s42001-019-00060-w
7. Hase, P., Xie, H., Bansal, M.: The out-of-distribution problem in explainability and search methods for feature importance explanations. In: NeurIPS (2021)
8. J., D.C., B., M.: How populist are parties? measuring degrees of populism in party manifestos using supervised machine learning. Political Analysis **30** (2022)
9. Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., Vachudova, M.A.: Chapel hill expert survey trend file, 1999–2019. Electoral Studies **75** (February 2022)

10. Jullien, M., Valentino, M., Freitas, A.: Do transformers encode a foundational ontology? probing abstract classes in natural language. CoRR **abs/2201.10262** (2022)
11. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
12. Lyu, Q., Apidianaki, M., Callison-Burch, C.: Towards faithful model explanation in nlp: A survey. Computational Linguistics **50**(2), 657–723 (2024)
13. M, C., J., D.C.: Populism and emotions: a comparative study using machine learning. Italian Political Science Review **53**, 351–366 (2023)
14. Mandler, H., Weigand, B.: A review and benchmark of feature importance methods for neural networks. ACM Computing Surveys **56**(12), 1–30 (2024)
15. Manigrasso, F., Schouten, S.F., Morra, L., Bloem, P.: Probing llms for logical reasoning. Lecture Notes in Computer Science, vol. 14979. Springer (2024)
16. Meijers, M., Zaslove, A.: Populism and political parties expert survey 2018 (poppa) (2020). https://doi.org/10.7910/DVN/8NEL7B
17. Meijers, M.J., Zaslove, A.: Measuring populism in political parties: Appraisal of a new approach. Comparative Political Studies **54**(2), 372–407 (2021)
18. N., H.: A natural language processing based text analysis of populist rhetoric in social media text messages. LingUU Journal **3(2)**, 57–72 (2019)
19. R., N.: A scoping review on the use of natural language processing in research on political polarization: trends and research prospects. J Comput Soc Sc **6**, 289–313 (2023). https://doi.org/https://doi.org/10.1007/s42001-022-00196-2
20. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T.P., Alayrac, J., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., Antonoglou, I., Anil, R., Borgeaud, S., Dai, A.M., Millican, K., Dyer, E., Glaese, M., Sottiaux, T., Lee, B., Viola, F., Reynolds, M., Xu, Y., Molloy, J., Chen, J., Isard, M., Barham, P., Hennigan, T., McIlroy, R., Johnson, M., Schalkwyk, J., Collins, E., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Meyer, C., Thornton, G., Yang, Z., Michalewski, H., Abbas, Z., Schucher, N., Anand, A., Ives, R., Keeling, J., Lenc, K., Haykal, S., Shakeri, S., Shyam, P., Chowdhery, A., Ring, R., Spencer, S., Sezener, E., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. CoRR **abs/2403.05530** (2024)
21. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
22. Rooduijn, M., Pirro, A.L., Halikiopoulou, D., Froio, C., van Kessel, S., de Lange, S.L., Mudde, C., Taggart, P.: The populist: A database of populist, far-left, and far-right parties using expert-informed qualitative comparative classification (eiqcc). British Journal of Political Science pp. 1–10 (2023)
23. Y., W., W., Q., X., Y.: Selecting between bert and gpt for text classification in political science research (2024), `https://arxiv.org/abs/2411.05050`