PARTISANLENS: A Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media

Michele Joshua Maggini[1]*, Paloma Piot[2]*, Anxo Pérez[2], Erik Bran Marino[3], Lúa Santamaría Montesinos[4], Ana Lisboa[5], Marta Vázquez Abuín[1], Javier Parapar [2], Pablo Gamallo[1],
[1]Centro Singular de Investigación en Tecnoloxías Intelixentes da USC,
[2]IRLab, CITIC Research Centre, Universidade da Coruña, [3]Universidade de Évora,
[4]Universidad de La Rioja, [5]GESIS Leibniz Institute for the Social Sciences

Correspondence: michelejoshua.maggini@usc.es and paloma.piot@udc.es

Abstract

Detecting hyperpartisan narratives and Population Replacement Conspiracy Theories (PRCT) is essential to addressing the spread of misinformation. These complex narratives pose a significant threat, as hyperpartisanship drives political polarisation and institutional distrust, while PRCTs directly motivate real-world extremist violence, making their identification critical for social cohesion and public safety. However, existing resources are scarce, predominantly English-centric, and often analyse hyperpartisanship, stance, and rhetorical bias in isolation rather than as interrelated aspects of political discourse. To bridge this gap, we introduce PARTISANLENS, the first multilingual dataset of 1617 hyperpartisan news headlines in Spanish, Italian, and Portuguese, annotated in multiple political discourse aspects. We first evaluate the classification performance of widely used Large Language Models (LLMs) on this dataset, establishing robust baselines for the classification of hyperpartisan and PRCT narratives. In addition, we assess the viability of using LLMs as automatic annotators for this task, analysing their ability to approximate human annotation. Results highlight both their potential and current limitations. Next, moving beyond standard judgments, we explore whether LLMs can emulate human annotation patterns by conditioning them on socio-economic and ideological profiles that simulate annotator perspectives. At last, we provide our resources and evaluation; PARTISANLENS supports future research on detecting partisan and conspiratorial narratives in European contexts.

1 Introduction

The spread of misinformation threatens democratic societies by distorting public understanding and reducing trust in institutions. This issue is particularly serious in debates about immigration, where misleading stories can shape public opinion, amplify fear, and increase social divisions (Marino et al., 2024). In Europe, this risk is growing as false or biased content influences migration discourses (Komendantova et al., 2023). Two main harmful forms are hyperpartisan narratives and Population Replacement Conspiracy Theories (PRCTs), which fuel hostility toward migrants (Dochow-Sondershaus and Teney, 2024). Hyperpartisan content promotes extreme views through one-sided, emotional language (Kiesel et al., 2019; Maggini et al., 2025b), while PRCTs spread the false idea that native populations are being deliberately replaced by migrants (Marino et al., 2024; Sedgwick, 2024), reinforcing fear and xenophobia (Vogel and Jiang, 2019).

Despite growing research in this area, important gaps remain. Most available datasets are focused on English and U.S.-centric, which limits their relevance for studying European contexts (Maggini et al., 2025b). Moreover, few works connect hyperpartisanship with PRCTs or relate them to stance, and none provide a multilingual benchmark focused on immigration in Europe. Existing efforts examine general propaganda (Da San Martino et al., 2020), expand to broader categories (Piskorski et al., 2023), or target specific issues such as climate change (Maggini et al., 2025a), but the combined, multilingual perspective on European migration debates is still missing.

To address these gaps, we introduce PARTISANLENS, a dataset of 1617 news headlines on European immigration in Spanish, Italian, and Portuguese, collected from 329 outlets between 2020 and 2024 via Media Cloud[1]. Each headline is annotated for hyperpartisanship, Population Replacement Conspiracy Theories (PRCT), stance, and three types of rhetorical bias: loaded language,

*Michele Joshua Maggini and Paloma Piot contributed equally to this work.

[1]https://www.mediacloud.org/

appeal to fear, and name-calling. We conduct experiments to assess the classification performance of transformer-based models, including LLMs, establishing new baselines for this task. In addition, we evaluate the potential of LLMs as surrogate annotators, analysing their ability to replicate human judgments. Finally, we design a novel *persona simulation* experiment, in which LLM annotators are conditioned on socio-economic and ideological profiles (derived from political compass and prism tests) to emulate different annotator backgrounds. We compare the agreement between simulated and real annotators, offering new insights into how ideological perspectives shape annotation.

Our main contributions are threefold. (i) We release PARTISANLENS, the first multilingual dataset on hyperpartisan and conspiratorial immigration narratives including languages that are underrepresented in political-narrative detection research. (ii) We propose a joint annotation scheme that integrates hyperpartisanship, PRCTs, stance, and rhetorical bias, capturing interconnected dimensions of political discourse. (iii) We conduct extensive LLM-based experiments, including the study of ideology-conditioned user-personas, showing both the promise and limitations of LLMs for subjective and context-dependent tasks. All data, code, and prompts are released publicly to support further research on multilingual misinformation detection[2].

2 Related Work

Resources for multilingual political misinformation detection The growing presence of misinformation and polarised political discourse has driven extensive efforts to curate datasets for detection tasks. Prior work has focused on identifying conspiracy theories mostly on tweets covering the COVID-19 outbreaks and following episodes. Pogorelov et al. (2021) collected tweets relating to COVID-19 with 5G wireless network for misinformation detection. By expanding the topics in relation with COVID-19, Langguth et al. (2023) introduced 12 categories like Satanism, Behaviour control, Intentional Pandemic, collecting conspiracy narrative tweets across different countries from 2020 to 2021. They also provided a list of actors, or perpetrators, of the pandemic narrative. With a particular focus on middle-income countries like Brazil, Indonesia and Nigeria, Kim et al. (2023) investigated the tweets on COVID-19 vaccine misinformation from 2020 to 2022. Lastly, Miani et al. (2021) released the LOCO topic-matched corpus, fathered via ready-made lists of conspiracy and mainstream websites.

In the related domain of hyperpartisan detection, Kiesel et al. (2019) is a fundamental work, since it was a highly participated shared task working on English news articles covering different U.S. based topics (e.g. presidency, and gun shootings). Given the nuanced nature of hyperpartisan language, related work often includes the detection of Fake News and Satire. Golbeck et al. (2018) collected a dataset of satirical (203) and fake news (283) stories on American politics. Despite these advances, many existing resources remain limited in both linguistic and topical coverage. Most high-quality datasets are in English and focus on U.S.-centric themes. Indeed, few works focus on scarce resource languages like Bangla, German and Italian (Hossain et al., 2020; Vogel and Jiang, 2019; Maggini et al., 2025a).

Stance detection, which involves classifying support, opposition or neutrality toward a controversial or polarising topic (Mohammad et al., 2017), has also seen significant resource creation. Mohammad et al. (2016) used tweets to track people's utterances towards Climate Change, Feminism, Donald Trump and other topics. Luo et al. (2020) curated a dataset of 56k news articles on global warming seen from a U.S. perspective. Lastly, Mets et al. (2024) annotated Estonian articles challenging the immigration stance detection, comparing GPT-3.5 and BERT-based classifiers.

While recent studies have begun to explore extensive multilingual misinformation like Nielsen and McConville (2022) and Leite et al. (2024) incorporating more than 40 languages, important gaps persist. In particular, there is still a shortage of high-quality multilingual datasets that reflect the European context like Mohtaj et al. (2024) and jointly address hyperpartisanship, stance, and rhetorical bias as interconnected elements of political discourse.

LLM as annotators Recent advances show that political science has increasingly adopted supervised machine learning as a powerful tool for large-scale analysis of political text (Horych et al., 2025; Stromer-Galley et al., 2025). Within this context, LLM-based classification offers a promising way

[2]Dataset and code are available here: https://github.com/MichJoM/PartisanLens.

to reduce manual annotation effort and costs while maintaining high levels of accuracy.

Törnberg (2024) adopted LLMs to annotate messages from the social media platform *X* based on the political belonging of the poster across 11 languages (only one from the Romance group) using the Twitter Parliamentarian Database (Van Vliet et al., 2020). When comparing their results with the gold-standard labels, they found that LLMs outperformed expert coders and fine-tuned baselines like BERT. However, results on public corpora may be inflated by pretraining contamination (data leakage) if models were exposed to the same sources during pretraining (Sainz et al., 2023). Heseltine and von Hohenberg (2024) report that GPT-4 performance drops for more abstract tasks (e.g., ideology detection) and for longer inputs (full articles), while simple binary decisions yield higher human–model agreement; they also note lower accuracy and increased latency for non-English texts.

Therefore, further investigation is needed to understand how incorporating additional inputs like sociodemographic data might enhance the performance of LLMs as annotators in subjective NLP tasks (Beck et al., 2024).

Finally, the choices made when using LLMs as annotators can introduce biases and errors (Baumann et al., 2025); recent works show that LLMs are not fully reliable at the instance level in sensitive tasks, yet capture overall trends (Piot et al., 2025b). Therefore, further investigation is needed to see how additional inputs, such as sociodemographic data, might improve LLM annotation in subjective NLP tasks (Beck et al., 2024).

LLM persona Variation among human annotators has long been recognised as a source of noise against a presumed *ground truth* (Plank, 2022; Schäfer et al., 2025). This issue is especially evident in subjective tasks such as toxic or racist language detection (Sap et al., 2022) and natural language inference (Biester et al., 2022), where sociodemographic differences strongly influence interpretation. Recent research examines how real or synthetic demographic profiles (e.g., PersonaHub (Ge et al., 2024)) correlate with annotation behaviour. In this direction, Fröhling et al. (2025) showed that injecting persona descriptions into LLM prompts enhances diversity and control in annotation, while Mukherjee et al. (2024) explored cultural variation through persona-based elicitation. Complementing these studies, Piot et al. (2025a) demonstrated that geographic bias affects how state-of-the-art LLMs classify hate speech, revealing the influence of user-personas on model outputs. Understanding such variations is essential when assessing whether LLMs can replicate or complement human annotation in politically sensitive domains, where models remain highly sensitive to design choices and vulnerable to manipulation or *LLM hacking* (Baumann et al., 2025).

3 Dataset Construction

In this section, we present the data collection scheme, the annotation protocol and the resulting dataset statistics.

3.1 Data collection and preparation

We focused on collecting immigration-related headlines in three underrepresented Romance languages: Spanish, Portuguese and Italian. These represent major linguistic areas along the Mediterranean migration routes into Europe (Aragall, 2018), and yet remain largely absent from resources for misinformation detection. Our data collection followed a hybrid strategy that combined keyword-based retrieval with LLM-assisted filtering. Using the Media Cloud platform[3], we extracted headlines from national and regional outlets in Spain, Italy, and Portugal. For each language, we used targeted keyword queries (e.g., equivalents of *migration* or *immigration policy*, *population replacement*) to capture a broad range of immigration narratives[4]. After removing duplicate URLs and headlines, this process resulted in approximately 400.000 unique entries.

Given this scale, we employed `Llama 3.1 8B` to perform an initial classification along two dimensions: hyperpartisan and PRCT. This automatic step ensured coverage of low-frequency but socially important categories. In particular, PRCT-related content represented less than 1% of the initial corpus. Guided by this filtering, we selected a balanced and diverse sample of 1617 headlines published between 2020 and 2024 from 329 distinct sources.

3.2 Annotation protocol

The annotation protocol was designed to capture both ideological content and rhetorical framing in

[3] https://www.mediacloud.org/

[4] Full keyword lists are available in our repository.

immigration-related headlines. Each item was annotated along *six* complementary dimensions that reflect the stance, narrative strategy, and emotional tone of the text[5]. At the content level, we considered: (i) **Hyperpartisan**, indicating whether the headline expressed a one-sided or ideologically extreme view; (ii) **PRCT**, marking conspiratorial framings such as references to population replacement or coordinated demographic change; and (iii) **Stance** (categorical), capturing the overall attitude toward immigration as *pro*, *neutral*, or *against*.

In addition, three rhetorical strategies were annotated at the span level: (iv) **Loaded language**, where emotionally charged wording was used to provoke reaction; (v) **Appeal to fear**, framing immigration as a threat or danger; and (vi) **Name calling**, involving derogatory or stereotyping expressions. These rhetorical cues are known to polarise debate by appealing to emotion rather than reasoned argumentation (Maggini et al., 2025a).

Annotation was conducted through a custom Web interface that masked outlet names to avoid source bias. Nine native speakers, 3 males and 6 females (three per language, all with at least a master's degree) carried out the task over four iterative rounds, beginning with pilot annotations and guideline refinement, followed by progressively larger batches. In total, the process required approximately 66 hours. The final dataset includes both individual annotations and majority labels, enabling not only supervised training but also future research on annotator disagreement and uncertainty.

To assess the annotation reliability, we computed Fleiss' κ across the three main annotation tasks (Hyperpartisan, PRCT, Stance). We did not compute agreement for the rhetorical strategies, as these were annotated at the span level. Their purpose was primarily to guide annotators by highlighting cues associated with hyperpartisan phenomena. Results are presented in Table 1 and indicate substantial to almost perfect agreement across all languages (Landis and Koch, 1977). Hyperpartisan labels were most consistent in Portuguese and Spanish, while PRCT achieved slightly higher agreement in Italian. Stance annotations were somewhat lower than the other tasks but still showed substantial reliability.

3.3 Dataset statistics

PARTISANLENS includes 527 headlines in Portuguese (PT), 565 in Italian (ITA), and 525 in Spanish (SPA). Table 2 summarises the majority voting distribution across all labels. Hyperpartisan headlines are prevalent in every language (range of 59-70%), while PRCT cases remain relatively rare (under 10% in PT and SPA, but higher in ITA at 28%). Stance distribution varies considerably: anti-immigration headlines dominate in Italian (59%) and Portuguese (40%), whereas Spanish headlines are mostly neutral (80%). Pro-immigration stances are scarce across the three languages.

[5]Full annotation guidelines are available in our repository.

	SPA	ITA	PT
Hyperpartisan	0.876	0.706	0.770
PRCT	0.880	0.774	0.721
Stance	0.837	0.744	0.737

Table 1: Inter-Annotator Agreement (Fleiss' κ).

Category	Label	SPA	ITA	PT
Hyperpartisan	True	336	337	367
	False	189	228	160
PRCT	True	46	160	9
	False	479	405	518
Stance	Pro	4	78	130
	Neutral	419	156	185
	Against	102	331	212
Loaded lang.	Present	346	234	213
	None	179	331	314
Appeal to fear	Present	107	168	35
	None	418	397	492
Name calling	Present	35	44	33
	None	490	521	494

Table 2: Dataset label distribution.

Stance-conditioned patterns. Figure 1 shows how PRCT and hyperpartisan content correlate with stance. The strongest signal comes from anti-immigration headlines: 81% contain PRCT claims and 56% are hyperpartisan. Neutral headlines are minor (14% PRCT, 27% hyperpartisan), while pro-immigration headlines are the least polemical (3% PRCT, 12% hyperpartisan). These results confirm that conspiratorial narratives cluster much more around negative stances than hyperpartisan style, which is more broadly distributed.

Language-level variation Figure 2 highlights how the two labels distribute across the languages' headlines. Italian outlets display the strongest conspiratorial signal: 27% of their immigration

headlines contain PRCT language, roughly three times the Spanish rate (8%) and nine times the Portuguese rate (3%). Hyperpartisan style shows a different trend: it peaks in Portuguese (70%) and remains consistently high in Spanish and Italian (≈60%). Taken together, these findings suggest that while hyperpartisan framing is common across languages, conspiratorial narratives concentrate more on Italian media.

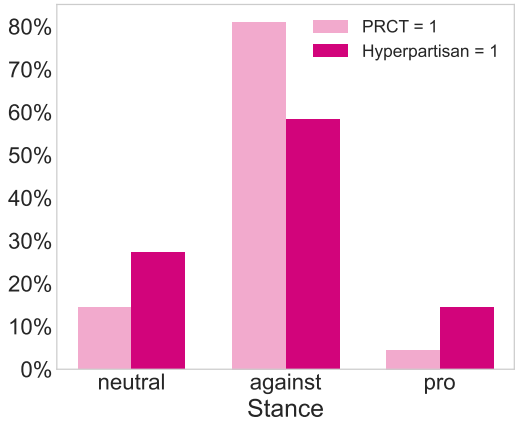

Figure 1: Hyperpartisan and PRCT stance proportions.

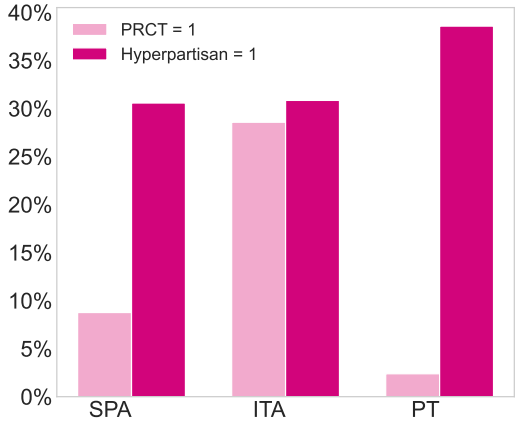

Figure 2: Hyperpartisan and PRCT content by language.

4 Benchmark Experiments: Setup and Results

To assess the utility of PARTISANLENS as a benchmarking resource, we compare state-of-the-art multilingual models under increasing supervision levels, from zero-shot prompting to supervised fine-tuning. Our goals are: (i) to establish baselines for the three target tasks (Hyperpartisan, PRCT, Stance) across Spanish, Italian and Portuguese (SPA, ITA, PT), and (ii) to quantify how supervision type and rationale-augmented learning affect accuracy and robustness.

4.1 Models

We evaluate multilingual transformers covering encoder and generative architectures. As a strong non-generative baseline, we include `mBERT` (Devlin et al., 2019) (110M parameters), widely used for cross-lingual classification. To compare against more recent instruction-tuned systems, we explore three generative LLMs of increasing scale: `Llama 3.1-8B`, `Mistral-Nemo-Base-2407` (12B) (Mistral AI team, 2025) and `Llama 3.3-70B` (Dubey et al., 2024).

4.2 Classification strategies

We design four complementary strategies to study supervision effects over the three target tasks (hyperpartisan, PRCT and stance):

Zero-shot prompting. Models receive a concise task description defining the three task categories. We set `temperature` $= 0$ for deterministic decoding and instruct models to output only a valid label (see Appendix, Table 9).

Few-shot prompting. To provide minimal task-specific context, we include ten balanced demonstrations (≈2 per label) selected via a Determinantal Point Process (DPP) (Kulesza, 2012) to maximise diversity and reduce redundancy. We interleave one positive instance and one negative in our few-shot. Prompts match the zero-shot setup except for the in-context examples. Further DPP details appear in Appendix E.

Supervised fine-tuning (labels). LLMs are fine-tuned for two epochs with learning rate $2e^{-5}$, linear decay, warm-up ratio 0.03, batch size 12, and `adamw_8bit`; inputs are truncated at 4096 tokens. These hyperparameters follow common practice in instruction-tuning and were selected after small validation sweeps to balance stability and avoid overfitting. For `mBERT`, we train for 5 epochs with learning rate $3e^{-4}$, batch size 16, and dropout 0.23.

Supervised fine-tuning (labels + rationales). In this variant, models are fine-tuned not only on classification labels but also on rationales, which are short, natural-language explanations derived from the rhetorical bias spans annotated during corpus creation. These rationales express how a headline carries persuasion, such as: *"The headline appeals to fear by suggesting a national threat"*. We generate fluent, context-aware rationales from anno-

Model	Setting	Hyperpartisan			PRCT			Stance		
		P	R	F1	P	R	F1	P	R	F1
mBERT	fine-tuned	0.6896	0.6312	0.6299	0.8589	0.8148	**0.8345**	0.6142	0.6047	0.5974
Llama 8B	zero-shot	0.6483	0.6483	0.6483	0.7419	0.6056	0.6306	0.5795	0.4569	0.4264
	few-shot	0.7117	0.7165	**0.7129**	0.7274	0.8041	0.7540	0.6227	0.5768	0.5761
	fine-tuned	0.6863	0.6662	0.6689	0.7751	0.6980	0.7263	0.6121	0.4846	0.4639
	rationales	0.6973	0.6680	0.6706	0.7513	0.6762	0.7028	0.6491	0.4674	0.4320
Llama 70B	zero-shot	0.6806	0.6885	0.6654	0.7814	0.8537	0.8090	0.6273	0.6042	0.5945
	few-shot	0.7054	0.6903	0.6938	0.8281	0.8175	0.8226	0.6457	**0.6112**	0.5964
	fine-tuned	0.7047	0.7085	0.6911	0.7607	**0.8654**	0.7943	0.6413	0.6086	0.6062
	rationales	**0.7179**	**0.7246**	0.7118	0.7295	0.8471	0.7598	0.6709	0.6051	0.6068
Nemo	zero-shot	0.6818	0.6839	0.6827	**0.9275**	0.7111	0.7668	0.5580	0.4750	0.4433
	few-shot	0.7039	0.7056	0.6851	**0.9325**	0.7237	0.7814	0.6717	0.5995	**0.6138**
	fine-tuned	0.7061	0.7109	0.7073	0.8875	0.7004	0.7512	0.5977	0.4849	0.4625
	rationales	0.6843	0.6474	0.6475	0.8117	0.6735	0.7126	**0.6877**	0.4642	0.4232

Table 3: Macro-averaged precision (P), recall (R), and F1 across tasks and all our considered models. Best metrics per task are marked in **bold**.

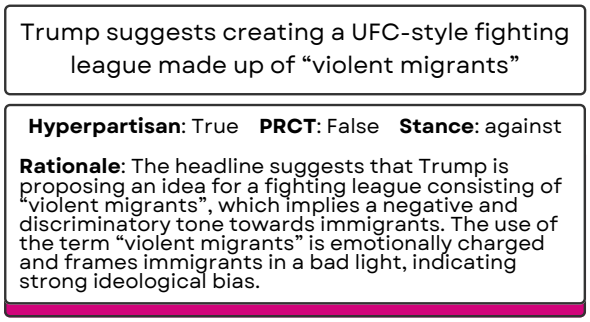

Figure 3: Example of a rationale-enhanced annotation.

tated templates using `Llama-3.3-70B-Instruct`, and concatenate them to the input text during fine-tuning. This encourages interpretable learning and alignment between evidence and prediction (Zhang et al., 2024; Piot and Parapar, 2025). Example of a rationale-enhanced instance is shown in Figure 3.

4.3 Evaluation metrics

We report macro-averaged precision, recall, and F1 across the three tasks and all our proposed models. These metrics are robust to class imbalance and ensure that minority classes contribute equally to the overall score. This follows standard practice in hyperpartisan detection, misinformation classification, and abusive-language research, where label distributions are typically skewed (Liu et al., 2025; Essahli et al., 2025; Assenmacher et al., 2025).

4.4 Results

Table 3 reports macro-averaged precision (P), recall (R), and F1 across the three tasks and all our proposed models.

Hyperpartisan. Performance is solid but not the easiest task (F1≈0.62–0.71). Top scores come from `Llama 8B` (few-shot), `Llama 70B` (rationales), and `Nemo` (fine-tuned), all around 0.71 with balanced precision and recall. Scale helps, but real demonstrations let small models match larger ones. Rationale inputs benefit `Llama 70B`, suggesting that style-sensitive evidence can be exploited at large capacity, but the other architectures do not share this pattern. Overall, hyperpartisan detection benefits most from light supervision, with few-shot prompts often matching, or even replacing, the need for large-scale fine-tuning.

PRCT. Detecting PRCT is *easier* than hyperpartisan related content. The best F1 is 0.82 (`Llama 70B` few-shot), and even zero-shot `Llama 70B` is strong (F1=0.81). `Nemo` achieves the highest precision (0.93) at the cost of recall (≈0.71–0.72), pointing to a conservative detector that captures clear conspiracy markers but misses subtler instantiations. Notably, `mBERT` fine-tuned also reaches F1=0.82. Rationale augmentation does not help here, aligning with the fact that our rationales encode rhetorical style rather than conspiracy semantics. The limiting factor is recall, not interpretability alignment.

Stance. Stance is the hardest task (best F1=0.61), with `Llama 70B` (fine-tuned / rationales) and `Nemo` (few-shot) tying at the top. Gains mostly come from modest recall (0.61), while smaller models stay below (F1=0.58). The pronounced difficulty likely reflects both class distributions (e.g., very rare pro stance in Spanish) and the need for contextual inferences beyond surface style, since

rationale inputs can even hurt (with Nemo obtaining a F1=0.42). The overall difficulty suggests that stance requires finer-grained semantic understanding than hyperpartisan or PRCT, and remains an open challenge even for very large models.

Three main insights emerge from these results. (*i*) Supervision matters: few-shot and fine-tuning typically add 4–7 F1 points over zero-shot. (*ii*) Scale helps but is not enough: `Llama 70B` is most consistent, yet `Llama 8B` few-shot matches top hyperpartisan scores. (*iii*) Task complexity varies: *PRCT* > *hyperpartisan* > *stance*. PRCT benefits from distinctive lexical/conceptual signatures, hyperpartisan is stylistic and diffuse, and stance seems to demand finer pragmatic cues. A detailed per-language analysis, including $F1_{MACRO}$, P/R, and error profiles, is provided in Appendix G.

5 LLMs as Potential Data Annotators

To evaluate whether LLMs can serve as reliable annotators, we compare their agreement against the reference level of human agreement across languages. The core question is: *does an LLM approach human agreement levels on our three targets (Hyperpartisan, PRCT, Stance) in the three languages?*

5.1 Setup

To answer this research question, we measure mean pairwise Cohen's κ for human–human (H–H) and LLM–human pairs using two models: (*i*) `Llama 3.3-70B` fine-tuned with labels + rationales (L–H), and (*ii*) `gpt-5-mini` in inference mode (G–H). We chose the 70B model with rationales due to its robustness across tasks, and include `gpt-5-mini` to study whether a recent inference model narrows the human gap. The agreement results for all the annotation pairs (human pairs and LLM-human pairs) are summarised in Table 4[6].

5.2 Results

Human agreements (H-H) Among the three tasks, human agreement is overall higher for Spanish headlines ($\kappa \approx 0.86$–0.88), followed by Italian ($\kappa \approx 0.59$–0.74), with more variability in Portuguese ($\kappa \approx 0.66$–1.00). We attribute the perfect agreement for PCRT in Portuguese due to extreme class sparsity (only n=9 positives). The Spanish annotators higher agreement, which may be related to the country's generally higher level of immigration integration and a more informed understanding of immigration issues, as reported by the OECD (2023)'s Indicators of Immigrant Integration 2023. This pattern suggests that both linguistic factors and local media influence how easy it is to judge a headline.

Label	Spanish			Italian			Portuguese		
	H–H	L–H	G–H	H–H	L–H	G–H	H–H	L–H	G–H
Hyperpartisan	0.864	0.424	0.594	0.698	0.524	0.452	0.666	0.228	0.219
PRCT	0.893	0.483	0.549	0.743	0.632	0.541	1.000	0.124	0.662
Stance	0.858	0.364	0.633	0.598	0.377	0.567	0.682	0.392	0.480

Table 4: Mean Cohen's κ by language and label. The agreement pairs are: **H–H** (human vs. human), **L–H** (`Llama 3.3-70B` vs. human), **G–H** (`gpt-5-mini` vs. human).

LLM vs. human agreements (L–H / G–H) Models remain below H–H across languages, but `gpt-5-mini` (G–H) is closer to human agreement than `Llama 3.3-70B` (L–H). In Spanish, `gpt-5-mini` improves over `Llama 3.3-70B` on all tasks, especially on the Stance category (0.633 vs. 0.364). In Italian, however, `Llama 3.3-70B` obtains better results in Hyperpartisan and PRCT than `gpt-5-mini`, but `gpt-5-mini` still obtains way better performance on Stance, being close to human agreement (0.567 vs. 0.598). In Portuguese, `gpt-5-mini` yields the large PRCT difference in comparison to `Llama 3.3-70B` (0.662 vs. L–H 0.124), but, in this case, for the three labels, the LLMs are still way below compared to human agreement.

In sum, although we evaluated two large and competitive LLMs, its limited agreement with experts shows that fully unsupervised annotation remains out of reach for these nuanced categories. Human supervision is therefore still indispensable, making resources in this domain crucial for advancing robust detection methods.

6 User-Persona LLMs

Persona-Based prompting intends to condition a models' output to reflect the characteristics of specific personas, enabling researchers to simulate opinions, values, and attitudes more effectively (Beck et al., 2024). In this experiment, we examine whether LLMs can emulate patterns of human annotation behaviour when conditioned on realistic user profiles. Our motivation is to assess whether user-persona LLMs reflect the same socio-

[6]Each H–H value averages across all human pairs. L–H and G–H average the corresponding LLM vs. each human.

demographic and ideological variability observed among human annotators. In other words, if human disagreement is partially driven by background traits, an LLM simulating those traits should align more closely with annotators sharing the same profile. We therefore ground LLM personas using real socio-demographic and political data from our annotators, and compare agreement levels between persona-conditioned LLMs and their human counterparts across Spanish, Italian, and Portuguese.

6.1 Persona prompt

We build structured persona prompts representing each annotator's political and demographic background. To do so, each human annotator completed two anonymous instruments: the Political Compass test,[7] which places respondents on Left–Right and Authoritarian–Libertarian axes, and the PRISM quiz,[8] which captures social and economic orientations. We standardise test scores within each dimension and combine them with self-reported sociodemographics (e.g., age range, gender) to construct concise, non-identifying textual profiles. Personas are generated from predefined templates that map score ranges to qualitative descriptors (e.g., "economically left-leaning, socially moderate, libertarian-leaning"). Full sociodemographic summaries and persona templates are included in the Tables 8 and 10 (Appendix).

6.2 Setup

We implement persona prompting with two models: `Llama 3.1-70B` and `gpt5-mini`. For both, we extend the zero-shot classification prompt from Section 4 with the same user-persona context, with all other prompt components remaining identical. The persona context states the annotator's traits (gender, age range, and economic/social/authority–liberty positions derived from Compass/PRISM) before the task instructions. For evaluation, we compare each persona-LLM to: (i) its matched human counterpart (the annotator whose profile it simulates) and (ii) non-matched annotators in the same language. We also contrast these results with the human–human agreement baseline (Table 4).

6.3 Results

Agreement user-persona vs. corresponding human Table 5 reports Cohen's κ between each persona-conditioned LLM and the human it simulates, using `Llama-70B` and `gpt5-mini`. If we compare the overall results to the generic L–H baselines in Table 4, LLM persona conditioning helps in several settings, and `gpt5-mini` often amplifies these gains, especially for PRCT and Stance. Regarding Spanish annotators, `Llama-70B` personas improves PRCT over the generic baseline (0.599 vs. 0.483) and slightly Hyperpartisan (0.459 vs. 0.424), but lower Stance (0.252 vs. 0.364). This suggests that persona conditioning may help capture content focus and priorities rather than polarity judgments. With `gpt5-mini`, agreement rises across all three: Hyperpartisan (0.598 vs. 0.635), PRCT (0.711 vs. 0.754), and Stance (0.663 vs. 0.767).

Persona	Hyperpartisan		PRCT		Stance	
	L70B	gpt5-m	L70B	gpt5-m	gpt5-m	gpt5-m
SPA_1	0.5331	0.6334	0.5531	0.6252	0.3441	0.7672
SPA_2	0.4151	0.6351	0.5889	0.7540	0.2162	0.6192
SPA_3	0.4276	0.5255	0.6550	0.7540	0.1969	0.6020
ITA_1	0.4908	0.5068	0.7409	0.7274	0.3758	0.6316
ITA_2	0.5024	0.5278	0.6217	0.7060	0.3049	0.5374
ITA_3	0.4658	0.4567	0.7392	0.6320	0.3629	0.5882
PT_1	0.2822	0.2143	0.3151	0.6622	0.4367	0.4284
PT_2	0.2435	0.1397	0.3151	0.6622	0.4313	0.3989
PT_3	0.2136	0.1806	0.2268	0.6622	0.4860	0.5599

Table 5: Cohen's κ for each user-persona vs. its corresponding human annotator, comparing `Llama 3.1-70B` (L70B) and `gpt5-mini` (gpt5-m) inference.

For Italian annotators, `Llama-70B` personas shows the same trend, with clear gains on PRCT (0.701 vs. 0.632), are near parity for Stance (0.348 vs. 0.377), and slightly below for Hyperpartisan (0.486 vs. 0.524). This pattern reinforces the idea that persona conditioning is most effective for tasks involving content relevance (PRCT), while stylistic or bias judgments are less influenced by the persona information. `gpt5-mini` keeps PRCT competitive (0.689 vs 0.727) and even delivers gains in Stance (0.586), clearly above both the baseline (0.377). Hyperpartisan remains a bit below baseline. Overall, Italian sees the strongest persona-driven PRCT alignment and a notable stance improvement with `gpt5-mini`.

Finally, if we analyse the Portuguese results, despite extreme PRCT sparsity, `Llama-70B` personas improve the generic baseline on all tasks. `gpt5-mini` produces a large PRCT increase (0.662), a modest Stance performance (0.462), and a decrease for Hyperpartisan (0.178). This pattern suggests persona cues align well with distinctive

[7] https://www.politicalcompass.org/
[8] https://prismquiz.github.io/

Original Language	English Translation
Migranti, il M5s: "Lollobrigida parla di sostituzione etnica? E' ignorante, mica razzista"	Migrants: M5s: "is Lollobrigida talking about etnic replacement? He is ignorant, not racist"
El PP prepara toda su artillería contra la "independencia fiscal" de Cataluña	PP is preparing the artillery against the "financial independence" of Catalonia
Immigrazione, Trump: "Avvelena il sangue del nostro Paese"	"Immigration, Trump: It poisons our Country's blood"

Table 6: Selected headlines with human and user-personas disagreements.

conspiratorial markers even under sparsity, while stylistic judgments remain fragile in PT.

Looking at the overall results, persona conditioning increases LLM–human PRCT agreement in all languages (largest gains in ITA and PT). Using `gpt5-mini`, it clearly improves Stance in SPA/ITA. Hyperpartisan performance remains mixed (SPA gains, ITA near parity, and PT drops). Overall levels are still way down compared to the H–H reference (Table 4), but grounding in real profiles recovers meaningful human-like variability for PRCT and, with stronger inference, also for stance.

6.4 Error analysis

We explore LLM's deficiencies and attribute them to several factors. One important issue is the limited understanding of sarcastic and metaphorical sentences. In Table 6, we can see some examples of headlines with disagreements among annotators and user-personas. A second limitation concerns the model's insufficient understanding of semantics and context. For example, the Portuguese term *Imigrantes* can denote both "immigrants" and the name of a Brazilian highway. The model struggled to disambiguate the meaning based on context. Similarly, the word "invading" carries different connotations across Portuguese, Italian, and Spanish, and the LLM's rigid interpretation suggests a lack of sensitivity to cultural context. Another illustrative shortcoming is the model's failure to interpret irony, sarcasm, or implicit critique. Such use of rhetorical devices is a subtle but powerful tool in political discourse, where it frequently conveys unstated or indirect opinions.

7 Conclusions and Future Work

This paper introduced PARTISANLENS, a multilingual dataset of 1617 news headlines for analysing hyperpartisan content, population replacement conspiracy theories, and stance towards immigration news in Spanish, Italian, and Portuguese. A robust protocol with nine native speakers yielded substantial inter-annotator agreement on the main tasks, with clear language effects: Italian shows the strongest PRCT signal, while Portuguese has higher hyperpartisan prevalence. As a benchmark, PARTISANLENS shows that even light supervision (few-shot) improves over zero-shot and model scale helps, but can be matched by well-chosen demonstrations. Moreover, the three main categories are separated in difficulty, with PRCT showing the easiest to detect, hyperpartisan next, and stance remaining the most challenging.

We also evaluated LLMs as annotators. Strong language models such as `Llama 3.3-70B` remain below human agreement across tasks and languages. Conditioning LLMs on realistic annotator profiles led them to better reflect human-like variability, improving alignment with individual annotation patterns for PRCT across languages, showing mixed effects for hyperpartisan content, and leaving stance differences largely unchanged. We release data, code and all materials to support reproducibility and future work on multilingual partisan narratives in European media.

Future work will (i) enlarge language coverage to other European languages, (ii) add document-level context to enable discourse-aware models, and (iii) conduct cross-lingual transfer experiments to better isolate language and domain effects.

Computational Resources

Experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO2eq/kWh. A cumulative of 21 hours of computation was performed on hardware of type RTX A6000 (TDP of 300W). Total emissions are estimated to be 1.94 kgCO2eq of which 0 per cent were directly offset. Estimations were conducted using MachineLearning Impact calculator (Lacoste et al., 2019).

Limitations

Our dataset is necessarily small: constructing a fine-grained, multi-task resource required substantial expert effort (approximately 66 hours of annotation by nine native speakers, experts in the matter). Because Media Cloud provides headlines but not article bodies, we restricted the collection to headlines to avoid copyright issues around full-text scraping; as a result, models and analyses operate without broader article context. The set of rhetorical biases is also limited by design. Following Maggini et al. (2025a), we focused on the phenomena most influential in shaping hyperpartisan style (loaded language, appeal to fear, name-calling). Expanding to additional techniques would have significantly increased annotator training time and burden, which we judged inappropriate for a volunteer cohort.

Cultural and contextual sensitivity is another constraint. Italian, Spanish, and Portuguese political discourses are embedded in specific historical and social frames that LLMs may not fully capture. That is, what appears hyperpartisan to a model may be mischaracterised relative to local norms. Finally, we acknowledge that labelling content as "hyperpartisan" contains normative and political judgments that extend beyond purely linguistic analysis. Our goal is analytical clarity and reproducibility, but these categories should be interpreted with care and with awareness of their social implications. However, we caution that the dataset could be misused by malicious actors to train models capable of generating hyperpartisan or discriminatory content.

Ethical Statement

All annotations were performed by volunteers who participated freely, without time pressure or coercion. Annotators could pause or withdraw at any point, and they received compensation for their work commensurate with the expected effort. Socio-economic and political self-reports used to derive personas were fully anonymised. We collected only coarse, non-identifying attributes (e.g., age range, self-reported gender) and ideology scores that were standardised before use. Raw responses and any potentially identifying information were not retained or released; persona prompts were generated from these de-identified summaries, and the public release includes only anonymised labels and headlines.

Acknowledgments

The authors thank the funding from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors thank the financial support supplied by the grant PID2022-137061OB-C21 funded by MICIU/AEI/10.13039/501100011033 and by "ERDF/EU". The authors also thank the funding supplied by the Consellería de Cultura, Educación, Formación Profesional e Universidades (accreditations ED431G 2023/01 and ED431C 2025/49) and the European Regional Development Fund, which acknowledges the CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

Special thanks go to Marta Muñoz González and Amelia Fresa for their contribution during the annotation process.

References

Xavier Aragall. 2018. 2018: Migration in the Mediterranean; Situation, Context and Evolution. https://www.iemed.org/publication/2018-migration-in-the-mediterranean-situation-context-and-evolution.

Dennis Assenmacher, Paloma Piot, Katarina Laken, David Jurgens, and Claudia Wagner. 2025. Beyond the explicit: A bilingual dataset for dehumanization detection in social media. *Preprint*, arXiv:2510.18582.

Joachim Baumann, Paul Röttger, Aleksandra Urman, Albert Wendsjö, Flor Miriam Plaza del Arco, Johannes B. Gruber, and Dirk Hovy. 2025. Large language model hacking: Quantifying the hidden risks of using llms for text annotation. *Preprint*, arXiv:2509.08825.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. 2024. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of*

the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2589–2615, St. Julian's, Malta. Association for Computational Linguistics.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Laura Biester, Vanita Sharma, Ashkan Kazemi, Naihao Deng, Steven Wilson, and Rada Mihalcea. 2022. Analyzing the effects of annotator gender across NLP tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 10–19, Marseille, France. European Language Resources Association.

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online). International Committee for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Stephan Dochow-Sondershaus and Céline Teney. 2024. Opinion polarization of immigration and eu attitudes between social classes–the limiting role of working class dissensus. *European Societies*, 26(5):1363–1394.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, ..., and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Soufiane Essahli, Oussama Sarsar, Ahmed Bentajer, Anas Motii, and Imane Fouad. 2025. Fakezero: Real-time, privacy-preserving misinformation detection for facebook and x. *Preprint*, arXiv:2510.25932.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. 2025. Personas with attitudes: Controlling LLMs for diverse data annotation. In *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 468–481, Vienna, Austria. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, and 12 others. 2018. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA. Association for Computing Machinery.

Michael Heseltine and Bernhard Clemm von Hohenberg. 2024. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1):20531680241236239.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2025. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico. Association for Computational Linguistics.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. 2020. BanFakeNews: A dataset for detecting fake news in Bangla. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France. European Language Resources Association.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. 2023. COVID-19 vaccine misinformation in middle income countries. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3903–3915, Singapore. Association for Computational Linguistics.

Nadejda Komendantova, Dmitry Erokhin, and Teresa Albano. 2023. Misinformation and its impact on contested policy issues: The example of migration discourses. *Societies*, 13(7).

Alex Kulesza. 2012. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter

dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484.

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2024. Euvsdisinfo: A dataset for multilingual detection of pro-kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5380–5384, New York, NY, USA. Association for Computing Machinery.

Xuannan Liu, Zekun Li, Peipei Li, Huaibo Huang, Shuhan Xia, Xing Cui, Linzhi Huang, Weihong Deng, and Zhaofeng He. 2025. Mmfakebench: A mixed-source multimodal misinformation detection benchmark for lvlms. In *ICLR*.

Yiwei Luo, Dallas Card, and Dan Jurafsky. 2020. Detecting stance in media on global warming. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3296–3315, Online. Association for Computational Linguistics.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 2025a. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 168–187, Vienna, Austria. Association for Computational Linguistics.

Michele Joshua Maggini, Davide Bassi, Paloma Piot, Gaël Dias, and Pablo Gamallo Otero. 2025b. A systematic review of automated hyperpartisan news detection. *PLOS ONE*, 20(2):1–39.

Erik Bran Marino, Jesus M Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(11):603.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLOS ONE*, 19(4):1–16.

Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2021. Loco: The 88-million-word language of conspiracy corpus. *Behavior research methods*, pages 1–24.

Mistral AI team. 2025. Mistral NeMo | Mistral AI. https://mistral.ai/news/mistral-nemo. Accessed: 09/01/2025.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Salar Mohtaj, Ata Nizamoglu, Premtim Sahitaj, Vera Schmitt, Charlott Jakob, and Sebastian Möller. 2024. Newspolyml: Multi-lingual european news fake assessment dataset. In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, pages 82–90.

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. 2024. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA. Association for Computational Linguistics.

Dan S. Nielsen and Ryan McConville. 2022. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA. Association for Computing Machinery.

OECD. 2023. Indicators of Immigrant Integration 2023 — oecd.org. https://www.oecd.org/en/publications/indicators-of-immigrant-integration-2023_1d5020a6-en.html. [Accessed 06-06-2025].

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2025a. Personalisation or prejudice? addressing geographic bias in hate speech detection using debias tuning in large language models. *Preprint*, arXiv:2505.02252.

Paloma Piot, David Otero, Patricia Martín-Rodilla, and Javier Parapar. 2025b. Can llms evaluate what they cannot annotate? revisiting llm reliability in hate speech detection. *Preprint*, arXiv:2512.09662.

Paloma Piot and Javier Parapar. 2025. *Towards Efficient and Explainable Hate Speech Detection via Model Distillation*, page 376–392. Springer Nature Switzerland.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. 2023. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. 2021. Wico text: A labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, OASIS '21, page 21–25, New York, NY, USA. Association for Computing Machinery.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wuehrl, Sabine Weber, and Roman Klinger. 2025. Which demographics do LLMs default to during annotation? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17331–17348, Vienna, Austria. Association for Computational Linguistics.

Mark Sedgwick. 2024. The great replacement narrative: fear, anxiety and loathing across the west. *Politics, Religion & Ideology*, 25(4):548–562.

Jennifer Stromer-Galley, Brian McKernan, Saklain Zaman, Chinmay Maganur, and Sampada Regmi. 2025. The efficacy of large language models and crowd annotation for accurate content analysis of political social media messages. *Social Science Computer Review*, 0(0):08944393251334977.

Petter Törnberg. 2024. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 0(2024):08944393241286471.

Livia Van Vliet, Petter Törnberg, and Justus Uitermark. 2020. The twitter parliamentarian database: Analyzing twitter politics across 26 countries. *PLoS one*, 15(9):e0237073.

Inna Vogel and Peter Jiang. 2019. Fake news detection with the new german dataset "germanfakenc". In *Digital Libraries for Open Knowledge - 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings*, pages 288–295.

Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. 2024. Efficient toxic content detection by bootstrapping and distilling large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21779–21787.

A Data Statement

We report the Data Statement using the schema proposed by Bender and Friedman (2018).

Curation rationale The selected texts consist of news headlines collected via API from the platform Media Cloud, an open-source media research project, enabling the study of news and information flow globally. The goal in selecting texts is to cover and analyse the spread, the linguistic patterns and the stance of headlines on European Immigration, detecting the hyperpartisan and neutral content.

Language variety Referring to the language tags contained in BCP-47, we list the language variety in our dataset: it-IT, es-ES, pt-BR, pt-PT.

Annotator demographic Table 7 presents the demographics of the annotators who participated in the curation of this dataset.

Text characteristics The dataset is composed of news headlines on European Immigration collected from a wide range of media outlets covering the whole political leaning spectrum. For this reason, the texts are short. Due to the nature of the task and the topic, some headlines may refer to immigration using racist and toxic language, spreading xenophobic messages, which could result in the adoption of a conspiracist tone with specific linguistic and topic patterns (PRCT). The primary goal of the dataset is the detection of extremely polarised news to understand how they address these themes.

B Sociodemographic Data

Table 8 illustrates the sociodemographic data we collected for each annotator. All of those dimensions have been used to craft the Persona Prompts.

C Annotation Tool

The custom platform allows annotators to annotate one headline at a time using the labels introduced in 3.1. The output file is a csv file.

D Prompts

Tables 9 and Table 10 show the prompts used in the experiments.

	Age	Gender	Race/Ethnicity	Native language	Socioeconomic status	Training
SPA 1	27	Male	Hispanic	Spanish	employed	STEM (PhD)
SPA 2	28	Female	Caucasian	Spanish	PhD student	STEM (MSc)
SPA 3	28	Female	White	Galician	employed	Humanities (MA)
ITA 1	28	Male	Italian	Italian	PhD student	Humanities (PhD)
ITA 2	27	Male	Caucasian	Italian	PhD student	Humanities (PhD)
ITA 3	26	Female	White	Italian	unemployed	Humanities (MA)
PT 1	55	Female	Caucasian	Portuguese	employed, married	Humanities (MA)
PT 2	26	Female	Caucasian	Portuguese	Student	Humanities (BA)
PT 3	24	Female	White	Portuguese	employed	STEM (MSc)

Table 7: Participant demographic overview

Attribute	**Value**
Gender:	Male
Education:	Doctorate
Background:	Humanities
Age:	28
COMPASS Economic:	-5.25
COMPASS Social:	-6.21
PRISM Government:	100 Technocracy
PRISM Economy:	84 Mixed Markets, 16 Ordoliberalism
PRISM Society:	100 Freedom
Native Language:	Italian
Native or Migrant:	Migrant
Ethnicity:	Italian
Employment Status:	Student
Marital Status:	Partnered
Religion:	Atheist
Disability Status:	No

Table 8: Example of sociodemographic data. *COMPASS Economic* ranges from negative to positive values, where lower scores indicate a left-leaning economic stance, and higher scores indicate a right-leaning stance. *COMPASS Social* follows the same logic, with lower values reflecting a libertarian orientation and higher values an authoritarian one.

You are a strict JSON generator. Analyze the following news headline and output a JSON object with this exact format:

```
{
  "hyperpartisan": "<Boolean>",
  "prct": "<Boolean>",
  "stance": "<pro|against|neutral>",
}
```

Hyperpartisan: Does it show strong ideological bias or use emotionally charged language? (True/False)
PRCT: Does it contain Population Replacement Conspiracy Theory content claiming a deliberate plan to replace native populations? (True/-False)
Stance: What is its stance toward immigration policies? (pro/against/neutral).
DO NOT include any commentary or explanation. Only return valid JSON.
Headline: <headline>

Table 9: Zero-shot and few-shot prompt.

You are a <*gender*> individual with a <*age*> years old, holding a master's degree in <*subject*>. Your COMPASS test values indicate you are leaning towards <*political leaning*>.
You advocate for <*PRISM_Government*>, support <*PRISM_Economic*> matters, and align with <*PRISM_Society*>.
Your native language is <*language*>, and you identify as <*religion*>.
You are a strict JSON generator. Analyze the following news headline and output a JSON object with this exact format:

```
{
  "hyperpartisan": "<Boolean>",
  "prct": "<Boolean>",
  "stance": "<pro|against|neutral>",
}
```

Hyperpartisan: Does it show strong ideological bias or use emotionally charged language? (True/False)
PRCT: Does it contain Population Replacement Conspiracy Theory content claiming a deliberate plan to replace native populations? (True/-False)
Stance: What is its stance toward immigration policies? (pro/against/neutral).
DO NOT include any commentary or explanation. Only return valid JSON.
Headline: <headline>

Table 10: Zero-shot Persona prompts.

E Determinantal Point Process

Determinantal Point Processes (DPP) are probability distributions over sets of points, used in physics, statistics, and machine learning for selecting diverse and representative subsets.

Given a large set:

$$A = \{1, 2, \ldots, N\}$$

with items

$$I_A = \{x_1, x_2, \ldots, x_N\},$$

subset selection is computationally expensive since it involves evaluating 2^M possibilities. DPP addresses this by first embedding data points (here, with Sentence-BERT), then constructing a kernel matrix:

$$K_{ij} = k(x_i, x_j)$$

based on similarity. The probability of selecting a subset $Y \subset A$ is:

$$P(Y) = \frac{\det(K_Y)}{\det(K + I)},$$

where K_Y is the restriction of K to the indices in Y.

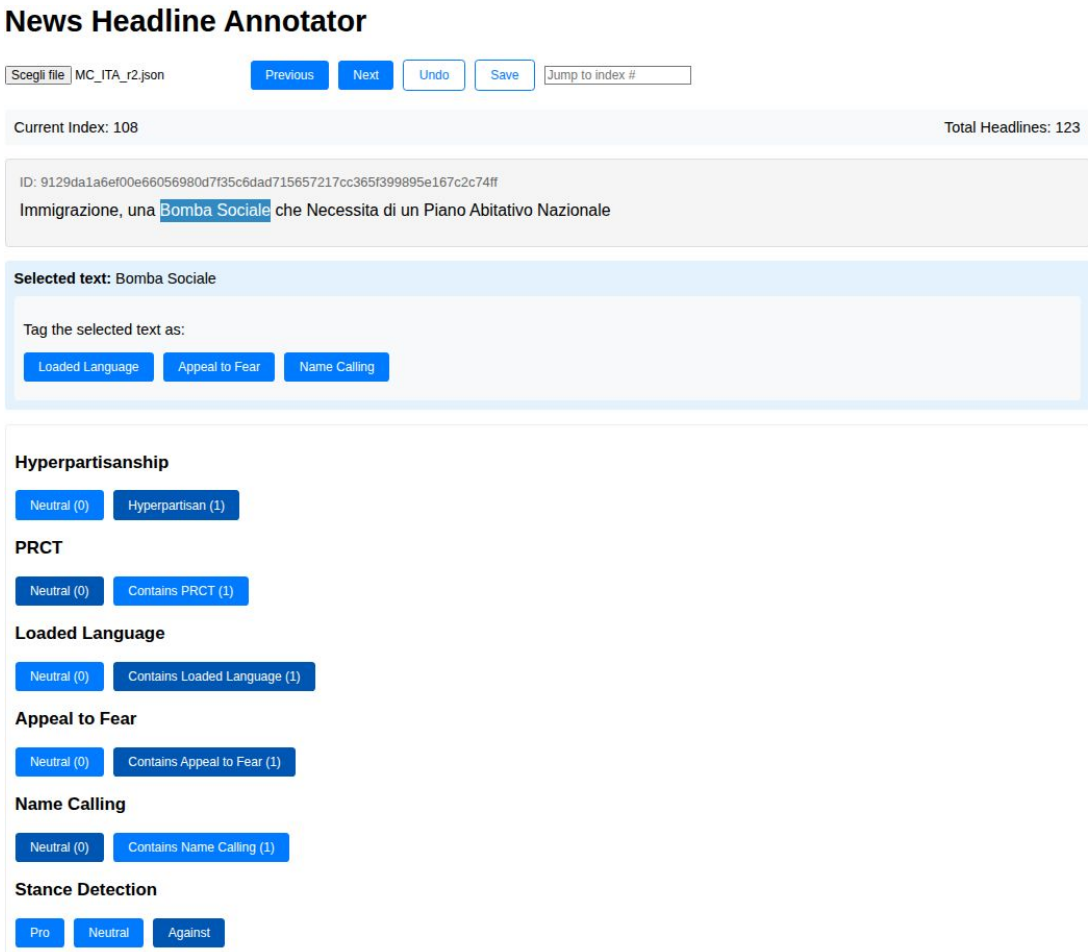

Figure 4: Example of the custom HTML platform used for the annotation process.

The goal is to select the subset:

$$Y_{\text{best}} = \arg \max_{Y \subset A, |Y|=k} \det(L_Y),$$

of fixed size k, which maximises diversity in the representation space. Overall, DPP sampling ensures that selected datapoints are both representative and diverse.

F Classification Errors

Across models and prompting strategies, we observe consistent disparities in how negative and positive instances are classified. All systems show a strong bias toward predicting the negative class, achieving high true-negative rates but struggling to detect positives reliably. Zero-shot Llama 8B exemplifies this imbalance, with very few positive predictions despite strong performance on negatives. Few-shot prompting generally improves recall for the positive class, though often at the cost of misclassifying more negatives. Fine-tuning mitigates this trade-off for some models, particularly Llama 70B, leading to more balanced predictions and a notable increase in positive recall without severe degradation of negative precision. Rationales provide mixed results: while they can improve reasoning transparency, their effect on classification accuracy is unstable and model-dependent. Nemo models remain highly conservative, showing the strongest skew toward negative predictions in all configurations. Overall, the results highlight the difficulty of achieving balanced detection in low-resource or imbalanced settings, emphasising the importance of both fine-tuning and prompting strategies to counteract default negative bias.

G Per-language Analysis of Classification Results

Table 11 presents the performance by language across Spanish (SPA), Italian (ITA), and Portuguese (PT) under the fine-tuning with rationales setting. Looking at the results, hyperpartisan is robust in all three languages, with the best scores in Italian ($F1_{MACRO}$= 0.77, `Llama 3 70B`) and a

Model	Lang.	Hyperpartisan			PRCT			Stance		
		P	R	F1	P	R	F1	P	R	F1
Llama 8B	SPA	0.75	0.75	0.75	0.68	0.64	0.66	**0.65**	**0.71**	0.53
	ITA	0.72	0.63	0.62	0.84	0.73	0.75	0.57	0.44	0.45
	PT	0.66	0.65	0.65	0.59	0.73	0.63	0.57	0.47	0.42
Llama 70B	SPA	0.67	0.68	0.65	0.67	0.82	0.71	0.46	0.53	0.43
	ITA	0.78	**0.77**	**0.77**	0.85	**0.88**	**0.86**	0.54	0.50	0.50
	PT	0.72	0.66	0.60	0.57	0.72	0.59	**0.65**	0.62	**0.62**
Nemo	SPA	0.75	0.76	0.75	0.97	0.79	0.85	0.43	0.47	0.34
	ITA	**0.81**	0.75	0.75	0.83	0.68	0.70	0.60	0.44	0.44
	PT	0.61	0.61	0.60	**0.99**	0.75	0.83	0.53	0.45	0.41

Table 11: Fine-tuning with rationales results by language (macro-averaged).

tie at the top in Spanish ($F1_{MACRO}= 0.75$, Llama 3 8B and Nemo). Portuguese is lower overall (best 0.65, Llama 3 8B). PRCT shows more extreme language effects depending on the model: Italian peaks ($F1_{MACRO}= 0.86$, Llama 3 70B), while Nemo is strongest in Spanish (0.85) and Portuguese (0.83), but lower recall (R≈0.79/0.75), indicating conservative detection that favours salient conspiracy markers.

Stance remains the most challenging and language-sensitive task: Portuguese sees the best result ($F1_{MACRO}= 0.62$, Llama 70B), Italian is moderate (0.50, Llama 70B), and Spanish lags (best $F1_{MACRO}= 0.53$, Llama 8B). Overall, Italian headlines favour higher PRCT and hyperpartisan detectability (consistent with a stronger signal), whereas Spanish stance is notably difficult, likely reflecting a skew toward neutral headlines and subtler polarity cues. Regarding Portuguese, it benefits most for stance from large-scale fine-tuning.