IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News

Michele J. Maggini, Davide Bassi, Angelo Valente, Gaël Dias, Pablo Gamallo
Centro Singular de Investigación en Tecnoloxías Intelixentes da USC
Universidade de Santiago de Compostela, Santiago de Compostela, Spain
University of Padova, Padova, Italy
UNICAEN, ENSICAEN, CNRS, GREYC, Normandie Univ, GREYC UMR 6072, F-14000 Caen, France
{author1name.surname}@usc.es

Abstract

This study investigates the linguistic composition of hyperpartisan paragraphs in Italian news on climate change, Ukraine war, and immigration by publicly disclosing the dataset to ensure reproducibility. We introduce a new corpus, **IHPP**, of 356 articles, for a total of 4,861 paragraphs annotated for hyperpartisan news detection at the paragraph level and enriched with span-level annotations of six semantic-pragmatic linguistic traits: *figurative speech*, *irony/sarcasm*, *epithet*, as well as *hyperbolic* and *loaded language*. We hypothesized that these traits, while violating Gricean maxims, are key mechanisms of hyperpartisan rhetoric. To test this, we fine-tuned a set of mono- and multilingual BERT models for hyperpartisan detection and evaluated their incorporation in the embedding space. Then, we applied explainable techniques, e.g. Integrated Gradients and SHAP to analyze how models allocate attribution to normal and linguistic-trait tokens. Our result show that loaded language is the most discriminative trait. The dataset is released: https://github.com/MichJoM/IHPP-Climate.

Keywords: hyperpartisan detection, dataset, low-resource language, explainability

1. Introduction

Among the many applications of NLP, the automatic detection of hyperpartisan (HP) language has emerged as a critical research area with significant societal implications (Potthast et al., 2018). HP news is characterized by extreme bias favoring one political side, contributing to polarization and misinformation (Kiesel et al., 2019; Baly et al., 2020), particularly in anti-establishment media narratives (Ernesto de León and Adam, 2024). Effectively identifying such content is crucial to maintaining healthy information ecosystems. Existing datasets predominantly focus on English (Horne et al., 2018; Lyu et al., 2024a; Kiesel et al., 2019), while underrepresented languages like Italian require attention, as partisan expression differs across linguistic and political systems (Roberts, 2022). A second major limitation in current approaches is the oversimplification of the HP detection task into a binary classification problem, with inadequate consideration of the underlying linguistic mechanisms. Most of the literature treats it as a standard binary classification task, focusing on lexical characteristics or sentiment (Kiesel et al., 2019), while neglecting the role of pragmatics, how language is used in context to construct meaning beyond the literal level (Maggini and Gamallo Otero, 2024; Lyu et al., 2024b).

There were attempts to address the rhetorical dimension with the introduction of logical fallacies and their persuasive function (Martino et al., 2019; Jin et al., 2022), but other pragmatic elements that shape implied meaning in HP detection, such as figurative speech, irony and sarcasm, epithets remain underexplored in computational methods (Yang et al., 2023; Ziems et al., 2024). In this regard, the function played by the linguistic trait (LT) proxies with regard to HP language is underexplored in the NLP field.

Since HP news targets an already polarized audience (Barkho, 2021), the usage of LT reinforces existing beliefs through a shared vocabulary and implications (Waltinger, 2009). These mechanisms would operate through what pragmatics theory identifies as a violation of the Gricean maxims (Grice, 1975): loaded language corrupts the maximum of quality and manner by introducing diverting evaluative distortion; hyperboles results in the exaggeration of quality, resulting in over-representing information flowing into caricatural language; figurative speech, neologism, irony and sarcasm break the maxim of manner, with the injection of ambiguity with abstract references (e.g. poetic, metasemantic) beyond the literal meaning or a non-obvious vocabulary.

To address these gaps, we investigate selected LT that characterize HP language in Italian news, focusing on different polarizing topics, like Ukraine war and climate change, susceptible to partisan framing (Falkenberg et al., 2022). This paper makes three contributions. First, we introduce IHPP, a paragraph-level Italian corpus annotated for HP detection and enriched with span-level LT annotations across six categories. Second, we incorpo-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1998–2011
11–16 May 2026.

rate LT into BERT-based classifiers via embedding fusion and conduct an ablation study to quantify each trait's discriminative contribution. Third, we apply Integrated Gradients and SHAP to examine whether models rely on LT tokens in ways consistent with their actual predictive value.

2. Related Work

Hyperpartisan News Detection. HP news detection has been widely studied as a downstream task for fine-tuned models (Kiesel et al., 2019; Potthast et al., 2018; Halterman and Keith, 2025). Potthast et al. (2018) found that stylistic characteristics outperform topic-based features for HP detection. Recent work with LLMs shows that while In-Context Learning benefits from rule-based instructions, fine-tuning remains computationally efficient (Maggini et al., 2025c), and encoder models generally outperform generative LLMs for this task in underrepresented languages like Persian (Omidi Shayegan et al., 2024). While treating HP detection task as a binary classification, these works neglect pragmatic dimensions, which could give further insights on how the neural networks learn HP representation.

Domain-specific Datasets. Several datasets address polarizing topics across languages (Kurfali et al., 2025). Piskorski et al. (2023)'s SemEval 2023 Task 3 covers COVID-19, climate change, migration, and the Ukraine war across six languages, including Italian, and it is specific for logical fallacy detection. For climate change specifically, datasets range from contrarian claim detection (Coan et al., 2021) to stance detection (Vaid et al., 2022) and multilingual news analysis (Wasi et al., 2024; Maggini et al., 2025a). Immigration discourse has been studied through large-scale corpus analysis (Blokker et al., 2023), while the Ukraine war's polarization has been examined through partisan news sharing patterns (Zhu et al., 2024; Khairova et al., 2023) and cross-source semantic similarity (Khairova et al., 2024; Pavlichenko, 2022). With regards to Italy's political panorama, Cignarella et al. (2020) collected tweets for stance detection to investigate the phenomenon of "*Movimento delle Sardine*" expanding on previous works based on the referendum of the Italian Constitution (Lai et al., 2019, 2018). Despite the effort, these works largely overlook Italian mainstream and independent news, relying on a narrow range of sources and focusing on specific topics.

Linguistic Trait Detection. The semantic-pragmatic detection of linguistic cues has proved to be a challenging task. Shutova et al. (2016) investigated the field of metaphor identification by simultaneously drawing knowledge from linguistic and visual data. Chakrabarty et al. (2022) released the FLUTE dataset spanning across Metaphor, Sarcasm, Simile and Idioms collected through human-AI collaboration. Sarcasm has been further explored by Oprea and Magdy (2020), who released a dataset of English tweets addressing the distinction between intended and perceived sarcasm. Lastly, Tsvetkov et al. (2014) focused on the discrimination between literal and metaphorical meaning, using lexical semantic features of the words that participate in their construction. Rodríguez et al. (2023) demonstrated that metaphorical framing plays a role in propaganda detection through multi-task learning. During EVALITA 2018, one of the main task was Irony detection and in this occasion Giudice (2018) adopted a variation of LSTM architecture to detect the irony within Italian tweets. Inspired by Dinu et al. (2021), who deepened the field of pejorative language in multilingual setting, efforts in detecting Epithets to express misogyny in Italian on Twitter have been done by Muti et al. (2024). In contrast, we simultaneously annotated and modeled different pragmatic features and jointly assessed their contribution towards HP detection for an underrepresented language.

Explainability and Embedding-Based Enhancements. XAI methods, such as SHAP for feature selection (Bangerter et al., 2023) and Integrated Gradients (IG) for sentiment/sociopsychological markers (Aghababaei et al., 2025), enhance propaganda detection, with LIME and Anchors improving BERT classifier interpretability (Szczepański et al., 2021). Atanasova (2024) found gradient-based techniques, including IG, excel in faithfulness to model internals, while SHAP optimizes confidence in identifying key tokens. Embedding-based methods using n-grams, emotion lexicons, and sentiment features boost HP detection (Mohan and Chen, 2025; Sourati et al., 2023; Kiesel et al., 2019). We advance this by comparing IG and SHAP attributions for LT versus non-LT tokens in HP classification.

3. Dataset

Theoretical Motivation The linguistic traits annotated in this work are not merely stylistic features; they operate at two interconnected levels that are particularly consequential in news discourse. At the discourse level, their introduction corrupts the logical flow of the journalistic message: a news paragraph that begins with neutral reportage on a given topic and incorporates slurs, hyperbolic comparisons, or ironic reframings undergoes a shift in discourse function, violating the Gricean maxim of relation (Grice, 1975) by introducing content that

Bias Type	Definition	Distribution
Hyperpartisan Classification		
Hyperpartisan Language	Text that displays extreme bias favoring one particular political side, often employing pronounced use of rhetorical biases.	HP 1594 N 3267
Linguistic Traits		
Loaded Language (LL)	Parts of speech that are emotionally loaded depending on the context, slurs, attacks against ideologies/parties/institutions.	3578
Figurative Speech (FS)	Usage of metaphors, similes, analogies to depict something or someone with other semantic fields.	1331
Irony/Sarcasm (IS)	Concealment of one's thoughts behind words that have an opposite or different meaning from the literal one. Sarcasm is insincere speech aiming to hurt.	721
Hyperbolic Language (HL)	Adoption of hyperboles and exaggerated comparisons.	1243
Epithet (EP)	Adjectives or phrases expressing a quality or attribute regarded as characteristic of the person or thing mentioned.	1091
Neologism (NEO)	Newly coined words or expressions.	108

Table 1: Taxonomy of LT and HP language detection used in our annotation scheme. We applied the definitions of: LL by Weston (2018); HL by Carston and Wearing (2015) with the consequence of being an independent category than IS and FS.

serves a rhetorical rather than informative purpose. At the semantic level, LT reconstruct the meaning of the information itself. Journalistic language conventionally operates with a relatively denotative mapping between sign and referent (Ogden and Richards, 1930): its primary communicative function is informative. LT, particularly figurative speech and loaded language, disrupt this mapping by substituting or overlaying the conceptual node of the semantic triangle with content drawn from affectively or ideologically charged domains. When a political statement is, for instance, framed as "dropping a bomb" (see 1), the referent (a ministerial remark) is reconstructed through a conceptual metaphor (Lakoff and Johnson, 2008) that imports connotations of violence, surprise, and destabilization into the reader's mental model of the event. In this sense, LT are not incidental to hyperpartisan rhetoric, they are its primary mechanism for reshaping how readers conceptualize political reality.

HP language differs from purely biased language in both intent and degree. Biased language can emerge from structural or architectural choices, such as source selection, framing, or scope of coverage, and may be unintentional, operating across a wide range of topics and domains without a specific polarizing goal (Blodgett et al., 2020). HP language, by contrast, deploys these rhetorical strategies deliberately and specifically in politically polarized contexts, with the explicit intent of reinforcing pre-existing partisan beliefs rather than informing (Maggini et al., 2025b; Walker et al., 2025). It is precisely this intentionality, operating through the semantic-pragmatic mechanisms described above, that distinguishes HP rhetoric from incidental bias and makes the automatic detection of LT a meaningful proxy for hyperpartisanship.

3.1. Source Selection

To investigate HP language in Italian, we combined data from different sources. We re-annotated Italian datapoints from Semeval 2023 Piskorski et al. (2023)'s and De Mattei et al. (2020)'s datasets, which cover different topics such as the Ukraine war and Italian/European politics from mainstream and "alternative" media sources. After removing duplicates, we strategically sampled articles to balance dataset composition: given the typically lower frequency of hyperpartisan content in mainstream sources, we included a higher proportion of articles from nicolaporro.it[1] to achieve sufficient coverage of HP instances for robust analysis. This yielded 192 articles from Piskorski et al. (2023), 45 from De Mattei et al. (2020), and 119 from nicolaporro.it. We acknowledge that with this decision we are introducing source bias and that other italian sources should be studied to generalize our findings.

3.2. Data Preparation

SemEval 2023 data required no additional processing, being already paragraph-segmented. In contrast, De Mattei et al. (2020)'s dataset needed paragraph splitting. Textual segments with partial sentences or verbless clauses (<20 tokens) were merged to ensure discourse coherence and adequate analytical context for LT detection. Remaining texts were then segmented into paragraphs using the HTML <p> tags.

3.3. Annotation Framework

Our annotation framework addresses two core constructs: hyperpartisanship at the paragraph level, and LT at the span level (see Figure 1).

Hyperpartisanship is characterized as extremely one-sided content marked by strong bias (Maggini et al., 2025b). We model it as a binary paragraph-level classification to track the distribution of HP content throughout paragraphs (see Figure 2). LT are grounded in pragmatics theory, which is the

[1] https://www.nicolaporro.it/articoli/ambiente-sostenibilita/

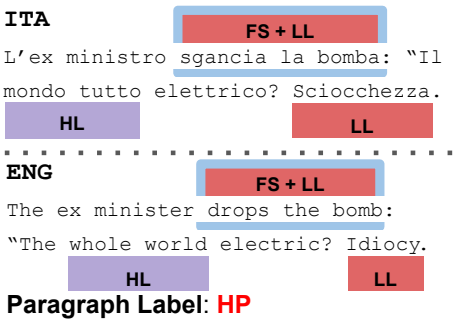

Figure 1: Example of the annotation scheme for one paragraph. FS: Figurative Speech, LL: Loaded Language, HL: Hyperbolic Language.

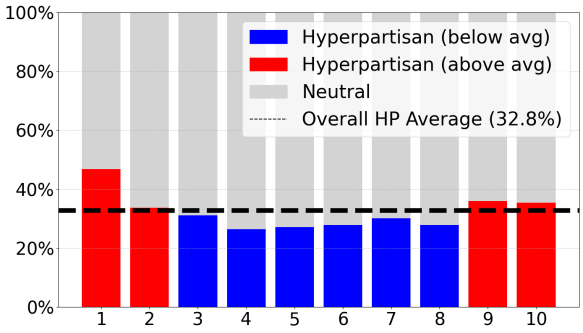

Figure 2: Distribution of HP paragraphs across article deciles.

Statistic	Train	Test
Par. in Climate Change	1254	318
Par. in Ukraine War	1077	250
Par. in EU Institutions	801	214
Par. in Immigration	503	123
Par. in COVID	253	68
Avg. words per paragraph	47.54	48.00
Avg. words per article	520.62	154.65
Total words	184,819	46,705
# Loaded Language	2861	717
# of Figurative Speech	1060	271
# of Epithet	846	245
# of Neologism	87	21
# of Irony/Sarcasm	573	148
# of Hyperbolic Language	984	259

Table 2: Dataset statistics.

study of language in context (Scott, 2022) and they potentially undermine the Gricean maxims of quality, relevance, quantity, and manner (Grice, 1975). We compiled a list of 10 LT and conducted a pilot annotation on 15% of our data. This exploratory phase allowed us to homogenize LT names and unify those with similar definitions under the same label (e.g., Irony and Sarcasm → Irony/Sarcasm, Metaphors and Similarities → Figurative Speech), leading to a total of 6 LT types. Table 1 presents our taxonomy and the LT's distribution.

3.4. Annotation Process

Two native Italian Ph.D. students with expertise in NLP and disinformation analysis annotated the dataset using a custom web interface that masked outlet names to prevent source-related bias. The annotation process consisted of three phases: **(1) Training Phase**: annotators studied guidelines, conducted pilot annotations, and participated in interactive sessions to clarify edge cases; **(2) Annotation Phase**: each paragraph was independently annotated for hyperpartisanship (paragraph-level) and LT (word/span-level). Although annotators had access to full articles, paragraphs were annotated consecutively to preserve local narrative flow and discursive coherence; **(3) Curation Phase**: all disagreements were collaboratively resolved through discussion and review of previous annotations. Inter-annotator reliability (IAR) was high, with Krippendorff's α of 0.937 for HP detection, and an average Jaccard Similarity of 0.814 for LT, indicating strong agreement across all dimensions.

3.5. IAR Error Analysis

Jaccard similarity scores for LT reveal high annotation reliability: Neologism (0.949), Epithet (0.88), Irony/Sarcasm (0.847), and Figurative Language (0.801) show strong agreement. Only Loaded and Hyperbolic Language scored slightly lower (0.709 and 0.701), due to their similar characteristics.

3.6. Data Statistics and Analysis

Italian **H**yper**P**artisan and **P**ragmatics (IHPP) in News is, to our knowledge, the largest dataset on Italian HP news detection focused on the following main topics: Ukraine war, climate change, Immigration and European Institutions (see Table 2) and is annotated considering pragmatic layers. The final dataset comprises 356 articles containing 4,861 paragraphs, with 8,104 annotated LT spans divided into train and test with 80-20 proportions after randomly shuffling it. We applied a stratified split to maintain consistent proportions of HP and neutral paragraphs across training and test sets. The dataset is slightly imbalanced, with more neutral paragraphs (3,267) than hyperpartisan ones (1,594). This distribution is expected and, in our view, representative: even highly biased news must include neutral and functional paragraphs to ensure clarity and coherence in communication, rather than consisting solely of hyperpartisan content.

By looking at Figure 2, HP paragraphs concentrate in the first decile (46.8%), drop to about one third in the second decile—a proportion that is also observed at the end of the articles—and remain below average (32.8%) throughout the main

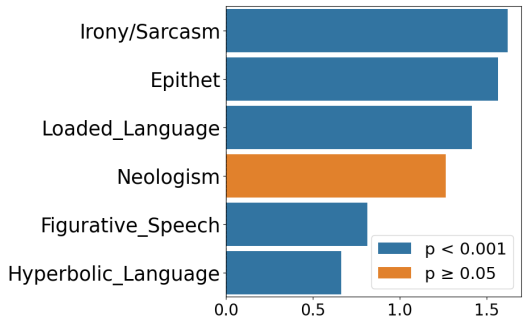

Figure 3: We measured the effect sizes using Log Ratio (LR): the logarithm base 2 of the ratio of the frequencies between the two groups. LR = 0: equal frequency in both groups; LR>0: higher frequency in the hyperpartisan group; LR<0: higher frequency in the neutral group.

body. All LT categories peak sharply in the articles' first decile (titles and leads), aligning with clickbait and persuasive strategies used to influence reader engagement and shape initial narratives (Chakraborty et al., 2016), a tactic common in sensationalist media (Munger, 2020; Blom and Hansen, 2015). However, while overall HP frequency decreases after the opening, the core persuasive function is sustained internally. Specifically, Hyperbolic Language, Irony/Sarcasm, and Figurative Speech maintain stable frequencies throughout the article, suggesting these traits serve not only to attract readers but to reinforce bias and pushing particular agendas within the article's main body.

Quantitatively, all identified LT categories demonstrate their value in distinguishing HP content, as evidenced by their positive Log Ratio (LR) (see Figure 3). This positive LR indicates that every LT is more frequent in the HP group compared to the neutral group. We confirmed statistical significance using a chi-squared test; however, only Neologism shows a non-significant p-value, likely due to its comparatively low frequency within the corpus.

4. Experimental Setup

4.1. Models, Fine-Tuning and Ablation

We selected mono and multilingual BERT language models with different sizes and training strategies: `dbmdz/bert-base-italian-uncased` (ita-uncased), `dbmdz/bert-base-italian-xxl-uncased` (ita-xxl-uncased), `nickprock/sentence-bert-base-italian-uncased` (sent-ita), `nickprock/sentence-bert-base-italian-xxl-uncased` (sent-ita-xxl), and `google-bert/bert-base-multilingual-cased` (multi-uncased).

Firstly, we fine-tuned (FT) the selected models for HP detection. Then, to verify the effectiveness of introducing LT (see Table 3), we incorporated the LT by appending their textual spans to the input text as a paired sentence, processing them through the model to obtain pooled embeddings, and fusing these with a gating mechanism (FT+Embeddings). Following Sun et al. (2019), our model extracts hidden states from the BERT encoder and applies average pooling separately over the text segment and traits segment, excluding special tokens. The pooled text and traits features are concatenated and passed through a fusion layer with a gating mechanism, implemented as a linear layer followed by a sigmoid activation, to combine the features dynamically. The fused representation is then fed into a classifier, consisting of a linear layer, ReLU activation, dropout (0.2), for HP classification. Lastly, we performed an ablation study to investigate the most contributive features and compared this setting against FT+Embeddings. We selected the optimal hyperparameters for each model using Optuna[2] and ran the models for 5 runs each and 2 epochs to avoid overfitting, given the limited size of the dataset.

4.2. Integrated Gradient and SHAP

To understand the LT token-level contributions to our classification task, we employed two post-hoc explanation methods for local attribution: Integrated Gradient (IG), and SHAP (Lundberg and Lee, 2017). Unlike simpler methods such as saliency maps, IG computes the integral of gradients along a linear path from a baseline input to the actual input. Following Sundararajan et al. (2017), we computed LT and non-LT token's importance by IG along this path from a zero-embedding baseline to the input embedding. We respectively calculated the attribution ratio A_{ratio} of LT and non-trait tokens as defined in Equation 1, where $\hat{a}_i$ is the result of applying IG.

$$A_{\text{ratio}} = \frac{\text{median}\big(\{\hat{a}_i \mid i \in \text{span}\}\big)}{\text{median}\big(\{\hat{a}_j \mid j \in \text{all valid tokens}\}\big)} \quad (1)$$

Thus, values of 1 mean the trait span's importance is average (similar to the text's median); >1: the trait span is more important than average (i.e. model attributes higher scores to it); <1: the trait span is less important than average.

Then, we applied SHAP to show feature contribution directionality. It calculates the contribution of each feature to a specific prediction by considering all possible combinations of features, assigning a unique value to each feature that represents its fair share of the prediction output. This makes it particularly effective at identifying how features push the prediction from the baseline value.

[2] https://github.com/optuna/optuna?tab=readme-ov-file

In particular, we first mapped token span positions removing special tokens and computed IG and SHAP for tokens annotated with LT. Their attribution scores were then compared against those of other tokens in the same text. This method allows us to identify which tokens most significantly influence HP classification.

5. Experiments and Results

5.1. Fine-Tuning and Ablation Study

Fine-Tuning. As shown in Table 3, FT models achieved F1 scores ranging from .673 (multi-uncased) to .729 (sent-ita-xxl). In this setting, models pretrained on NLI datasets and with bigger vocabulary outperformed those linear-probed on full-text corpora (ita-uncased and its xxl variant), as sentence-based models optimize for sentence-level semantics rather than word-level predictions. Sent-ita-xxl achieved the highest F1 (.867), precision (.866) and accuracy (.867), demonstrating that sentence-BERT architecture is particularly suited for document classification tasks like HP detection. XXL variants, with larger vocabularies and more parameters, improved precision at a slight recall cost, yielding higher F1 scores through more confident but selective predictions. Lastly, although being trained on Italian, the multi-uncased model performed poorly (F1 .673) compared to the other models, likely because its cross-lingual generalization capability compromises performance in specialized domains.

FT Error Analysis. Ita-xxl-uncased predicted on the test set 220 true positive and 605 true negative, one less than sent-ita-xxl, which, in contrary, predicted 591 true negative. The multilingual-uncased labeled 126 false positive cases. Across all the models, the paragraphs containing Loaded Language were the most classified as false negative and false positive, but they are also the most frequent. However, the pattern stays the same across the models, highlighting a hierarchy of task complexity which also follows the distribution of the LT: False Negative: LL > HL> FS> IS> EP> NEO; False Positive: LL> FS> EP> HL> IS> NEO.

Fine-Tuning with Embeddings. All the models benefitted from FT+EMB configuration. Particularly, the multilingual variant's performance F1 increased by 18 points, and its Precision by 28. This enhancement indicated that the introduction of LT embeddings improved the detection of positive instances while incorporating this kind of information during FT, and confirmed the utility of linguistic cues. Particularly, the multi-uncased showed that the task input helped compensate for the lack of strong dataset-specific representations in its pretrained embeddings.

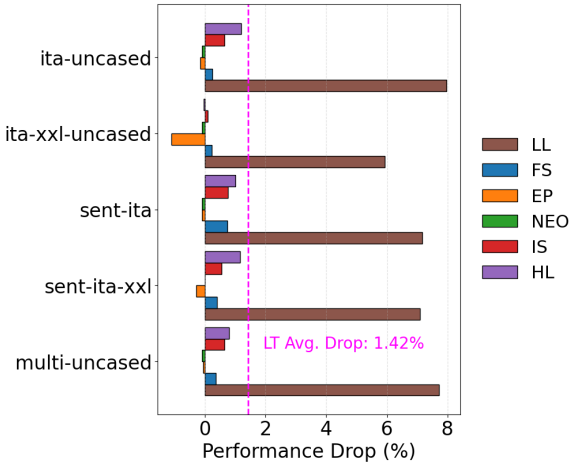

Figure 4: LT average drop by model in ablation.

FT+EMB Error Analysis. This setting confirms that the trend in misclassification remains consistent with previous results, following the established hierarchy. The only exception is Neologism, which now appears exclusively in paragraphs annotated as false positives.

Ablation Study. Ablating LT categories revealed different impacts (see Figure 4) and an average F1 drop of 1.42%. Removing Loaded Language caused the largest drop (7.17%), while Figurative Speech, Irony/Sarcasm, and Hyperbolic Language had minimal negative effects (0.39%, 0.54%, 0.82%). Paragraphs containing at least Loaded Language account for 73.3% of all errors, confirming the difficulty of the task and the positive contribution of converting this specific LT into an embedding representation. Instead, removing Epithets and Neologism slightly improved F1 by 0.35% and 0.08%. Despite their positive ablation impact, they do not introduce confusion, since their error rates are low (see Table 4), suggesting they may suffer from redundancy with co-occurring LT like Loaded Language, and they are insufficient as standalone predictive features in current model architectures.

To investigate this redundancy, we analyzed co-occurrence patterns between LT categories (see Table 5). The high conditional probabilities reveal that most LT tokens co-occur with Loaded Language: 82.5% of Epithets instances and 76.5% of Neologism instances appear alongside Loaded Language. However, their Pearson correlations with Loaded Language differ substantially (0.41 for Epithet vs. 0.10 for Neologism). Epithets appear in 2.8% of HP paragraphs, nearly half the time with Loaded Language, while Neologisms are rare (0.17%) and only 4 times appear isolated.

Epithet and Neologism demonstrate higher error rates when isolated from LL (23.5% and 50.0% respectively), but their statistical insignificance (p >

Model	FT				FT+Embeddings			
	Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
ita-uncased	.808 ± .017	.677± .024	.749 ± .022	.711± .018	.863±.003	.863±.002	.864±.003	.863±.002
ita-xxl-uncased	.818 ± .012	.731± .022	.706 ±.024	.718± .033	.863±.007	.862±.006	.863±.007	.862±.006
sent-ita	.790± .009	.654± .008	.765± .008	.705± .014	.859±.006	.859±.010	.859±.006	.858±.008
sent-ita-xxl	.828± .014	.753± .012	.706± .009	.729 ± .012	**.867**±.008	**.866**±.009	**.867**±.008	**.867**±.009
multi-uncased	.739 ± .011	.571 ±.024	.819± .086	.673±.041	.854 ± .002	.852 ± .003	.854± .002	.853 ±.002

Table 3: Performance of models on HP detection. FT best scores are underlined. FT+EMB are bolded.

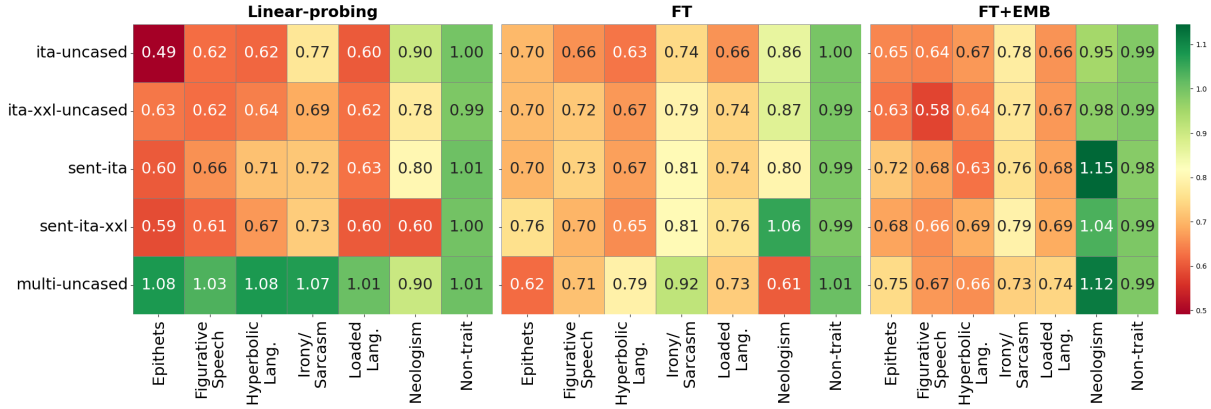

Figure 5: Attribution ratios values A_{ratio} by Models × LT × Learning variants.

Linguistic Trait	Error (%)	FP (%)	FN (%)
Figurative Speech	23.72	19.07	4.65
Neologism	23.53	23.53	0.00
Loaded Language	21.90	12.66	9.23
Hyperbolic Language	18.43	12.90	5.53
Irony/Sarcasm	18.03	13.93	4.10
Epithet	17.49	16.39	1.09

Table 4: FT+EMB error rates for LT. Error (&) represents the percentage of paragraphs containing that trait that are misclassified. Because paragraphs can contain multiple traits, errors are counted toward each trait present.

Condition	Expression	Value
EP → LL	$P(\text{LL} \mid \text{EP})$	0.8251
EP → NEO	$P(\text{NEO} \mid \text{EP})$	0.0492
NEO → LL	$P(\text{LL} \mid \text{NEO})$	0.7647
NEO → EP	$P(\text{EP} \mid \text{NEO})$	0.5294
LL → EP	$P(\text{EP} \mid \text{LL})$	0.3984
LL → NEO	$P(\text{NEO} \mid \text{LL})$	0.0343

Table 5: Co-occurrences of EP, LL, NEO.

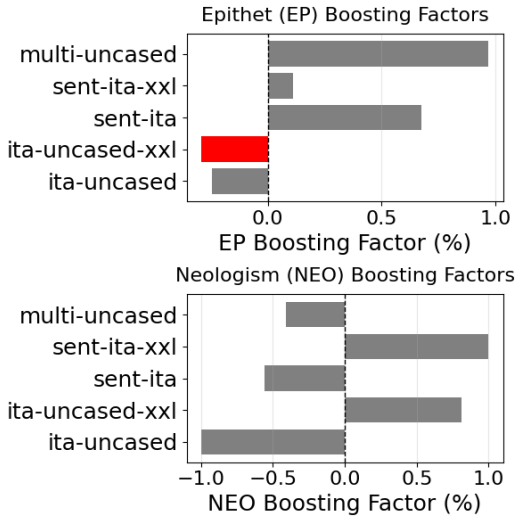

Figure 6: Cross-model comparison of boosting effects (%) in IG attribution for LT, showing context-dependent patterns influenced by LL.

0.45) and limited occurrence, especially for Neologism, make them unsuitable as robust features given overfitting concerns.

The combination of (1) proportional error rates, (2) high P(Loaded Language|LT), and (3) negative ablation impact suggests Epithet and Neologism function as Loaded Language proxies rather than independent discriminative features. In fact, when Loaded Language is present, these features provide redundant information; when ablated, models rely more heavily on Loaded Language's robust signal, improving performance. The ablation impact was not clearly tied to vocabulary size or language, as Irony/Sarcasm affected all models negatively, but ita-xxl-uncased reduced its impact compared to ita-uncased, while sentence models showed the opposite trend. Learning polarized vocabulary (i.e., Neologisms) via FT+EMB aided HP detection, and its ablation increased false positive for this LT, underscoring LTs' role as hyperpartisanship proxies.

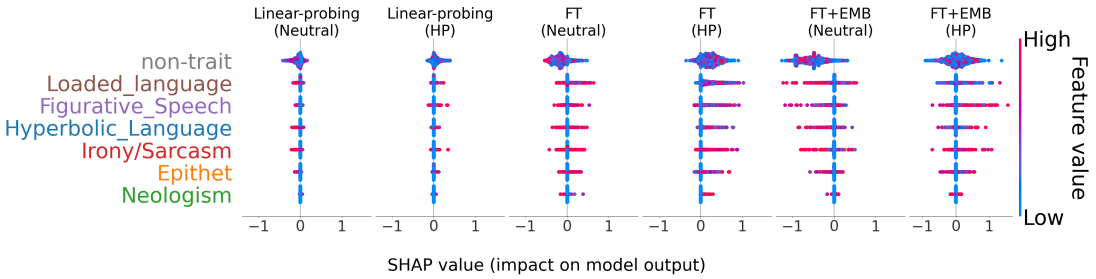

Figure 7: Comparison of SHAP values x learning strategies ordered by magnitude and split by predicted label. Samples predicted as HP have positive SHAP value toward label HP, viceversa for Neutral.

5.2. IG and SHAP Analysis

Integrated Gradient. As shown in Figure 5, IG analysis revealed that models frequently over-attribute importance to Neologism and Epithets, despite their minimal contribution to predictive performance. Their high attribution scores stem from the co-occurrence with Loaded Language, rather than genuine discriminative power, indicating that these LTs function as redundant, correlation-driven cues rather than independent features. Figure 5 also shows IG attribution ratios for LTs, with non-trait tokens averaging 1 as a baseline (per Equation 1).

We compared three model types to analyze their attribution of LT in detecting HP text: linear-probing (out-of-the-box models), FT and FT+EMB, finding that fine-tuned models achieved higher attribution scores, indicating robust trait detection after task-specific training.

Across models, a consistent hierarchy of LT emerged: Figurative Speech received the highest attribution scores in FT models (except for the multi-uncased variant), followed by Irony/Sarcasm, Loaded Language, and Epithets. This suggests models prioritize emotionally charged or persuasive traits as hyperpartisan signals, regardless of architecture.

Neologism attribution improved significantly in FT+EMB (+31.2% vs. linear-probed, +24.9% vs. FT), exceeding non-trait baselines (e.g., 1.15 for non-XXL models). However, ablation studies revealed removing Neologism improved F1 by 0.08% while removing Loaded Language caused a 7.17% drop, indicating its far stronger influence.

To disentangle whether high attribution stems from genuine discriminative power or Loaded Language co-occurrence, we conducted a conditional attribution analysis comparing isolated vs. co-occurring instances (see Equation 2). First, we verified whether vocabulary size differences could explain model variation. Tokenization analysis revealed all models split neologisms into subwords (0% single-token rate), confirming neologisms lack dedicated representations across all architectures.

$$\frac{\text{median}(A_{ratio}\ \text{LL}) - \text{median}(A_{ratio}\ \text{isolated})}{|\text{median}(A_{ratio}\ \text{isolated})\,|} \quad (2)$$

Conditional IG scores showed Neologism boosting factors ranging from +81.4% to +100% in some models and -41.2% to -100% in others, yet δF1 $\leq$ 0.08% (p>0.05) across all architectures (see 6. Given only 4 isolated Neologism instances in the test set, this ±100% variance reflects statistical noise rather than meaningful architectural differences. Combined with high Loaded Language co-occurrence (76.5%), the attribution likely stems from spurious correlations rather than genuine discriminative value.

Epithets showed similar patterns: boosting factors from 67.7% (sent-ita) to 97.1% (multi-uncased) with non-significant ablation impact (p>0.05), indicating Loaded Language dependency. Despite its attribution scores, Epithets provide minimal independent discriminative value due to high co-occurrence with Loaded Language (P(Loaded Language|Epithets) = 0.825).

Our analysis reveals that Neologism and Epithets exhibit correlation-driven attribution: models assign high importance scores due to co-occurrence with Loaded Language (P(LL|NEO)=0.765, P(LL|EP)=0.825) rather than independent discriminative power, as evidenced by minimal ablation impact. This demonstrates that high attribution scores indicate model reliance but not necessarily feature utility and IG reveals what models *do* learn, not what they *should* learn.

SHAP analysis. Fig. 7 represents SHAP values toward HP predictions across different models and learning strategies (see also Table 6). Each dot represents a paragraph, with red indicating frequent trait presence and blue indicating infrequent occurrence.

The linear-probed models show SHAP values concentrated near zero across both neutral and HP predictions, indicating it does not rely on LT to distinguish between classes. This suggests the model lacks domain-specific knowledge about which linguistic features signal hyperpartisanship, treating all LT as approximately equal in importance regard-

Feature	LP		FT		FT+EMB	
	N	H	N	H	N	H
EP	-.0013	.0008	.0088	.0290	-.0048	.0068
FS	-.0012	.0013	.0040	.0198	-.0067	.0133
HL	-.0012	.0010	-.0003	.0087	-.0099	.0014
IS	-.0011	.0008	.0018	.0116	-.0062	.0034
LL	-.0013	.0011	.0104	.0318	-.0050	.0111
NEO	.0003	.0008	.0094	.0118	-.0023	.0025
NT	-.0015	.0012	-.0048	.0095	-.0142	-.0004

Table 6: Avg SHAP by LT, Label and Config.

less of their actual discriminative power.

In contrast, FT enables trait-based discrimination. Indeed, the FT model learns to identify HP language from contextual patterns without explicit trait embeddings. Loaded Language, Figurative Speech, and Irony/Sarcasm demonstrate clear discriminative power by contributing positively to HP predictions (predominantly positive SHAP values in the HP column) and negatively or near-zero to neutral predictions. Instances with low trait occurrence (i.e. blue dots) in the neutral column show negative or near-zero SHAP values, confirming that trait absence contributes to neutral classification. However, the relatively tight clustering suggests that their contribution is primarily determined by their presence versus absence, with limited contextual modulation. On the other hand, FT+EMB model exhibits substantially increased SHAP value variability, with Loaded Language and Figurative Speech spanning both positive and negative ranges even within HP predictions. This context-dependent behavior indicates that the same trait can reinforce HP classification in some paragraphs while contributing weakly or negatively in others, depending on surrounding context (e.g., Loaded Language in quoted speech may be interpreted differently than when used directly by the author). This contrasts sharply with FT, where traits show consistent directionality: negative for neutral, positive for HP. The embeddings enable the model to weight trait importance based on broader context, leading to more nuanced but less linearly separable feature contributions. The non-trait feature evidences this contextual sophistication showing minimal spread in linear-probing and moderate activity in FT, and exhibiting the widest SHAP value distribution in FT+EMB. This indicates the embeddings enable richer extraction of contextual signals from surrounding text, with the model learning complex relationships between trait-bearing and non-trait tokens rather than treating non-annotated text as background noise. However, not all LT prove equally discriminative. Neologism and Epithet consistently show weaker SHAP magnitudes and tighter clustering across all models. For Neologism, this stems from its rarity combined with high Loaded Language co-occurrence (76.5%). For Epithet, despite moderate frequency, its high correlation with Loaded Language (82.5% co-occurrence) limits independent discriminative power.

6. Conclusion

This paper offers a threefold contribution. First, it presents a novel Italian corpus comprising 4,861 paragraphs sourced from diverse sources and different political leanings, focusing on polarizing topics. Second, it proposes an innovative approach by integrating semantic-pragmatic LT at the span level, enhancing HP detection beyond traditional methods. This technique addresses prior limitations of linguistics-based models by offering a perspective on how models interpret and process such information, proving that not all LT carries equal discriminative power (e.g. Neologism) depending on their frequency or high co-occurence with other, more impactful LT. This highlights a broader challenge in computational pragmatics: theoretically valid linguistic categories may not translate to effective machine learning features. Third, it evaluates the combined impact of LT using explainability techniques and reveals their varying contributions.

Limitations

Annotation. The annotation process, which proved to be highly time- and labor-intensive, was conducted ensuring the well-being of annotators and took 80 hours. Since they volunteered, no compensation was required. Annotators gave informed consent, were fully briefed on the study's objectives, and retained the right to withdraw at any time.

Scope. We recognize that labeling content as "hyperpartisan" carries deep normative and political implications that extend beyond linguistic analysis. This classification process involves making value judgments about the boundaries of acceptable political discourse, which can have significant consequences for democratic participation. Moreover, our dataset could be misused with malicious intent like censoring political discourse and silencing legitimate viewpoints. We stress that any automated model trained on this data should serve only as a decision-support tool. Human moderators must retain ultimate oversight, providing contextual judgment that mitigates systematic biases against specific groups or communication styles.

SemEval 2023 comparison. We reannotated the SemEval 2023 task 3 dataset to be suitable for HP detection. The logical fallacy detection is out of the scope of the paper.

Infrastructure. The current version of the paper does not contain either the computational infrastruc-

tures and the optuna parameters we used, because they are intended to be in the Appendix.

Cultural and Contextual Sensitivity. Italian political discourse operates within specific cultural, historical, and social contexts that may not be fully captured by our linguistic analysis. Content that appears hyperpartisan to our models may represent normative political expression within Italian democratic traditions, risking the inappropriate pathologization of culturally embedded political styles.

Source Bias. While incrementing the dataset, we selected only a specific with right-wing media outlet. In this way, we recognize that we are limiting the efficiency of our study and that further works should consider investigating multiple sources, comprising extremist left-wing sources, to generalize our findings.

Ethics Statement

The dataset contains harmful content such as slurs, and loaded language against political actors and institutions.

We clarify that this study is not intended to assess or critique the author's viewpoints, but to explore linguistic and stylistic markers typical of hyperpartisan narratives. All textual data were used solely for academic, non-commercial research purposes. No personal or sensitive information was collected or analyzed.

We recognize that labeling content as "hyperpartisan" can carry normative and political implications. To mitigate this, we rely on established definitions in prior work and focus strictly on the linguistic and rhetorical features of the text, not on truthfulness or misinformation per se. Furthermore, we are committed to avoiding any misuse of this research, including its potential deployment for censorship or political targeting.

We aim for full transparency by documenting dataset sources, annotation procedures, and model behavior, and we encourage further studies using a broader range of sources to minimize ideological bias.

We caution that the fine-grained annotations provided in this work could be misused. In particular, they may be exploited to develop more sophisticated forms of hyperpartisan language or to fine-tune LLMs in ways that facilitate the spread of misinformation. Researchers and developers should remain vigilant about these potential risks and consider appropriate safeguards when applying these annotations.

7. Acknowledgements

This paper was funded from the Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1: HYBRIDS Project. It was also funded by MCIU/AEI (PID2024-161928OB-I00 and AIA2025-163322-C62) and by the Galician Government (Research Center of Galicia accreditation 2024-2027 ED431G-2023/04 and GPC ED431B 2025/16).

8. Bibliographical References

Ali Aghababaei, Jan Nikadon, Magdalena Formanowicz, Maria Laura Bettinsoli, Carmen Cervone, Caterina Suitner, and Tomaso Erseghe. 2025. Application of integrated gradients explainability to sociopsychological semantic markers.

Pepa Atanasova. 2024. A diagnostic study of explainability techniques for text classification. In *Accountable and explainable methods for complex reasoning over text*, pages 155–187. Springer.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Dinkov, Ahmed Ali, James Glass, and Preslav Nakov. 2020. What was written vs. who read it: News media profiling using text analysis and social media context. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3364–3374, Online. Association for Computational Linguistics.

Micaela Bangerter, Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Alberto Volpe, Carmen De Maio, and Claudio Stanzione. 2023. Unisa at semeval-2023 task 3: A shap-based method for propaganda detection. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 885–891.

Leon Barkho. 2021. Without fear or favor? the social reality of partisan language. In Stephen J. A. Ward, editor, *Handbook of Global Media Ethics*, pages 459–478. Springer Verlag.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5454–5476.

Nico Blokker, André Blessing, Erenay Dayanik, Jonas Kuhn, Sebastian Padó, and Gabriella Lapesa. 2023. Between welcome culture and border fence: A dataset on the european refugee

crisis in german newspaper reports. *Language Resources and Evaluation*, 57(1):121–153.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100.

Robyn Carston and Catherine Wearing. 2015. Hyperbolic language and its relation to metaphor and irony. *Journal of Pragmatics*, 79:79–92.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 9–16.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2020. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. In *CEUR Workshop Proceedings*, pages 1–10. Ceur.

Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific reports*, 11(1):22320.

Lorenzo De Mattei, Michele Cafagana, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 2020. Change-it@ evalita 2020: Change headlines, adapt news, generate. In *Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*. European Language Resources Association (ELRA).

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. 2021. A computational exploration of pejorative language in social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mykola Makhortykh Ernesto de León and Silke Adam. 2024. Hyperpartisan, alternative, and conspiracy media users: An anti-establishment portrait. *Political Communication*, 41(6):877–902.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. 2022. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780.

Valentino Giudice. 2018. Aspie96 at ironita (evalita 2018): Irony detection in italian tweets with character-level convolutional rnn. *Proceedings of EVALITA*, pages 160–165.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Andrew Halterman and Katherine A. Keith. 2025. Codebook llms: Evaluating llms as measurement tools for political science concepts.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. 2018. Assessing the news landscape: A multi-module toolkit for evaluating the credibility of news. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 235–238, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection.

Nina Khairova, Andrea Galassi, Fabrizio Lo Scudo, Bogdan Ivasiuk, Ivan Redozub, et al. 2024. Unsupervised approach for misinformation detection in russia-ukraine war news. In *CEUR WORKSHOP PROCEEDINGS*, volume 3722, pages 21–36. CEUR-WS.

Nina Khairova, Bogdan Ivasiuk, Fabrizio Lo Scudo, Carmela Comito, and Andrea Galassi. 2023. A first attempt to detect misinformation in Russia-Ukraine war news through text similarity. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 559–564, Vienna, Austria. NOVA CLUNL, Portugal.

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.

Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. 2025. ClimateEval: A comprehensive benchmark for NLP tasks related to climate change. In *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207, Vienna, Austria. Association for Computational Linguistics.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2018. Stance evolution and twitter interactions in an italian political debate. In *International conference on applications of natural language to information systems*, pages 15–27. Springer.

Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 2019. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions.

Hanjia Lyu, Jinsheng Pan, Zichen Wang, and Jiebo Luo. 2024a. Computational assessment of hyperpartisanship in news titles. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 999–1012.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024b. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Michele Maggini and Pablo Gamallo Otero. 2024. Leveraging advanced prompting strategies in LLaMA3-8B for enhanced hyperpartisan news detection. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 531–539, Pisa, Italy. CEUR Workshop Proceedings.

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.

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.

Michele Joshua Maggini, Erik Bran Marino, and Pablo Gamallo Otero. 2024. Leveraging Advanced Prompting Strategies in Llama-8b for Enhanced Hyperpartisan News Detection.

Michele Joshua Maggini, Dhia Merzougui, Rabiraj Bandyopadhyay, Gaël Dias, Fabrice Maurel, and Pablo Gamallo. 2025c. Are llms enough for hyperpartisan, fake, polarized and harmful content detection? evaluating in-context learning vs. fine-tuning.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-Grained Analysis of Propaganda in News Articles. ArXiv:1910.02517 [cs].

Karthik Mohan and Pengyu Chen. 2025. Embedding-based approaches to hyperpartisan news detection. *arXiv preprint arXiv:2501.01370*.

Kevin Munger. 2020. All the news that's fit to click: The economics of clickbait media. *Political Communication*, 37(3):376–397.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Lorenzo Musetti, Samuel Algherini, Silvia Ronchi, Gianmarco Saretto, Caterina Zapparoli, and Alberto Barrón-Cedeño. 2024. Pejorativity: Disambiguating pejorative epithets to improve misogyny detection in italian tweets. *arXiv preprint arXiv:2404.02681*.

Charles Kay Ogden and Ivor Armstrong Richards. 1930. *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt, Brace.

Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. 2024. An evaluation of language models for hyperpartisan ideology detection in Persian Twitter. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 51–62, Torino, Italia. ELRA and ICCL.

Silviu Oprea and Walid Magdy. 2020. iSarcasm: A dataset of intended sarcasm. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online. Association for Computational Linguistics.

L Pavlichenko. 2022. Polarization in media political discourse on the war in ukraine: critical discourse analysis. *Alfred Nobel University Journal of Philology*, 2(24):214–223.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 2018. A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 231–240, Melbourne, Australia. Association for Computational Linguistics.

Kenneth M. Roberts. 2022. Populism and polarization in comparative perspective: Constitutive, spatial and institutional dimensions. *Government and Opposition*, 57(4):680–702.

Daniel Baleato Rodríguez, Verna Dankers, Preslav Nakov, and Ekaterina Shutova. 2023. Paper bullets: Modeling propaganda with the help of metaphor. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 472–489.

Kate Scott. 2022. *Figurative Language*, Cambridge Introductions to the English Language, page 165–188. Cambridge University Press.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. 2023. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 380–385, Minneapolis, Minnesota. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Mateusz Szczepański, Marek Pawlicki, Rafał Kozik, and Michał Choraś. 2021. New explainability method for bert-based model in fake news detection. *Scientific reports*, 11(1):23705.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Alexander C Walker, Jonathan A Fugelsang, and Derek J Koehler. 2025. Partisan language in a polarized world: In-group language provides reputational benefits to speakers while polarizing audiences. *Cognition*, 254:106012.

Ulli Waltinger. 2009. Polarity reinforcement: Sentiment polarity identification by means of social semantics. In *AFRICON 2009*, pages 1–6. IEEE.

Azmine Toushik Wasi, Wahid Faisal, Taj Ahmad, Abdur Rahman, and Mst Rafia Islam. 2024. Dhoroni: Exploring bengali climate change and environmental views with a multi-perspective news dataset and natural language processing.

Anthony Weston. 2018. *A rulebook for arguments*. Hackett Publishing.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Yiming Zhu, Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Yuyang Wang, and Pan Hui. 2024. A study of partisan news sharing in the russian invasion of ukraine. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1847–1858.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024.

Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.