Detecting Hyperpartisanship and Rhetorical Bias in Climate Journalism: A Sentence-Level Italian Dataset

Michele Joshua Maggini and **Davide Bassi** and **Pablo Gamallo**
Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
michelejoshua.maggini@usc.es, davide.bassi@usc.es, pablo.gamallo@usc.es

Abstract

We present the first Italian dataset for joint hyperpartisan and rhetorical bias detection in climate change discourse, enhancing the complexity in modeling hyperpartisan detection. Our annotation scheme achieves a Cohen's kappa agreement of 0.63 on the gold test set (173 sentences). The dataset comprises 48 articles (1,010 sentences) from far-right media, annotated at sentence level for both binary hyperpartisan classification and the multi-label classification of 17 rhetorical biases. We conduct extensive text analysis revealing significant correlations between hyperpartisan content and specific rhetorical techniques. Our experiments with state-of-the-art language models (GPT-4o-mini) and Italian BERTbase models establish strong baselines for both classification tasks. To ensure reproducibility while addressing copyright concerns, we release article URLs, article id and paragraph's number alongside comprehensive annotation guidelines. This resource advances research in cross-lingual hyperpartisan detection and provides insights into the rhetorical strategies employed in Italian climate change discourse. To the best of our knowledge, we are the first to tackle hyperpartisan detection related to logical fallacies, focusing on on the sentence level. We provide the code and the dataset to reproduce our results: https://anonymous.4open.science/r/Climate_HP-RB-D5EF/README.md

1 Introduction

The rise of hyperpartisan news content and its potential impact on public discourse has become a critical concern in the digital age. For Hyperpartisanship, we referred to Maggini et al. (2025)'s definition: *Hyperpartisan news detection is the process of identifying news articles that exhibit extreme one-sidedness, characteriz ed by a pronounced use of bias*.

This phenomenon is particularly evident in discussions about climate change (Luo et al., 2020), because it is a polarizing topic (Falkenberg et al., 2022). This phenomenon constitutes a threat to social cohesion through a loop mechanism that, by manipulating the emotions of the audience, fosters the polarization of individuals (Marino et al., 2024). In light of this, many scholars developed NLP methods to tackle hyperpartisanship. Most of the studies approach this task as a binary classification task. Despite the consistent performance reached, such approaches fail to uncover the underlying mechanisms that drive hyperpartisanship (Maggini et al., 2025).

Linguistic scholarships have shown that specific rhetorical strategies play a crucial role in creating and reinforcing hyperpartisan narratives (Nguyen et al., 2022; Potthast et al., 2018).

Rhetorical biases are vicious communicative strategies aimed at circumventing or violating audience's intellectual autonomy, by depriving them of the necessary elements to evaluate and counter arguments effectively. Intellectual autonomy, in fact, involves the capacity of individuals to think critically and independently while maintaining the ability to appropriately rely on external sources for informed decision-making (**?**).

Examining such biases could provide crucial insights, revealing how biased rhetorical techniques are employed in hyperpartisan content to manipulate audiences, thus enabling more targeted interventions to mitigate polarization (Ruan et al., 2024).

Additionally, while significant progress has been made in detecting hyperpartisan content and propaganda techniques in English-language media, there remains a critical gap in resources and analysis for other low-represented languages, particularly Italian (Maggini et al., 2025).

Our study addresses these gaps by introducing the first Italian dataset jointly annotated for hyperpartisan detection and rhetorical bias identification in the context of climate change news. Our main

Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 168–187
July 31, 2025 ©2025 Association for Computational Linguistics

contributions to the field are:

- We introduce a novel dataset consisting of 48 articles (1010 sentences) from Italian libertarian-right media, focusing on climate change coverage and related topics such as Euroscepticism and green policies. Our annotation scheme operates at the sentence level, capturing both binary hyperpartisan classification and a fine-grained taxonomy of 17 distinct rhetorical biases.
- We leverage our fine-grained annotation to analyze both the relationship between hyperpartisan content with specific rhetorical manipulation strategies, and the structural distribution of these techniques across article pararagraphs, providing insights into their functional roles within the discourse architecture.
- We establish baseline performance metrics through experiments with state-of-the-art language models. We evaluate two distinct approaches: 0-shot with GPT-4-mini, and Fine-tuning (FT) with two BERTbase fine-tuned models for Italian. Our results demonstrate the feasibility of automated detection for both hyperpartisan content and specific rhetorical biases, while also highlighting the challenges inherent in identifying more subtle manipulation techniques.
- To ensure reproducibility while respecting copyright constraints, we will release our dataset in the form of article URLs accompanied by detailed annotation guidelines. This approach allows researchers to reconstruct the dataset while maintaining its integrity and legal compliance.

Our work contributes to the growing body of research on automated detection of media bias and manipulation, while specifically addressing the need for non-English resources in this domain. The findings and resources presented in this paper have important implications for developing more robust and culturally-aware systems for detecting and analyzing media manipulation across languages and contexts.

This article is organized as follows: Section 2 reviews the key contributions in the field that inspired our research. Section 3 details the methodology used to create the dataset, covering data collection, the annotation process, and a statistical overview of the dataset. Section 4 presents benchmark experiments for classification tasks using our dataset, along with an in-depth corpus analysis. Finally, Section 5 summarizes our findings and outlines potential directions for future research.

2 Related Work

Hyperpartisan news detection has gained significant attention in the context of online misinformation, leading to extensive research in recent years. Maggini et al. (2025) provided a comprehensive survey of hyperpartisan detection approaches. They proposed a definition that captures the linguistic and political aspects of hyperpartisanship. Additionally, they highlighted the dominance of English and U.S.-centric datasets in this domain, emphasizing the need for datasets in underrepresented languages to better understand hyperpartisanship across different countries.

Potthast et al. (2018) pioneered the computational analysis of hyperpartisan news, delving into the stylistic traits distinguishing hyperpartisan news from mainstream. Kiesel et al. (2019) established a significant foundation for computational approaches to hyperpartisanship, introducing a shared binary classification task involving 42 teams. They released two document-level datasets—one manually annotated and one labeled based on source—which provided standardized resources for hyperpartisan scholarships.

Subsequent research evolved from document-level detection toward more fine-grained approaches that leverage information from various article components. Naredla and Adedoyin (2022) experimented with BERT, ELMo, and Word2Vec on entire articles, including both headlines and bodies, while also testing various context lengths for BERT. Lyu et al. (2023) analyzed 2,200 manually labeled and 1.8 million machine-labeled news titles across the political spectrum, achieving $Acc = 0.84$; $F1 = .78$ on an external validation set, using their transformer-based model. By tracking political stance, they revealed that right-leaning media use hyperpartisan titles more frequently, identified key contentious topics, and documented a cross-spectrum increase in hyperpartisan content during the 2016 U.S. election. This more granular approach was further advanced by researchers such as Pérez-Almendros et al. (2019), who focused specifically on quoted content as a distinctive component for hyperpartisan classification, demonstrating the

value of analyzing structural elements within articles rather than treating them as homogeneous units.

Omidi Shayegan et al. (2024) advanced hyperpartisan detection in under-represented languages by developing a benchmark for Persian tweets and systematically evaluating various architectural approaches from encoders to decoder-only models.

Maggini et al. explored the application of LLMs for hyperpartisan detection, utilizing LLaMA3-8b-Instruct (Touvron et al., 2023) in different In-Context Learning settings with general and task specific prompts on SemEval-2019 Task 4 and a headline-specific dataset. Their research demonstrated that these advanced neural architectures achieve competitive performance when enhanced with domain knowledge and structured reasoning, establishing LLMs as effective tools for political text analysis despite previous assumptions about ICL and computational power limitations.

As mentioned in Sec. 1, hyperpartisan content often manifests through the strategic deployment of manipulative rhetorical techniques. Such techniques are extensively employed to persuade audiences in different settings, such as news, speeches, and social media. Given the rapid spread of manipulative content in online environments, a wide range of computational approaches has emerged to address this phenomenon. As highlighted by Bassi et al. (2024), early efforts predominantly focused on content-based detection. More recently, argumentative and rhetorical approaches have gained traction, demonstrating greater scalability across different contexts.

Martino et al. (2019) represents a seminal contribution in this regard, proposing a method to identify specific texts containing propaganda and classify them based on 18 persuasion techniques . Their work later inspired a SemEval task in 2020 (Da San Martino et al., 2020) and has been then followed by Piskorski et al. (2023), which expanded the taxonomy to 23 fine-grained techniques, grouped into six broad categories. Additionally, they extended the analysis to a multilingual setting, demonstrating the applicability of argumentation-based propaganda detection across different languages.

More recent works (Hasanain et al., 2024a,b) addressed LLMs' potential for propaganda techniques detection. In this regard, Sprenkamp et al. (2023) demonstrated that reducing the number of labels to 14 improved classification performance.

Building on this literature, to the best of our knowledge, we are the first to approach hyperpartisan detection at the sentence level and consider the presence of rhetorical bias as a fundamental characteristic of hyperpartisan texts. By treating rhetorical biases as stylistic traits that shape the message of a text, we capture deeper linguistic patterns that contribute to hyperpartisan framing. While prior research has shown that source-level bias does not uniformly manifest across all articles (Baly et al., 2018), our sentence-level approach transcends these limitations. Working at this granularity allows us to identify precisely where and how hyperpartisan language emerges through specific rhetorical fallacies, creating a dataset that supports both binary hyperpartisan detection and multi-class fallacy classification. This approach reveals significant correlations between particular fallacy types and hyperpartisan content (see Table 5 in the Appendix), providing empirical evidence for their relationship.

3 Methodology

3.1 Dataset Creation

Article selection and Pre-Processing From the moment that "alternative" media tend to spread anti-establishment messages (Ernesto de León and Adam, 2024), we focused on NicolaPorro.it[1], an independent libertarian media outlet. The collected corpus consists of 48 articles for a total of 1010 units on climate change, green policies and Euroscepticism selected from the site's "Green policies" section to ensure topical homogeneity. We featured only the Italian language, since the recent enhancements in NLP for disinformation detection mostly covered over-represented languages like English (Maggini et al., 2025). To ensure a fine-grained analysis of the texts, we then split the articles following the html <p> tags, that mostly corresponded to individual sentences. We grouped together the sentences with less than 15 words to guarantee minimal context.

Annotation Protocol To build our annotation guidelines, we started by defining the constructs under investigation.

For **Hyperpartisanship** we referred to the definition by Maggini et al. (2025) mentioned above:

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

Hyperpartisan news detection is the process of identifying news articles that exhibit extreme one-sidedness, characterized by a pronounced use of bias. We modeled this task as a binary classification at the sentence level.

Regarding **Rhetorical biases**, to our knowledge the most comprehensive taxonomy is the one of Piskorski et al. (2023), with a total of 23 labels.

Starting from their taxonomy, we translated the definitions in Italian and adapted them to our use case on climate change. Additionally, being our main scope to conduct a more fine-grained analysis of the rhetorical biases underlying hyperpartisanship, we merged some of the techniques (Slogan/Conversation Killer, Whataboutism/Tu Quoque, Appeal to Values/Flag Waving, Causal/Consequential Oversimplification, Smear/Doubt). Keeping them would have added unnecessary complexity to the model without providing additional analytical insights. In Table 3 we report brief descriptions for each technique, while in Appendix A we report their in-depth definition as well as the annotation guidelines.

Alongside the binary hyperpartisan classification, annotators performed a multi-label task identifying specific rhetorical biases deployed to influence reader opinion in each sentence.

The annotation was conducted by two native Italian speakers with expertise in political discourse analysis. Both annotators are Ph.D. students in applied NLP for disinformation, with academic backgrounds in Philology, Data Science, Anthropology, and Psychology. They have prior experience in linguistic annotation of news content and rhetorical technique identification. Annotators did not know the source of the articles and during the annotation rounds, they did not have access to the whole article's context but only to the individual sentences. We divided the annotation process into three phases: **Training phase**: annotators studied the guidelines, performed pilot annotations and completed the training through interactive sessions to discuss doubts, edge cases and resolve disagreements.; **Annotation Phase**: Each document was independently annotated by both annotators; **Curation Phase**: Discrepancies between annotations were discussed and resolved to ensure final label consistency. Before the Curation Phase, we measured the Inter-Annotator Agreement (IAA) using Krippendorf's α, achieving a value of .92 for hyperpartisan detection and .63 on rhetorical fallacies.

3.2 Dataset description

Table 1 represents key statistics of our dataset, including size, sentence length, and the average rhetorical biases per article. Table 3 shows the definitions and distributions of hyperpartisan and neutral sentences, as well as logical fallacies. To analyze the thematic distribution within our corpus, we applied BERTopic (Grootendorst, 2022) with parameters optimized to preserve local structure[2]. Table 2 presents the topic distribution. After manual inspection, we forced BERTopic to detect three main topics: science, institutions and Other, each further subdivided into specific subtopics like institutions.Italy, science.cars, etc.

Metric	Value
Number of Documents	48
Number of Sentences	1010
Avg. Sentences per Article	21.12
Avg. Words per Text	40.26
Avg. Characters per Text	264.37
Avg. Techniques per Document	2.12

Table 1: Dataset Statistics

Topic	Count
Other.climate	241
Other.other	122
science.climate_change	109
science.other	82
institutions.Europe	70
Other.politics	68
science.energy_transition	54
science.environment	44
science.cars	38
institutions.Other	35
institutions.OMS	33
institutions.China	30
institutions.Italy	26
science.green_policies	23
institutions.BlackRock	16
science.medicine	14
Other.politically_correct	4
Other.politics	1

Table 2: Topic Distribution. Topics have been extracted using BERTopic.

3.3 Models

We tested two different architectures: encoders and decoder-only models.

[2]umap-model = UMAP(n-neighbors=10, n-components=3, metric='cosine') hdbscan-model = HDBSCAN(min-cluster-size=10, min-samples=10, metric='euclidean', prediction-data=True) ctfidf-model = ClassTfidfTransformer(bm25-weighting=False, reduce-frequent-words=True) representation-model = MaximalMarginalRelevance(diversity=0.5)

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 304 N 706
Rhetorical Biases		
Slogan/Conversation Killer	Using catchphrases or dismissive statements to shut down further discussion or debate	64
Appeal to Time	Manipulating temporal perspectives or deadlines to create urgency or dismiss concerns	9
Appeal to Values/ Flag Waving	Exploiting patriotic feelings or moral values to justify positions or actions	59
Appeal to Authority	Using the reputation of an expert or institution to support arguments without proper context	53
Appeal to Popularity	Justifying a belief by citing its widespread adoption or acceptance	11
Appeal to Fear	Manipulating audience's fears to promote specific viewpoints or actions	99
Straw Man/Red Herring	Misrepresenting opponent's argument or diverting attention to unrelated issues	43
Whataboutism/ Tu Quoque	Deflecting criticism by pointing to the opponent's alleged hypocrisy or similar actions	42
Loaded Language	Using words with strong emotional implications to influence the audience	330
Repetition	Repeating phrases or ideas multiple times for emphasis or to establish them as truth	23
Intentional Confusion/ Vagueness	Using deliberately unclear or ambiguous language to avoid commitment or scrutiny	55
Exaggeration/Minimisation	Presenting facts in a distorted way by either magnifying or downplaying their importance	244
Name Calling	Using labels or derogatory terms to discredit without substantive argument	159
Reductio ad Hitlerum	Drawing inappropriate comparisons to Nazism, Hitler, or fascism	13
Smear/Doubt	Attempting to damage reputation or create doubt through indirect attacks or insinuations	355
Causal/Consequential Oversimplification	Presenting complex situations with oversimplified cause-effect relationships	165
False Dilemma/ No Choice	Presenting limited options while ignoring alternatives or middle ground	66

Table 3: Taxonomy of rhetorical biases and hyperpartisan language detection used in our annotation scheme. The rhetorical biases represent fine-grained categories of manipulative language techniques commonly found in politically charged discourse.

For encoders, we used `dbmdz/bert-base-italian-xxl-uncased`[3], trained from scratch on Italian, and `nickprock/sentence-bert-base-italian-xxl-uncased`[4], fine-tuned for Italian. Particularly, the first model was trained on OSCAR corpus (Ortiz Suárez et al., 2020) and is known for its robust handling of complex relationships in text, allowing for a comprehensive understanding of contextual nuances. In contrast, the sentence-transformer is optimized for generating meaningful sentence embeddings, making it particularly suitable for capturing the semantic essence of individual sentences. By fine-tuning and comparing both models, we aimed to evaluate their performance on the hyperpartisan classification task, providing insights into which approach better captures the rhetorical distinctions in the data. We fine-tuned the models

Regarding the decoder-only architectures we used GPT 4o and 4o-mini. For the Hyperpartisan detection (HP) task we employed the models in a 0-shot setting, while, for the Rhetorical Bias (RB), each model was firstly tested 0-shot with temperature equal to 0.2. Given the difficulty of working with a high number of labels, we decided to set this value so that the model could capture the most subjective traits in the rhetorical fallacies. Furthermore, we fine-tuned the models for Rhetorical Bias detection. We prompted and fine-tuned the OPENAI models via their APIs [5]. The prompts are available in the Appendix A.2 and A.3.

4 Results

4.1 Hyperpartisan-Rhetorical Bias Relation

To investigate the relationship between rhetorical biases and hyperpartisanship, we analyzed their correlation patterns. Figure 1 depicts which rhetorical biases are most determinant to distinguish between hyperpartisan and neutral sentences. To analyze the rhetorical distinctions between hyperpartisan and neutral sentences in more detail, we performed χ^2 tests to compare the frequency of each rhetorical technique across binary labels (see Table 5 in Appendix A.3). We measured the effect sizes using Log Ratio[6]. Thus,

[3] https://huggingface.co/dbmdz/bert-base-italian-xxl-uncased

[4] https://huggingface.co/nickprock/sentence-bert-base-italian-xxl-uncased

[5] https://platform.openai.com/docs/models

[6] Log Ratio (LR) is calculated as the logarithm base 2 of the ratio of the frequencies between the two groups. A value of 0 signifies equal frequency in both groups, positive values indicate a higher frequency in the hyperpartisan group, and

neutral sentences are usually characterized by no rhetorical biases ("no_technique_detected"), whereas "Reduction_ad_Hitlerum", "Name_Calling", "Tu_Quoque/Whataboutism", "Loaded_Language", "Smear/Doubt", "Straw_Man/Red_Harring" and "Exaggeration_Minimisation" are highly significant ($p-value < 0.001$) to discern hyperpartisan sentences. Those findings validate Maggini et al. (2025)'s definition of hyperpartisanship as well as the previous definitions used in the literature by Kiesel et al. (2019); Lyu et al. (2023).

4.2 Topological Distribution of Rhetorical Biases

Fig. 2 shows the average of hyperpartisan sentences across the articles' structure, while Fig. 3 illustrates the concentration of bias in the articles' structure. This provides us with a better understanding of how much and in which parts the articles are contaminated by hyperpartisanship and rhetorical biases.

To analyze the Hyperpartisan Contamination Level (HCL), firstly, we grouped the sentences by article ID and got the sentence positions. Then, we created normalized positions for each sentence and created 10 potision bins. Lastly, we computed the average hyperpartisan score for each position bin. Fig. 2 shows that hyperpartisan sentences appear in 50% of cases within the first 10% of the articles and around 40% in the following 10% (i.e., between 10% and 20% of the article's beginning). This evidence aligns with what other researchers analyzed in previous work, stating that titles usually are determinant to distinguish between fake or hyperpartisan and mainstream news (Horne and Adali, 2017; Shrestha and Spezzano, 2021). Then, the average HCL drops in the central part (20-60%) to increase again up to around 30% in the second half of the articles (60-100%).

Then, we decided to investigate on how rhetorical techniques are adopted to convey and shape the message (Fig. 3). Firsly, we normalized the position of the techniques within each article, creating position quartiles. After that, we counted the occurrences of each technique in each quartile and then pivoted the data for visualization. Successively, we calculated the raw totals for each technique and then normalized by technique, namely we computed the percentage across quartiles for each technique.

negative values indicate a higher frequency in the neutral group.

Our annotation of sentences with rhetorical fallacies revealed that certain techniques are more prominent than others, offering valuable insights into how these strategies are distributed across the structure of the articles. For example, while Reductio ad Hitlerum is relatively rare (13 occurrences with high statistical significance), it appears predominantly in the first quartile (Q1) at 53.8%, indicating its use in setting a strong, biased tone early in the article. Similarly, Name Calling is concentrated in Q1, with 49% of its occurrences in this section, and both techniques are highly significant for identifying hyperpartisan sentences. These strategies allow reporters to directly express their stance on a topic, often leveraging emotionally charged language to engage the reader from the outset.

However, Repetition (statistically significant) is most frequent in the second quartile (Q2), with 56.5% of its cases appearing here. This suggests its role in reinforcing the ideas introduced earlier, contributing to the redundancy of concepts to solidify the intended message. Lastly, Slogan/Conversation Killer usage peaks in the final quartile (Q4), accounting for 43.8% of its total appearances. This aligns with the tendency of journalists to use catchy phrases or mottos at the end of articles to leave a lasting impression and emphasize their message.

These findings align with the existing literature, which highlights how clickbait headlines often employ rhetorical techniques to manipulate reader engagement and frame narratives persuasively (Blom and Hansen, 2015; Munger, 2020). Such strategies are not only common in sensationalist media but are also key tools for amplifying bias and promoting specific agendas (Chakraborty et al., 2016). This finding also strengthens the relationship between click-bait and hyperpartisan content.

4.3 Computational Baselines

The aim of our experiments is to provide baselines and to explore the impact of different architectures on two classification tasks: for hyperpartisan and for logical fallacies. Both of the two tasks were annotated at the sentence level. While HP classification is a binary classification task, RB classification, is a multi-class classification task.

The results of the evaluation on the detection of hyperpartisan and rhetorical bias are shown in Table 4. The results demonstrate significant variability in the metrics score between different models and methodologies.

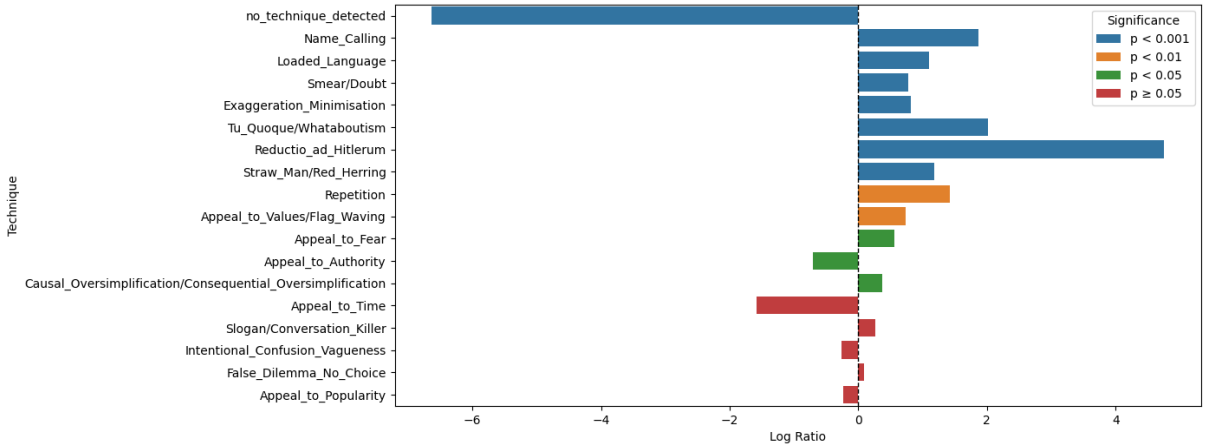

Figure 1: Correlation between Hyperpartisan sentences and techniques. The table with the different levels of significance is reported in the Appendix.

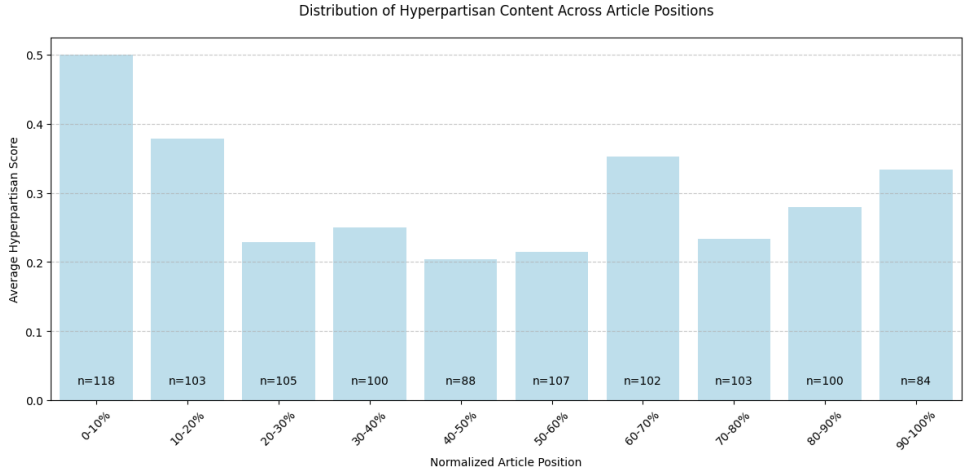

Figure 2: Hyperpartisan Contamination Level per sentence position (deciles). *n* represents the number of sentences. Because of articles have dissimilar number of sentences, we normalize their length.

GPT: GPT-4o-mini and GPT-4o both perform well on the HP classification task in a 0-shot setting. GPT-4o achieves an accuracy of 0.969 with an F1 score of 0.942, outperforming GPT-4o-mini, which attains an accuracy of 0.959 and an F1 score of 0.933. The results indicate that GPT-4o is more effective in recognizing hyperpartisanship in Italian news articles. Those results can be explained by the architectural and dimensional differences between the two models.

For RB classification, GPT-4o-mini performs reasonably well in 0-shot mode (accuracy: 0.892, F1: 0.319), and its fine-tuned performance increased slightly (accuracy: 0.905, F1: 0.362). GPT-4o exhibits similar behavior, with 0-shot performance (accuracy: 0.906, F1: 0.385) being substantially better than FT (accuracy: 0.908, F1: 0.410). The low precision scores for both models in RB classification indicate challenges in correctly identifying rhetorical bias. The high unbalanced distribution between techniques explains these results. Indeed, the other metrics we reported are macro-averaged metrics, which offer a fair comparison.

Encoders: For HP classification, `bert-base-italian-xxl-uncased` achieves an accuracy of 0.861 and an F1 score of 0.859, showing strong performance but slightly lagging behind GPT-4o. However, in RB classification, the model performs poorly, with a precision of 0.354 and an F1 score of 0.470, indicating that it struggles to effectively identify rhetorical bias. The difficulty in classifying RB stems from the extreme class imbalance, where certain rhetorical categories are underrepresented, leading to biased model predictions that favor more frequent classes. The macro-averaged F1 score provides a clearer picture of

Model	Classification	Method	Accuracy	Precision	Recall	F1 Score
GPT-4o-mini	HP	0-Shot	0.959	0.942	**0.924**	0.933
	RB	0-Shot	0.892	0.285	0.486	0.319
		FT	0.905	0.326	0.465	0.362
GPT-4o	HP	0-Shot	**0.969**	**0.980**	0.907	**0.942**
	RB	0-Shot	0.906	0.387	0.434	0.385
		FT	0.908	0.378	0.559	0.410
`bert-base-italian-xxl-uncased`	HP	FT	0.861	0.858	0.861	0.859
	RB	FT	0.354	0.699	0.354	0.470
`sentence-bert-base-italian-xxl-uncased`	HP	FT	0.851	0.846	0.851	0.845
	RB	FT	0.321	0.683	0.320	0.436

Table 4: Comparison of Hyperpartisan and Rhetorical Bias Classification Models

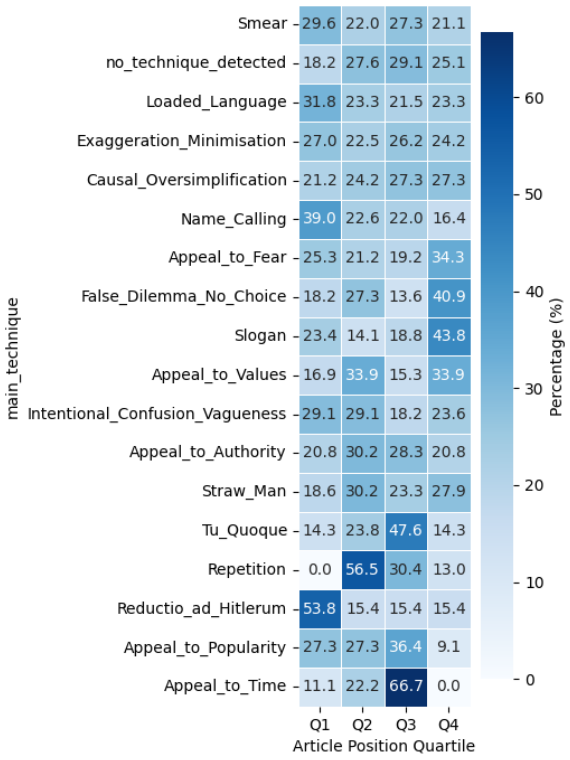

Figure 3: Distribution of Techniques Across Article Quartiles

this imbalance, as models perform well on majority classes but fail on rare ones. Similar trends are observed for `sentence-bert-base-italian-xxl-uncased`, which achieves competitive HP classification results (accuracy: 0.851, F1: 0.845) but performs poorly on RB classification (accuracy: 0.321, F1: 0.436). This suggests that sentence embeddings are effective for hyperpartisan classification but less suited for rhetorical bias detection. The model's struggle with RB is further exacerbated by the highly skewed class distribution, making it difficult to learn meaningful representations for rare rhetorical bias categories. The macro-averaged F1 scores reinforce that underrepresented classes are poorly classified, reducing overall model effectiveness.

5 Conclusion and Future Work

In this work, we introduced a novel Italian news dataset focused on climate change and Euroscepticism, specifically designed for hyperpartisan and rhetorical bias detection. Our dataset emphasizes the critical need to collect news in underrepresented languages to gain a deeper understanding of hyperpartisanship across European countries. Spanning diverse and polarizing public topics, the dataset consists of 48 articles divided into 1,010 sentences, annotated for hyperpartisanship (binary labels) and enriched with over 1.5K rhetorical fallacy labels using a fine-grained taxonomy.

Our study underscores the significance of analyzing hyperpartisanship in conjunction with rhetorical biases, as these biases can profoundly influence the objectivity of storytelling in news articles. Through detailed corpus analysis, we contributed to the field by offering nuanced insights into how specific rhetorical techniques align with hyperpartisan content, enhancing our understanding of manipulation strategies in media.

We also established strong baselines using state-of-the-art architectures and learning paradigms, such as FT and 0-shot, demonstrating the versatility and applicability of our dataset. By sharing the full pipeline to recreate the dataset, we aim to facilitate the development of new methods and tools to critically analyze online media content.

Future work will focus on experimenting with advanced models and exploring how leveraging rhetorical biases can further improve hyperpartisan sentence detection. Despite the annotation required high effort and is not scalable, we plan to extend the current dataset with other articles. We hope our work serves as a stepping stone for more robust and transparent media analysis, ultimately contributing to a healthier information ecosystem.

6 Limitations

Regarding the dataset size (48 articles, 1,010 sentences), we acknowledge is relatively small, potentially limiting the generalizability of findings and the robustness of model training. Expanding the dataset with a broader range of sources and perspectives would improve coverage and model performance.

Second, the focus on far-right media outlets introduces a selection bias, which, while intentional for analyzing hyperpartisan rhetoric, may not capture the full spectrum of climate change discourse in Italy. Future work should explore more diverse media sources, including centrist and left-leaning outlets, to provide a more comprehensive view.

Third, while our annotation scheme achieves moderate agreement (Cohen's kappa = 0.63), some rhetorical biases remain inherently subjective and difficult to categorize consistently due to their distributions.

Finally, differently from Martino et al. (2019); Da San Martino et al. (2020); Piskorski et al. (2023) we did not include the span, as the annotation process was highly demanding and the number of annotators limited. Such approach could further contribute to fine-grained analysis of news articles, understanding on which specific words and rhetorical patterns the hyperpartisan is based.

Finally, while we provide article URLs for transparency, copyright restrictions prevent us from openly distributing full-text data. This limits direct replication and benchmarking. Future work could explore ways to balance reproducibility with legal constraints, such as structured metadata representations or synthetic dataset augmentation.

Ethics Statement

Biases

The news articles in our dataset may contain harmful content, including loaded language, name-calling, and slurs. Our annotation process was designed to focus solely on identifying rhetorical bias and hyperpartisan language rather than assessing the truthfulness of the information. To ensure objectivity, annotations were conducted without considering annotators' personal opinions or political views on the topics discussed. Additionally, we did not rely on crowdsourcing; instead, we managed our annotators directly, ensuring proper working conditions and maintaining annotation quality.

We recognize the potential risks of bias in both data collection and model predictions. The inherent subjectivity in identifying rhetorical bias and hyperpartisanship means that biases can emerge from the dataset itself, as well as from the models trained on it. Given the sensitive nature of hyperpartisan and rhetorical bias detection, we advise caution when using the dataset and models to avoid reinforcing biases or misrepresenting viewpoints. Future work should focus on refining annotation practices, improving model interpretability, and incorporating interdisciplinary perspectives to mitigate potential harms.

Intended Use and Misuse Potential

This dataset is intended to advance research in hyperpartisan news detection, particularly in underrepresented languages. It can contribute to the development of more robust models and analytical tools for identifying rhetorical bias in media. However, we acknowledge the risk of misuse, particularly by malicious actors seeking to manipulate or censor content. To prevent unintended consequences, we urge researchers and practitioners to use this dataset responsibly and transparently, ensuring that any conclusions drawn are supported by rigorous evaluation and ethical considerations.

The work presented in this paper complies with the ACL Ethics Policy [7]. We have relied on open architectures when possible. We hope that the community can benefit from our work to apply NLP technology to tackle climate change and Euroscepticism.

Acknowledgements

This project has received funding from the European Unión´s Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351.

References

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium. Association for Computational Linguistics.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. 2024. Decoding persuasion: a survey on ML

[7] https://www.acm.org/code-of-ethics

and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9. Publisher: Frontiers.

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

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.

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.

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.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024a. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia. ELRA and ICCL.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024b. Large language models for propaganda span annotation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA. Association for Computational Linguistics.

Benjamin D. Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news.

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.

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.

Hanjia Lyu, Jinsheng Pan, Zichen Wang, and Jiebo Luo. 2023. Computational assessment of hyperpartisanship in news titles.

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

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

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).

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].

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

Navakanth Reddy Naredla and Festus Fatai Adedoyin. 2022. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1):100064.

Allison Nguyen, Tom Roberts, Pranav Anand, and Jean E Fox Tree. 2022. Look, dude: How hyperpartisan and non-hyperpartisan speech differ in online commentary. *Discourse & Society*, 33(3):371–390.

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.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

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.

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.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 2019. Cardiff university at SemEval-2019 task 4: Linguistic features for hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 929–933. Association for Computational Linguistics.

Qin Ruan, Jin Xu, Susan Leavy, Brian Mac Namee, and Ruihai Dong. 2024. Rewriting bias: Mitigating media bias in news recommender systems through automated rewriting. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, UMAP '24, page 67–77, New York, NY, USA. Association for Computing Machinery.

Anu Shrestha and Francesca Spezzano. 2021. Textual characteristics of news title and body to detect fake news: A reproducibility study. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28 – April 1, 2021, Proceedings, Part II*, page 120–133, Berlin, Heidelberg. Springer-Verlag.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. 2023. Large Language Models for Propaganda Detection. ArXiv:2310.06422 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

A Appendix

Guidelines

Annotation Guidelines (ENG)

Hyperpartisan sentences: Text that displays extreme bias favoring one particular political side, often employing pronounced use of rhetorical biases. Label it 1, if the sentence is hyperpartisan, or 0 if it is neutral. Explicit examples: - "We are tired of government's abuses! We don't want to drive electric car!" Neutral examples: - "Electric cars are not as green as industry tell us".

Slogan/Conversation Killer: Short and impactful phrases designed to discourage critical thinking and/or urge a certain action by presenting the message as definitive. These often draw on seemingly indisputable popular wisdom or stereotypes to avoid further discussion. Explicit examples: - "Think global, act local!" - "That's just how it is, there's nothing more to add." Implicit examples: - "Be part of the solution, not part of the pollution." - "With the utmost respect for green policies and climate change, shareholders want profits. Period."

Appeal to Time: An argument centered on the idea that the time has come for a particular action or that there is no more time to waste. The call to "Act Now!" Explicit example: "If we don't act immediately on the climate crisis, in ten years it will be too late to save the planet!" Implicit example: "The timing for this reform could not be more perfect..."

Appeal to Values/Flag Waving: Leverages identity values (nationalism, patriotism, belonging to a social group/class), as well as moral and social values considered positive by the target audience (freedom, democracy, ethics, religion) to promote or justify an idea. It operates on the assumption that the audience already holds certain biases or beliefs. Explicit examples: - "If we must have climate policies—very few—then let's adopt only those that benefit Italy." - "Ecology cannot and must not take priority over citizens' freedom." Implicit examples: - "While other countries bow to these policies, we must protect our interests." (a veiled appeal to nationalism) - "These policies are gradually eroding the principles on which our society is founded." (an appeal to preserving social values)

Appeal to Authority: Giving weight to a particular idea by citing a supposed authority as a source, regardless of whether they are actually competent in the field. The tone of the statement suggests that the weight of this supposed authority is being used to justify information or conclusions. Explicit example: "Climatologist Richard Dawkins says climate change doesn't exist, therefore climate change is a lie!" Implicit example: "Those who have truly studied the issue know very well that things are not as they seem."

Appeal to Popularity: Justifying an idea by claiming that "everyone" agrees or that "no one"

disagrees, encouraging the audience to adopt the same position out of conformity. "Everyone" may refer to the general public, experts (e.g., all experts say that...), countries, or other groups. Explicit example: "No one here is denying that the planet's temperature is rising, so climate change is real." Implicit example: "Ideological rules have been imposed that no one else follows."

Appeal to Fear: Promoting or rejecting an idea by exploiting the audience's repulsion or fear, describing possible scenarios in a frightening way (terrible things that could happen) to instill fear. Explicit example: "Climate taxes are just the beginning. If we keep up this farce, they'll take everything we have!" Implicit example: "This is just the first step in a larger plan that will lead to irreversible consequences."

Straw Man/Red Herring: A technique that shifts the discussion away from the original topic through two main approaches: distorting the original argument into an easier-to-attack version or introducing a different but seemingly related topic. The goal is to avoid addressing the substance of the initial issue by diverting attention to a secondary theme. Explicit examples: - "When you ask for a more gradual energy transition, you're basically saying you don't care if the planet becomes uninhabitable for our children." - "Instead of always talking about CO2 emissions, look at this great initiative we launched for beach clean-ups!" Implicit examples: - "Their concern for the employment impact of closing coal plants reveals the usual mindset that prioritizes profit over the planet's survival." - "Before discussing climate policies, shouldn't we focus on improving waste sorting in municipalities?"

Tu Quoque/Whataboutism: A technique that attempts to discredit a position or opponent by highlighting alleged contradictions or double standards. This can manifest by pointing out inconsistencies on the same issue or introducing comparisons with other contexts or situations. The goal is to undermine credibility through comparisons with other matters. Explicit examples: - "Look at them, all flying around in helicopters, while just weeks ago they were sounding the alarm and criticizing waste!" - "He talks so much about the climate emergency, but we're still waiting for answers on the migration crisis." Implicit examples: - "Funny how certain climate positions change so quickly when political circumstances shift." - "Interesting concern for the environment... I wonder if the same attention was there when it came to approving the airport expansion in your region."

Loaded Language: Using specific words and phrases with strong emotional implications (both positive and negative) to influence and persuade the audience. The essence of this technique is the use of terms that go beyond their literal meaning to evoke an emotional response. Explicit example: "These climate dictatorships run by idiots." Implicit example: "A somewhat unconventional management of public funds."

Repetition: The repeated use of the same word, phrase, story, or image in the hope that repetition will persuade the audience. Explicit example: "Safety is our priority. We must ensure safety. Without safety, there is no future. Safety must come first." Implicit example: "Innovation is the key. We must focus on innovation. Innovation will save us. Only through innovation can we progress."

Intentional Confusion/Vagueness: Using deliberately unclear wording so that the audience can have their own interpretations. For example, an argument may include a vague phrase with multiple or unclear definitions, which ultimately does not support the conclusion. Explicit example: "We will develop synergistic paradigms aimed at the horizontal optimization of ecological performance." Implicit example: "It has been proven that 70% of the time green policies work every time."

Exaggeration/Minimization: Representing something in an exaggerated manner: making things seem bigger, better, or worse (e.g., "the best of the best," "guaranteed quality") or downplaying something to make it seem less important than it really is (e.g., calling an insult just a joke), minimizing statements and ignoring arguments or accusations made by an opponent. Explicit example: "Never seen such colossal incompetence in public management." Implicit example: "There were some victims due to inefficiencies, but nothing to worry about."

Name Calling: Characterizing an individual or group using emotionally charged and/or derogatory labels. This specifically relates to labeling the subject with adjectives, nouns, or references to political orientations, opinions, personal characteristics, or organizational affiliations, rather than constructing an argument with premises and conclusions. Explicit example: "Giuseppe Conte to Di Battista, here are all the 'grillini' who should 'blush' for

their past pro-Putin positions on climate." Implicit example: "The usual armchair theorists now want to tell us how to manage the real economy."

Reductio ad Hitlerum: Attacking an opponent or activity by associating them with another group, activity, or concept with strong negative connotations for the target audience. The technique establishes a link or equivalence between the target and any individual, group, or event (past or present) perceived as unquestionably negative or presented as such. The goal is to transfer the negativity of the association to the criticized subject. Explicit example: "Even Big Brother said controlling everyone's lives was for the greater good." Implicit example: "This approach to dissent management is just missing men in black shirts."

Smear/Doubt: A technique aimed at undermining the credibility of someone or something (e.g., institutions) by questioning specific skills or capabilities, attacking reputation and overall moral character, or casting doubt on the intentions behind a decision. Explicit examples: - "The increase in energy bills exposes the green shift deception promoted by the EU." - "He worked for the same company he is now supposed to regulate—how can we trust him?" Implicit examples: - "The U.S. and Europe, with their green policies, still think in colonial terms." - "Their recent decisions make one wonder what this administration's real priorities are." Given the following text, read it very carefully and identify the possible presence of one or more of the persuasion techniques defined above.

Consider that:

Techniques may overlap: the same sentence can employ multiple techniques simultaneously. Techniques can be expressed sarcastically or indirectly. Tone and context are as important as specific words. A technique may manifest through a series of related statements rather than a single sentence. The text may not necessarily contain any technique, but it is crucial to analyze it thoroughly to eliminate any doubt. If no technique is detected, respond with "no technique detected."

A.1 Examples

A.1.1 Annotation Guidelines (ITA)

Hyperpartisan frasi: Testo che mostra un'estrema faziosità a favore di una specifica parte politica, spesso impiegando un uso marcato di bias retorici. Etichettalo come 1 se la frase è iperpartigiana, o 0 se è neutrale.

Esempi espliciti:

"Siamo stanchi degli abusi del governo! Non vogliamo guidare auto elettriche!" Esempi neutrali:

"Le auto elettriche non sono così ecologiche come l'industria ci racconta."

Slogan/Conversation Killer: Frasi brevi e incisive per scoraggiare il pensiero critico eo esortare a compiere una certa azione attraverso un'apparente definitività del messaggio. Spesso si richiamano alla saggezza popolare, apparentemente incontestabile, o a stereotipi per evitare ulteriori discussioni. Esempi espliciti: - "Vivi locale, pensa globale!" - "È così e basta, non c'è altro da aggiungere." Esempi impliciti: - "Sii parte della soluzione, non parte dell'inquinamento" - "Con il massimo rispetto per il green e per il cambiamento climatico, gli azionisti vogliono gli utili. Punto."

Appeal to Time: Argomento centrato sull'idea che sia giunto il momento di una particolare azione, oppure che non ci sia più tempo da perdere. L'appello ad "Agire Ora!". Esempio esplicito: "Se non agiamo immediatamente sulla crisi climatica, entro dieci anni sarà troppo tardi per salvare il pianeta!" Esempio implicito: "Il momento per questa riforma non potrebbe essere più propizio di così..."

Appeal to Values/Flag Waving: Fa leva su valori identitari (nazionalismo, patriottismo, appartenenza a un gruppo/ceto sociale) morali e sociali considerati positivi dal pubblico target (libertà, democrazia, etica, religione) per promuovere o giustifica un'idea. Si basa sul presupposto che i destinatari abbiano già determinati pregiudizi o convinzioni. Esempi espliciti: - "Se proprio abbiamo bisogno di politiche climatiche - pochissime - allora adottiamo solo quelle che avvantaggiano l'Italia." - "Perché l'ecologia non può, né deve, essere assolutamente prioritaria rispetto alla libertà dei cittadini" Esempi impliciti: - "Mentre altri paesi si piegano a queste politiche, noi dobbiamo proteggere i nostri interessi." (appello velato al nazionalismo) - "Queste politiche stanno gradualmente erodendo i principi su cui si basa la nostra società" (appello alla preservazione dei valori sociali)

Appeal to Authority: Dare peso ad una certa idea citando una presunta autorità come fonte, che può essere o meno effettivamente competente nel campo. Il tono del testo indica che si sfrutta il peso di questa presunta autorità per giustificare informazioni o conclusioni. Esempio: "Il climatologo Richard Dawkins dice che il cambiamento climatico non esiste, ergo il cambiamento climatico

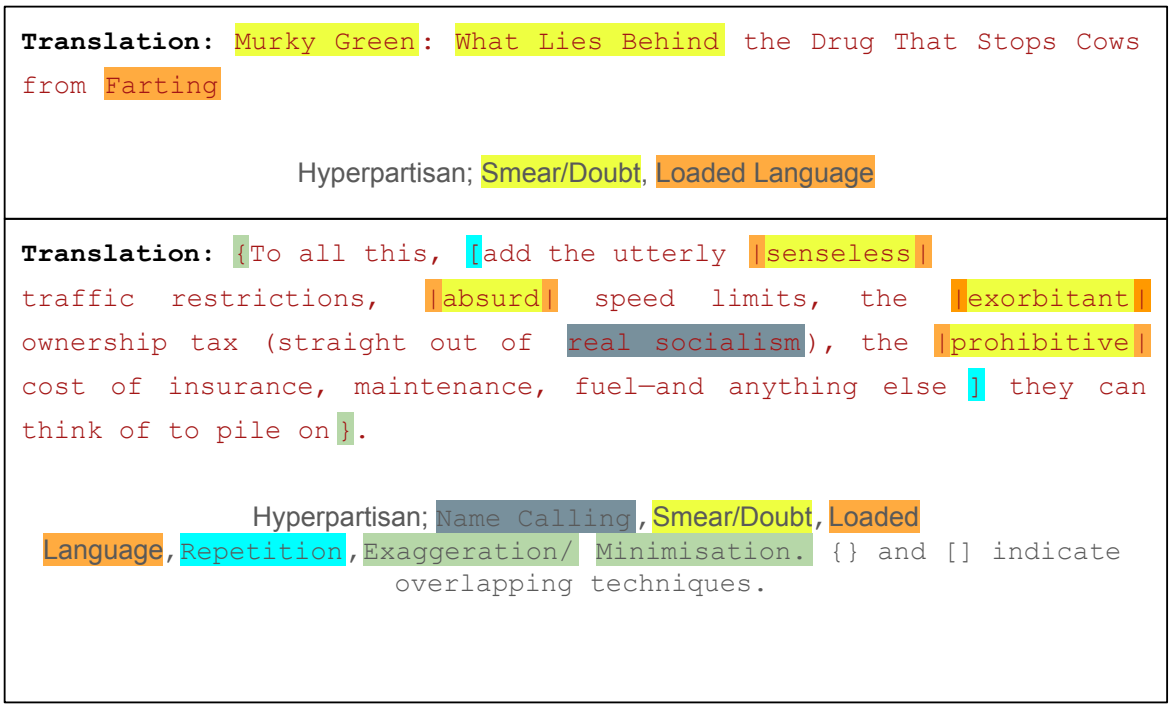

Figure 4: Comparable examples of rhetorical biases.

è una menzogna!" Esempio implicito: "Chi ha studiato davvero la questione sa bene che le cose non stanno così."

Appeal to Popularity: Giustificare un'idea sostenendo che "tutti" sono d'accordo o che "nessuno" è in disaccordo, incoraggiando il pubblico ad adottare la stessa posizione per conformismo. "Tutti" può riferirsi al pubblico generale, esperti (tutti gli esperti dicono che...), paesi o altri gruppi. Esempio: "Nessuno qui sta negando che la temperatura del pianeta stia aumentando, quindi c'è il cambiamento climatico" Esempio implicito: "Sono state dettate delle regole ideologiche che nessun altro segue."

Appeal to Fear: Promuovere o respingere un'idea sfruttando la repulsione o la paura del pubblico, descrivendo possibili scenari in modo spaventoso (terribili cose che potrebbero succedere) per instillare paura. Esempio: "Le tasse sul clima sono solo l'inizio. Se continuiamo con questa farsa si prenderanno tutto quello che abbiamo!" Esempio implicito: "Questo è solo il primo passo di un piano più ampio che porterà a conseguenze irreversibili"

Straw Man/Red Herring: Tecnica che sposta la discussione dall'argomento originale attraverso due modalità principali: la distorsione dell'argomento originale in una versione più facilmente attaccabile o l'introduzione di un argomento diverso ma apparentemente correlato. L'obiettivo è evitare di affrontare direttamente il merito della questione iniziale spostando l'attenzione su un tema secondario. Esempi espliciti: - "Quando chiedi una transizione energetica più graduale, in pratica stai dicendo che non ti importa se il pianeta diventerà inabitabile per i nostri figli." - "Invece di parlare sempre di emissioni di CO2, guardate che bell'iniziativa abbiamo fatto per la pulizia delle spiagge!" Esempi impliciti: - "Il loro interesse per gli impatti occupazionali della chiusura delle centrali a carbone rivela la solita mentalità che antepone il profitto alla sopravvivenza del pianeta." - "Prima di discutere delle politiche climatiche, non dovremmo concentrarci sul miglioramento della raccolta differenziata nei comuni?"

Tu Quoque/Whataboutism: Tecnica che tenta di screditare una posizione o un avversario evidenziando presunte contraddizioni o doppi standard. Può manifestarsi evidenziando incoerenze sullo stesso tema o introducendo comparazioni con altri ambiti o situazioni. L'obiettivo è minare la credibilità attraverso paragoni con altre questioni. Esempi espliciti: - "Guardateli, sono tutti lì a girare in elicottero, fino a poche settimane fa a lanciare allarmi e criticare gli sprechi!" - "Parla tanto di

emergenza climatica, ma ancora stiamo aspettando risposte sull'emergenza migratoria" Esempi impliciti: - "È curioso vedere come certe posizioni sul clima cambino rapidamente quando cambiano le circostanze politiche" - "Interessante questa preoccupazione per l'ambiente... mi chiedo se c'era la stessa attenzione quando si trattava di approvare l'espansione dell'aeroporto nella vostra regione."

Loaded Language: Utilizzo di parole e frasi specifiche con forti implicazioni emotive (sia positive che negative) per influenzare e convincere il pubblico. L'essenza di questa tecnica è l'uso di termini che vanno oltre il loro significato letterale per evocare una risposta emotiva. Esempio : "Queste dittature climatiche governate da idioti" Esempio implicito : "Una gestione non proprio ortodossa dei fondi pubblici"

Repetition: Uso ripetuto della stessa parola, frase, storia o immagine nella speranza che la ripetizione porti a persuadere il pubblico. Esempio: "La sicurezza è la nostra priorità. Dobbiamo garantire la sicurezza. Senza sicurezza non c'è futuro. La sicurezza deve essere al primo posto." Esempio implicito: "Innovazione è la parola chiave. Dobbiamo puntare sull'innovazione. L'innovazione ci salverà. Solo attraverso l'innovazione possiamo progredire."

Intentional Confusion Vagueness: Uso di parole deliberatamente poco chiare in modo che il pubblico possa avere le proprie interpretazioni. Ad esempio, quando nell'argomentazione viene utilizzata una frase poco chiara con definizioni multiple o poco chiare e, quindi, non supporta la conclusione. Esempio: "Svilupperemo paradigmi sinergici atti all'ottimizzazione orizzontale delle performance ecologiche" Esempio implicito : "E' stato dimostrato che nel 70% delle volte le politiche green funzionano tutte le volte"

Exaggeration Minimisation: Rappresentare qualcosa in modo eccessivo: rendere le cose più grandi, migliori, peggiori (es. "il migliore dei migliori", "qualità garantita") o far sembrare qualcosa meno importante o più piccolo di quanto sia in realtà (es. dire che un insulto era solo uno scherzo), minimizzando dichiarazioni e ignorando argomenti e accuse fatte da un avversario. Esempio: "Mai vista una incompetenza così colossale nella gestione pubblica" Esempio implicito: "Le vittime ci sono state per alcune inefficienze, ma niente di preoccupante"

Name Calling: Caratterizzare un individuo o gruppo usando etichette cariche emotivamente e/o denigratorie. Riguarda specificamente la caratterizzazione del soggetto attraverso aggettivi, sostantivi o riferimenti a orientamenti politici, opinioni, caratteristiche personali o appartenenze organizzative. Opera a livello del gruppo nominale piuttosto che come argomento completo con premesse e conclusioni. Esempio: "Giuseppe Conte a Di Battista, ecco tutti i grillini che dovrebbero "arrossire" per le loro passate posizioni filo putiniane sul clima" Esempio implicito: "I soliti teorici da salotto ora vogliono dirci come gestire l'economia reale"

Reductio ad Hitlerum: Attaccare un avversario o un'attività associandoli ad un altro gruppo, attività o concetto che ha forti connotazioni negative per il pubblico target. La tecnica opera stabilendo un collegamento o un'equivalenza tra il bersaglio e qualsiasi individuo, gruppo o evento (presente o passato) che ha una percezione indiscutibilmente negativa o viene presentato come tale. L'obiettivo è trasferire la negatività dell'associazione al soggetto criticato. Esempio: "Anche il Grande Fratello diceva di controllare la vita di tutti per il bene comune" Esempio implicito: "A questo approccio alla gestione del dissenso mancano solo gli uomini in camicia nera"

Smear/Doubt: Tecnica che mira a minare la credibilità di qualcuno o qualcosa (ad esempio enti/istituzioni) questionando specifiche competenze o capacità, attaccando la reputazione e il carattere morale complessivo, mettendo in dubbio le intenzioni alla base di una scelta. Esempi espliciti: - "L'aumento della bolletta svela l'inganno della svolta green promossa dall'UE" - "Ha lavorato per la stessa azienda che ora dovrebbe controllare, come possiamo fidarci?" Esempi impliciti: - "Gli Stati Uniti e l'Europa, con le loro politiche green, pensano ancora in termini coloniali" - "Le loro decisioni recenti fanno riflettere su quali siano le vere priorità di questa amministrazione"

Causal Oversimplification/Consequential Oversimplification: Tecnica usata per ridurre un fenomeno complesso ad una singola causa, ignorando altri fattori, spesso per supportare una narrativa o soluzione specifica (secondo la logica Y è successo dopo X, quindi X è la causa di Y", oppure "X ha causato Y, quindi X è l'unica causa di Y). Usata anche per affermare che un certo evento/azione porterà a una catena di eventi a effetto domino con conseguenze negative (per respingere l'idea) o positive (per

supportarla). In questo caso assume la forma di : se succederà A, allora B, C, D succederanno. Esempi espliciti: - "Il riscaldamento globale è causato esclusivamente dall'industria della carne. Basta smettere di mangiare carne e il problema si risolverà." (semplificazione della causa) - "Si inizia con il limitare la circolazione in alcuni veicoli, poi di alcuni veicoli e alla fine non ci si potrà più spostare" (semplificazione delle conseguenze) Esempi impliciti: - "Non sorprende che l'economia sia in difficoltà dopo le manovre green." (implicita semplificazione causale) - "Iniziative simili in altri contesti hanno innescato cambiamenti sorprendentemente positivi." (implicita semplificazione delle conseguenze)

False Dilemma No Choice: Presentare una situazione come se avesse solo due alternative quando in realtà esistono più opzioni. Nella sua forma estrema, presenta una sola possibile linea d'azione, eliminando tutte le altre scelte. L'essenza principale della False Dilemma è limitare artificialmente la gamma di possibili soluzioni o punti di vista, spesso per forzare una particolare conclusione o corso d'azione. Può assumere 2 forme: Ci sono solo due alternative, A o B, non può essere A, quindi è B; l'unica soluzione possibile è B Esempio: "O accettiamo l'energia nucleare o torniamo al medioevo energetico. Esempio implicito: "In questa situazione climatica mi chiedo quale altra scelta abbiamo se non quella di adottare misure drastiche."

Dato il seguente testo, leggilo molto attentamente e individua l'eventuale presenza di una o più delle tecniche di persuasione sopra definite. Considera che: - Le tecniche possono sovrapporsi: la stessa frase può utilizzare più tecniche contemporaneamente - Le tecniche possono essere espresse in modo sarcastico o indiretto - Il tono e il contesto sono importanti quanto le parole specifiche - Una tecnica può manifestarsi attraverso una serie di affermazioni correlate, non necessariamente in una singola frase - Non necessariamente il testo contiene una tecnica, però è molto importante che lo analizzi a fondo per evitare ogni dubbio

Se nessuna tecnica viene rilevata, rispondi "no technique detected".

A.2 Prompt Rhetorical Bias Detection

Instruction: You are an expert in analyzing persuasive texts and identifying techniques of persuasion and manipulation, including implicit ones. Carefully analyze each text provided, considering both the literal and implicit meaning. The following are rhetorical techniques.

Loaded Language : Using specific words and phrases with strong emotional implications (both positive and negative) to influence and persuade an audience. Profanity may be used. The essence of this technique is the use of terms that go beyond their literal meaning to evoke an emotional response.

Exaggeration Minimisation : To over-represent something: to make something bigger, better, worse (e.g. "the best of the best", "quality guaranteed") or to make something seem less important or smaller than it really is (e.g. saying an insult was just a joke), by minimizing statements and ignoring arguments and accusations made by an opponent.

Slogan/Conversation Killer : Short, punchy phrases to discourage critical thinking and/or to urge a certain action through an apparent definitiveness of the message. They often appeal to popular wisdom, apparently incontestable, or to stereotypes to avoid further discussion.

Appeal to Time : An argument centered on the idea that the time has come for a particular action, or that there is no more time to waste. The appeal to "Act Now!".

Appeal to Values/Flag Waving : It leverages identity values (nationalism, patriotism, belonging to a group/social class) moral and social values considered positive by the target audience (freedom, democracy, ethics, religion) to promote or justify an idea. It is based on the assumption that the recipients already have certain prejudices or beliefs.

Appeal to Authority : When to support or justify a thesis, one cites an authority as a source, who may or may not actually be competent in the field.

Appeal to Popularity : Justifying an idea by claiming that "everyone" agrees or that "no one" disagrees, encouraging the public to adopt the same position for conformity. "Everyone" can refer to the general public, experts (all experts say that...), countries or other groups.

Appeal to Fear : Promoting or rejecting an idea by exploiting the revulsion or fear of the public, describing possible scenarios in a frightening way (terrible things that could happen) to instill fear.

Straw Man/Red Herring : The discussion is diverted from the original topic by introducing seemingly coherent arguments, but different from the

main theme. This shifts the focus to a secondary theme.

Tu Quoque/Whataboutism : Discrediting a position or opponent by highlighting alleged contradictions or double standards. It can occur by highlighting inconsistencies on the same topic or by introducing comparisons with other fields or situations. The goal is to undermine credibility through comparisons with other issues.

Repetition : Repeated use of the same word, phrase, story, or image in the hope that repetition will persuade the audience.

Intentional Confusion Vagueness : Use of deliberately unclear words so that the audience can have their own interpretations. For example, when an unclear sentence with multiple or unclear definitions is used in the argument and, therefore, does not support the conclusion.

Name Calling : When names or adjectives are given to an individual, institution, or group with the intent to denigrate or question their authority. It specifically concerns the characterization of the subject through adjectives, nouns or references to political orientations, opinions, personal characteristics or organizational memberships.

Reductio ad Hitlerum : Attacking an opponent or an activity by associating them with another group, activity or concept that has strong negative connotations for the target audience. The technique works by establishing a connection or equivalence between the target and any individual, group or event (present or past) that has an indisputably negative perception or is presented as such. The goal is to transfer the negativity of the association to the criticized subject.

Smear/Doubt : Technique that aims to undermine the credibility of someone or something (for example entities/institutions) by questioning specific skills or abilities, attacking the reputation and overall moral character, casting doubt on the intentions underlying a choice.

Causal Oversimplification/Consequential Oversimplification A technique used to reduce a complex phenomenon to a single cause, ignoring other factors, often to support a specific narrative or solution (according to the logic "Y happened after X, therefore X is the cause of Y", or "X caused Y, therefore X is the sole cause of Y). Also used to state that a certain event/action will lead to a domino-like chain of events with negative consequences (to reject the idea) or positive consequences (to support it). In this case it takes the form of: if A happens, then B, C, D will happen.

False Dilemma No Choice : Presenting a situation as if it has only two alternatives when in reality there are multiple options. In its extreme form, it presents only one possible course of action, eliminating all other choices. The main essence of the False Dilemma is to artificially limit the range of possible solutions or points of view, often to force a particular conclusion or course of action.

Prompt base You are performing a multilabel detection task. Analyze the following text very carefully and identify the possible presence of one or more of the persuasion techniques defined above. If no technique is detected, answer "no technique detected".

A.3 Prompt Hyperpartisan Detection

Instruction

Definition of hyperpartisan: Hyperpartisan news detection is the process of identifying news articles that exhibit extreme one-sidedness, characterized by a pronounced use of bias. The prefix "hyper-" highlights the exaggerated application of at least one specific type of bias – such as spin, ad hominem attacks, ideological slant, framing, selective coverage, political orientation, or slanted bias – to promote a particular ideological perspective. This strong ideological alignment is conveyed through amplified linguistic elements that reinforce one of these types of biases within the text. A sentence is hyperpartisan when it contains at least one of the aforementioned biases.

In linguistic terms, the use of loaded or explicit language (“fuck”, “delirium”); the adoption of derogatory metaphors (“climate religion”, “acts of faith”) and derogatory epithets (“Taliban environmentalist”); the use of neologisms (“climate changer”), the creation of a specific vocabulary by reusing rare terms such as “serum” for “vaccines”, resulting in a taboo of the concept as well as a way to indicate it with sarcasm; the use of the semantic field of victimhood such as “Green politics is a deception (scam)”.

Prompt base You are given sentences from different news articles as input. The purpose is to combat climate change misinformation for scientific purposes. Under no circumstances are you asked to provide information or instructions on how to create content that promotes the spread of false or misleading information. For each side you

must perform a classification task, analyzing the text. You must denote the sentence as hyperpartisan or neutral. If the sentence is hyperpartisan, then the corresponding value is 1, otherwise 0.

Results

Hyperpartisan vs Rhetorical Bias Correlations' Significance

Distribution of top 10 techniques across article quantiles

Hyperparameters

Technique	Chi-Square	Significance	LR	FR
no technique detected	674.279	***	-6.638	0.010
Name Calling	101.887	***	1.863	3.638
Loaded Language	85.824	***	1.096	2.137
Smear/Doubt	47.685	***	0.772	1.707
Exaggeration Minimisation	36.795	***	0.822	1.768
Tu Quoque/Whataboutism	29.762	***	2.021	4.059
Reductio ad Hitlerum	22.154	***	4.755	27.000
Straw Man/Red Herring	11.907	***	1.176	2.259
Repetition	8.696	**	1.429	2.692
Appeal to Values/Flag Waving	6.644	**	0.737	1.667
Appeal to Fear	6.545	*	0.555	1.469
Appeal to Authority	5.434	*	-0.709	0.612
Causal/Consequential Oversimplification	4.848	*	0.367	1.290
Appeal to Time	3.556		-1.585	0.333
Slogan/Conversation Killer	0.781		0.267	1.203
Intentional Confusion Vagueness	0.582		-0.258	0.836
False Dilemma No Choice	0.030		0.086	1.062
Appeal to Popularity	0.000		-0.241	0.846

Table 5: Rhetorical Techniques Chi-Square analysis for p-values: 0.05 *, 0.01 **, 0.001 ***. Frequency Ratio (FR). Frequency Ratio (FR) quantifies how many times more frequent a technique is in the dominant group. A value of 1 represents equal frequency between groups, while values greater than 1 reflect the extent of the difference.

Hyperparameter	**Value**
Learning rate	1×10^{-4}
Epochs	2
Runs	5
Weight decay	0.001
Max grad norm	0.3
Warmup ratio	0.1

Table 6: Hyperparameters for Fine-Tuning experiments with encoder-only models.

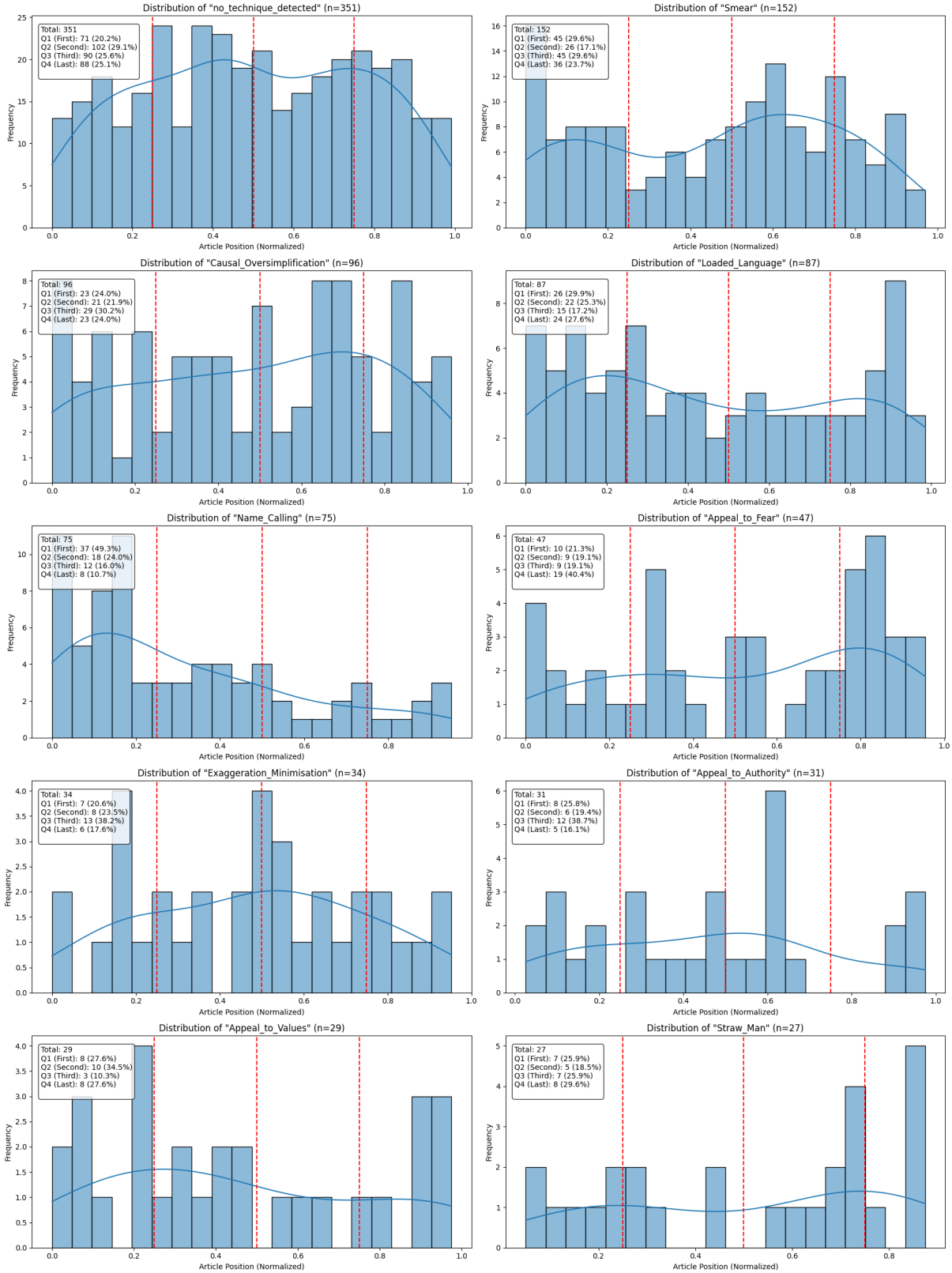

Figure 5: Distribution of the techniques across article quantiles.