

TESE DE DOUTORAMENTO

HYPERPARTISAN NEWS DETECTION

Michele Joshua Maggini

ESCOLA DE DOUTORAMENTO INTERNACIONAL DA UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

PROGRAMA DE DOUTORAMENTO EN INVESTIGACIÓN EN TECNOLOXÍAS DA INFORMACIÓN

SANTIAGO DE COMPOSTELA
2026



Declaración do autor da tese
HYPERPARTISAN NEWS DETECTION

Don Michele Joshua Maggini

Presento a miña tese, seguindo o procedemento adecuado ao Regulamento, e declaro que:

1. *A tese abarca os resultados da elaboración do meu traballo.*
2. *De ser o caso, na tese faise referencia ás colaboracións que tivo este traballo.*
3. *Confirmo que a tese non incorre en ningún tipo de plaxio doutros autores nin de traballos presentados por min para a obtención doutros títulos.*
4. *A tese é a versión definitiva presentada para a súa defensa e coincide a versión impresa coa presentada en formato electrónico.*

E comprométome a presentar o Compromiso Documental de Supervisión no caso de que o orixinal non estea na Escola.

En Santiago de Compostela, X

Asdo. Michele Joshua Maggini



Autorización do director/titor da tese
HYPERPARTISAN NEWS DETECTION

Don Pablo Gamallo Otero, Profesor Catedrático da Área de Sistemas Intelixentes da Universidade de Santiago de Compostela

Don Gael Dias, Profesor Catedrático da Área XXX

INFORMAN:

Que a presente tese correspóndese co traballo realizado por ***Don Michele Joshua Maggini****, baixo a nosa dirección/titorización, e autorizamos a súa presentación, considerando que reúne os requisitos esixidos no Regulamento de Estudos de Doutoramento da USC, e que como directores/titores desta non incorre nas causas de abstención establecidas na Lei 40/2015.*

En Santiago de Compostela, X

Asdo. Pablo Gamallo Otero
Director/a tese

Asdo. Gael Dias
Director/a tese

A todas la cousas invisible and visibili.

To the immense old hands posed on my forehead that mangled my name as a child. I would call to you from downstairs, and I grew accustomed to waiting for your whistle, the one that the blackbird in the garden learned and still sings, disturbing the sleep of the cat laying in the sun, unable to find you. But you are with me, and I am with you.

Acknowledgments

Firstly, the gratitude goes to my supervisors, Pablo Gamallo and Gaël Dias, for their immense support thoughout these year, their time and knowledge have been essential to raise me as an independent researcher and make this work possible. Then, I would like to thank all the people of the HYBRIDS network for their work and for the journey we shared together, especially the other DCs candidates with whom deep bonds born, while relying on each other during both chaotic and calm periods. My deepest thank also go to my parents and my family, who took an important place in this journey to welcome their son every time he came back home and let him bring home anywhere. It will be summer forever for you who passed away during these years. Moving to my friends, I will always be grateful to have you in my life, because distance is just a geometric construct: Silvia and Mattia, Belin, Michele, Marco, Lorenzo, Donata and Mario. Thanks to everyone I ever met for your existence. Lastly, Amelia, for whom words are incapable of describing the love I feel. And even if they were, what could be said would merely be a description, much like replying to Nisargadatta's question: "What is Bombay?". I only say that all the poetry is no longer in the poems, but in the days we tell that there may be other days to tell, the ones starting with: "Do you remember when?".

This thesis was funded from the Horizon Europe research and innovation programme under the Marie Sk lodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee - Grant Number: EP/X036758/1: HYBRIDS Project.

Contents

List of Abbreviations **17**

List of Publications included in this Ph.D. Dissertation **19**

List of Figures **25**

List of Tables **27**

Resumen **31**

Resumen **33**

Abstract **35**

1 Introduction **1**

1.1 Thesis Topic and Motivations 1

1.2 The Hyperpartisan News Crisis in Europe 3

1.3 From Traditional to Digital Political Communication 5

1.3.1 The Common Ground of Hyperpartisan News 5

1.3.2 The Evolution of Political Parallelism in European Media 6

1.3.3 Internet, Populism, and the Transformation of Political Communication 7

1.3.4 The Alternative Media Landscape in Europe 7

1.3.5 From Populism to Hyperpartisanship: The Radicalization Dynamic . 8

1.3.6 Defining Hyperpartisan as a per-se phenomenon: disentangling differences with Fake News and similar research fields 9

1.4 The regulation of Disinformation in Europe 15

1.5 Research Positioning and Thesis Structure 16

2 Hypothesis and Objectives **19**
2.1 Data and Annotation 19
2.1.1 Learning Strategies and Architectures 20
2.1.2 Linguistic Analysis 21
2.2 Research Questions 21
2.3 Objectives and Contributions 25
2.3.1 Systematic Review of Hyperpartisan News Detection 25
2.3.2 Benchmarking Learning Paradigms for Misinformation Detection . . 25
2.3.3 Multilingual Hyperpartisan and Conspiracy Detection in European Media 26
2.3.4 Italian Hyperpartisan Detection with Rhetorical Bias Taxonomy . . . 27
2.3.5 Pragmatic Features for Italian Hyperpartisan Detection 27
2.3.6 Italian Hyperpartisan Neutralization 28

3 Methodology **29**
3.1 Defining Hyperpartisan News 30
3.1.1 Where can hyperpartisanship be detected? Methodological operationalizations and their implications 35
3.1.2 How is hyperpartisanship annotated and operationalized? The measurement dilemma 37
3.1.3 Datasets in Literature 39
3.2 Literature Review 47
3.2.1 Literature Review: Hyperpartisan and Misinformation Detection . . . 47
3.2.2 In-Context Learning 63
3.2.3 Rhetorical Bias and Persuasion Techniques 64
3.2.4 Linguistic Trait Detection 64
3.2.5 Domain-Specific Datasets 65
3.2.6 Explainability and Embedding-Based Enhancements 66
3.2.7 LLMs as Annotators and Persona Prompting 66
3.2.8 Text Style Transfer: Neutralization of Biased Texts 67
3.3 Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning 68

3.3.1 Models . . . 70
3.3.2 Prompt design . . . 70
3.3.3 Determinantal Point Process . . . 71
3.3.4 ICL Setting . . . 72
3.4 PartisanLens: A Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media . . . 77
3.4.1 Data collection and preparation . . . 79
3.4.2 Benchmark Experiments: Setup . . . 81
3.4.3 Persona prompt . . . 86
3.5 Detecting Hyperpartisanship and Rhetorical Bias in Climate Journalism: A Sentence-Level Italian Dataset . . . 87
3.6 IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News . . . 92
3.6.1 Experimental Setup . . . 97
3.7 Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization . . . 99
3.7.1 Design Principle . . . 99
3.7.2 Dataset Construction . . . 101
3.7.3 Three-Stage Annotation Pipeline . . . 102
3.7.4 Benchmark Experiments: Setup . . . 104
3.8 Summary of Methodology . . . 110

4 Results and Discussion **111**
4.1 Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning . . . 111
4.1.1 Fine-Tuning . . . 112
4.1.2 In-Context Learning . . . 113
4.1.3 Summary for Zero-shot configurations . . . 116
4.1.4 FS: DPP-selected vs Random examples . . . 117
4.1.5 Chain of Thought . . . 118
4.1.6 Insights from Fine-Tuning vs. In-Context Learning in LLMs . . . 119
4.1.7 Final Remarks . . . 120
4.2 PartisanLens: A Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media . . . 122

4.2.1 Annotation Process Results and Dataset Statistics 122
4.2.2 Classification Results 124
4.2.3 LLMs as Potential Data Annotators 126
4.2.4 127
4.2.5 Results 128
4.2.6 Error analysis 130
4.2.7 Classification Errors 130
4.2.8 Final Remarks 131
4.3 Detecting Hyperpartisanship and Rhetorical Bias in Climate Journalism: A Sentence-Level Italian Dataset 132
4.3.1 Dataset 132
4.3.2 Hyperpartisan-Rhetorical Bias Relation 133
4.3.3 Topological Distribution of Rhetorical Biases 138
4.3.4 Computational Baselines 141
4.3.5 Final Remarks 144
4.4 IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News 144
4.4.1 Dataset 145
4.4.2 Fine-Tuning and Ablation Study 149
4.4.3 Integrated Gradient and SHAP Analysis 153
4.4.4 Final Remarks 155
4.5 Par-ITA: Seq2seq vs LLMs on neutralizing capabilities 156
4.5.1 Data Generation and Pipeline Evaluation 156
4.5.2 Seq2seq Model Results 159
4.5.3 LLM Results 161
4.5.4 Human Evaluation and Metric Alignment 162
4.5.5 Final Remarks 162

5 Conclusions and Future Work **165**

Conclusions **165**
5.1 Contributions 165
5.2 Conclusion 168
5.3 Future Work 169

Bibliography **173**

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
AI Act	Artificial Intelligence Act (EU regulation 2024)
BERT	Bidirectional Encoder Representations from Transformers
CoT	Chain-of-Thought
CNN	Convolutional Neural Network
DPO	Direct Preference Optimization
FT	Fine-Tuning
GPT	Generative Pre-trained Transformer
ICL	In-Context Learning
LSTM	Long Short-Term Memory
LLM/LLMs	Large Language Model(s)
ML	Machine Learning
NLP	Natural Language Processing
RoBERTa	Robustly Optimized BERT Pretraining Approach
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
HP/HN	Hyperpartisan/Hyperpartisan News
IAA	Inter-Annotator Agreement
PRCT	Population Replacement Conspiracy Theories
RQ	Research Question
IHPP	Italian Hyperpartisan Paragraph-level Dataset

Continued on next page

List of abbreviations continued from previous page

Abbreviation	**Definition**
Par-ITA	Italian Hyperpartisan Neutralization Parallel Corpus
PartisanLens	Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media
PLOS One	Public Library of Science One
DSA	Digital Services Act (EU regulation 2024)
EU	European Union
UKRI	UK Research and Innovation
VLOP/VLOPs	Very Large Online Platform(s)
VLOSE/VLOSEs	Very Large Online Search Engine(s)
COVID-19	Coronavirus Disease 2019

List of Publications included in this Ph.D. Dissertation

This Ph.D. dissertation, particularly chapters 3.8 and 4.5.5 utilize the contents of the following publications:

1. Maggini MJ, Bassi D, Piot P, Dias G, Otero PG (2025) A systematic review of automated hyperpartisan news detection. PLoS ONE 20(2): e0316989. https://doi.org/10.1371/journal.pone.0316989.

    - Type: Journal paper.
    - **Quality indicators:** Scimago Q1 (Multidisciplinary).
    - **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
    - **Co-Author contribution**: Davide Bassi: conceptualization, writing and revision of the paper; Paloma Piot: designing tables; Gäel Dias: review of the paper; Pablo Gamallo: conceptualization, design, review of the paper.
    - **Reproduction authorization:** This is an open access article distributed under the terms of the CreativeCommonsAttributionLicense, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See: https://doi.org/10.1371/journal.pone.0316989.

2. Michele Joshua Maggini, Paloma Piot, Anxo Pérez, Erik Bran Marino, Lúa Santamaría Montesinos, Ana Lisboa Cotovio, Marta Vázquez Abuín, Javier Parapar, and

Pablo Gamallo. 2026. PartisanLens: A Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media. In Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1171–1186, Rabat, Morocco. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.53, https://aclanthology.org/2026.eacl-long.53/.

- Type: Conference paper.
- **Quality indicators:** Conference ranking (GRIN-SCIE): A.
- **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
- **Co-Author contribution**: Paloma Piot: implementation of the algorithms, conceptualization, writing, data creation, and review; Anxo Pérez: implementation of the algorithms, writing, data creation, and review; Erik Bran Marino: conceptualization, writing, data creation; Lúa Santamaría Montesinos, Ana Lisboa Cotovio, Marta Vázquez Abuín: data creation; Javier Parapar: review of the paper; Pablo Gamallo: review of the paper.
- **Reproduction authorization:** ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

3. Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 2025. Detecting Hyperpartisanship and Rhetorical Bias in Climate Journalism: A Sentence-Level Italian Dataset. In Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025), pages 168–187, Vienna, Austria. Association for Computational Linguistics. doi: 10.18653/v1/2025.climatenlp-1.11, 10.18653/v1/2025.climatenlp-1.11.

   - Type: Workshop paper.

- **Quality indicators:** Conference ranking (GRIN-SCIE): A*.
- **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
- **Co-Author contribution**: Davide Bassi: conceptualization, conduction of the experiments; Pablo Gamallo: conceptualization, review of the paper.
- **Reproduction authorization:** ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

4. Maggini, M. J., Merzougui, D., Bandyopadhyay, R., Dias, G., Maurel, F., Gamallo, P. (2026). Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning. Proceedings of the International AAAI Conference on Web and Social Media, 20(1), 1551–1572. `https://doi.org/10.1609/icwsm.v20i1.42712`.

    - Type: Conference paper.
    - **Quality indicators:** Conference ranking (GRIN-SCIE): A.
    - **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
    - **Co-Author contribution**: Merzougui, Dhia: implementation of the algorithms, design and conduction of the experiments; Bandyopadhyay, Rabiraj: implementation of the algorithms, design; Dias, G.: conceptualization, review; Maurel, F.: review; Pablo G.: conceptualization, review.
    - **Reproduction authorization:** AAAI policies: `https://mas.cs.umass.edu/icmas/copyright.html`. In return for these rights, AAAI hereby grants to the above authors, and the employers for whom the work was performed, royalty-free permission to:Retain all proprietary rights other than copyright (such

as patent rights). Personal reuse of all or portions of the above article/paper in other works of their own authorship. Reproduce, or have reproduced, the above article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. Make limited distribution of all or portions of the above article/paper prior to publication. In the case of work performed under U.S. Government contract, AAAI grants the U.S. Government royalty-free permission to reproduce all or portions of the above article/paper, and to authorize others to do so, for U.S. Government purposes.

5. Maggini, M. J., Bassi, D., Valente, A., Dias, G., Gamallo, P. (2026). IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News. In Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026) (pp. 1998–2011). European Language Resources Association (ELRA). `https://doi.org/10.63317/2sabeivqu9dt`.

   - Type: Conference paper.
   - **Quality indicators:** Conference ranking (GRIN-SCIE): B.
   - **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
   - **Co-Author contribution**: Bassi Davide: design, data creation; Valente Angelo: data creation; Dias G.: review, writing of the paper; Gamallo P.: conceptualization, review of the paper.
   - **Reproduction authorization:** ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

6. Maggini, M. J., Søren Fomsgaard, Michele Maestroni, Gaël Dias, Pablo Gamallo. (2026). Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization. In Proceedings of the 64rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). San Diego, California, USA. Association for Computational Linguistics.

    - Type: Conference paper.
    - **Quality indicators:** Conference ranking (GRIN-SCIE): A*.
    - **Author's contribution:** research of the topic, conceptualization, design and implementation of the algorithms, design and conduction of the experiments, and writing of the article.
    - **Co-Author contribution**: Søren Fomsgaard: implementation of the algorithms, conduction of the experiments; Michele Maestroni: data creation; Gaël Dias: conceptualization, review; Pablo Gamallo: conceputalization, design, review.
    - **Reproduction authorization:** ACL materials are Copyright © 1963–2026 ACL; other materials are copyrighted by their respective copyright holders. Materials prior to 2016 here are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License. Permission is granted to make copies for the purposes of teaching and research. Materials published in or after 2016 are licensed on a Creative Commons Attribution 4.0 International License.

List of Figures

Fig. 3.1 The bar chart illustrates the trend of the selected publications over time. . . . 31
Fig. 3.2 The pie chart shows the amin publishers for the selected papers. 32
Fig. 3.3 Datasets for Hyperpartisan News detection task. 41
Fig. 3.4 PRISMA Flow Diagram. 42
Fig. 3.5 Datasets' language distribution. 43
Fig. 3.6 Example of the custom HTML platform used for the annotation process. . . 80
Fig. 3.7 Example of a rationale-enhanced annotation. 83
Fig. 3.8 Example of the annotation scheme for one paragraph for the sentence-level dataset on climate change. 90
Fig. 3.9 Example of the annotation scheme for one paragraph for the HIPP dataset. FS: Figurative Speech, LL: Loaded Language, HL: Hyperbolic Language. . . 96
Fig. 3.10 The three-stage annotation and evaluation pipeline of Par-ITA. 99

Fig. 4.1 Results for zero-shot and CoT grouped by models. 113
Fig. 4.2 Results zero-shot and CoT grouped by configuration. 114
Fig. 4.3 FS DPP vs Random results. 117
Fig. 4.4 Hyperpartisan and PRCT stance proportions. 123
Fig. 4.5 Hyperpartisan and PRCT content by language. 123

Fig. 4.6 Correlation between Hyperpartisan sentences and techniques. 135
Fig. 4.7 Log Ratio of the Rhetorical Techniques with regards to Hyperpartisan/Neutral Label. 136
Fig. 4.8 Distribution of the Rhetorical Techniques across article quantiles. 138
Fig. 4.9 Distribution of Rhetorical Techniques across article quartiles 139
Fig. 4.10 Hyperpartisan Contamination Level per sentence position (deciles). 141
Fig. 4.11 Distribution of Rhetorical Techniques across articles' bodies normalized. . . 142
Fig. 4.12 Distribution of HP paragraphs across article deciles. 148
Fig. 4.13 Log Ratio of the LT with regards to Hyperpartisan/Neutral paragraphs. . . . 149
Fig. 4.14 LT average drop by model in ablation. 151
Fig. 4.15 Attribution Ratios: Models × LT × Variants 151
Fig. 4.16 Comparison of SHAP values across learning strategies ordered by magnitude. 153
Fig. 4.17 Correlation between human consensus and GPT-4o-mini scores (averaged Syntax and Semantics) during data generation. 157
Fig. 4.18 Cohen's κ between human experts and GPT-4o-mini on binary label for the Rewritten task across models. 157
Fig. 4.19 Inter-annotator agreement (Cohen's κ) between humans across evaluated LLMs. 159

List of Tables

Tab. 1.1 Definitions of hyperpartisanship in the selected papers. 10
Tab. 1.2 Examples of statements for specific biases and hyperpartisan statements for that bias. 14

Tab. 3.1 Queries performed with advanced search for each database and the number of papers retrieved. 33
Tab. 3.2 Datasets collection and description. Part 1. 44
Tab. 3.3 Datasets collection and description. Part 2. 45
Tab. 3.4 Datasets collection and description. Part 3. 46
Tab. 3.5 Traditional Machine Learning algorithms used in the selected literature. . . 48
Tab. 3.6 Feature-based approaches for hyperpartisan news detection. 50
Tab. 3.8 Collection of the best performative DL models used in the literature. 53
Tab. 3.10 Metrics Results Literature Review Part 1 61
Tab. 3.11 Metrics Results Literature Review Part 2 62
Tab. 3.12 Description of the datasets used in our experiments. 70
Tab. 3.13 Table showing different examples of instruction for PL Detection. 73
Tab. 3.14 Dataset label distribution. 82
Tab. 3.15 Example of sociodemographic data. 84

Tab. 3.16 Participant demographic overview 85
Tab. 3.17 Zero-shot and few-shot prompt. 86
Tab. 3.18 Zero-shot Persona prompts. 87
Tab. 3.19 Taxonomy of rhetorical biases and hyperpartisan language detection for sentence-level dataset. 92
Tab. 3.20 Taxonomy of LT and HP language detection for HIPP. 96
Tab. 3.21 Par-ITA: Examples of model-generated and human-supervised neutral rewrites for hyperpartisan input phrases. 101
Tab. 3.22 Statistics of the human-curated Par-ITA corpus. 103
Tab. 3.23 Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. 107
Tab. 3.24 Prompt used for zero-shot neutralization with big size LLMs 107
Tab. 3.25 Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. Part 1. . . . 108
Tab. 3.26 Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. Part 2 . . . 109

Tab. 4.1 Performance of models in the FT setting. 112
Tab. 4.2 Comprehensive comparison of F1 and accuracy across all models, configurations, and datasets. 121
Tab. 4.3 IAA across annotation tasks 123
Tab. 4.4 Macro-averaged metrics across tasks and models: PartisanLens. 124
Tab. 4.5 Fine-tuning with rationales results by language (macro-averaged). 125
Tab. 4.6 Mean Cohen's κ by language and label: PartisanLens. 126
Tab. 4.7 PartisanLens:Cohen's κ between user-persona and corresponding human annotator. 128
Tab. 4.8 Selected headlines with human and user-personas disagreements. 129
Tab. 4.9 Dataset Statistics: sentence-level. 133
Tab. 4.10 Topic Distribution of sentence-level dataset. 134
Tab. 4.11 Rhetorical Techniques' Chi-Square. 137
Tab. 4.12 Comparison of Hyperpartisan and Rhetorical Bias Classification Models . . 141
Tab. 4.13 Hyperparameters for Fine-Tuning experiments with encoder-only models. . . 143
Tab. 4.14 IHPP topic-wise statistics for train and test splits. 147
Tab. 4.15 Comparison of Italian language models on HP detection task. 150
Tab. 4.16 LT FT+EMB error statistics. 152
Tab. 4.17 Par-ITA: IAA during preliminary LLM selection phase. 156
Tab. 4.18 Generation quality of texts. 158

Tab. 4.19 Seq2seq model performance on Par-ITA test set. 160
Tab. 4.20 LLM performance on Par-ITA test set. 161
Tab. 4.21 Correlation between mean human ratings and automatic metrics across LLM configurations. 162

Resumen

No ámbito da comunicación política, as noticias modelan as percepcións dos votantes e actúan como nexo entre os actores democráticos. Este tipo de transmisión da mensaxe representa como os diferentes actores e partes comparten e experimentan o espazo político. Garantir o dereito á liberdade de expresión é fundamental para que as opinións e perspectivas transmitidas polos medios reflictan diferentes ideoloxías, contribuíndo así ao fortalecemento do debate público.

Porén, as noticias non reciben unha cobertura homoxénea nin perseguen os mesmos obxectivos. O grao dos seus efectos está relacionado co tema tratado, xa que a difusión tende a intensificarse nas comunidades onde determinados asuntos son percibidos como especialmente sensibles. Nalgúns casos, como nas cámaras de eco, a manipulación dos contidos reforza o sentimento de pertenza a un mesmo grupo que comparte marcos narrativos comúns sobre un determinado acontecemento. As formas de polarización tamén dependen de como os lectores, influenciados polas súas crenzas epistemolóxicas, interactúan coas noticias, xa que estas interaccións poden producir efectos de retroalimentación sobre os estilos de escrita xornalística, reflectíndose en cambios de ton e de veracidade percibida. Deste xeito, cada medio desenvolve a súa propia maneira de reinterpretar e presentar as realidades políticas.

A finalidade deste traballo consiste en analizar o panorama europeo das noticias hiperpartidistas e a diferenciación dos temas especialmente relevantes para as políticas da Unión Europea sobre inmigración, cambio climático e euroescepticismo. Estes asuntos constitúen elementos centrais das campañas da extrema dereita e da extrema esquerda para obter apoio político mediante a exacerbación da polarización, reproducindo dinámicas características do modelo estadounidense de comunicación política.

Para acadar este obxectivo desenvolvéronse catro corpus, un multilingüe e tres en italiano, contribuíndo á consolidación dos recursos textuais sobre noticias hiperpartidistas en linguas con escaseza de recursos. Estes recursos abranguen italiano, castelán e portugués, linguas

faladas no sur de Europa en países especialmente expostos aos efectos da polarización política. Os corpus proporcionan evidencia sobre a relación entrelazada entre o construto do hiperpartidismo e os seus compoñentes psicolingüísticos. The *Climate Change Sentence-Level Dataset* analiza como os sesgos retóricos modelan e polarizan a mensaxe; *HIPP* ofrece información sobre as dimensións pragmáticas da linguaxe hiperpartidista; *Par-ITA* é un corpus paralelo para a neutralización deste tipo de discurso, desenvolvido mediante unha metodoloxía human-and-machine-in-the-loop supervisada por Large Language Models actuando como xuíces. Finalmente, *PartisanLens* é un recurso multilingüe no que se avaliou a capacidade dos Large Language Models para emular o comportamento dos anotadores humanos.

Todos os procesos de anotación e revisión foron realizados por expertos humanos, convertendo estes recursos en corpus de referencia (gold standard) para este dominio de aplicación. A través deles avaliáronse diversas arquitecturas neuronais, incluíndo encoders, decoders e modelos sequence-to-sequence, mediante experimentos de benchmarking baseados en standard fine-tuning, estratexias avanzadas de aprendizaxe como Direct Preference Optimization (DPO) e In-Context Learning (ICL).

Os resultados destes experimentos establecen referencias e recursos valiosos para a comunidade de Procesamento da Linguaxe Natural que traballa na detección da desinformación e das noticias hiperpartidistas.

Palavras chaves

Hyperpartisan News Detection, Idiomas con recursos minoritarios, Creacción de recursos, Persona-Prompting, LLMs, Explainability.

Resumen

En el campo de la comunicación política, las noticias modelan el imaginario de los votantes y actúan como nexo entre los actores democráticos. Este tipo de transmisión del mensaje representa cómo los diferentes actores y partes comparten y viven el espacio político. Garantizar el derecho a la libertad de expresión es fundamental para que las opiniones y las perspectivas expresadas por los medios reflejen diferentes ideologías, contribuyendo a alimentar el debate público. Sin embargo, las noticias ni reciben una cobertura homogénea, ni tienen los mismos objetivos. El grado de sus efectos está relacionado con el tema, en cuanto la difusión se alimenta entre las comunidades donde esto se considera más sensible. En algunos casos, como en las echo chambers, la manipulación del contenido fortalece el sentimiento de pertenencia a una misma familia que comparte los nudos narratológicos fundamentales sobre un hecho. Las formas de polarización dependen también de cómo los lectores, influenciados por sus creencias epistemológicas, interactúan con las noticias, porque pueden tener un efecto de feedback en el estilo de escritura de ellas, reflejándose en un cambio de tono y veracidad. De esta manera, cada medio evoluciona su manera de reinterpretar y presentar las condiciones políticas. La finalidad de este trabajo consiste en analizar el panorama europeo de las noticias hiperpartidistas y la diferenciación de los temas que son particularmente relevantes para las políticas de la UE sobre la inmigración, el cambio climático y el euroescepticismo: o sea, los temas en que se basan las campañas de extrema derecha e izquierda para ganar consentimiento exacerbando la polarización, reproduciendo dinámicas características del modelo de comunicación política estadounidense. Para lograr este objetivo hemos desarrollado y creado 4 corpora, 1 multilingüe y 3 en italiano, consolidando el repertorio textual de las noticias hiperpartidistas para idiomas con escasez de recursos, cubriendo italiano, castellano y portugués: idiomas hablados en el sur de Europa en países amenazados por la polarización. Los corpora proporcionan datos sobre la relación entrelazada entre el constructo del hiperpartidismo y sus componentes psi-

colingüísticos. *The Climate Change Sentence-Level Dataset* analiza cómo los sesgos retóricos moldean el mensaje polarizándolo; *HIPP* revela más información sobre la parte pragmática del lenguaje hiperpartidista; mientras *Par-ITA* es un corpus paralelo para la neutralización de este tipo de lenguaje y fue creado con una metodología human-&machine-in-the-loop supervisada con Large Language Models como jueces. Por último, *PartisanLens* es un recurso multilingüe donde se evaluó la capacidad de los Large Language Models de emular a los anotadores. Los humanos condujeron todas las anotaciones y la parte de revisión de las mismas, así que estos recursos son gold standard para este dominio aplicativo. De hecho, a través de estos evaluamos varias arquitecturas (encoders, decoders, seq2seq) haciendo benchmarking con standard fine-tuning y varias estrategias de aprendizaje (e.g. DPO), y In-Context-Learning. Los resultados de estos experimentos establecen benchmarks útiles para la comunidad de PLN que trabaja sobre la desinformación.

Palabras claves

Hyperpartisan News Detection, Low-resource languages, Resource creation, Persona-Prompting, LLMs, Explainability.

Abstract

In the field of political communication, news media shape voters' perceptions and act as a bridge among democratic actors. This form of message transmission reflects how different actors and stakeholders share and experience the political sphere. Ensuring the right to freedom of expression is essential for the opinions and perspectives conveyed by media outlets to reflect diverse ideologies, thereby contributing to the vitality of public debate.

However, news coverage is neither homogeneous nor driven by the same objectives. The extent of its effects is closely related to the topic being addressed, as dissemination tends to intensify within communities where certain issues are perceived as particularly sensitive. In some cases, such as echo chambers, content manipulation reinforces a sense of belonging among individuals who share common narrative frameworks regarding a given event. Forms of polarization also depend on how readers, influenced by their epistemological beliefs, interact with news content, as their responses may generate feedback effects on journalistic writing styles, ultimately influencing tone and perceived truthfulness. Consequently, each media outlet develops its own way of reinterpreting and presenting political realities.

The aim of this work is to analyze the European landscape of hyperpartisan news and the differentiation of topics that are particularly relevant to European Union policies on immigration, climate change, and Euroscepticism. These issues constitute key themes in far-right and far-left political campaigns seeking to gain public support by exacerbating polarization and reproducing dynamics characteristic of the U.S. political communication model.

To achieve this objective, we developed four corpora, one multilingual and three in Italian, contributing to the consolidation of textual resources for hyperpartisan news in low-resource languages. These resources cover Italian, Spanish, and Portuguese, languages widely spoken in Southern European countries increasingly affected by political polarization. The corpora provide evidence on the intertwined relationship between the construct of hyperpartisanship

and its psycholinguistic components. The *Climate Change Sentence-Level Dataset* investigates how rhetorical biases shape and polarize messages; *HIPP* provides insights into the pragmatic dimensions of hyperpartisan language; *Par-ITA* is a parallel corpus designed for the neutralization of hyperpartisan discourse and was developed through a human-and-machine-in-the-loop methodology supervised by Large Language Models acting as judges. Finally, *PartisanLens* is a multilingual resource used to evaluate the ability of Large Language Models to emulate human annotators.

All annotation and revision processes were conducted by human experts, making these resources gold-standard datasets for this application domain. Using these corpora, we evaluated several neural architectures, including encoders, decoders, and sequence-to-sequence models, through benchmarking experiments based on standard fine-tuning, advanced learning strategies such as Direct Preference Optimization (DPO), and In-Context Learning (ICL).

The results establish valuable benchmarks and resources for the Natural Language Processing community working on disinformation and hyperpartisan news detection.

Keywords

Hyperpartisan News Detection, Low-resource languages, Resource creation, Persona-Prompting, LLMs, Explainability.

CHAPTER 1

INTRODUCTION

1.1 Thesis Topic and Motivations

The primary objective of this thesis is to present the work conducted within the HYBRIDS MSCA project, focused on investigating the impact of disinformation in Europe and the means to counter it. In particular, the thesis analyzes the linguistic properties of hyperpartisan news — that is, extremely polarizing news content — and models them as effective proxies in computational linguistics. Current approaches to hyperpartisan detection lack systematic feature engineering, often reducing the task to a binary classification of content without providing interpretive insight into the underlying mechanisms. By introducing linguistic features that cover both the rhetorical and pragmatic dimensions of language, this work aims to shed light on how models leverage these properties, moving beyond surface-level classification toward a more explanatory framework. A further motivation concerns the availability of resources. The dominance of U.S.-based institutions in NLP research has led to a concentration of datasets and benchmarks in English, embedding socio-political biases that reflect the U.S. context rather than the European one. When such resources are used in pre-/post-training, fine-tuning, or inference, they risk misrepresenting the political and social landscape of European countries. This concern is particularly acute for Southern European languages, spoken in regions at heightened risk of polarization, where low-resource conditions remain a critical gap. The creation of multilingual corpora in these languages is therefore not only a methodological contribution but a necessary step toward more representative and reliable disinformation detection systems.

To support a fine-grained analysis of hyperpartisan language, the resources developed in

this thesis cover multiple textual units, ranging from headlines to paragraphs. Rather than relying on full articles for classification, the research focuses on smaller discourse segments, allowing for a more precise examination of the linguistic realizations of hyperpartisanship. While modern deep-learning architectures are capable of automatically extracting complex textual patterns, this thesis argues that linguistic analysis remains essential for ensuring transparency and interpretability. Integrating computational and linguistic approaches makes it possible to uncover the underlying properties of hyperpartisan language, providing insights that can both improve feature-based modeling and inform the development of novel algorithms capable of exploiting such features in a principled manner.

The datasets developed throughout this thesis reflect this methodological perspective. *PartisanLens* focuses exclusively on headlines, a design choice largely determined by the data collection pipeline based on *MediaCloud*[1]. The *Sentence-Level Italian Dataset* constitutes the preliminary study and methodological foundation for the subsequent Italian resources presented in this work. Its fine-grained annotation process enabled the refinement of the hyperpartisan language construct within a specific European political and linguistic context. Although the dataset focuses exclusively on a libertarian right-wing news source, this decision was motivated by the findings of Potthast et al. [2018b], who demonstrated that extremist sources from opposing sides of the political spectrum often exhibit similar stylistic characteristics associated with hyperpartisan communication.

Building upon this preliminary investigation, the *IHPP* dataset expanded the scope of analysis by incorporating both right- and left-wing sources, thereby providing a more balanced representation of the political spectrum. In addition, the unit of analysis was extended from individual sentences to short paragraphs containing one to three sentences. This shift enabled a stronger focus on the pragmatic dimensions of hyperpartisan language, which require longer syntactic and discourse structures to be adequately interpreted. The paragraph-level perspective also made it possible to investigate the internal composition of news articles, identifying the distribution and interaction of neutral and hyperpartisan content within the same text. The decision to restrict the analysis to short paragraphs was further motivated by technical constraints, as the autoencoder architectures employed in this study operated under limited context-window capacities for the identification of hyperpartisan and neutral segments.

Finally, *Par-ITA* was developed following the same methodological principles and research objectives as the previous datasets. In this case, however, the paragraph-level unit of analysis

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

was additionally motivated by the annotation process itself. The dataset required the manual inspection, validation, and revision of paragraphs rewritten by Large Language Models (LLMs), a task that demanded careful human evaluation. Restricting the context to paragraph-sized units ensured the feasibility and reliability of the annotation process while preserving sufficient contextual information for the linguistic analysis of hyperpartisan discourse.

1.2 The Hyperpartisan News Crisis in Europe

The proliferation of hyperpartisan news represents one of the most pressing challenges to democratic discourse in contemporary Europe. Hyperpartisan media outlets are characterized by extremely one-sided coverage, emotional framing, and the prioritization of ideological alignment over factual accuracy [Kiesel et al., 2019c, Potthast et al., 2018a]. Such news has flourished in the digital ecosystem, fundamentally altering the landscape of political communication. Unlike traditional partisan journalism, which typically maintained some commitment to factual reporting within an ideological framework, hyperpartisan content often blurs the boundaries between news, opinion, and propaganda [Benkler et al., 2018]. This phenomenon has intensified across European democracies, where diverse media systems and rich linguistic spaces create particularly complex challenges for identifying, monitoring, and mitigating its effects, especially in Southern Europe [Humprecht et al., 2020].

The consequences of hyperpartisan news extend beyond individual media outlets, affecting the very fabric of democratic society. Research demonstrates that exposure to hyperpartisan content correlates with increased political polarization [Levendusky, 2024], decreased trust in mainstream institutions [Fletcher et al., 2018], and reduced capacity for cross-partisan dialogue [Sunstein, 2018]. In Europe, these dynamics have manifested in contentious public debates surrounding climate change, migration policy, and responses to European policies (Euroscepticism), whereas hyperpartisan narratives have often dominated social media discourse and shaped public opinion in ways that transcend national boundaries [Fusiek et al., 2022]. The 2019 European Parliament elections, for instance, witnessed unprecedented levels of hyperpartisan content on social media platforms, with right-wing alternative media networks successfully mobilizing audiences through emotionally charged, identity-based messaging [Alonso-Muñoz and Casero-Ripollés, 2020, Bene et al., 2025].

Europe's particular vulnerability to hyperpartisan disinformation stems from several structural factors. First, the continent's linguistic diversity creates fragmented information environ-

ments where cross-border fact-checking and content moderation prove exceptionally difficult [Humprecht et al., 2020]. A hyperpartisan narrative emerging in Hungarian media, for example, may circulate extensively before being identified and debunked by fact-checkers operating primarily in English, French, or German. Second, European media systems exhibit considerable variation in their resilience to disinformation, with countries characterized by high levels of political parallelism (particularly in Southern and Eastern Europe) showing greater susceptibility to hyperpartisan influence than those with stronger professional journalism traditions [Hallin and Mancini, 2011, Brüggemann et al., 2014]. Third, the digital transformation has enabled alternative media outlets to bypass traditional gatekeepers, reaching audiences directly through social media platforms whose algorithmic curation often amplifies emotionally resonant, partisan content over balanced reporting [Holt et al., 2019].

The crisis of trust in established media institutions has created fertile ground for hyperpartisan actors. Across Europe, public confidence in news media has declined substantially over the past decade, with the 2021 Reuters Institute Digital News Report documenting that fewer than half of respondents in Western and Southern European countries rely on news most of the time [Newman et al., 2021]. This erosion of trust does not affect all outlets equally; while confidence in legacy media has fluctuated depending on the region, with a general 2% increase, certain alternative and hyperpartisan sources have gained credibility among specific audience segments, particularly those holding anti-establishment or populist attitudes [Aalberg et al., 2017], which go under the name of Euroscepticism. The Edelman Trust Barometer has consistently shown that traditional media rank among the least trusted institutions in several European democracies, trailing behind businesses, NGOs, and even government in some contexts [E., 2021]. This trust deficit creates what Wardle and Derakhshan [2017] term an ”information disorder”, constituted by three types of information in relation to harm and falseness: Mis-information, Dis-information, and Mal-information.

Particularly, the exposure to hyperpartisan content in Southern European countries is structurally determined as analyzed by Hallin and Mancini [2004b] and Brüggemann et al. [2014]. By defining this geographical area as *Mediterranean* or *Polarized Pluralist Model* in contrast with the Democratic Corporatist Model of the northern countries, the *Mediterranean* is characterized by low reach of the daily press, high political parallelism, weak professionalization, and strong state intervention in the media sector. As a result, the media sector developed as an extension and an institution of the political and literary worlds more than of the market. For instance, in Italy and in France, in the 20th century, we witness the rise of

several strong party presses (*L'Unità*, the main paper of the Italian Communist Party, *Il Popolo* of the Christian Democrats, l'*Avanti* of the Socialists) in parallel with the crystallization of the political democracy. These conditions created an environment predisposed to partisan information flows. Additionally, Humprecht et al. [2020] clustered the Southern countries like Italy, Spain and Portugal as polarized and based on low-resilience to online disinformation. It is worth noting that they define resilience as: "resilience refers to a structural context in which disinformation does not reach a large number of citizens". The principal subjects identified are populism and far-right movements, which have proven particularly effective in exploiting emotional manipulation and conspiratorial narratives: in Italy, Lega and Fratelli d'Italia; in Spain, Vox; and in Portugal, Chega, have all contributed to measurable increases in societal polarization through their digital communication strategies [Marino et al., 2024].

1.3 From Traditional to Digital Political Communication

1.3.1 The Common Ground of Hyperpartisan News

Disinformation, misinformation, and mal-information primarily occupy the digital sphere, directly impacting the quality of the media environment and shaping user perception. While these phenomena are not inherently new, they have proliferated alongside rapid technological advancement. Despite varying political outcomes across regions, the communicative language of hyperpartisanship reveals cross-border traits that transcend physical boundaries.

Because these dynamics exist within the digital environment—a political space defined by inherent power relationships—the influence of certain actors is disproportionate. Historically, this manifested as "Americanization" during the 1970s and 1980s [Hallin and Mancini, 2004a], a process of homogenization that shaped the global media ecosystem. Today, global social media platforms designed for U.S. markets create structural incentives for engagement-driven, emotionally charged content regardless of the local context [Chadwick, 2017].

Consequently, while European countries and the U.S. might be expected to manifest distinct methods of news expression and sharing, the digital nature of the internet often bypasses these contextual differences. Common denominators emerge in both topic and style. The content focuses on high-stakes, polarizing issues of public interest where voter beliefs inherently conflict. The linguistic framing used to report this news remains stylistically consistent across borders.

However, European hyperpartisanship is not merely an imitation of the American model.

European political communication is rooted in distinct traditions of political parallelism, public service broadcasting, and stringent state media regulation [Hallin and Mancini, 2011, Brüggemann et al., 2014]. This creates a heterogeneous landscape where hyperpartisan dynamics are shaped by local political cleavages, media ownership structures, and specific regulatory frameworks. For the purpose of this thesis, the focus remains specifically on the engineered textual features that define these shared communicative strategies.

1.3.2 The Evolution of Political Parallelism in European Media

European journalism developed within frameworks of political parallelism (i.d. the systematic alignment of media outlets with political parties or ideological tendencies) that contrasted with Anglo-American ideals of press independence [Hallin and Mancini, 2011, Seymour-Ure, 1974]. In Southern European countries (Italy, Spain, France), newspapers traditionally maintained explicit party affiliations, while Northern European democracies featured pillarization creating distinct media spheres aligned with religious or political communities [Hallin and Mancini, 2011].

From the mid-20th century onward, declining party membership, television's rise, and media commercialization drove a partial "Americanization" of European journalism toward professional autonomy and objectivity norms [Hallin and Mancini, 2004a]. Public service broadcasters institutionalized ideals of political balance, leading observers to anticipate convergence toward liberal media models [Blumler and Gurevitch, 2002].

This convergence proved illusory. Rather than eliminating political parallelism, recent transformations produced new forms of media-politics alignment [Selvik and Høigilt, 2021]. Contemporary parallelism operates through market segmentation, selective news consumption, and audience partisanship rather than formal party-press ties [Selvik and Høigilt, 2021]. The digital transformation fundamentally altered these mechanisms: where traditional parallelism involved stable institutional connections, contemporary digital parallelism functions through fluid, networked relationships based on ideological affinity [Chadwick, 2017]. Alternative media outlets cultivate relationships with political movements through shared ideology rather than organizational integration, enabling more extreme positions freed from traditional institutional constraints [Holt et al., 2019, Heft et al., 2020].

1.3.3 Internet, Populism, and the Transformation of Political Communication

Blumler and Gurevitch [2002] points out that the dissolution of institutions' role in providing an organized meaning, identity and authoritative information has been replaced by a more fragmented and accelerated society, leading to a loss of the government's influence where popular support is more contingent and effective communication is more vital.

The rise of digital platforms has coincided with a populist resurgence across Europe. Populism—understood as politics framed as a struggle between virtuous "people" and corrupt "elite" [Mudde and Kaltwasser, 2017, Hameleers and Goldberg, 2022, Bene et al., 2025]—finds effective expression through digital technologies that enable direct communication with supporters while bypassing traditional gatekeepers [Engesser et al., 2017]. Platform affordances, including algorithmic amplification of emotional content and metrics rewarding engagement, align closely with populist communication strategies [Gerbaudo, 2018].

European populist movements have proven adept at digital mobilization. Italy's Five Star Movement pioneered online grassroots organizing [Mosca and Quaranta, 2016], while Germany's AfD achieved extraordinary Facebook engagement despite limited organizational resources [Schroeder, 2019]. These movements adapted communication styles to platform logics, producing content designed for viral circulation [Hameleers et al., 2017].

Populist movements and hyperpartisan media operate symbiotically: populists generate demand for validating content while hyperpartisan outlets amplify populist messages and provide platforms for populist politicians [Waisbord, 2018, Moffitt, 2016]. This relationship proved particularly visible during COVID-19, when populist actors used social media to contest public health narratives while hyperpartisan outlets amplified these messages with conspiracy theories [Katsambekis, 2022, Bruns et al., 2020].

1.3.4 The Alternative Media Landscape in Europe

Alternative media outlets position themselves in opposition to mainstream journalism while claiming to provide systematically excluded perspectives [Holt et al., 2019]. They have proliferated across Europe, particularly on the far-right but also in the far-left, showing common traits [de León et al., 2024]. These range from relatively professional operations to small-scale blogs and social media pages [Nygaard, 2020].

Alternative media share characteristic features that distinguish them from traditional out-

lets. They adopt adversarial stances toward established institutions, accusing mainstream journalists of bias and censorship [Figenschou and Ihlebæk, 2019]. They blur boundaries between news, opinion, and activism, rejecting objectivity norms for explicit partisan advocacy [Atton, 2001]. They cultivate close relationships with political movements while maintaining nominal independence [Holt, 2019].

Alternative media content exhibits consistent patterns of inflammatory language, emotional intensification, and identity-based framing [KhosraviNik and Esposito, 2018, de León et al., 2024]. Immigration coverage employs threat narratives and dehumanizing language [Krzyżanowski, 2020], while political opponents are characterized as traitorous or malevolent [Hameleers and Schmuck, 2017]. The alternative media sector has emerged as a primary site for hyperpartisan production, combining ideological extremity with formats optimized for social media circulation [Heft et al., 2020, de León et al., 2024].

1.3.5 From Populism to Hyperpartisanship: The Radicalization Dynamic

The trajectory from populist politics to hyperpartisan media involves mutual radicalization in which political actors and outlets progressively adopt more extreme positions.

However, de León et al. [2024] distinguishes hyperpartisan media and alternative media, since the latter is not necessarily partisan. Populist communication strategies mirror hyperpartisan media characteristics [Hameleers and Schmuck, 2017]. As populist parties normalize these styles, media outlets serving populist audiences intensify partisan rhetoric to maintain engagement [Aalberg et al., 2017].

This radicalization operates through several mechanisms. Algorithmic amplification rewards emotionally charged content [Tucker et al., 2018], while audience polarization creates feedback loops reducing exposure to moderating influences [Prior, 2007]. Market fragmentation enables niche outlets to survive by serving highly partisan audiences, eliminating competitive pressures toward moderation [Stroud, 2011].

The result has been progressive blurring between partisan journalism and propaganda. Even in countries with stronger journalistic traditions, alternative media ecosystems have camouflaged extreme content with a normalized appearance while spreading extremist content[Heft et al., 2020]. This fusion of populist politics with hyperpartisan media creates information environments where democratic deliberation becomes difficult: opponents are enemies, factual disagreements become identity markers, and compromise appears as betrayal [Mudde and Kaltwasser, 2017].

1.3.6 Defining Hyperpartisan as a per-se phenomenon: disentangling differences with Fake News and similar research fields

Despite the described overlaps and links between Hyperpartisan News and Populism, we need to clarify the specific nature of this task, disentangling it from other similar ones in the Natural Language Processing field. Hence, we are going to integrate the preliminary definition of Hyperpartisan News in contrast with its parent phenomena. Wardle and Derakhshan [2017] defines the characteristics of Mal-/Mis-/Dis-Information depending on the intent and the epistemic status of the communication, which could be Harmful or False.

Misinformation is predominantly constituted by false connections and misleading content, while Malinformation by leaks, harassment and hate speech. The overlapping area between these categories forms Disinformation, predominantly spreading false context, imposter, manipulated and fabricated content.

For instance, a source could spread fake news containing some instances of hyperpartisanship, like hate speech and/or misleading and manipulated content. Nonetheless, it is possible to find online content which is extremely partisan without relying on false information.

Indeed, hyperpartisan news detection often overlaps or is confused with other disinformation tasks, such as fake news detection [Garg and Sharma, 2022, Ross et al., 2021, Mourão and Robertson, 2019, Sousa-Silva, 2022, Nguyen et al., 2019, Agerri, 2019], and stance detection [Bourgonje et al., 2017].

Specifically, hyperpartisanship might be conveyed through elements of fake news, aimed at propagating a specific agenda and manipulating readers to adopt a particular position on a given topic [Agerri, 2019] and with the adoption of populist rhetoric leveraging hate speech and conspiracy theories, for instance on immigration policies [Maggini et al., 2026].

This news presents specific linguistic cues. Considering the lexicon, articles exhibit a high count of adjectives and adverbs [Pérez-Almendros et al., 2019, Dumitru and Rebedea, 2019], depicting political adversaries and specific agenda settings against their moral values; extensive use of pronouns, and words of disgust [Knauth, 2019] because they often reflect subjective point of view while deviating from professional journalistic norms and code of practices.

Furthermore, concerning their structure, longer paragraphs written in a sensationalist style compose the bodies, full of emotional language and rare terms [Sengupta and Pedersen, 2019], or even the creation of an ad-hoc vocabulary [Walker et al., 2025]. Also, hyperpartisan news articles display hyper-polarized linguistic traits in their titles, and they are used to immediately

catch the reader's attention [Ross et al., 2021]. The polarized content usually focuses on the first and last paragraphs, reflecting the best practices to generate high engagement while repeating the content of the title and concluding the article with a call to action [Maggini et al., 2025c]. Although all political leanings can produce such content, right-wing media, in particular, often employ hyperpartisan headlines [Lyu et al., 2024a, Amason et al., 2019].

Definitions and Analogue biases

Hyperpartisan news detection is a task in which certain textual features indicated above suggest that the writer is expressing an extremist, one-sided opinion. Moreover, various degrees with which typologies of bias occur contribute to make the text hyperpartisan. Firstly, by gathering the definitions of the phenomenon of our interest - when explicited by the authors -, it is notable that they share common traits (e.g. this kind of news refer to one-sided political agenda, and is the primary cause of polarization), but also diverge spanning similar domains This information is synthesized in Table 1.1, where the numeric codes in the **Char** columns refer to the following feature taxonomy.

Table 1.1: Definitions of hyperpartisanship in the selected papers.

Reference	Claim	Char
Barnidge and Peacock [2019]	Social media also have facilitated the rise of hyperpartisan news. Hyperpartisan news: (1) has an obviously one-sided political agenda, which makes no effort to balance opposing views; (2) pushes anti-system messages that are critical of both mainstream media and establishment politics, often relying on misinformation to do so; and (3) relies heavily on social media as a platform for dissemination. Thus, hyperpartisan news can be situated squarely at the intersection of partisan and alternative news, and considerable overlap exists between hyperpartisan news and "fake" news.	1, 6, 7
Kiesel et al. [2019d]	Hyperpartisan articles mimic the form of regular news articles but are one-sided in the sense that opposing views are either ignored or fiercely attacked.	1, 2, 3, 4, 5, 7, 8

Continued on next page

Table 1.1 – continued from previous page

Reference	**Claim**	**Char**
Potthast et al. [2018b]	Prone to misunderstanding and misuse, the term "fake news" arose from the observation that, in social media, a certain kind of "news" spreads much more successfully than others, and this kind of "news" is typically extremely one-sided (hyperpartisan), inflammatory, emotional, and often riddled with untruths.	1, 3, 4, 5, 7, 8
Lyu et al. [2024a]	We think that a better understanding of hyperpartisanship can be achieved by considering not only (1) the news that contains one-sided opinions but also (2) the news that describes conflicts and the underlying politically polarized climate because both of them could lead to an increase in the public's perceived polarization (Yang et al. 2016; Fiorina, Abrams, and Pope 2005; Levendusky and Malhotra 2016). Additionally, coverage quantity itself can be considered as a particular form of bias (Lin, Bagrow, and Lazer 2011). In particular, we seek to extend previous studies' definitions of hyperpartisan news to include news that covers partisan conflicts and confrontations.	1, 3, 4, 7
Gangula et al. [2019]	Media bias can be observed and defined through various factors. In the political domain, it ranges from selectively publishing articles to specifically choosing to highlight some events, parties and leaders. We also come across articles where bias can be detected by observing the unclear assumptions, loaded language, or lack of proper context.	1, 6, 7, 8

Continued on next page

Table 1.1 – continued from previous page

Reference	Claim	Char
Pierri et al. [2020]	Researchers use different terms to indicate the same issue, namely disinformation, misinformation, propaganda, junk news and click-bait. In this work we use the word disinformation, rather than the more popular "fake news", to refer to a variety of low credibility content which comprises false news intended to harm, misleading and non-factual reporting, hyperpartisan news and propaganda, and unverified rumours	1, 4, 7
Huang and Lee [2019]	Hyperpartisan news is news riddled with untruth and twisted statements of information. This type of news spread more successfully than others. Hyperpartisan news not only can mislead readers but also cause polarisation within a community or society.	1, 4, 7
University Politehnica of Bucharest et al. [2021]	Although there are more subcategories related to fake news such as satire, parody or clickbait, a general definition of the term could be the following: fake news represents a way to spread false information to mislead the public, damage the reputation of an entity or have a political or financial gain. The idea of misleading and influencing the public is also linked to the notion of hyperpartisan news that has the role of presenting extremist or conspiratorial opinions with intentional misconceptions.	2, 3, 4, 5, 7
Garg and Sharma [2022]	The aim is to classify news as real or fake. Fake news is news that is intentionally generated to misguide people. It may exist in various forms, including misleading content, biased news, satirical news, rumours, hyperpartisan news, deceptive news, disinformation, clickbait, and hoax.	1, 4, 7

However, considering the literature analyzed, there are several taxonomies proposals for

junk news like Zannettou et al. [2019], Kapantai et al. [2021]. We will use the bias categories collected by Oxford[2] and Rodrigo-Ginés et al. [2024] to discuss the founding biases of hyperpartisan articles.

Spin bias, or *rhetoric bias* [Rodrigo-Ginés et al., 2024], strictly concerns the linguistic structure of the article, its persuasion. The deliberate or inadvertent misrepresentation of research outcomes, leading to unjustified indications of positive or negative results, potentially could result in misleading conclusions. Written language is the product of the conscious application of strategic discursive and persuasive patterns to interest the readers. The words contribute to giving a particular meaning to the entire text, especially if they leverage an emotional lexicon with superlatives.

Ad hominem bias is a rhetorical strategy in which one moves away from the topic of the controversy by contesting not the statement of the interlocutor, but the interlocutor themselves and his personal characteristics or traits [Walton, 1998]. This rhetorical strategy was frequently used in sophistry and is still widely used today in political discussions and journalistic controversies.

Presence bias or opinion statement involves the inclusion of subjective opinions within news articles, influencing readers' perceptions. It occurs when factual reporting is mingled with subjective viewpoints or opinions [Tran, 2020]. In other words, it reflects the degree of agreement and statement sharing of an entity, i.e. users or publishers [Anand et al., 2007].

Ideological bias occurs when news reporting or content is influenced by a particular ideological stance or viewpoint, impacting the presentation and selection of news topics. Ideological detection is different from political bias because some ideologies can be shared even by opposite parties. Ideologies often contrast each other, but to be classified they need this comparison [Sharma et al., 2020].

Framing bias involves presenting information to shape or influence people's perceptions of an issue or event by emphasizing certain aspects while downplaying others [Baumer et al., 2015, Roy and Goldwasser, 2020]. In this case, using linguistics and rhetorical figures helps the author partially present the selected information. Therefore, framing expresses a publishers leaning towards an ideology. Frames are tools that emphasize specific information while potentially favoring one aspect over another, with or without being slanted [Kong et al., 2018]. It is performed in moral content and style used [Xu et al., 2020].

Coverage bias, is not present in Table 1.2 since it is not a textual bias. It refers to the

[2] `https://catalogofbias.org/biases/spin-bias/`

disproportionate attention or neglect of topics or events in news reporting, leading to an imbalance in coverage across different subjects [Leeson and Coyne, 2011].

Political bias could be easily confused with ideological bias. Since a party is a combination of both an ideology and a political leaning, this bias is related to the inclination of news media or information sources or people to favor one political party's agenda [Honeycutt and Jussim, 2023].

Importantly, these biases are not binary: a text may exhibit one or more of them to a moderate degree without crossing into hyperpartisan territory. Table 1.2 illustrates this asymmetry — hyperpartisan language necessarily involves bias, but biased language does not necessarily imply hyperpartisanship." In that regard, table 1.2 provides examples showing the difference between biased text and hyperpartisan language, stressing the hypothesis for which a sentence could be biased but not necessarily implying a hyperpartisan framing, while vice versa is not true.

Table 1.2: Examples of statements for specific biases and hyperpartisan statements for that bias.

Bias	**Biased Example**	**Hyperpartisan Example**
Spin bias	How can we trust their solutions when they fail to understand the basic principles of economics?	Their economic policies are a disaster, proving yet again their ignorance and incompetence.
Ad hominem	The proposal is flawed because it comes from someone who has no experience in the field.	We can't expect anything good from someone who has never done anything worthwhile!
Opinion statement bias	It's clear that this is the best approach for our society.	This is unequivocally the only right path forward for our nation; anyone who disagrees is simply blind to the truth.
Ideology bias	Socialist policies always lead to inefficiency and economic downfall.	The leftist agenda destroys economies and personal freedoms every time it's implemented.
Framing bias	The data support that this policy will decrease crime rates.	The indisputable facts confirm that only this policy can save us from spiralling into a crime-infested nightmare.
Political bias	This party's proposal will bankrupt the nation.	The opposition's plan is a surefire way to plunge our country into insurmountable debt and chaos.

1.4 The regulation of Disinformation in Europe

Once we have defined the domain of hyperpartisan news, in order to have a full picture of the phenomenon, we should present what the European Commission has introduced to legally tackle it. Hyperpartisan news can be harmful due to its content against minorities, the involvement and the influence on voting processes with consequences on the democratic fabric. Especially, with the spread of LLM-Chatbots (e.g. ChatGPT) the rapidity and the amount of such algorithmic curated content have increased [Naja Bentzen, 2025].

As a result, to regulate the practices of disinformation, three different regulations have been introduced since 2018 by the EU Commission: the EU's Digital Services ACT (DSA) [European Parliament and Council, 2020], AI Act [Commission, 2024b], and the Code of Practice on Disinformation [Commission, 2022].

First of all, the EU AI Act was formally adopted in 2024, representing the regulatory framework against the opportunities and risks posed by AI systems. It introduces a risk-based approach to AI governance while classifying the developed systems into four tiers: unacceptable risk (prohibited), high-risk, limited risk, and minimal risk. The prohibited adoption of AI concerns the deployment of subliminal techniques and manipulative systems exploiting minority groups' vulnerabilities, Article 5(1). Considering disinformation, this prohibition could put guardrails on disinformation campaigns covering minors or marginalized groups. This framing of disinformation has consequences in its regulation, because political deepfakes or emotionally charged narratives aimed at a broad audience could fall outside the scope of this prohibition [Ó Fathaigh et al., 2025]. However, AI Act takes care of the users' welfare with Article 50(4), which introduces watermarks as a requirement to indicate AI origin while respecting the freedom of expression and artistic forms [Gavriil and Pavlidis, 2025]. While being applied to both providers and deployers, the AI Act does not explicitly target disinformation, and most of the obligations fall on providers, with limited regulations on the deployers' side [Gavriil and Pavlidis, 2025]. In any case, the providers must respect the transparency principles when providing AI-generated content (visual, audio).

In this matter, the DSA, whose adoption and application started in 2024, gains particular significance. Whereas the AI Act describes the regulation of practices about the design, development, and deployment of AI systems, focusing on those classified as high-risk, the DSA mostly concerns how content is distributed and moderated across digital platforms. Indeed, it represents the main EU legal instrument to "protect European citizens from targeted disinformation" [Commission, 2024a]. In the context of the EU regulatory approach, the DSA

addresses the amplification and circulation of content in online spaces, while the AI Act governs how such content might be synthetically generated. Under the DSA, Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs), namely those with more than 45 million users in the EU, are subject to specific due diligence obligations, including mandatory risk assessments, algorithmic transparency, data access for vetted researchers, and independent audits, especially in relation to systemic risks. Amongst the included risks are disinformation (Recital 83) and algorithmic manipulation. Articles 34 and 35 focus on threats to elections, vulnerable groups, and systemic harms enabled by proprietary algorithms. The text argues that the DSA's mandatory, recurring risk assessments for Very Large Online Platforms (VLOPs) demonstrate the urgent need to impose similar obligations on AI developers. As generative AI models grow exponentially more powerful, developers must thoroughly document and proactively evaluate the risks of creating addictive, biased, or manipulative systems. Purely voluntary self-assessments are no longer sufficient in the face of looming threats such as AI-driven fake accounts, spam, phishing, and massive disinformation campaigns. The AI Act complements the DSA by strengthening transparency and audit requirements for synthetic content. Together, the two regulations underscore the value of sustained external audits to oversee risks like generative models fueling "information disorders" at scale [Nannini et al., 2025].

In addition to the DSA, the Code of Practice on Disinformation [Commission, 2022] constitutes the specific instrument with the purpose of combating online disinformation. Established in 2018, the Code of Practice was significantly strengthened in 2022, with the aim of becoming recognised as a Code of Conduct under the Digital Services Act (DSA). In February 2025, the Commission and the European Board for Digital Services endorsed the integration of the Code into the framework of the DSA. The Code brings together a wide range of signatories, including online platforms and industry associations, as well as advertisers, fact-checkers, civil society organisations, and researchers. The Code contains numerous commitments and measures, covering various areas, from the demonetisation of disinformation and the transparency of political advertising to the empowerment of users and cooperation with fact-checkers [Ó Fathaigh et al., 2025].

1.5 Research Positioning and Thesis Structure

In this paragraph, we are providing the overall structure of the rest of the manuscript.

Hypothesis and objectives summarize the hypothesis guiding our analysis as research questions. There, they have been clustered with respect to the main areas where we have contributed to the investigation. General Methodology describes the adopted processes for the different experiments and the literature review, and it is organized by topic. Results and Discussion contains the results of the principal experiments and their analysis, as well as a meaningful review of the main findings. General Conclusion wraps the main findings together, representing an attempt to find a narrative fitting our work, and summarizes the main contributions that we have shared during this period.

CHAPTER 2

HYPOTHESIS AND OBJECTIVES

This chapter articulates the guiding hypotheses, research questions, and objectives that structure the experimental work presented in this thesis. The hypotheses emerge from the theoretical foundations established in the introduction and literature review, while the research questions operationalize these hypotheses into empirically testable inquiries. The objectives, organized by the publications that constitute the core contributions of this dissertation, translate these questions into concrete methodological and analytical goals.

The overarching hypothesis of this thesis posits that hyperpartisan language constitutes a distinct register within political journalism, characterized by identifiable linguistic, pragmatic, and rhetorical features that transcend national and linguistic boundaries. While hyperpartisan content adopts the structural conventions of standard news reporting, it systematically deploys specific communicative strategies that distinguish it from balanced journalism. The digital environment in which this content circulates facilitates cross-border dissemination, suggesting a degree of stylistic convergence that enables multilingual and comparative investigation.

From this foundational premise, we derive specific hypotheses organized around three thematic areas: data and annotation practices, learning strategies and model architectures, and linguistic analysis.

2.1 Data and Annotation

The NLP community has predominantly focused on English-language resources for hyperpartisan detection, leaving mid- and low-resource European languages substantially understudied. We hypothesize that this imbalance has limited our understanding of how hyperpartisan

rhetoric manifests across different media systems and political cultures. Consequently, the development of multilingual datasets covering Romance languages such as Italian, Spanish, and Portuguese represents a necessary step toward more generalizable detection methods.

Regarding annotation methodology, we posit that fine-grained linguistic annotation enriches hyperpartisan detection beyond binary classification. Specifically, marking semantic-pragmatic markers such as figurative speech, irony, loaded language, and hyperbolic expressions during human annotation should improve model performance when these features are embedded during fine-tuning. Furthermore, we hypothesize that human annotators maintain advantages over Large Language Models in detecting complex pragmatic phenomena, particularly irony and sarcasm, where contextual understanding proves essential.

The potential of LLMs as annotation tools nevertheless warrants investigation. We hypothesize that LLMs can approximate human annotation patterns when provided with sociodemographic information reflecting annotator backgrounds, though their reliability remains contingent on task complexity and cultural specificity. For data augmentation purposes, LLMs may effectively assist human annotators in generating controlled variations of hyperpartisan content while preserving the original bias characteristics.

2.1.1 Learning Strategies and Architectures

Model performance on hyperpartisan and related misinformation tasks depends critically on architecture, parameter scale, and pre-training data composition. We hypothesize that In-Context Learning strategies, while offering flexibility, generally underperform compared to Supervised Fine-Tuning, which adapts model weights to task-specific distributions. Within ICL approaches, we expect rule-based methods such as codebooks to outperform Chain-of-Thought prompting for classification tasks, as explicit classification criteria provide clearer decision boundaries than step-by-step reasoning processes.

The effectiveness of different learning paradigms likely varies across task types. Encoder-only architectures may excel at linguistically-oriented tasks such as hyperpartisan and harmful content detection, where bidirectional attention captures nuanced stylistic features. Decoder-only models, conversely, may demonstrate advantages in tasks requiring world knowledge, such as fake news verification and political bias identification.

For text generation tasks such as neutralization, we hypothesize that Direct Preference Optimization can enhance output quality beyond standard fine-tuning by explicitly optimizing for neutrality while maintaining semantic fidelity.

2.1.2 Linguistic Analysis

Hyperpartisan content systematically violates Gricean conversational maxims through rhetorical strategies that prioritize persuasion over information. We hypothesize that specific rhetorical biases, including loaded language, appeals to fear, and ad hominem attacks, show statistically significant associations with hyperpartisan classification. Moreover, the distribution of hyperpartisan content within articles follows predictable patterns, with higher concentrations in opening paragraphs designed to capture reader attention and concluding sections that reinforce partisan framings.

The pragmatic dimension of hyperpartisan language extends beyond rhetorical techniques to encompass figurative language, irony, and neologisms that signal in-group membership and reinforce polarized worldviews. We hypothesize that models attending to these linguistic traits achieve better explainability and that explainability techniques such as Integrated Gradients and SHAP can reveal how neural networks allocate attention to pragmatic markers during classification.

2.2 Research Questions

The hypotheses enumerated above translate into eight research questions that guide the empirical investigations across the publications constituting this thesis.

The research questions are structured into four interconnected thematic clusters. The initial phase (RQ1–RQ4) establishes the theoretical and technical foundations by defining the linguistic boundaries of hyperpartisanship and systematically evaluating model architectures, adaptation paradigms, and prompting strategies across diverse linguistic contexts. Building upon these technical benchmarks, the second cluster (RQ5–RQ6) shifts to the socio-technical dimension, examining how sociodemographic biases influence model behavior and assessing the viability of LLMs as substitutes for human expertise in subjective annotation.

A specialized linguistic and architectural analysis follows (RQ7–RQ8), which isolates specific semantic-pragmatic markers and investigates how neural models internalize these features during the learning process. Finally, these insights converge in an applied and ethical framework (RQ9), which explores the potential for models to not only detect but also actively neutralize hyperpartisan bias while maintaining the semantic integrity of the original discourse.

RQ1: *Is it possible to define hyperpartisan language, disentangling it from related tasks (e.g. misinformation)? If so, what are its linguistic and rhetorical properties? Which one in*

the literature resulted in being efficient for its detection?

This question is motivated by the absence of a shared definition of hyperpartisanship in NLP [Maggini et al., 2025e] and by evidence that hyperpartisan content is stylistically distinct from both mainstream journalism and outright fake news [Potthast et al., 2018b]. Scholars have noted that hyperpartisan outlets systematically violate journalistic norms of balance [Barkho, 2021] and deploy specific rhetorical strategies that distinguish them from mainstream reporting [Bassi et al., 2024]. The SemEval-2019 shared task [Kiesel et al., 2019d] provided the first systematic benchmark for operationalizing the concept computationally, yet the boundary with propaganda [Da San Martino et al., 2019] and political bias [Rodrigo-Ginés et al., 2024] remains contested.

RQ2: *How do different LLM adaptation paradigms, specifically Fine-Tuning versus In-Context Learning, compare for hyperpartisan, fake news, harmful content, and political bias detection, considering variations in model architecture, parameter scale, and pre-training data?*

The rise of instruction-tuned decoder-only models [Brown et al., 2020, Ouyang et al., 2022] has renewed interest in whether In-Context Learning (ICL) can match supervised fine-tuning [Devlin et al., 2019, Liu, 2019] for classification tasks. Prior work has shown that fine-tuned encoders remain competitive or superior across a range of text classification benchmarks [Edwards and Camacho-Collados, 2024], yet the relative strengths of encoder-only versus decoder-only architectures for misinformation-related tasks have not been systematically compared [Aman, 2024, Chen and Shu, 2024].

RQ3: *What is the impact of various In-Context Learning strategies, including zero-shot, few-shot, codebook-based, and Chain-of-Thought prompting, on classification performance and stability across misinformation detection tasks?*

Lu et al. [2022] demonstrated few-shot prompting being sensitive to the example selection and ordering. Determinantal Point Process (DPP) methods offer a principled approach to selecting diverse and representative few-shot examples [Yang et al., 2023b, Wang et al., 2024b], limiting the introduction of noise in the prompt and could optimize their ordering. Chain-of-Thought [Wei et al., 2022], defined as a series of intermediate reasoning steps, has shown mixed results on different tasks, eliciting zero-shot latent reasoning capabilities of LLMs [Kojima et al., 2022]. On the other hand, codebook-based prompting, which provides explicit decision criteria following a rule-based approach, has been proposed as a more principled alternative for political science concepts [Halterman and Keith, 2025] and shows promise in zero-shot

political relation classification [Hu et al., 2024]. However, the comparative effectiveness of these strategies for hyperpartisan and misinformation detection remains an open question [Maggini and Gamallo Otero, 2024].

RQ4: *How do model performance and optimal detection strategies vary across different languages, particularly for mid- and low-resource European languages?*

The NLP community has predominantly produced literature and created resources on hyperpartisan news for English [Maggini et al., 2025e], leaving other families of languages understudied. The SemEval-2023 multilingual propaganda task [Piskorski et al., 2023b] demonstrated substantial cross-lingual performance gaps, while stance detection research on immigration in lower-resource languages has highlighted the specific challenges posed by domain and cultural specificity [Mets et al., 2024].

RQ5: *To what extent do sociodemographic and ideological profiles influence LLM behavior when detecting hyperpartisan content in multilingual contexts?* During the annotation process demographics and cultural background of annotators influence the labels they assign in text annotation [Schäfer et al., 2025]. In NLP, there is a branch called perspectivism, which leverage data annotated by different individuals in order to model varied perspectives that influence their opinions and world view, and it revealed to be useful in subjective tasks to capture the variance of the annotations [Frenda et al., 2025]. With this regard, for instance, Sarumi et al. [2025] showed that LLMs selectively use demographic attributes from the personas. Beck et al. [2024] explored sociodemographic prompting as a method for steering LLM outputs toward specific annotator perspectives though its robustness and reliability vary substantially across models, tasks, and languages [Mukherjee et al., 2024]. Persona-based prompting has emerged as a related strategy to inject diversity into LLM-generated annotations [Fröhling et al., 2025], but the degree to which ideological conditioning affects outputs specifically in multilingual hyperpartisan contexts has not been studied.

RQ6: *Can LLMs effectively replace or augment human annotators for subjective annotation tasks in political discourse analysis?*

This question is rooted in the emerging practice of adopting LLMs as annotators, given their reasoning capabilities. In this field, the results are contradictory. In the political domain, Heseltine and von Hohenberg [2024], Törnberg [2024] tested GPT-4 for this purpose, finding out that despite being cost-efficient and reducing the annotation time [Horych et al., 2025], the task complexity undermines its efficacy and reliability [Ziems et al., 2024], representativeness, bias, sensitivity to prompt variations and English language preference [Pavlovic and Poesio,

2024, Plank, 2022].

RQ7: *What linguistic and rhetorical features define hyperpartisan language, especially in Southern Europe, and how do semantic-pragmatic markers compare with logical fallacies as discriminative features?*

Because of Hyperpartisan content correlate with extreme political communication [Buarque, 2023] and relies on strategic communication based on logical and content short circuit consisting in adopting rhetorical techniques like loaded language and ad hominem attacks [Walton, 1998], it violates Gricean maxims. In terms of pragmatics, figurative language [Lakoff and Johnson, 2008], irony and hyperbole [Scott, 2022] are also constituents of the speech act and form a further layer of pragmatic complexity that standard classification approaches often fail to capture [Chakrabarty et al., 2022]. Their contribution in shaping hyperpartisan news remains uncovered. Rather, a dominant pragmatic approach focused on logical fallacies. They have been proposed as an orthogonal taxonomy of manipulative argumentation [Piskorski et al., 2023c, Sourati et al., 2023], but their relationship to hyperpartisanship, particularly in the Italian media context, has not been empirically examined [Maggini et al., 2025b].

RQ8: *How do neural models perceive and process semantic-pragmatic features, and do these features contribute meaningfully to hyperpartisan detection when embedded during training?*

Explainability techniques such as Integrated Gradients [Sundararajan et al., 2017] and SHAP [Lundberg and Lee, 2017] have been applied to understand how transformer models allocate attention to specific linguistic spans during classification [Aghababaei et al., 2025, Atanasova, 2024], with evidence that annotated linguistic features can guide model attention toward more interpretable decision regions [Szczepański et al., 2021]. Their application to the hyperpartisan domain remains limited [Bangerter et al., 2023b]. Whether embedding fine-grained pragmatic annotations, such as span-level markers of irony, loaded language, or hyperbole, during fine-tuning improves both performance and explainability for hyperpartisan detection remains an open empirical question [Lyu et al., 2024b].

RQ9: *Which textual features and training strategies prove most effective for neutralizing hyperpartisan language while preserving semantic content?*

Text style transfer and detoxification research has demonstrated that parallel corpora are essential for training effective neutralization models [Logacheva et al., 2022], and that supervised fine-tuning on parallel data consistently outperforms unsupervised approaches [Dale et al., 2021]. Direct Preference Optimization [Rafailov et al., 2024] has emerged as a promising

method for aligning generative models with human preferences, extending beyond standard fine-tuning by explicitly contrasting preferred and dispreferred outputs. However, the application of these methods to the specific case of hyperpartisan language neutralizationand the automated evaluation of neutralization quality remain largely unexplored [Pryzant et al., 2020, Toshevska and Gievska, 2025] and needs a careful approach to avoid biasing the hyperpartisan source (e.g. completely changing the wording).

2.3 Objectives and Contributions

The research questions are addressed through five interconnected studies, each contributing novel resources, methodological advances, or empirical insights to the field. The objectives and contributions are organized by publication.

2.3.1 Systematic Review of Hyperpartisan News Detection

In the work *A systematic review of automated hyperpartisan news detection*, we presented the first literature review on hyperpartisan news detection. Its primary objective constituted the core of our research, that is, gathering and systematizing all the works done so far to find possible gaps as a starting point to develop new research lines. This paper helped us grasp a comprehensive understanding of the field and, from this knowledge, lay the foundations of the research questions. Using the PRISMA methodology, we reviewed and systematized approaches and datasets from 81 articles published from January 2015 to 2024. Our analysis includes differentiating hyperpartisan news detection from similar tasks, identifying text sources, labeling methods, and evaluating models. Thus, the contributions of this literature review lie in providing guidance to new and old researchers in the misinformation domain, especially by providing the datasets used across the literature, and highlighting several gaps that could serve as a stimulus for future research lines.

2.3.2 Benchmarking Learning Paradigms for Misinformation Detection

The following study, *Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning*, provides a comprehensive evaluation of LLM capabilities for hyperpartisan, fake news, harmful tweets, and political bias detection across ten datasets and five languages (English, Spanish, Portuguese, Arabic, and Bulgarian). This work addresses RQ1, RQ2, and RQ3 through a systematic comparison

of Fine-Tuning and In-Context Learning approaches. Furthermore, it serves as a preliminary model analysis to investigate the potential of these architectures in future tasks.

The primary objective is to evaluate whether In-Context Learning with state-of-the-art LLMs can match or exceed the performance of fine-tuned smaller models. Secondary objectives include assessing the effectiveness of different prompting strategies (zero-shot, few-shot with random and DPP-selected examples, codebook-based, and Chain-of-Thought) and examining how optimal strategies vary across languages and task types.

The contributions are threefold: (i) the first comprehensive benchmark comparing encoder-only and decoder-only architectures across multiple misinformation detection tasks; (ii) empirical evidence that fine-tuning consistently outperforms In-Context Learning, with fine-tuned encoders excelling at linguistically-oriented tasks and fine-tuned decoders performing better on knowledge-intensive tasks; (iii) demonstration that codebook-based prompting generally outperforms Chain-of-Thought for classification; and (iv) state-of-the-art results on four benchmark datasets.

2.3.3 Multilingual Hyperpartisan and Conspiracy Detection in European Media

The second study, *PartisanLens: A Multilingual Dataset of Hyperpartisan and Conspiratorial Immigration Narratives in European Media* introduces PartisanLens, a multilingual dataset of 1,617 hyperpartisan news headlines on immigration in Spanish, Italian, and Portuguese, addressing RQ3, RQ4, and RQ5. The dataset uniquely annotates hyperpartisan content alongside Population Replacement Conspiracy Theories (PRCT), stance toward immigration, and rhetorical biases.

The objectives span across releasing the first multilingual European dataset connecting hyperpartisanship with conspiratorial narratives; evaluating LLM classification performance and establishing baselines for these interconnected tasks; and investigating whether LLMs can emulate human annotation patterns when conditioned on sociodemographic and ideological profiles.

The contributions of the paper are the following: (i) PartisanLens as a novel resource filling the gap in non-English hyperpartisan detection research; (ii) a joint annotation scheme capturing the multidimensional nature of political discourse; and (iii) experimental evidence on LLM-based annotation, including a persona simulation methodology that reveals how ideological conditioning affects model outputs.

2.3.4 Italian Hyperpartisan Detection with Rhetorical Bias Taxonomy

The third study, *Detecting Hyperpartisanship and Rhetorical Bias in Climate Journalism: A Sentence-Level Italian Dataset* presents the first Italian dataset for joint hyperpartisan and rhetorical bias detection in climate change discourse, directly addressing RQ6. The corpus comprises 48 articles (1,010 sentences) annotated at the sentence level for binary hyperpartisan classification and 17 rhetorical bias categories.

The objectives are to create a resource enabling fine-grained analysis of how hyperpartisan content correlates with specific rhetorical techniques; to establish baselines using both encoder models and GPT-4 variants; and to analyze the distribution of rhetorical biases across hyperpartisan versus neutral content.

This work contributed in (i) releasing an italian dataset simultaneously modeling hyperpartisan language and rhetorical biases on polarizing topics (e.g. climate change); (ii) an empirical validation that certain rhetorical techniques, particularly loaded language, appeals to fear, and name-calling, correlate significantly with hyperpartisan classification; and (iii) baseline performance metrics demonstrating that GPT-4o achieves strong zero-shot performance on hyperpartisan detection while struggling with fine-grained rhetorical bias classification.

2.3.5 Pragmatic Features for Italian Hyperpartisan Detection

The fourth study, *IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News* introduces IHPP, a paragraph-level dataset of 356 Italian articles (4,861 paragraphs) enriched with span-level annotations for semantic-pragmatic linguistic traits: figurative speech, irony/sarcasm, hyperbolic language, loaded language, epithets, and neologisms. This work addresses RQ6 and RQ7 while providing insights on how the Encoder-only models embed and leverage the linguistic features during the classification task.

In this work, we investigate whether pragmatic features provide discriminative power for hyperpartisan detection; we assess the contribution of individual linguistic traits through ablation studies; and we analyze how models allocate attention to pragmatic markers using explainability techniques.

We contributed to the NLP community by releasing (i) IHPP, the largest Italian hyperpartisan dataset with linguistic pragmatic annotations; (ii) this work provides empirical demonstration that embedding linguistic trait information during fine-tuning improves classification performance by approximately 15 percentage points; (iii) additionaly, the ablation

analysis reveals that loaded language provides the strongest discriminative signal while epithets and neologisms show redundancy with co-occurring traits; and (iv) Integrated Gradients and SHAP analyses show that models do attend preferentially to annotated pragmatic spans when classifying hyperpartisan content.

2.3.6 Italian Hyperpartisan Neutralization

The fifth study, *Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization* presents Par-ITA, the first human-supervised parallel corpus for Italian hyperpartisan neutralization, comprising 2,475 paragraph pairs, addressing RQ5 and RQ8.

In this case, we developed a rigorous methodology for creating high-quality parallel data through humans-in-the-loop. The pipeline consisted of human supervision at different steps and applying distinct score scales to LLM textual generations. We benchmarked neutralization performance across seq2seq and decoder-only architectures using standard fine-tuning, Direct Preference Optimization, and In-Context Learning; furthermore, GPT-4o-mini has been evaluated as an automated judge for neutralization quality, comparing its judgements with those given by humans.

The contributions encompass: (i) Par-ITA as a foundational resource for non-English text neutralization research; (ii) a three-stage annotation pipeline (preliminary LLM selection, human-supervised generation, post-hoc validation) applicable to other subjective NLP tasks; (iii) evidence that DPO effectively maximizes neutrality scores for seq2seq models; and (iv) analysis revealing systematic biases in automated evaluation, with GPT-4o-mini overpenalizing content on sensitive political topics compared to human judgments.

CHAPTER 3

METHODOLOGY

This chapter presents the methodological framework underlying our research, organized around five complementary papers that together form a comprehensive investigation into hyperpartisan language detection and its linguistic underpinnings. Following a funneling architecture, each paper progressively narrows focus from foundational principles to specific linguistic phenomena, while building on the insights of its predecessors.

We begin with a systematic review of hyperpartisan news detection, which establishes the landscape of existing research and methodologies and identifies critical gaps that motivate the subsequent work. This literature foundation reveals that current approaches often overlook low-resource languages and lack sufficient interpretability in their decision-making processes. Leveraging these insights, we propose a detailed benchmarking study of Transformer architectures and training techniques, evaluating their effectiveness for hyperpartisan detection in both high- and low-resource settings. This systematic comparison allows us to select optimal models for downstream applications. Building on this foundation, we develop novel linguistic resources for Italian, Spanish and Portuguese that capture the semantic dimension of partisan language. Simultaneously, we explore the potential of LLMs not merely as classifiers but as active participants in the annotation process. Indeed, we assess LLM alignment in annotation tasks and investigate how Persona Prompting can emulate politically motivated behavior, providing the machine with the human's socio-demographic data.

Furthermore, we focused on developing resources for the Italian language for hyperpartisan news detection, focusing on the relationship of this phenomenon with different linguistic layers (semantics, pragmatics, and logic). At the same time, while creating such resources,

we adopted LLMs in data augmentation pipelines—using human-in-the-loop strategies to accelerate the creation of neutralized versions of hyperpartisan content.

Finally, we complement these resource development and evaluation efforts with explainability studies that probe the linguistic and pragmatic knowledge encoded within Transformer models, unveiling what these architectures have learned about partisan language.

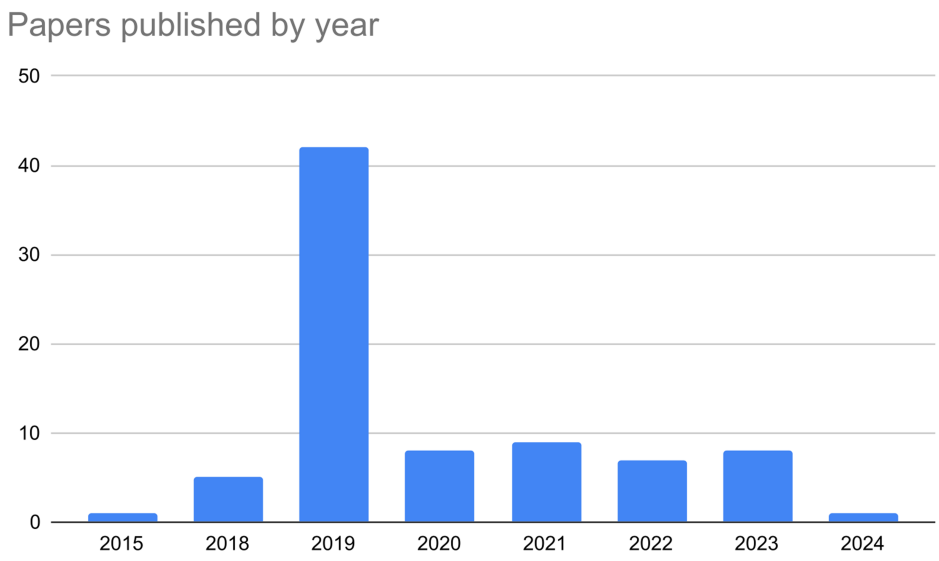
Together, these contributions advance the state of hyperpartisan detection through improved model selection, richer multilingual resources, enhanced data augmentation strategies, and deeper interpretability, providing both practical tools and theoretical insights for the community.

3.1 Defining Hyperpartisan News

In this section, we will present and describe the methodology adopted to conduct the systematic review following Keele et al. [2007]'s guidelines. The intent of the systematic review we conducted in 2023/2024, being the first in this domain, was to establish the foundation of this applied NLP task, separating it from collateral and similar tasks (e.g. disinformation detection). Considering the time window selected and corresponding to the first appearance of the term "hyperpartisan" and the beginning of the doctoral studies of the candidate, we collected quantitative data depicting the trend of published papers in this research area. Especially 3.1 and 3.2 provide informative insights on the popularity of this research line as well as the main publishers. Clearly, in 2015, that was the breakthrough of this research area, only ¡5 works covered the topic, which resonated in the community and attracted researchers' interest quickly. In 2019, Kiesel et al. [2019d] organized shared task 4 for SemEval-2019, focusing on Hyperpartisan News Detection. This competition, SemEval, is a series of international NLP research workshops hosted by ACL. This event was so successful that 42 teams participated, and this explains the peak recorded in that year. This high number of participants is reflected in Fig. 3.2, showing that 50% of papers on hyperpartisan detection are distributed by ACL. In that context, Bert [Devlin et al., 2019] was already popular enough, since it was released the year after and it contributed to stimulating the community. Later, the trend of this topic has been stabilized across the years with an average of 8 papers per year. The data referring to 2024 is scarce because it was collected in the first quarter of the year during the very first R&R. Lastly, 3.2 indicates that aside from ACL, another top conference, ACM, stores in its proceedings 16.7% of the papers. All the other publishers account for on average 5.6% each.

Considering the severe impact of hyperpartisan news on democracy and the distorted narration it spreads, its investigation suffers from the fact that researchers consider it marginal, possibly a subtopic of the broader field of disinformation detection.

Figure 3.1: The bar chart illustrates the trend of the selected publications over time.

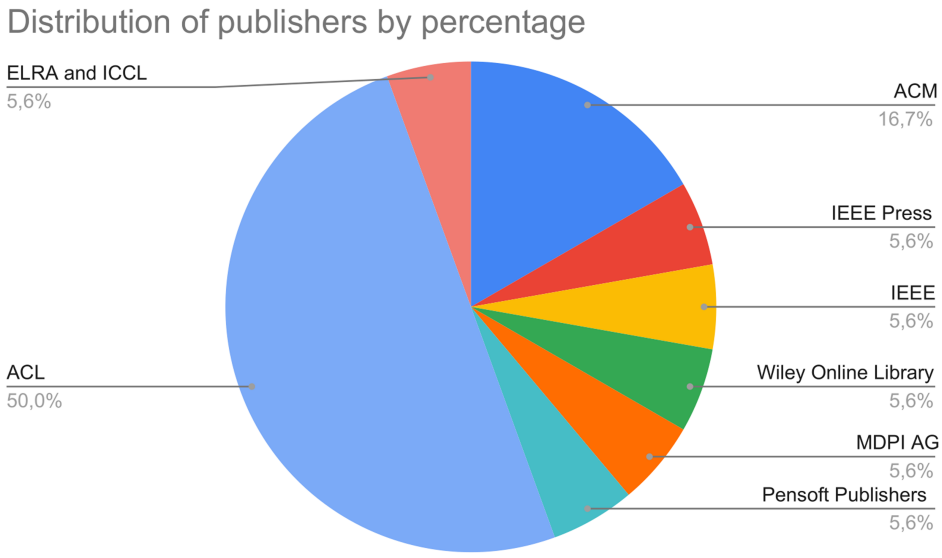


Search strategy

To do the literature review, we adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) guidelines [Moher et al., 2011], consisting of a checklist[1] and a flow diagram in Fig. 3.4 to illustrate in a simplified and clear way the steps taken. To retrieve papers, five different academic online databases were used to overcome their respective limitations [Gusenbauer and Haddaway, 2020] in terms of topic coverage and papers available: ACM Digital Library, Google Scholar, Scopus, ProQuest, and IEEE Xplore. Our query archetype was: `((hyperpartisan OR ``political bias'' OR ``hyper-partisan'' OR partisanship OR hyperpartisanship OR ``politica`

[1] `http://www.prisma-statement.org/documents/PRISMA_2020_checklist.pdf`

Figure 3.2: The pie chart shows the amin publishers for the selected papers.



`polarization'') AND (news OR bias OR articles) AND (detection OR classification))`. The first set of words contains the different homographs. We also searched in all subject fields to capture as many semantically similar papers as possible, including potentially miscategorized papers. We selected the 2015-2024 timeframe to analyze trends before the term was coined, considering a period in which studies on this topic grew and increasingly powerful models were employed.

For the purpose of obtaining pertinent papers related to our questions, the queries reported in Table 3.1 were a refined result of a structured process based on different steps introduced by Rodrigo-Ginés et al. [2024]. Queries within each database were structured to match titles, abstracts, and keywords. We extended this pipeline, introducing the following step: "Network visualization and exploration". It concerned the usage of the research software ResearchRabbit[2] to exhaustively capture possibly omitted papers through the citation links structure.

[2] `https://researchrabbitapp.com`

Table 3.1: Queries performed with advanced search for each database and the number of papers retrieved.

Database	Query	Docs
ACM Digital Library	`[[[Title: hyperpartisan] OR [[Title: political polarization] AND [[Title: news] OR [Title: article]]]] AND [[Abstract: hyperpartisan] OR [Abstract: media bias] OR [Abstract: or news] OR [All: article]]] OR [[[Keywords: hyperpartisan] OR [[Keywords: media bias] AND [[Keywords: news] OR [Keywords: article]]]] AND [[Title: hyperpartisan] OR [Title: hyper-partisan] OR [Title: partisan] OR [Title: political polarization] OR [[Title: media bias] AND [[Title: news] OR [Title: article]]]]] AND [E-Publication Date: (01/01/2015 TO 12/31/2024)]`	723
Google Scholar	`(hyperpartisan OR hyper-partisan OR hyperpartisanship OR polarization) AND (news OR bias OR articles) AND (detection OR classification)`	1800
IEEE Xplore	`(All Metadata:hyperpartisan OR All Metadata:hyper-partisan OR All Metadata:hyperpartisanship OR All Metadata:hyper-partisanship) AND (All Metadata:detection)`	1
ProQuest	`hyperpartisan news` + filters	159
ScienceDirect	`(hyperpartisan OR hyper-partisan OR political polarization OR media bias) AND NLP`	2
Scopus	`(hyperpartisan OR hyper-partisan OR partisanship OR hyperpartisanship OR political polarization) AND (news OR articles OR bias) AND (classification OR detection)`	97

- **Keywords domain extrapolation:** Initial reviews on similar topics helped us identify the keywords used in this domain. We noticed a lack of scientific agreement on writing "hyperpartisan". To cover all these morphologically diverse forms (hyper-partisan, hyperpartisan, hyper-partisan), we included them in our queries, treating them as synonyms;

- **Iterative searches:** This process allowed us to select the most appropriate terms by comparing the results retrieved using different keywords combinations. We examined how much title, titles and abstracts related with the queries;

- **Verifying against established literature:** To ensure the efficiency of our search terms, we compared the results to a list of papers in the domain of hyperpartisan detection;

- **Network visualization and exploration:** To further validate our verification, we used ResearchRabbit, a tool to visualize the citation links between papers in the same collection. It suggested similar papers written by the same or different authors, highlighting stored publications in the user's folder. This tool helped us in gauging the coherence of our results.

Selection criteria

Before describing the screening process, we illustrate the criteria employed for the paper selection.

- Inclusion criteria
  - Papers primarily focused on automated hyperpartisan news detection;
  - Papers that used the related task (e.g. fake news detection) as a synonym of hyperpartisan news detection;
  - Publications from 2015 to 2024;
- Exclusion criteria
  - Exclusion of sources that either address the hyperpartisan news detection problem from a theoretical perspective, namely theory papers, or manual detection;
  - Studies discussing only related topics, such as fake news detection, stance detection, or political bias;
  - Findings that do not use news domain datasets as the main source for hyperpartisan news detection, i.e. social network analysis, comments analysis, and tweets detection-based approaches;
  - Literature reviews, books, thesis and posters.

Screening and selection process

The following search strategy and procedures for study selection and analysis were used. The study selection, quality assessment of the included studies, and thematic analysis were performed by one author (PP). However, the procedures and findings were discussed by all authors, and potential disagreements were resolved by consensus. To manage the screening and selection processes, we utilized *Rayyan*[3] for its AI-powered capabilities, which allowed the two reviewers to conduct a blinded selection process, preventing any mutual influence. Specific eligibility criteria were established to ensure the reliability of the study. These criteria were applied independently by each reviewer to maintain objectivity and consistency. The criteria included: relevance to the predefined inclusion criteria, evaluation of models using

[3] https://www.rayyan.ai/

both accuracy and F1 score, and comprehensive reporting of the dataset used. Only papers that met the criteria and were accepted by both reviewers were selected. In cases where there was disagreement, a third reviewer was consulted to assess the paper's eligibility. The initial dataset consisted of 723 papers from ACM Digital Library[4], 571 from Google Scholar[5], 1 from ScienceDirect[6], 97 from Scopus[7], 159 from ProQuest[8], and 1 from IEEE Xplorer[9]. Notably, Google Scholar initially retrieved 1800 results, but we noted that, after the threshold of 500 results, it did not produce relevant documents. That led us to manually collect only the first 571 papers.

We conducted the entire selection process using Rayyan, as described in 3.4. It automatically detected 112 duplicates. After manual checks, we removed them. Left with 1441 studies, screening titles and abstracts was the initial step. Following thorough evaluations, 70 papers were retained from a curated pool of 110, eliminating 40 papers that did not meet specific focus or dataset criteria.

Additionally, to examine the cohesion and coherence of our references, we used ResearchRabbit[10] to visualize the citation network, identifying two prominent clusters with centers in Potthast et al. [2018b] and Kiesel et al. [2019d]. 15 additional papers were included after exploring similar works and citations, thanks to this procedure.

3.1.1 Where can hyperpartisanship be detected? Methodological operationalizations and their implications

The systematic review reveals that the operationalization of hyperpartisanship detection—namely, *where* and *at what granularity* researchers seek to identify partisan signals—remains highly fragmented and theoretically underdetermined. This fragmentation reflects both the nascent state of the field and a fundamental unresolved question: whether hyperpartisanship is an intrinsic property of a news article in its entirety, or whether it manifests unevenly across linguistic and structural components.

The prevailing approach prioritizes textual content within online news outlets, grounded in the assumption that linguistic markers encode partisan bias. Within this paradigm, researchers

[4] https://dl.acm.org/
[5] https://scholar.google.com/
[6] https://www.sciencedirect.com/
[7] https://www.scopus.com/home.uri
[8] https://www.proquest.com/index
[9] https://ieeexplore.ieee.org/Xplore/home.jsp
[10] https://www.researchrabbit.ai/

have adopted diverse detection scopes, ranging from document-level analysis to fine-grained section-specific approaches. Some studies examine entire articles [Kiesel et al., 2019d], while others focus on particular regions hypothesized to carry concentrated partisan signals: headlines [Lyu et al., 2024a, Amason et al., 2019], individual sentences [jeong Lim et al., 2018], or quoted material [Pérez-Almendros et al., 2019]. The rationale underlying these choices is intuitive: headlines are crafted to capture reader attention and may therefore embody stronger rhetorical positioning [Poudel et al., 2025], while quotes anchor claims to sources and may reveal selective framing, consisting in emphasizing certain aspects of reality, shaping interpretation, evaluation and response [Goffman, 1974]. However, the literature provides scant evidence systematically comparing the predictive power of these granularities or examining whether hyperpartisanship detection at the headline level correlates meaningfully with article-level bias. This represents a significant gap: without such comparative analysis, researchers risk optimizing models to local linguistic patterns that may not generalize to the phenomenon of interest.

Beyond purely textual analysis, the field has begun exploring source-level and network-based approaches, recognizing that hyperpartisanship may be as much a product of institutional positioning as of linguistic choice. Publishers, interlinked through economic and political networks [Hermann and Chomsky, 1994], constitute a polarized ecosystem. Some studies infer hyperpartisanship from media outlet bias or journalist ideology [M. Alzhrani, 2022, Kiesel et al., 2019d], leveraging external metadata such as hyperlinks [Hrckova et al., 2021, Kulkarni et al., 2018, Alabdulkarim and Alhindi, 2019, Joo and Hwang, 2019]. Yet this source-centric approach introduces a confounding problem: while an article originating from a highly partisan outlet tends toward hyperpartisanship, it is not guaranteed to manifest it [Tran, 2020, Jiang et al., 2019]. This discrepancy was explicitly highlighted by Jiang et al. [2019], suggesting that source bias and article-level partisanship are distinct phenomena that may co-occur but do not perfectly align. The implications are substantial: relying exclusively on publisher metadata as a proxy for hyperpartisanship risks conflating institutionalized bias with rhetorical extremism, potentially misclassifying mainstream outlets' coverage of polarizing topics as hyperpartisan or, conversely, normalizing consistently extreme discourse from ideologically aligned publishers.

Sentiment represents another instrumental signal. Empirical findings suggest that hyperpartisan articles exhibit more aggressive and negative sentiment profiles [Palić et al., 2019] compared to mainstream news. Multiple studies leveraged sentiment as a feature [Baly et al.,

2019, Chen et al., 2019, Anthonio and Kloppenburg, 2019], employing tools like TextBlob[11] and VADER[12]. While sentiment correlates with hyperpartisanship, it is crucial to recognize that strong sentiment is not synonymous with hyperpartisan framing: legitimate investigative journalism, incisive opinion pieces, and passionate coverage of injustice may all exhibit high sentiment without necessarily qualifying as hyperpartisan. This suggests that affect-based features are useful but insufficient signals, and that their interpretation depends critically on the precise definition of hyperpartisanship adopted.

Recently, multimodal approaches have emerged, reflecting growing recognition that news artifacts are rarely text-only. Spezzano et al. [2021] demonstrated that incorporating visual elements alongside textual analysis enhances detection performance and, remarkably, that automated classifiers outperformed human annotators. This finding invites intriguing speculation: do images convey partisan framing more potently than text, or do multimodal models simply capture redundant signals more robustly? And if machines succeed where humans struggle, does this indicate that hyperpartisanship manifests in low-level statistical patterns that exceed human perceptual capacity, or rather that human judgment integrates contextual and background knowledge that remains inaccessible to models? These questions underscore the theoretical ambiguity surrounding the phenomenon itself.

3.1.2 How is hyperpartisanship annotated and operationalized? The measurement dilemma

Perhaps more troubling than the spatial fragmentation of detection sources is the fundamental heterogeneity in how hyperpartisanship itself is conceptualized, measured, and labeled. This heterogeneity betrays a field still grappling with basic definitional clarity—a problem that threatens the comparability and reproducibility of empirical findings.

The measurement challenge emerges at the intersection of social science and computational linguistics. In political science, polarization has been formalized through instruments like the CSES Polarization Index (PI) [Dalton, 2021], which maps political parties onto a Left-Right continuum while accounting for party size and vote share, providing a macro-level assessment of ideological dispersion. Such indices offer rigor precisely because they are grounded in institutional structures and aggregate-level phenomena. In stark contrast, computational approaches to hyperpartisan news detection have largely converged on label schemas derived

[11] `https://textblob.readthedocs.io/en/dev/`
[12] `https://github.com/cjhutto/vaderSentiment`

ex post from annotation campaigns whose guidelines are rarely disclosed, limiting access to the construct assumed in their research.

The earliest influential shared task, SemEval-2019 [Kiesel et al., 2019d], adopted a binary schema distinguishing merely hyperpartisan from mainstream (non-hyperpartisan). While this binary framing provides simplicity for supervised learning, it obscures considerable variation: a moderately slanted article, a quasi-factual piece with pointed omissions, and an overtly inflammatory screed populated with logical fallacies all collapse into hyperpartisanship. Other studies adopted complementary orthogonal distinctions, labeling articles as Left/Right without necessarily marking intensity [Kim and Johnson, 2022, M. Alzhrani, 2022]. Such choices reflect explicit research questions—if the goal is to map ideological orientation, Left/Right is appropriate; if the goal is to flag manipulative rhetoric, intensity and the identification of the presence of rhetorical devices matter more than direction. However, Potthast et al. [2018b] argued persuasively that coarse Left/Right distinctions obscure meaningful heterogeneity, particularly in multi-party democracies where center, progressive, conservative, libertarian, and other ideological positions cannot be collapsed onto a single axis without profound information loss.

A minority of studies have pursued finer-grained scales, treating hyperpartisanship detection as a multi-class or regression problem. Aksenov et al. [2021] employed 5-point and 7-point numerical scales, anchoring them to ideological positions: for instance, the 7-point scale mapped 1.0-2.5 to ”far-left,” 2.5-3.5 to ”center-left,” and so forth. Similarly, Baly et al. [2019] and Azizov et al. [2023] attempted to capture granularity by distinguishing finer categories (e.g., far-left, left, center, right, far-right). These approaches acknowledge that hyperpartisanship is not a binary property but a continuum or a polytypic space. Yet they introduce new complications: what ensures that ”center-left” at 2.5-3.5 on a 7-point scale corresponds meaningfully across datasets, languages, or political contexts? Do these ordinal positions carry comparable semantic content in a leftist democracy versus a rightist one? How should cultural and temporal variation in partisan expression be integrated?

This measurement discord has profound implications for model development and evaluation. Models trained on binary labels learn to segment a one-dimensional space into two regions; those trained on continuous scales learn to regress positions onto a continuum; those trained on multi-class schemes partition the space into discrete categories. These are distinctly different learning problems, and a model fine-tuned for binary classification will likely perform poorly when tasked with fine-grained assessment. Crucially, without standardized annotation

guidelines and inter-rater agreement metrics, comparing performance across datasets is perilous: a 90% accuracy score on one corpus may not be commensurable with such a score on another if the underlying label distributions, annotation protocols, or theoretical definitions differ. This recommendation emerges strongly from the review: the field would benefit from establishing shared annotation standards and transparent documentation of labeling decisions, enabling reproducibility and meta-analytic comparison. Moreover, future work must grapple explicitly with the question of whether hyperpartisanship is fundamentally unidimensional (a spectrum from balanced to extreme) or multidimensional (encompassing intensity, direction, and perhaps other aspects such as emotional tone or factual accuracy).

3.1.3 Datasets in Literature

In the previous sections, we discussed the modeling of the phenomenon of our interest and its operationalization. On the other hand, here we gathered the datasets built on these assumptions and we are going to analyze them, illustrating the language distributions, accessibility, and other important aspects. Effective models depend on top-notch data quality to function optimally. However, constructing a high-quality, well-balanced dataset can be both time-consuming and resource-intensive. This challenge is compounded by shifts in data policies across social networks since the Cambridge Analytica scandal, leading to potential difficulties or cost changes in obtaining data. Additionally, a trend has emerged within news sources where access to data is restricted due to its previous utilization in training models like GPT [13]. Consequently, news sources have implemented paywalls and crawler restrictions[14], making it exceedingly challenging to gather suitable information for this and similar tasks.

To support upcoming studies on identifying hyperpartisan news and related tasks, we have created the extensive tables: 3.3 and 3.4, which outline key attributes of datasets relevant to hyperpartisan news detection. This table includes datasets referenced in different papers. Some are not primarily used for hyperpartisan detection but could be. It is important to note that when subsets or extended versions of earlier datasets exist, we consider them separate entities denoted by *. Additionally, datasets marked with ** signify merged collections. The column labeled *Data* indicates the number of articles gathered by the researchers.

[13] https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/

[14] https://ilmanifesto.it/termini-e-condizioni

To provide comprehensive insights into the table, we will give a philological explanation of the datasets marked with the symbols * and **. *Framing Triplet Dataset* is a combination of the following datasets: SemEval-2019 task 4 along with Roy and Goldwasser [2020]'s data. Furthermore, Roy and Goldwasser [2020] expands the *SemEval-2019 task 4* dataset by incorporating articles collected from polarized sources and then labeled through mediabiasfactcheck.com. Regarding BIGNEWS, collected by Liu et al. [2022a], it has two subsets, respectively: *BIGNEWSBLN*, a downsampled corpus maintaining an equal distribution of ideologies, and *BIGNEWSALIGN*, which clusters news stories from opposing sources but on the same topic. In their research, Tran [2020] utilized a subset of *All-the-news*[15]. Furthermore, Potthast et al. [2018b] worked with a subset of articles crawled from the URLs contained in *The BuzzFeed-Webis Fake News Corpus* collected by Silverman et al. [2016]. By cleaning Potthast et al. [2018b]'s dataset, Sánchez-Junquera et al. [2021] obtained a new dataset. The same researchers created *StereoImmigrants*, a collection of Spanish news about immigrants, for Sánchez-Junquera [2021].

Labeling and retrieving processes relied upon platforms like Allsides[16], Factcheck[17], Politifact[18], as ground truth for establishing the bias of an article and as a source for collecting data. Indeed, in these contexts, experts assign news to the political orientation.AllSides.com is a media company that specializes in providing balanced news coverage by collecting and comparing news stories from various sources with different political leanings. The platform categorizes news articles based on their political bias—whether left, center, or right—and scores them according to the level of partisanship they contain.

Since Baly et al. [2020] noted that training models with big datasets reduces the performance due to their noise, researchers started to prefer the quality rather than the dimension. Indeed, Lyu et al. [2024a], after a deeper analysis of the SemEval 2019 dataset, revealed several issues with this widely used ground truth dataset: class imbalance, task-label unalignment, and distribution shift. As we can see from 3.5, there is an imbalanced distribution towards English data, leaving the context of minority languages understudied.

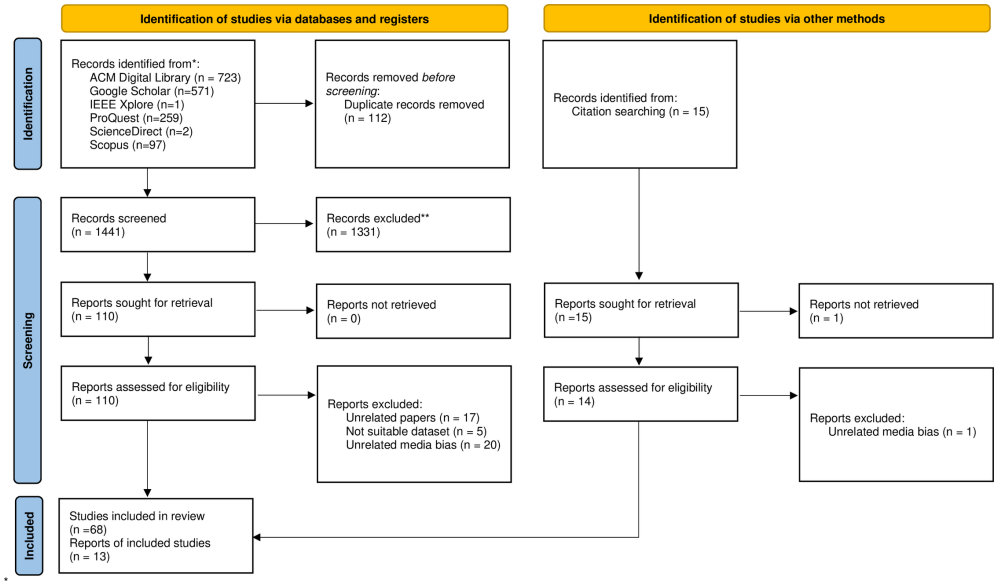
[15] https://www.kaggle.com/datasets/snapcrack/all-the-news
[16] https://allsides.com/
[17] https://mediabiasfactcheck.com/
[18] https://www.politifact.com/

Figure 3.3: Datasets for Hyperpartisan News detection task.

Dataset	Reference	Year	Size	Bias Label	Language	Media's country	Availability
No Name	[133]	2024	22628	U/A	Estonian	Estonia	No
No Name	[129]	2024	U/A	U/A	Persian	Iran	No
Task 3A	[81]	2023	55000	AllSides	English	U.S.	Yes
Task 3B	[81]	2023	8000	AllSides	English	U.S.	Yes
Allsides-L	[109]	2023	719256	Allsides	English	U.S.	Yes
No Name	[46]	2023	1824824	AllSides, Media Bias Factcheck	English	U.S.	No
Framing Triplet Dataset	[118]	2022	25627	Media Bias Factcheck	English	U.S.	Yes
No Name	[79]	2022	10000	Allsides	English	U.S.	No
TVP Info	[146]	2022	81694	U/A	Polish	Poland	Upon request
TVN 24	[146]	2022	128527	U/A	Polish	Poland	Upon request
BIGNEWS	[117]	2022	3689229	Allsides, adfontesmedia	English	U.S.	Upon request
*BIGNEWS BLN	[117]	2022	2331552	Allsides, adfontesmedia	English	U.S.	Upon request
*BIGNEWS ALIGN	[117]	2022	1060512	Allsides, adfontesmedia	English	U.S.	Upon request
GERMAN dataset	[80]	2021	47362	None	German	Germany	Yes
No Name*	[62]	2021	1555	BuzzFeed	English	U.S.	No
Stereo Immigrants	[132]	2021	3704	Manually labeled	Spanish	Spain	Yes
No Name*	[120]	2020	21645	Media Bias Factcheck	English	U.S.	Yes
The Annotated Data Dataset	[147]	2020	46	Media Bias Factcheck	English	U.S.	No
No Name	[59]	2020	37000	None	Italian	Italy	No
*PoliNews	[49]	2020	83000	U/A	English	U.S.	No
Presidential	[71]	2020	178572	Allsides, Media Bias Factcheck	English	U.S.	No
POLUSA	[141]	2020	9000000	None	English	U.S.	Yes
*Politifact	[61]	2020	18027	Politifact.com	English	U.S.	No
No Name	[51]	2020	4627	Manually labeled	English	India	No
No Name	[121]	2020	34737	Allsides	English	U.S.	Yes
All-Sides	[148]	2019	10385	None	English	U.S.	No
Telugu	[58]	2019	1327	Manually labeled	Telugu	India	Yes
NELA-2018	[142]	2019	713534	Allsides, Media Bias Factcheck, BuzzFeed et al.	English	U.S.	Yes
by-article	[18]	2019	1273	Manually labeled	English	U.S.	Yes
by-publisher	[18]	2019	754000	BuzzFeed news, Media Bias Factcheck	English	U.S.	Yes
BASIL	[143]	2019	300	Manually labeled	English	U.S.	Yes
The BuzzFeed-Webis Fake News Corpus 2016	[33]	2018	1627	BuzzFeed	English	U.S.	Yes
Reuter	U/A	U/A	18519	U/A	English	U.S.	Yes
No Name	[63]	2018	88	Crowd-sourcing	English	U.S.	No
MBCF	[144]	2018	1066	Media Bias Factcheck	English	U.S.	Yes
NELA-2017	[145]	2018	136000	U/A	English	U.S.	Yes
BuzzFeed 2016	buzzfeednews.com	2016	2282	BuzzFeed	English	U.S.	Yes
No Name	[52]	2015	74	Crowd-sourcing	English	U.S.	No

https://doi.org/10.1371/journal.pone..t008

Figure 3.4: PRISMA Flow Diagram.



The Flow Diagram illustrates the steps during document collection and evaluation. We skimmed more than 1551 papers and finally we selected a subset of 82.

Figure 3.5: Datasets' language distribution.

Dataset	Reference	Year	Size	Bias Label	Language	Media's country	Availability
No Name	[133]	2024	22628	U/A	Estonian	Estonia	No
No Name	[129]	2024	U/A	U/A	Persian	Iran	No
Task 3A	[81]	2023	55000	AllSides	English	U.S.	Yes
Task 3B	[81]	2023	8000	AllSides	English	U.S.	Yes
Allsides-L	[109]	2023	719256	Allsides	English	U.S.	Yes
No Name	[46]	2023	1824824	AllSides, Media Bias Factcheck	English	U.S.	No
Framing Triplet Dataset	[118]	2022	25627	Media Bias Factcheck	English	U.S.	Yes
No Name	[79]	2022	10000	Allsides	English	U.S.	No
TVP Info	[146]	2022	81694	U/A	Polish	Poland	Upon request
TVN 24	[146]	2022	128527	U/A	Polish	Poland	Upon request
BIGNEWS	[117]	2022	3689229	Allsides, adfontesmedia	English	U.S.	Upon request
*BIGNEWS BLN	[117]	2022	2331552	Allsides, adfontesmedia	English	U.S.	Upon request
*BIGNEWS ALIGN	[117]	2022	1060512	Allsides, adfontesmedia	English	U.S.	Upon request
GERMAN dataset	[80]	2021	47362	None	German	Germany	Yes
No Name*	[62]	2021	1555	BuzzFeed	English	U.S.	No
Stereo Immigrants	[132]	2021	3704	Manually labeled	Spanish	Spain	Yes
No Name*	[120]	2020	21645	Media Bias Factcheck	English	U.S.	Yes
The Annotated Data Dataset	[147]	2020	46	Media Bias Factcheck	English	U.S.	No
No Name	[59]	2020	37000	None	Italian	Italy	No
*PoliNews	[49]	2020	83000	U/A	English	U.S.	No
Presidential	[71]	2020	178572	Allsides, Media Bias Factcheck	English	U.S.	No
POLUSA	[141]	2020	9000000	None	English	U.S.	Yes
*Politifact	[61]	2020	18027	Politifact.com	English	U.S.	No
No Name	[51]	2020	4627	Manually labeled	English	India	No
No Name	[121]	2020	34737	Allsides	English	U.S.	Yes
All-Sides	[148]	2019	10385	None	English	U.S.	No
Telugu	[58]	2019	1327	Manually labeled	Telugu	India	Yes
NELA-2018	[142]	2019	713534	Allsides, Media Bias Factcheck, BuzzFeed et al.	English	U.S.	Yes
by-article	[18]	2019	1273	Manually labeled	English	U.S.	Yes
by-publisher	[18]	2019	754000	BuzzFeed news, Media Bias Factcheck	English	U.S.	Yes
BASIL	[143]	2019	300	Manually labeled	English	U.S.	Yes
The BuzzFeed-Webis Fake News Corpus 2016	[33]	2018	1627	BuzzFeed	English	U.S.	Yes
Reuter	U/A	U/A	18519	U/A	English	U.S.	Yes
No Name	[63]	2018	88	Crowd-sourcing	English	U.S.	No
MBCF	[144]	2018	1066	Media Bias Factcheck	English	U.S.	Yes
NELA-2017	[145]	2018	136000	U/A	English	U.S.	Yes
BuzzFeed 2016	buzzfeednews.com	2016	2282	BuzzFeed	English	U.S.	Yes
No Name	[52]	2015	74	Crowd-sourcing	English	U.S.	No

https://doi.org/10.1371/journal.pone..t008

Table 3.2: Datasets collection and description. Part 1.

Dataset	Reference	Year	Size	Bias Labeling	Lang.	Avail.
No Name	Omidi Shayegan et al. [2024]	2024	U/A	U/A	Persian	No
Task 3A [19]	Azizov et al. [2023]	2023	55000	AllSides	English	Yes
Task 3B [20]	Azizov et al. [2023]	2023	8000	AllSides	English	Yes
Allsides-L [21]	Ko et al. [2023]	2023	719256	Allsides	English	Yes
No Name	Lyu et al. [2024a]	2023	1824824	AllSides, Media Bias Factcheck	English	No
Framing Triplet Dataset [22]	Kim and Johnson [2022]	2022	25627	Media Bias Factcheck	English	Yes
No Name	Sridharan and S. [2022]	2022	10000	Allsides	English	No
TVP Info	Szwoch et al. [2022]	2022	81694	U/A	Polish	UR
TVN 24	Szwoch et al. [2022]	2022	128527	U/A	Polish	UR
BIGNEWS	Liu et al. [2022a]	2022	3689229	Allsides, adfontesmedia	English	UR
*BIGNEWS BLN	Liu et al. [2022a]	2022	2331552	Allsides, adfontesmedia	English	UR
*BIGNEWS ALIGN	Liu et al. [2022a]	2022	1060512	Allsides, adfontesmedia	English	UR
GERMAN dataset [23]	Aksenov et al. [2021]	2021	47362	None	German	Yes
No Name*	Sánchez-Junquera et al. [2021]	2021	1555	BuzzFeed	English	No
Stereo Immigrants[24]	Sánchez-Junquera [2021]	2021	3704	Manually labeled	Spanish	Yes
No Name* [25]	Roy and Goldwasser [2020]	2020	21645	Media Bias Factcheck	English	Yes

*: indicates datasets that are actually subsets or extended versions of earlier versions.

Table 3.3: Datasets collection and description. Part 2.

Dataset	Reference	Year	Size	Bias Labeling	Lang.	Avail.
The Annotated Data Dataset	Lim et al. [2020]	2020	46	Media Bias Factcheck	English	No
No Name	Pierri et al. [2020]	2020	37000	None	Italian	No
*PoliNews	Tran [2020]	2020	83000	U/A	English	No
Presidential	Alzhrani [2020]	2020	178572	Allsides, Media Bias Factcheck	English	No
POLUSA [26]	Gebhard and Hamborg [2020]	2020	9000000	None	English	Yes
*Politifact	University Politehnica of Bucharest et al. [2021]	2020	18027	Politifact.com	English	No
No Name	Sharma et al. [2020]	2020	4627	Manually labeled	English	No
No Name	Baly et al. [2020]	2020	34737	Allsides	English	Yes
All-Sides	Li and Goldwasser [2019]	2019	10385	None	English	No
Telugu[27]	Gangula et al. [2019]	2019	1327	Manually labeled	Telugu	Yes
NELA-2018 [28]	Norregaard et al. [2019]	2019	713534	Allsides, Media Bias Factcheck, BuzzFeed et al.	English	Yes
by-article [29]	Kiesel et al. [2019d]	2019	1273	Manually labeled	English	Yes
by-publisher [30]	Kiesel et al. [2019d]	2019	754000	BuzzFeed news, Media Bias Factcheck	English	Yes
BASIL[31]	Fan et al. [2019]	2019	300	Manually labeled	English	Yes
The BuzzFeed-Webis Fake News Corpus 2016 [32]	Potthast et al. [2018b]	2018	1627	BuzzFeed	English	Yes

U/A: unavailable.

Table 3.4: Datasets collection and description. Part 3.

Dataset	Reference	Year	Size	Bias Labeling	Lang.	Avail.
Reuter[33]	U/A	U/A	18519	U/A	English	Yes
No Name	jeong Lim et al. [2018]	2018	88	Crowd-sourcing	English	No
MBCF [34]	Baly et al. [2018]	2018	1066	Media Bias Factcheck	English	Yes
NELA-2017[35]	Horne et al. [2018]	2018	136000	U/A	English	Yes
BuzzFeed 2016[36]	buzzfeednews.com	2016	2282	BuzzFeed	English	Yes
No Name	Baumer et al. [2015]	2015	74	Crowd-sourcing	English	No

U/A: unavailable.

3.2 Literature Review

3.2.1 Literature Review: Hyperpartisan and Misinformation Detection

This section provides a comprehensive overview of the computational approaches developed to detect hyperpartisan content, misinformation, and related forms of media bias. We organize the literature thematically, beginning with detection methodologies and progressing toward emerging paradigms such as in-context learning, explainability, and text neutralization.

The detection of hyperpartisan content encompasses a range of methodologies, varying from traditional non-deep learning approaches to deep learning models, as well as mixed learning algorithms. Non-deep-learning methodologies often rely on traditional machine learning algorithms, leveraging handcrafted features and rule-based systems to identify linguistic patterns, stylistic markers, and network structures within textual and metadata sources. These approaches commonly include stylometric analysis and topic modeling methods to discern biased content. In contrast, deep learning methodologies harness the power of neural networks to automatically extract intricate features from raw data, enabling the identification of complex patterns and relationships in unstructured text or network data. These techniques, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and Transformers, excel in learning representations directly from the data.

In the following subsections, to encompass the risk of bias, we grouped and discussed the literature on hyperpartisan detection by model architecture. We will differentiate between ML, DL, and other methodologies adopted in the papers selected for the systematic review. Lastly, the DL section includes shallow Deep Learning models and the Transformers family.

Non Deep Learning Methods

Table 3.5 categorizes papers using the traditional machine learning approaches. The methodologies involved algorithms like Support Vector Machines, Random Forest, and Logistic Regression, followed by Linear Regression, Naive Bayes, Linear SVC, KNN, XGBoost, and Maxent. There are effective strategies adopted with the SVM model. For instance, [Alabdulkarim and Alhindi, 2019] used this algorithm combined with the sentiment analysis via the National Council Canada's Emotion Lexicon (NRC Emotion Lexicon) to analyze the emotional content in the article. Moreover, they extended the linguistic approach by applying Linguistic Inquiry and Word Count (LIWC). They also considered the articles' structure and metadata as features. [Nguyen et al., 2019], adopted n-grams, i.e. bi- and tri-grams, and

Table 3.5: Traditional Machine Learning algorithms used in the selected literature.

Algorithm	**Reference**
Support Vector Machine (SVM) [Hearst et al., 1998]	[Pérez-Almendros et al., 2019, Nguyen et al., 2019, Knauth, 2019, Alabdulkarim and Alhindi, 2019, Anthonio and Kloppenburg, 2019, Yeh et al., 2019, Dumitru and Rebedea, 2019]
XGBoost [Chen and Guestrin, 2016]	[Gupta et al., 2019]
Maxent [Merow et al., 2013]	[Agerri, 2019]
Random Forest [Breiman, 2001]	[Chakravartula et al., 2019, Cruz et al., 2019, Stevanoski and Gievska, 2019, Naredla and Adedoyin, 2022, Aksenov et al., 2021]
Naive Bayes	[Chen et al., 2019, Amason et al., 2019]
Copula Ordinal Regression	[Baly et al., 2019]
Logistic Regression	[Sengupta and Pedersen, 2019, Garg and Sharma, 2022, Saleh et al., 2019, Bestgen, 2019, Srivastava et al., 2019]
Linear Support Vector Classifier (SVC)	[Palić et al., 2019]
Linear Classifier	[Hanawa et al., 2019]

"Algorithm" specifies the method; "Reference" indicates the authors who used that method.

dependency sub-trees that impacted the performance. On the other hand, [Yeh et al., 2019] experimented several embeddings: Doc2Vec [Le and Mikolov, 2014], Glove [Pennington et al., 2014], ELMo [Peters et al., 2018]. They found out that "adding simple lexical and sentiment features hurts the performance". Dumitru and Rebedea [2019] studied the linguistic divergences between fake and hyperpartisan news employing an SVM. In this case, it emerged that hyperpartisan articles exhibit more sentences and a higher adjective count compared to unbiased news. When comparing the characteristics of extremely polarized articles against fake news, they noted that the former contains high usage of question/exclamation marks and adjectives. These sentence-related features delineate distinct linguistic patterns. Srivastava

et al. [2019] confirmed the robust potentialities of the Logistic Regression, ranking in the second place at the SemEval-2019. Srivastava et al. [2019] built representations with Universal Sentence Encoder (USE) [Cer et al., 2018] and combined both semantic and handcrafted features, paying attention to the grade of the adjectives and subjectivity, and distinguishing between two levels of polarity: sentence and article level. Garg and Sharma [2022] used the Reuter Dataset for the training and the test, combining the ELMo embedding with a logistic regression classifier as already done by Jiang et al. [2019] and Srivastava et al. [2019], confirming the effectiveness of this method. Srivastava et al. [2019] discovered that the most relevant features concern bias lexicon and polarity. Hanawa et al. [2019] placed third at the SemEval-2019 and found that article length was a distinctive trait of biased articles. By working at the phrase level, they created a set of phrases to discern the different types of articles, paying attention to removing n-grams containing publishers' style biases. Lyu et al. [2024a] focused on news titles with a topic-based approach. They also built a dataset considering two distinct typologies of news titles, augmenting the granularity of the detection. The first category pertains to descriptions of confrontations or conflicts between opposing parties, suggesting a deeply polarized political climate. The second set involves opinions that express a biased, inflammatory, and aggressive stance against a policy, a political party, or a politician. Baly et al. [2019] thought there was an interdependence between factuality and political ideology bias, so they introduced a multi-task learning setup with the Copula Ordinal Regression (COR) [Walecki et al., 2016]. They used the entire news outlet and considered diverse scales for measuring factuality (3-point scale) and political bias (7-point scale). Agerri [2019] used Maximum Entropy Modeling (MaxEnt) to bypass linguistic features to build a model capable of generalizing as much as possible; they devised a document classification system that combines clustering features with simple local features. They showcased the effectiveness of employing distributional features from large in-domain unlabeled data. Gupta et al. [2019] approached the task using n-gram embeddings with article and title polarity, implementing the XGBoost model with all of these scalar features, but it performed poorly. They derived their methodology of applying stylometric analysis from Potthast et al. [2018b]. This approach utilized n-grams, readability scores, and Part-of-Speech (PoS) followed by binary classification. Thanks to unmasking information, they simultaneously compared documents with opposite political leanings. In doing so, Potthast et al. [2018b] investigated the style variations depending on the political orientation and confronted it with topic-based bag-of-words models. This methodology highlighted the limited usefulness of integrating corpus characteristics when

performing a granular distinction among left, right, and mainstream styles. Indeed, both the political extremes show similarities and can produce confounding effects in the model. Hence, concerning the style analysis for the hyperpartisan detection, the categories should be limited to mainstream and hyperpartisan without considering the specific leaning. Furthermore, for a complete understanding of the approaches used in the literature, we summarized them in Table 3.6. In this case, ELMo, BERT, and Word2Vec embeddings were used as features of Non-Deep Learning algorithms. 3.6 describes only the features used with the best models proposed in Non-Deep Learning approaches in Table 3.5. We distinguish between features (Morpho-syntactic, Lexicon, Semantic, Sentiment and Metadata) and approaches (style-based and topic-based).

Table 3.6: Feature-based approaches for hyperpartisan news detection.

Ref.	**Morpho-syntactic**	**Lexicon**	**Semantic**	**Sentiment**	**Meta-data**	**Style-based**	**Topic-based**
Pérez-Almendros et al. [2019]	x	x	GloVe			x	
Nguyen et al. [2019]	x		*n*-grams			x	
Dumitru and Rebedea [2019]			TF-IDF	x	x	x	
Hanawa et al. [2019]			BERT embeddings			x	
Knauth [2019]	x	x				x	x

Continued on next page

Ref.	**Morpho-syntactic**	**Lexicon**	**Semantic**	**Sentiment**	**Meta-data**	**Style-based**	**Topic-based**
Alabdulkarim and Alhindi [2019]	x	x	TF-IDF unigrams	x	x	x	
Anthonio and Kloppenburg [2019]	x	x		x			
Yeh et al. [2019]			BoW				
Sengupta and Pedersen [2019]			*n*-grams				
Gupta et al. [2019]	x		*n*-grams	x		x	
Agerri [2019]	x		W2V		x		
Chakravartula et al. [2019]			BoW				

Continued on next page

Ref.	Morpho-syntactic	Lexicon	Semantic	Sentiment	Meta-data	Style-based	Topic-based
Garg and Sharma [2022]			ELMo embed-ding			x	
Cruz et al. [2019]	x		BoW			x	
Stevanoski and Gievska [2019]	x	x	W2V		x	x	
Naredla and Ade-doyin [2022]			ELMo embed-ding			x	
Chen et al. [2019]	x	x	BoW	x			
Amason et al. [2019]	x	x	BoW	x		x	
Baly et al. [2019]	x	x	W2V	x	x		
Saleh et al. [2019]	x	x	BoW			x	
Bestgen [2019]	x		BoW				
Srivastava et al. [2019]	x	x	USE Cer et al. [2018]	x		x	
Palić et al. [2019]	x	x	W2V	x	x		

The features are the following: Morpho-syntactic, Lexical, Semantic, and Sentiment. The

approaches are: Style-based and Topic-based. The style-based approach includes tools and methods like vocabularies, reading scores, count of pronouns and other parts of speech, n-grams of part of speech, punctuation, lemma n-gram similarity score, and article length features to analyze the writing style. Topic-based refers to methods like topic clusters and ad hoc vocabularies to compare topics. Metadata includes internal and external links, dates, and journalists' bios. "Morpho-syntactic" involves using Part-of-Speech tagging, dependency trees, and quantitative approaches like word or adjective distributions. "Lexicon" refers to the utilization of external vocabularies like LIWC, or handcrafted. "Semantic" regards the adoption of embeddings and frequency-based features: TF-IDF, FastText, GloVe, and BERT embeddings. "Sentiment" indicates the application of algorithms for sentiment analysis like Valence Aware Dictionary and sEntiment Reasoner (VADER) and National Research Council Canada Emotion Lexicon (NRCLex).

In the following paragraphs, we analyzed the Deep Learning methods adopted by diverse authors to solve the hyperpartisan detection task. In 3.8, we categorized papers using the traditional machine learning approaches. Lastly, at the end of the chapter, a comprehensive table, split into 3.10 and 3.11, collects and illustrates the results rounded to two decimals reported by all the authors studied in our systematic review.

Table 3.8: Collection of the best performative DL models used in the literature.

Algorithm	Description	Reference
Recurrent Neural Network-based (RNN) [Rumelhart et al., 1986]	Recurrent Neural Networks are a class of artificial neural networks designed to process sequential data by maintaining an internal memory. They possess loops that allow information persistence, enabling them to consider past inputs for present computations, making them adept at handling time series or sequential data in various applications.	Sharma et al. [2020]

Continued on next page

Algorithm	Description	Reference
Mesh Neural Network [Sridharan and S., 2022]	It is a neural network aiming at maximazing the weight of past examples. Its structure is composed by recurrent nodes that boost the inductive abilities of the machine.	Sridharan and S. [2022]
Catboost [Dorogush et al., 2017]	Catboost stands out as a specialized library crafted for boosting gradients. It employs a unique algorithm to prevent target leakage, ensuring high efficiency and precision.	Azizov et al. [2023]
Convolutional Neural Network	Convolutional Neural Networks were designed for image processing but were applied to Natural Language Processing, in which the input text is a one-dimensional vector of tokens, or like a matrix of embeddings. This model can learn both local and global dependencies as well as hierarchical features in the data.	Papadopoulou et al. [2019], Gerald Ki Wei and Jun Choi [2021], Pérez-Almendros et al. [2019], Jiang et al. [2019], Färber et al. [2019], Joo and Hwang [2019], Zehe et al. [2019], Ruan et al. [2021]
Hierarchical Attention Network (HAN) [Yang et al., 2016]	The HAN is a model capable of balancing the information in a current state, deciding whether to update it and how much the past information contributes to its new state.	Cruz et al. [2020], Moreno et al. [2019], M. Alzhrani [2022], Gangula et al. [2019], Ko et al. [2023]

Continued on next page

Algorithm	Description	Reference
Long Short Term Memory-based models [Hochreiter and Schmidhuber, 1997]	Unidirectional LSTM retain information solely from preceding inputs as it has only processed data from the past. Employing a bidirectional approach involves processing inputs in two directions: from past to future and from future to past. The distinctive aspect of this method, compared to unidirectional LSTM, lies in the LSTM running in reverse, which captures information from the future. By amalgamating the two hidden states, this technique enables the retention of information from both the past and future at any given time. In NLP, the main difference lies in the context captured. Bi-LSTM can discover hidden reversed semantic and syntactic structures.	Isbister and Johansson [2019], Li and Goldwasser [2021], Cramerus and Scheffler [2019], Zhang et al. [2019], Roy and Goldwasser [2020]
RoBERTa [Liu et al., 2019]	RoBERTa, an extension of BERT, refines language understanding by optimizing training techniques, removing sentence order prediction, and leveraging large-scale data. It enhances pre-training for improved performance across various language tasks, achieving state-of-the-art results in natural language processing.	Kim et al. [2023b], Liu et al. [2022a], Kim and Johnson [2022], Smădu et al. [2023]

Continued on next page

Algorithm	Description	Reference
BERT [Devlin et al., 2018]	BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model by Google, adept at understanding the context in both directions of a sentence. It utilizes Transformer architecture, enabling versatile language understanding for diverse tasks like question answering and natural language understanding.	Ruan et al. [2021], Lee et al. [2019], Kim and Johnson [2022], Sánchez-Junquera et al. [2021], Da Silva and Paraboni [2023], Baly et al. [2020], Shaprin et al. [2019], Ahmed et al. [2023b], Shaprin et al. [2019], Drissi et al. [2019], Lyu et al. [2024a], Tran [2020], Huang and Lee [2019], Ahmed et al. [2023a], Mutlu et al. [2019], Ning et al. [2019], Sánchez-Junquera [2021], Omidi Shayegan et al. [2024]

Deep Learning: Non-Transformer-based Architectures

Pérez-Almendros et al. [2019] employed a fusion of CNN and LSTM, utilizing quantitative linguistic features extracted through GloVe. In this way, they highlighted the crucial role of

incorporating linguistic features alongside representations based on word vectors. Additionally, they built a meta-classifier to filter noisy data to apply to the by-publisher dataset. Jiang et al. [2019] won the SemEval-2019 Task 4 by combining rich morphological and contextual representations by averaging the three vectors per word into ELMo embeddings. Their model was used for further studies by Ruan et al. [2021] for pseudo-labeling frameworks: Overlap-checking and Meta-learning. Overlap-checking consists of adding data, helping the model train, while Meta-learning allows the model to be continually trained on a clean dataset and a pseudo dataset. This last work inspired Cruz et al. [2020]. In their article, they used a HAN combined with ELMo embeddings. The HAN is a model capable of balancing the information in a current state, deciding whether to update it and how much the past information contributes to its new state. In this case, the information stems from the sentence level, confirming that richer article representations yield better performances. By encapsulating the articles' structure, connectors and paying attention to stylistic markers, handcrafted stylistic features and emotion lexicons, they reached the state-of-the-art in 2020 on the SemEval-2019 Task 4 dataset. Ko et al. [2023] improved the HAN standard model by introducing Knowledge Encoding (KE) components. The HAN segment functions to grasp word and sentence relationships within a news article, employing a structured hierarchy across three levels—word, sentence, and title. Meanwhile, the KE component integrates common and political knowledge associated with real-world entities into the prediction process for determining the political stance of the news article. Since the model is not language-based, it could work with diverse languages beyond English. Li and Goldwasser [2021] developed a pre-training framework encoding knowledge about entity mentions, namely masked tokens as frame indicators, and modeling the propagation between users with a social information graph. They noted that models pre-trained on general sources and tasks have limited ability to focus on biased text segments. Cramerus and Scheffler [2019] introduced a voting system of LSTMs to build a controlled dataset to train another LSTM. It was an example to demonstrate the importance of having a balanced and clean dataset to run experiments. Lastly, Roy and Goldwasser [2020] built a Hierarchical-LSTM applied to subframes (n-grams) to tackle the framing bias. In this paper, they introduced a pioneering framework aimed at pretraining text models utilizing signals derived from the abundant social and linguistic context available, encompassing elements such as entity mentions, news dissemination, and frame indicators.

Deep Learning: Transformer-based Architetures

Regarding the Transformers architectures, we observed a massive utilization of BERTbase and BERTlarge. BERTbase is a pre-trained BERT model trained on a smaller dataset than BERTlarge. BERTbase differentiates itself in cased and uncased, depending on whether to discern between cased and uncased words. Baly et al. [2020] wanted to remove the bias when modeling the medium. They observed that combining bias mitigation with triplet loss, Twitter bios, and media-level representations increased the model's efficacy. Kim and Johnson [2022] proposed a multi-task BERT-based model with contrastive learning to tackle framing bias in news articles. Da Silva and Paraboni [2023] with BERT and combinations of syntactic bigram counts and psycholinguistic features investigated the inference of political information and hyperpartisanship on the author and text level, starting from linguistic data. Shaprin et al. [2019] showed that fine-tuning the model entails better results. Ahmed et al. [2023b] introduced a semi-supervised framework trained using federated learning, namely, algorithms are trained independently across diverse datasets. Furthermore, textual data are tagged to extrapolate wh-questions replies and temporal lexicon information. The same author replicated this approach in Ahmed et al. [2023a]. In the quest for precise detection and data denoising, the same author replicated this approach with variations in Ahmed et al. [2023a]. Ahmed et al. [2023a] employed an attention-based strategy to learn text representation, aiming to identify target expressions accurately while extracting pertinent contextual information. They generated a BERT attention embedding query utilizing lexicon expansion, content segmentation and temporal event analysis. Ultimately, this approach enhances the understanding of consecutive news articles within a temporal framework. Drissi et al. [2019] experimented using BERTbase and BERTlarge, feeding them with embeddings of different lengths. They were interested in analyzing the parts of the articles, looking for a consistent level of hyperpartisanship that demonstrated its existence. Huang and Lee [2019] from the confrontation between BERT and ELMo models, confirmed that the inputs and embedding dimensions contributed to affecting the performance positively. Mutlu et al. [2019] performed domain adaptation, showing its efficacy. Kim et al. [2023b] operated in a low-resource scenario with prompt-based learning and employed masked political phrase prediction and a frozen pre-trained language model that relies on the transformer architecture, utilizing the robustly optimized BERT approach known as RoBERTa as a backbone for their own model, MP-tuning. Liu et al. [2022a] focuses on political ideology and stance detection, comparing triplets of documents on the same history to detect dissimilarities amongst them. They trained RoBERTa through continual learning. Whereas

Smădu et al. [2023] improved their model's performance with cross-domain contrastive learning and this work is noticeable that they used GPT-2 for augmenting hyperpartisan textual data. Lastly, Omidi Shayegan et al. [2024] faced the task of Persian hyperpartisan tweets by prompting GPT-3.5, a multi-language conversational generative LLM released in 2022, and an open-weight model like Llama2 Touvron et al. [2023b]. Omidi Shayegan et al. [2024] compared the capabilities of Large Language Models (LLMs) and BERT-based models like RoBERTa and ParseBERT to detect English and Persian tweets, providing instructions with different levels of specificity to the models. Despite the huge dimension and the extensive training of LLMs, fine-tuning ParseBERT and RoBERTa has proven to be more efficient and practical for certain tasks.

Other Methods

Within the vast landscape of computational frameworks, certain algorithms defy classification within the traditional realms of deep learning or non-deep learning. This chapter delves into the exploration of these unique frameworks—sophisticated combinations of diverse models, labeling techniques and graph approaches—that operate beyond the conventional boundaries of established categorizations.

Tran [2020] applies a framework for presentation bias, studying hyperpartisanship with a graph-based method. This three-step framework is so structured: collecting related-articles clusters on the same topic; applying Aspect-based Sentiment Analysis (ABSA) with BERTbase to rate and classify fine-grained opinions in the pairs of sentences; the variation in bias between news sources within similar categories is figured out by contrasting the scores of matching pairs of articles. This comparison is done for every combination of news sources within these categories, and the differences in bias are averaged across all article groups. This averaging process leads to the development of a bias matrix. Kulkarni et al. [2018] proposed a Multi-View Document Attention Model (MVDAM) capable of modeling at the same time title, structure and metadata like links in order to estimate the political ideology of a news article. This framework, based on the Bayesian approach, utilizes different models for creating the 3-D representation: a convolutional neural network for learning the title, Node2Vec for the network and HAN for the content. Afsarmanesh et al. [2019] worked mostly on manual features like metatopic, namely polarizing topics, using an end-to-end tool: The Gavagai Explorer, which performed poorly. The IMPED model from Bastos et al. [2021] utilizes language and time-based structures found in social media posts to predict the probability of content removal

online. It achieves this by employing a parametric survival model. This model operates under the assumption that poor quality content displays distinctive traits, enabling its classification on a large scale. Patankar et al. [2019] delved into three pivotal areas to assess a bias-ranking system's efficacy within news articles. Researchers initially analyzed the correlation between user-generated and algorithm-generated bias ranks to gauge the system's alignment with user perceptions. Subsequently, subjective user reviews indicated widespread potential for the news recommendation system, albeit with concerns about personal political awareness influencing biased judgment. Lastly, bias scores across diverse news article types and sources unveiled significant variations, spanning Wikipedia featured articles, non-featured articles, blogs, and current news pieces. These findings substantially deepen comprehension of user perceptions, system usability, and its effectiveness in delivering balanced news. They provide crucial insights for refining bias-aware news recommendation systems, aiming to offer more equitable information to users. Potthast et al. [2018b] performed a political orientation prediction and hyperpartisan classification task using an unmasking technique with binary classifiers. For the first task, they found that left-wing news tends to be easily misclassified. This study noticed that individual political orientation struggles to predict, and that a style-based approach overcomes the content-based one. Moreover, they discovered subtle differences in style between hyperpartisan news belonging to different political leanings. Sánchez-Junquera et al. [2021] using masking and transformer-based models proved that topic-based approaches lead to better results than style-based. Instead, Sánchez-Junquera [2021] made a comparative examination of BERT-based models and masking-based models, enriching comprehension regarding the strengths and constraints of varied approaches in bias detection, offering crucial insights for upcoming research and advancements in this domain. In essence, these models' contribution lies in their capacity to augment the precision, clarity, and comprehensibility of bias detection within political and social discussions. Consequently, they propel advancements in this pivotal research domain.

Table 3.10: Descriptive results for the papers examined, reporting the most used measures: accuracy and F1 score. Part 1.

Reference	Dataset	Acc.	F1
Potthast et al. [2018b]	The BuzzFeed-Webis Fake News Corpus 2016	0.75	0.78
Sridharan and S. [2022]	Own	0.45	U/A
Ruan et al. [2021]	Semeval-2019	0.87	0.81
Papadopoulou et al. [2019]	Semeval-2019	0.61	0.71
Pérez-Almendros et al. [2019]	Semeval-2019	0.74	0.71
Gupta et al. [2019]	Semeval-2019	0.55	0.28
Kim and Johnson [2022]	Framing Triplet Dataset	0.84	U/A
Lyu et al. [2024a]	Own	0.84	0.78
Naredla and Adedoyin [2022]	Semeval-2019	0.88	U/A
Isbister and Johansson [2019]	Semeval-2019	0.80	0.80
Agerri [2019]	Semeval-2019	0.74	0.73
Sengupta and Pedersen [2019]	Semeval-2019	0.70	0.68
Dumitru and Rebedea [2019]	Own	0.93	0.81
Chakravartula et al. [2019]	Semeval-2019	0.59	0.69
Aksenov et al. [2021]	GERMAN Dataset	U/A	0.79
Smădu et al. [2023]	Semeval-2019	0.64	0.69
Chen et al. [2019]	Semeval-2019 Task 4 by-article	0.74	0.74
Drissi et al. [2019]	Semeval-2019 Task 4 by-article	0.77	0.75
Amason et al. [2019]	Semeval-2019 Task 4 by-article	0.65	0.73
Tran [2020]	PoliNews	U/A	U/A
Huang and Lee [2019]	Semeval-2019	0.68	U/A
Gerald Ki Wei and Jun Choi [2021]	The BuzzFeed-Webis Fake News Corpus 2016	0.73	U/A
Sharma et al. [2020]	Own	0.84	0.87
Ko et al. [2023]	Semeval-2019	0.95	U/A
Sánchez-Junquera et al. [2021]	The BuzzFeed-Webis Fake News Corpus 2016	0.89	0.86
Kim et al. [2023b]	Semeval-2019	0.91	U/A
Kulkarni et al. [2018]	Own	0.80	0.79
Nguyen et al. [2019]	Semeval-2019	0.75	0.74
Azizov et al. [2023]	Task 3A	0.69	0.69
Cruz et al. [2020]	Semeval-2019	0.82	0.81
Sánchez-Junquera [2021]	Stereoimmigrants	0.86	0.83

U/A: unavailable.

Table 3.11: Descriptive results for the papers examined, reporting the most used measures: accuracy and F1 score. Part 2.

Reference	Dataset	Acc.	F1
Knauth [2019]	Semeval-2019	0.67	0.69
M. Alzhrani [2022]	Presidential	0.91	0.90
Da Silva and Paraboni [2023]	Semeval-2019	0.78	0.77
Liu et al. [2022a]	Semeval-2019	0.71	0.77
Garg and Sharma [2022]	Reuter	U/A	0.93
Moreno et al. [2019]	Semeval-2019	0.72	0.69
Ahmed et al. [2023a]	Semeval-2019	U/A	0.91
Alabdulkarim and Alhindi [2019]	Semeval-2019	0.74	0.71
Joo and Hwang [2019]	Semeval-2019	0.74	0.70
Palić et al. [2019]	Semeval-2019	0.79	0.76
Jiang et al. [2019]	Semeval-2019	0.82	0.81
Cruz et al. [2019]	Semeval-2019	0.72	0.67
Afsarmanesh et al. [2019]	Semeval-2019	0.56	0.68
Mutlu et al. [2019]	Semeval-2019	0.78	0.76
Shaprin et al. [2019]	Semeval-2019	0.64	0.64
Anthonio and Kloppenburg [2019]	Semeval-2019	0.62	0.69
Cramerus and Scheffler [2019]	Semeval-2019	0.58	0.68
Stevanoski and Gievska [2019]	Semeval-2019	0.77	0.74
Färber et al. [2019]	Semeval-2019	0.60	0.70
Ning et al. [2019]	Semeval-2019	0.50	0.61
Saleh et al. [2019]	Semeval-2019	0.73	0.73
Zehe et al. [2019]	Semeval-2019	0.67	0.74
Lee et al. [2019]	Semeval-2019	0.62	0.69
Gangula et al. [2019]	Telugu	0.89	U/A
Ahmed et al. [2023b]	Semeval-2019	0.83	U/A
Hanawa et al. [2019]	Semeval-2019	0.81	0.80
Bestgen [2019]	Semeval-2019	0.70	0.68
Yeh et al. [2019]	Semeval-2019	0.80	0.79
Zhang et al. [2019]	Semeval-2019	0.68	0.63
Li and Goldwasser [2021]	Semeval-2019	0.86	0.84
Srivastava et al. [2019]	Semeval-2019	0.82	0.82
Baly et al. [2020]	Own	0.72	U/A
Roy and Goldwasser [2020]	Own	U/A	0.80
Omidi Shayegan et al. [2024]	Own	0.85	0.85

U/A: unavailable.

3.2.2 In-Context Learning

The emergence of Large Language Models has introduced new paradigms for hyperpartisan detection that move beyond traditional fine-tuning approaches. With the advent of recent decoder-only LLMs, In-Context Learning (ICL) has emerged as a valuable technique in NLP. Users interact with models directly through prompts, specific textual templates containing instructions and optionally examples [Brown et al., 2020]. This approach allows models to perform tasks without prior task-specific fine-tuning [Efrat and Levy, 2020], leveraging a single pre-trained model for various downstream tasks and enabling desired behavior specification via natural language. ICL has shown remarkable performance on challenging reasoning tasks [Brown et al., 2020, Wei et al., 2022]. However, ICL is highly sensitive to input format and order [Lu et al., 2022, Min et al., 2022], and can lead to irreproducible outcomes as slight prompt changes significantly impact performance [Lu et al., 2024, Sun et al., 2023]. To overcome these limitations, our work comprehensively tested various prompt strategies and demonstration selection methods, including DPP for more stable performance, and introduced codebook prompting. Research on prompt design aims to elicit better performance and reasoning. Notable approaches include CoT prompting [Kojima et al., 2022], which encourages step-by-step reasoning, and its zero-shot variant [Wei et al., 2022]. Lu et al. [2022] also highlighted the importance of careful prompt format and example selection in few-shot learning. Regarding the comparison of ICL and fine-tuning, Labrak et al. [2024] and Edwards and Camacho-Collados [2024] demonstrated that fine-tuning smaller models often outperforms ICL in larger language models across various NLP and text classification tasks. Aligned with these findings, our study expands this comparison by evaluating a broader range of prompt strategies and models (including ModernBERT [Warner et al., 2024] beyond Llama models), and by focusing on specialized domains to rigorously compare the efficacy of these tools. An example of this specific strategy is providing the model with a codebook, which provides definitions of categories, including examples and classification instructions. For instance, Vincent and Mestre [2018] developed a codebook to classify hyperpartisan news on a 5-point scale, and Hughes et al. [2021] crafted one for content analysis of COVID-19 articles, classifying tropes and rhetorical strategies to detect misinformation. Codebooks offer a method to explicitly prompt LLMs with structured contexts, eliciting their rule-based reasoning capabilities, which go beyond standard ICL that often relies primarily on examples. Hu et al. [2024] explored codebook application in zero-shot settings with closed models for political phenomena classification, using the codebook as a structured framework for interpretation. Similarly,

Halterman and Keith [2025] utilized codebooks to evaluate open models, demonstrating how these guidelines facilitate assessing an LLM's adherence to predefined classification logic. While these specific codebooks were tailored to different political phenomena, NLP tasks, or modeled tasks differently from our dataset intentions, thus not directly applicable to our study, they still served as valuable inspiration for our experimental setup and underscored the potential of integrating structured rule sets into LLM prompts. This enhances adherence to specific classification schemes, aligning with broader prompt engineering efforts to elicit precise and controlled LLM outputs, especially in domains requiring nuanced categorization criteria.

3.2.3 Rhetorical Bias and Persuasion Techniques

Closely related to hyperpartisan detection is the identification of rhetorical strategies employed to persuade readers. Da San Martino et al. [2019] represents a foundational work in this domain, 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 was then followed by Piskorski et al. [2023c], 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.

3.2.4 Linguistic Trait Detection

Beyond explicit propaganda techniques, hyperpartisan content often employs subtler linguistic devices that require specialized detection methods. 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 tasks 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 a 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.

3.2.5 Domain-Specific Datasets

The development of robust detection systems depends critically on the availability of high-quality annotated datasets. The growing presence of misinformation and polarized political discourse has driven extensive efforts to curate datasets for detection tasks. Prior work has focused on identifying conspiracy theories [Pogorelov et al., 2021, Miani et al., 2021, Langguth et al., 2023, Kim et al., 2023a], propaganda, and hyperpartisan content [Golbeck et al., 2018, Kiesel et al., 2019b, Vogel and Jiang, 2019, Hossain et al., 2020]. A related and well-established area is stance detection, which classifies whether a text expresses support, opposition, or neutrality toward a specific claim or topic [Mohammad et al., 2016, 2017, Küçük and Can, 2022]. Despite these advances, many existing resources remain limited in both linguistic and topical coverage. Most high-quality datasets are in English and focus on U.S.-centric themes, such as domestic politics, the COVID-19 pandemic, or climate change. While recent studies have begun to explore multilingual misinformation [Nielsen and McConville, 2022, Leite et al., 2024, Mohtaj et al., 2024], important gaps persist. In particular, there is still a shortage of high-quality multilingual datasets that reflect the European context and jointly address hyperpartisanship, stance, and rhetorical bias as interconnected elements of political discourse. Several datasets address polarizing topics across languages [Kurfali et al., 2025]. Piskorski et al. [2023a]'s SemEval 2023 Task 3 covers COVID-19, climate change, migration, and the Ukraine war across six languages, including Italian, and it is specific to 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., 2025c]. 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.

3.2.6 Explainability and Embedding-Based Enhancements

As detection models grow increasingly complex, understanding their decision-making processes becomes paramount for both scientific validity and practical deployment. XAI methods, such as SHAP for feature selection [Bangerter et al., 2023a] 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., 2019b]. We advance this by comparing IG and SHAP attributions for LT versus non-LT tokens in HP classification.

3.2.7 LLMs as Annotators and Persona Prompting

The labor-intensive nature of manual annotation has motivated exploration of LLMs as automated annotators. Recent advances show that political science has increasingly adopted supervised machine learning as a powerful tool for large-scale analysis of political text [Horych et al., 2025, Stromer-Galley et al., 2025]. Within this context, LLM-based classification offers a promising way to reduce manual annotation effort and costs while maintaining high levels of accuracy. Törnberg [2024] adopted LLMs to annotate messages from the social media platform *X* based on the political belonging of the poster across 11 languages (only one from the Romance group) using the Twitter Parliamentarian Database [Van Vliet et al., 2020]. When comparing their results with the gold-standard labels, they found that LLMs outperformed expert coders and fine-tuned baselines like BERT. However, results on public

corpora may be inflated by pretraining contamination (data leakage) if models were exposed to the same sources during pretraining [Sainz et al., 2023]. Heseltine and von Hohenberg [2024] report that GPT-4 performance drops for more abstract tasks (e.g., ideology detection) and for longer inputs (full articles), while simple binary decisions yield higher human–model agreement; they also note lower accuracy and increased latency for non-English texts. Finally, the implementation choices made by researchers while using LLMs as annotators influence the results, introducing potential systematic biases and random errors [Baumann et al., 2025]. Therefore, further investigation is needed to understand how incorporating additional inputs like sociodemographic data might enhance the performance of LLMs as annotators in subjective NLP tasks [Beck et al., 2024].

3.2.8 Text Style Transfer: Neutralization of Biased Texts

Beyond detection, an emerging research direction focuses on transforming hyperpartisan content into neutral alternatives. Text debiasing and neutralization strategies range from fine-grained classification [Maggini et al., 2025b] and counter-example generation [Sermsri and Panboonyuen, 2025] to reinforcement learning using embedding or classifier rewards [Liu et al., 2021a,b]. Recently, LLM-based approaches have dominated the field [Fernández et al., 2024, Juvino Santos et al., 2025, Dale et al., 2021, Sermsri and Panboonyuen, 2025]. Political neutralization is closely related to Text-Style Transfer (TST), typically involving word removal or synonym replacement [De Ruvo et al., 2025]. However, political discourse involves complex linguistic layers—including semantics, syntax, and pragmatics [Maggini et al., 2025b]—and rhetorical strategies like irony or cherry-picking [Maggini et al., 2025d]. Despite progress in detoxification and depolarization [Toshevska and Gievska, 2025, Dementieva et al., 2024b], the task is hindered by resource scarcity. Existing parallel datasets include the Wiki Neutrality Corpus [Pryzant et al., 2020], `PARADETOX` for detoxification [Logacheva et al., 2022], and Reddit-based offensive language corpora [Atwell et al., 2022]. While multilingual efforts like MultiParaDetox [Dementieva et al., 2024a] aim to automate collection, Italian remains under-represented. Aside from `DETOXIFY-IT` for toxic comments [De Ruvo et al., 2025], Par-ITA is the first human-curated parallel corpus for Italian hyperpartisan neutralization.

3.3 Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning

Task Formulation

The core objective of this study is to evaluate the effectiveness of various architectures and models in hyperpartisan (HP), fake news, political bias and harmful tweet detection, by comparing two widely used approaches: FT and ICL, where models are tested as off-the-shelf tools without additional training. Specifically, we focus on tasks such as identifying hyperpartisan and fake news, harmful tweets and a news' political leaning, recognizing the distinct linguistic and contextual challenges each task presents. Our approach encompasses both binary and multi-class classification scenarios, leveraging a variety of datasets and multilingual contexts. While the ultimate goal is not to develop production-ready models, we prioritize thorough experimentation with various transformer-based architectures, prompt strategies, and learning techniques. This exploration serves to highlight the strengths and limitations of the tested architectures, contributing to the broader effort of refining misinformation detection methodologies within the NLP community. We acknowledge the fact that hyperpartisan news shows peculiar stylistic traits, rather than fake news [Potthast et al., 2018b].

Datasets

For our experiments, we selected datasets for both binary and multiclass classification tasks. The 10 datasets focus on articles (see Table 3.12), headlines, tweets on COVID-19 and political news and different topics including TV, politics, sports, and health. They cover two types of domains: news and Twitter, and include four specific-oriented classification tasks: hyperpartisan and fake news, harmful tweet and political bias detection. For hyperpartisan detection, we selected the **SemEval-2019 Task 4 by-article** dataset [Kiesel et al., 2019a], which contains articles from hyperpartisan and mainstream websites annotated by three annotators. They mostly cover the first term of Trump, Gun Control and other U.S.-centric related topics. The dataset's strength lies in its article-level annotations that allow for analysis of extended argumentative structures and narrative techniques typical of hyperpartisan content, rather than just isolated claims. The **VIStA-H** dataset [Lyu et al., 2023] includes hyperpartisan and neutral news headlines from right, center and left U.S. newspapers. Its focus on headlines rather than full articles

complements the SemEval dataset. The dataset's temporal coverage, from 2014 to 2023, is particularly valuable for tracking how hyperpartisan news evolved through multiple election cycles. The **Fake News Net** dataset [Shu et al., 2017] contains news articles shared on Twitter, allowing for analysis of how hyperpartisan content circulates in social media environments. The **Spanish Fake News Corpus** [Gómez-Adorno et al., 2021] gathers news from Spanish newspapers and media company websites, along with fact-checking websites. This corpus was selected to broaden the linguistic and cultural scope of the research beyond English-language and U.S.-centric media. The incorporation of fact-checking websites provides an additional layer of verification that strengthens the dataset's reliability. The **Fake.br Corpus** [Monteiro et al., 2018] focuses on Brazilian Portuguese manually collected and checked news. The inclusion of this Brazilian Portuguese corpus further expands the cross-cultural and multilingual dimensions of the research. Brazil's distinct political landscape and media environment provide an important comparative case for testing the generalizability of the detection approaches. The manual verification process strengthens the dataset's reliability as a benchmark for testing detection methods in a language where NLP resources might be less abundant than for English or Spanish. **CLEF 2022 CheckThat! Lab Subtask 1C** [Nakov et al., 2022] was selected for its focus on COVID-19 misinformation tweets across multiple languages (Arabic, Bulgarian, and English), providing an opportunity to study fake news content around a global crisis that transcended national boundaries. Regarding political bias detection, we considered the **Qbias** dataset [Haak and Schaer, 2023], which contains articles from AllSides, a news aggregator with an established methodology for evaluating political leaning. This provides a more nuanced and professionally curated ground truth for political bias than many other available resources. This nuanced approach helps move beyond binary classifications of political content and supports more sophisticated analysis of bias indicators. Lastly, **CLEF 2023 CheckThat! Lab Task 3A** dataset [Azizov and Nakov, 2023] provides contemporary examples that reflect the current state of political communication and media bias. This diverse collection of datasets provides a comprehensive foundation for developing and evaluating models across multiple languages, cultural contexts, media formats and misinformation tasks. To produce input for the classifier in the Spanish Fake News Corpus and Qbias datasets, we concatenated the headline and the body of the article.

Dataset	Abbr.	Lang.	Timeframe	Train Size	Test Size	Domain	Type	Task	Label Ratios (Train / Test)
VIStA-H [Lyu et al., 2023]	HV	en	2014–2023	1999	201	News	S	HP	HP: .39/.63; N: .50/.50
SemEval-2019 by-article [Kiesel et al., 2019a]	SH	en	2007–N/A	645	628	News	D	HP	HP: .50/.50; N: .50/.50
Spanish Fake News Corpus [Gómez-Adorno et al., 2021]	SFN	es	2020–2021	676	572	News	D	FN	T: .50/.50; F: .50/.50
Fake News Net [Shu et al., 2017]	FNN	en	N/A	18556	4640	News	D	FN	T: .74/.74; F: .26/.26
Fake.br Corpus [Monteiro et al., 2018]	FBC	pt	2016–2018	5760	1440	News	D	FN	T: .50/.50; F: .50/.50
CLEF 2022 1C [Nakov et al., 2022]	C1A	ar	N/A	3624	1201	Twitter	S	HT	NH: .81/.84; H: .19/.16
	C1B	bu	N/A	708	325	Twitter	S	HT	NH: .87/.97; H: .13/.03
	C1E	en	N/A	3323	251	Twitter	S	HT	NH: .91/.84; H: .09/.16
CLEF 2023 3A [Azizov and Nakov, 2023]	C3A	en	N/A	45066	5198	News	D	PB	R: .39/.13; C: .34/.38; L: .27/.50
Qbias [Haak and Schaer, 2023]	QB	en	2012–2022	17403	4351	News	D	PB	R: .39/.13; C: .34/.38; L: .27/.50

Table 3.12: Description of the datasets used in our experiments. Dataset types: S = Sentence, D = Document. Tasks: HP = Hyperpartisan News Detection; FN = Fake News Detection; HT = Harmful Tweet Detection; PB = Political Bias Detection. Labels: HP = Hyperpartisan, N = Neutral; T/F = True/Fake; NH/H = Non-harmful/Harmful; R/C/L = Right/Center/Left. Abbr. contains the abbreviation we will use in this paper to refer to the datasets.

3.3.1 Models

For our experiment, we compared two types of model architectures: encoder-only and decoder-only models.

Encoder-Only Masked Language Models: Following Edwards and Camacho-Collados [2024], we selected BERT-derived models such as RoBERTa-base (125 million parameters) and RoBERTa-large (354 million parameters) [Liu, 2019], XLM-RoBERTa [Conneau et al., 2019], POLITICS [Liu et al., 2022b], which has been adapted for the political domain using continuous pre-training, and mDeBERTaV3 [He et al., 2021]. RoBERTa is pre-trained on English data, while XLM-RoBERTa is trained on 100 different languages, making it suitable for evaluating the impact of multilingual training on performance.

Decoder-Only Large Language Models: We experiment using various small-size open-weight LLMs: Llama-3.1-8b and Llama-3.1-8b-Instruct, Mistral-Nemo-Instruct-2407 [Mistral, 2024], Qwen2.5-7B-Instruct [Qwen et al., 2025]. These models are decoder-only and testing them allows us to observe generalizable effects across model families and tasks. The temperature was set to 0 for all the experiments. Generally, for non-English datasets, we evaluated models exclusively trained on multilingual datasets to ensure appropriate language coverage and performance.

3.3.2 Prompt design

Earlier studies like Wei et al. [2022], Jung et al. [2022] and Mishra et al. [2022] have demonstrated the effectiveness of using task-specific prompts. Therefore, following Edwards and

Camacho-Collados [2024] and Labrak et al. [2024], we constructed the prompts by concatenating the following elements: 1) an instruction detailing the task, domain, and describing the meaning of the label; 2) the input argument, supplying essential information for the task; 3) the constraints on the output space, guiding the model during output generation. To improve the coherence, the specificity of the prompts, the instructions to follow in the codebook, and the fine-grained reasoning in CoT for the political domain, we collaborated with an expert in Political Science. In particular, to structure and develop our codebooks, we were inspired by Vincent and Mestre [2018] and Halterman and Keith [2025], which introduced clear task definitions, explicit and exhaustive rules to determine the label of a data point, as well as provided examples covering both correct and borderline cases. We tested different prompting and ICL strategies such as zero-shot, Few-Shot, codebook and a variant of guided CoT, interpreting the reasoning as a multi-task evaluation [Lee et al., 2024, Duan et al., 2024], to provide explainable results like in Yang et al. [2023a]. We compare the random selection of Few-Shot exemplars with a more diversified selection using DPP, for which we provide a brief introduction in the following section. We also compare the results given by prompting the models with instructions containing different levels of complexity: general instructions, specific definitions of political phenomena or specialized instructions with more context provided. During the prompt optimization phase, we placed particular emphasis on ensuring that the model adhered to a consistent label format. This was crucial to ensure the outputs were reliably parseable. For instance, we discovered that the models, when asked to provide string labels, generated a few unparseable outputs, considered wrong at the time of inference. This behavior occurred across all the models and configurations. Moreover, initially we tested Mistral-7B and it frequently not followed the instructions regarding the template. This is why we did not introduce it in the experimental setting. Lastly, in CoT, to ensure its proper functioning, we made sure the models generated the thoughts before the final output.

3.3.3 Determinantal Point Process

Determinantal Point Process is a probability distribution over a cloud of points that are used as a computational tool across the fields of physics, statistics and machine learning [Gautier et al., 2019]. DPP has been used to select a diverse and representative set of data points for in-context learning [Yang et al., 2023b], data annotation [Wang et al., 2024b], instruction tuning [Wang et al., 2024a] and pre-training [Yang et al., 2024]. DPP has been preferred for these tasks because it helps in promoting efficiency while maintaining diversity of the selected subset from

a large set. To illustrate, let us consider two sets called the index set $A = \{1, 2, \ldots N\}$ and its corresponding item set $I_A = \{x_1, x_2, \ldots, x_N\}$. Then the problem of subset selection becomes evaluating 2^M subsets, which is computationally intractable and combinatorially explosive as the size of the super set grows. In order to approximately solve this problem, DPP first uses the representation of the data $\mathbf{x_i}$. We use Sentence-BERT [Reimers and Gurevych, 2019] to calculate the representations or embeddings. After the representations are computed, DPP algorithm works by calculating a Kernel $\mathbf{K}_{ij} = k(\mathbf{x_i}, \mathbf{x_j})$ where kernel function k can be any similarity or distance metric between 2 points. Based on this, we select a subset $Y \in A$, the probability selection of Y is given by.

$$P(Y) = \frac{det(\mathbf{K}_Y)}{det(\mathbf{K} + \mathbf{I})} \tag{3.1}$$

Here $\mathbf{K}_Y$ is the subset of the matrix $\mathbf{K}$ and consists of $\mathbf{K}_{ij}$ for $i, j \in Y$. $\mathbf{I}$ is the identity matrix and $det(\cdot)$ represents the determinant of a matrix. Under these conditions, the selection of the best subset can be formulated as the optimization problem as follows:

$$Y_{best} = argmax_{Y \subset A, |Y|=k} det(L_Y) \tag{3.2}$$

Algorithms exist to select such subsets of size k from the superset by sampling from the posterior distribution. For the task of selection we use the exact sampler [Gautier et al., 2019, Mazoyer et al., 2020] which is the part of DPPy package by [Gautier et al., 2019] and is faster than Monte-Carlo sampler [Bardenet and Hardy, 2019]. Since the algorithm works on sampling at the kernel level it helps in selecting data points which are diverse in the representation space.

3.3.4 ICL Setting

We aim to compare the capabilities of different learning techniques, namely FT and ICL, and model architectures for hyperpartisan, fake news, harmful tweet and political bias classification. To investigate the ability of LLMs on those tasks in ICL, we used Llama-3.1-8b-Instruct, Mistral-Nemo-Instruct-2407 and Qwen2.5-7B-Instruct by prompting them with different setups: 0-shot with General Prompt, 0-shot with Specific Prompt, CoT and Few-Shot with k-shot where k>0. In the k-shot configuration, we adopted the General Prompt along with random examples and the respective labels of the dataset. To further test the stability of the prompting, we used Determinantal Point Process [Gautier et al., 2019] to select a diverse set

of data points for each of the k-shot settings. The general prompt template is: <Role> <Task description> <Definition or Instructions> <Text to classify *or* examples followed by text to classify> <Response format>. We provide examples of our different prompts for PL detection in Table 3.13.

Table 3.13: Table showing different examples of instruction for PL Detection.

Prompt	**Text**
zero shot generic	"You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner:\n\nFinal Answer ==>left\n\nor \n\nFinal Answer ==>center \n\nFinal Answer ==>right\n\n Here's the text: "
zero shot generic	"You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Left-wing ideologies and moral values often include support for social equality, environmentalism, social justice, labor rights, and progressive taxation. Center ideologies and moral values often focus on a balance between market and social justice, pragmatism, and incremental reform. Right-wing ideologies and moral values often emphasize tradition, free market capitalism, individualism, and a limited role of government. Detect the political leaning of the headline. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner:\n\nFinal Answer ==>left\n\nor \n\nFinal Answer ==>center \n\nFinal Answer ==>right. \n\n Here's the text: ",

Table 3.13 – continued from previous page

Prompt	**Text**
few shot	”You are an expert in political domain. You will be provided with a text and some labeled examples, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: ’left’ if the article is left-wing, ’center’ if the political leaning is from the center, ’right’ if the article is right-wing’. Make sure to structure your answer in the following manner:\n\nFinal Answer ==>left\n\nor \n\nFinal Answer ==>center \n\nFinal Answer ==>right\n\n Here’s the text: ”,
codebook	”You are an expert in political domain. You will be provided with a codebook to follow and a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. *Definition: Left-wing ideologies and moral values often include support for social equality, environmentalism, social justice, labor rights, and progressive taxation. Center ideologies and moral values often focus on a balance between market and social justice, pragmatism, and incremental reform. Right-wing ideologies and moral values often emphasize tradition, free market capitalism, individualism, and a limited role of government. *Task: The possible choices are: ’left’ if the article is left-wing, ’center’ if the political leaning is from the center, ’right’ if the article is right-wing’. Make sure to structure your answer in the following manner:\n\nFinal Answer ==>[left or center or right] \n\n Here’s the text: ”,

Table 3.13 – continued from previous page

Prompt	Text
cot	”You are an expert in political domain. You will be provided with a text, and your task is to determine, to the best of your judgment, whether the text is left-, right-wing or neutral. If you are unsure, make the most informed decision you can based on the content. Determine the political leaning of the headline. The possible choices are: 'left' if the article is left-wing, 'center' if the political leaning is from the center, 'right' if the article is right-wing'. Make sure to structure your answer in the following manner:\n\nFinal Answer ==>[left or center or right] Reason step by step: 1. Identify key words and phrases: list of words or phrases that might indicate political bias. 2. Sentiment analysis: Analyze the tone and the language to see if there are polarizing words, or emotional language. 3. Are there any rhetorical devices used? (e.g., hyperbole, loaded questions). 4. Look for policy implications: Does the headline suggest support for or opposition to specific policies? Are there any implicit assumptions about political or economic systems? 5. Identify target audience: Who seems to be the intended audience for this headline? 6. Left-wing hyperpartisan: Consider you are a left-wing reader. Would you consider this article as hyperpartisan to your political stance? [Follow the instructions from 1 to 5.] 7. Right-wing hyperpartisan: Consider you are a right-wing reader. Would you consider this article as hyperpartisan to your political stance? [Follow the instructions from 1 to 5.] 8. Consider omissions and emphasis: What information is emphasized, and what might be omitted? How might this emphasis or omission indicate bias? 9. Final prediction. The output will be in this format. Remember to use all reasoning steps followed by <end>to write a coincise explanation: Example output: ** Key words: EXPLANATION <end>** Sentiment analysis: EXPLANATION <end>** Policy implications: EXPLANATION <end>* Target audience: EXPLANATION <end>** Left-wing hyperpartisan: EXPLANATION <end>* Right-wing hyperpartisan: EXPLANATION <end>** Omissions and emphasis: EXPLANATION <end>*Final Answer ==>left\n\nor \n\nFinal Answer ==>center \n\nFinal Answer ==>right\n\n Here's the text: ”

In order to maintain a balanced pool of examples, for multi-class datasets, we sampled from 1 to 3 examples per label; otherwise, we extracted the same *k*-shot per class. Finally, for non-English data, the corresponding roles and instructions were provided in the respective language to ensure accuracy and contextual relevance. The types of prompts we used are the following:

General Prompt By providing the model with task-specific context (e.g., a headline, article,

or tweet), we prompted it to classify the input text with the appropriate task label. With this configuration, we leverage the internal knowledge of the model to predict the answer, while being aware that it can suffer from political bias [Bang et al., 2024]. We used it in 0-shot and Few-Shot. **Specific Prompt** We slightly changed the previous template, introducing in the instruction the political definition of the phenomenon analyzed and some knowledge regarding the biases in partisan texts and asked the model to classify the text with the correct label. These political definitions were provided by a domain expert. Thus, we insert external knowledge and introduce a political definition to maximize the model's understanding capability and improve the quality of its outputs. We tested its efficacy only in zero-shot. **Codebook** Based on previous works, with the help of a Political Scientist, we crafted a specific codebook for each task. These codebooks contain a definition of the phenomenon, a description of the task's characteristics considering several aspects (e.g., style, narrative) and particular linguistic features (e.g., use of hashtags, tone, source credibility). Furthermore, the detection criteria contain examples to help the LLM understand the specific rules for each task. Crucially, this detailed codebook information was directly embedded within the <Definition or Instructions> component of our prompt template, allowing the LLM to perform rule-based reasoning for classification. This approach enables the models to leverage explicit, structured knowledge during inference, directly addressing the complexities of the tasks. **Guided CoT Prompt** We guided the model to break down its reasoning step by step before making a final classification on the specified context, ensuring it produced explanations for all the steps before the final prediction. Specifically, we divided the hyperpartisan classification task into different sub-tasks: sentiment analysis, rhetorical bias, framing bias, and ideology detection [Maggini and Gamallo Otero, 2024]. Moreover, by asking the model to identify itself with a specific political leaning, we are introducing recursive thinking [Duan et al., 2024]. This method encourages the model to consider multiple factors and clearly articulate its reasoning, potentially leading to more robust and explainable classifications. Lastly, our approach allows us to consider the multidimensionality of each misinformation phenomenon analyzed. By guiding the model through this structured reasoning process, we aimed to reduce misclassification and promote a more nuanced analysis. This approach also enabled us to observe how the model weighs different textual elements in its decision-making process, which helps in identifying any inherent biases or limitations within the model's reasoning. We conducted preliminary tests with various prompts and configurations to refine the ones used in this experiment, ultimately selecting the configurations that yielded the best results on the training set. The optimization

of these prompts was done manually rather than through automated methods. As a result, our prompts vary in terms of length, complexity, task specificity, and domain relevance, providing a comprehensive range of settings for evaluation. This structured and manually optimized approach not only enhances the model's classification performance but also provides deeper insights into the model's interpretability and decision-making process across different political contexts.

Benchmark Experiments: Setup

With this pool of datasets, we tested both Encoder-only and Decoder-only architectures in a multilingual setting for different detection tasks related to mis-/dis-information. This work serves as a methodological backbone for model selection in the following experiments involved in the thesis project. In other words, this work allows us to understand what methods and configurations are the best approach for our research purposes, but also to highlight the limitations (e.g. languages, tasks) of the current tested systems.

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

Design Principle

PartisanLens constitutes the first multilingual resource developed within this thesis and reflects its broader objective of creating European-centered datasets for the study of polarized and potentially harmful online narratives on immigration. The dataset was designed to support the investigation of anti-immigration discourse across Southern Europe, with a particular focus on Spain, Portugal, and Italy. These countries represent a relevant case study due to their geographical position as major entry points for migratory flows into Europe and their consequent exposure to political narratives that exploit immigration as a source of social and ideological polarization [Hallin and Mancini, 2004b, Brüggemann et al., 2014].

The creation of this resource is motivated by two of the central themes of the thesis. First, it addresses the lack of multilingual datasets tailored to European socio-political contexts, reducing the dependence on English-centric resources that may inadequately represent local manifestations of polarization. Second, it adopts a linguistically informed approach to the study of hyperpartisan discourse by moving beyond simple classification labels and incorporating

additional semantic and rhetorical dimensions that contribute to the construction of polarized narratives.

To achieve these objectives, the dataset focuses on two interconnected constructs: *Population Replacement Conspiracy Theory* (PRCT) and hyperpartisanship. The former represents a specific ideological narrative frequently employed in anti-immigration discourse, while the latter captures the broader communicative style characterized by extreme political polarization and rhetorical bias. The simultaneous annotation of these constructs enables the investigation of both the ideological content of headlines and the linguistic strategies through which such content is communicated.

The choice of headlines as the unit of analysis is primarily motivated by the data collection process. The corpus was gathered through *MediaCloud*, which provides large-scale access to news metadata but does not systematically preserve the full content of articles across sources. Nevertheless, headlines constitute a particularly valuable object of study. As the first point of contact between readers and news content, headlines play a crucial role in framing events, attracting attention, and shaping interpretations before the article itself is read. In hyperpartisan and conspiratorial communication, headline writing often concentrates rhetorical strategies into highly compressed linguistic forms, making it possible to observe salient manifestations of polarization even within limited textual contexts.

Beyond the primary labels, the dataset incorporates three rhetorical bias dimensions. These additional annotations provide a richer semantic characterization of the data and allow for a more detailed examination of how hyperpartisan and conspiratorial narratives are linguistically realized across different languages and national contexts. This design choice is consistent with the broader methodological philosophy of the thesis, which seeks to complement predictive models with interpretable linguistic information capable of explaining why particular instances are perceived as hyperpartisan.

A further objective of *PartisanLens* concerns the role of LLMs in annotation workflows. Indeed, we investigate whether these models can reproduce human judgments when evaluating complex socio-political and subjective constructs. To this end, we employ persona prompting strategies that incorporate annotators' sociodemographic characteristics (see Tables 3.16 and 3.15) into the prompting process. This experimental setting allows us to explore the extent to which LLMs can approximate human annotation behavior and whether demographic information influences their decisions.

More specifically, we compare LLM-generated annotations against human agreement lev-

els across the three target dimensions (HP, PRCT, and Stance) in all three languages. Rather than treating LLM outputs as inherently reliable, the study evaluates their performance against the reference standard established by human annotators. The central research question is therefore whether contemporary LLMs can approach human agreement levels when annotating nuanced political and ideological phenomena across multilingual and culturally diverse contexts.

In summary, *PartisanLens* serves multiple purposes within the broader research agenda of this thesis. It provides a multilingual resource grounded in European political realities, supports the investigation of the relationship between hyperpartisanship and anti-immigration conspiracy narratives, enriches the analysis through rhetorical bias annotations, and offers an empirical assessment of the potential and limitations of LLMs as annotation assistants. As the first resource developed in this research program, it establishes several of the methodological principles that are subsequently expanded in the sentence-level and paragraph-level datasets presented in the following sections.

3.4.1 Data collection and preparation

We focused on collecting immigration-related headlines in three underrepresented Romance languages: Spanish, Portuguese and Italian. These represent major linguistic areas along the Mediterranean migration routes into Europe [Aragall, 2018], and yet remain largely absent from resources for misinformation detection. Our data collection followed a hybrid strategy that combined keyword-based retrieval with LLM-assisted filtering. Using the Media Cloud platform[37], we extracted headlines from national and regional outlets in Spain, Italy, and Portugal. For each language, we used targeted keyword queries (e.g., equivalents of *migration* or *immigration policy*, *population replacement*) to capture a broad range of immigration narratives[38]. After removing duplicate URLs and headlines, this process resulted in approximately 400.000 unique entries. Given this scale, we employed `Llama 3.1 8B` to perform an initial classification along two dimensions: hyperpartisan and PRCT. This automatic step ensured coverage of low-frequency but socially important categories. In particular, PRCT-related content represented less than 1% of the initial corpus. Guided by this filtering, we selected a balanced and diverse sample of 1617 headlines published between 2020 and 2024 from 329 distinct sources.

[37] `https://www.mediacloud.org/`
[38] Full keyword lists are available in our repository.

The custom platform allows annotators to annotate one headline at a time using the labels introduced in 3.4.1. The output file is a csv file.

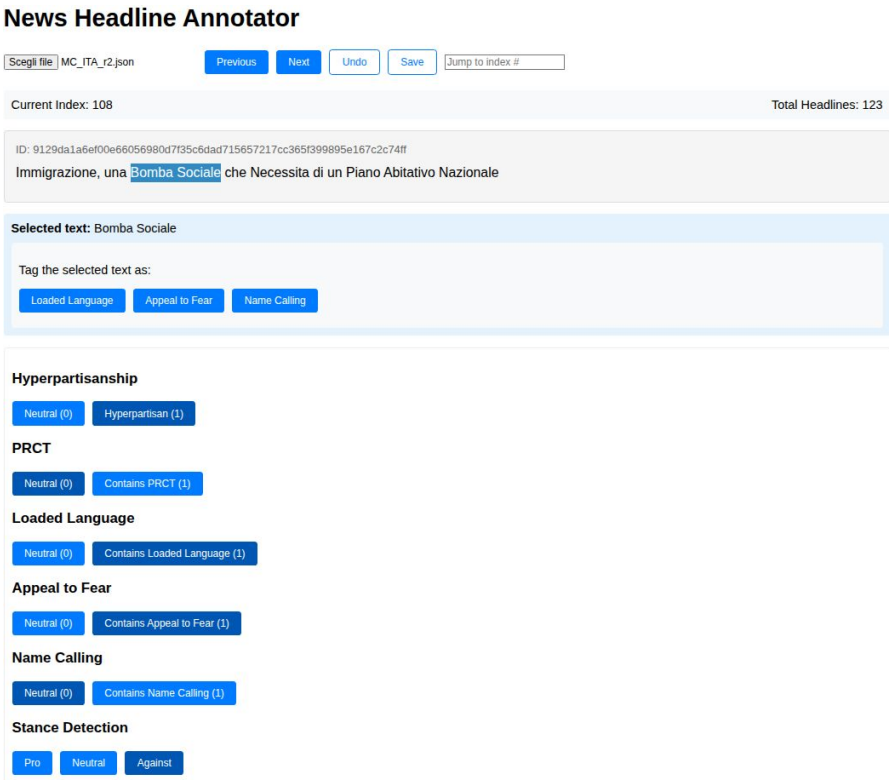


Figure 3.6: Example of the custom HTML platform used for the annotation process.

Annotation protocol

The annotation protocol was designed to capture both ideological content and rhetorical framing in immigration-related headlines. Each item was annotated along *six* complementary dimensions that reflect the stance, narrative strategy, and emotional tone of the text[39]. At the content level, we considered: (*i*) **Hyperpartisan**, indicating whether the headline expressed a one-sided or ideologically extreme view; (*ii*) **PRCT**, marking conspiratorial framings such as references to population replacement or coordinated demographic change; and (*iii*) **Stance**

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

(categorical), capturing the overall attitude toward immigration as *pro*, *neutral*, or *against*. In addition, three rhetorical strategies were annotated at the span level: (*iv*) **Loaded language**, where emotionally charged wording was used to provoke reaction; (*v*) **Appeal to fear**, framing immigration as a threat or danger; and (*vi*) **Name calling**, involving derogatory or stereotyping expressions. These rhetorical cues are known to polarise debate by appealing to emotion rather than reasoned argumentation [Maggini et al., 2025a]. Annotation was conducted through a custom Web interface that masked outlet names to avoid source bias. Nine native speakers, 3 males and 6 females (three per language, all with at least a master's degree) carried out the task over four iterative rounds, beginning with pilot annotations and guideline refinement, followed by progressively larger batches. In total, the process required approximately 66 hours. The final dataset includes both individual annotations and majority labels, enabling not only supervised training but also future research on annotator disagreement and uncertainty. To assess the annotation reliability, inter-annotator agreement was computed using Fleiss' κ across the three main annotation tasks (Hyperpartisan, PRCT, Stance). Rhetorical strategies were not evaluated for agreement as they were annotated at the span level and primarily served to guide annotators by highlighting cues associated with hyperpartisan phenomena.

Dataset statistics

PartisanLens includes 527 headlines in Portuguese (PT), 565 in Italian (ITA), and 525 in Spanish (SPA). Table 3.14 summarises the majority voting distribution across all labels. Hyperpartisan headlines are prevalent in every language (range of 59-70%), while PRCT cases remain relatively rare (under 10% in PT and SPA, but higher in ITA at 28%). Stance distribution varies considerably: anti-immigration headlines dominate in Italian (59%) and Portuguese (40%), whereas Spanish headlines are mostly neutral (80%). Pro-immigration stances are scarce across the three languages.

3.4.2 Benchmark Experiments: Setup

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

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

Table 3.14: Dataset label distribution.

Models

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

Classification strategies

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

Zero-shot prompting. Models receive a concise task description defining the three task categories. We set `temperature = 0` for deterministic decoding and instruct models to output only a valid label Table 3.17.

Few-shot prompting. To provide minimal task-specific context, we include ten balanced demonstrations (≈2 per label) selected via a Determinantal Point Process (DPP) [Kulesza, 2012] to maximise diversity and reduce redundancy. We interleave one positive instance and

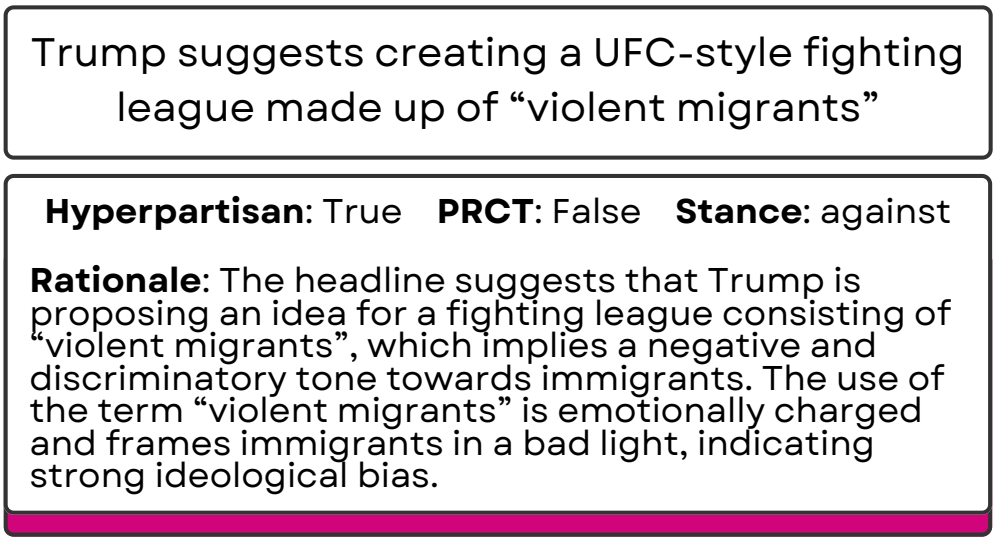


Figure 3.7: Example of a rationale-enhanced annotation.

one negative in our few-shot. Prompts match the zero-shot setup except for the in-context examples.

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

Supervised fine-tuning (labels + rationales). In this variant, models are fine-tuned not only on classification labels but also on rationales, which are short, natural-language explanations derived from the rhetorical bias spans annotated during corpus creation. These rationales express how a headline carries persuasion, such as *“The headline appeals to fear by suggesting a national threat”*. We generate fluent, context-aware rationales from annotated templates using `Llama-3.3-70B-Instruct`, and concatenate them to the input text during fine-tuning. This encourages interpretable learning and alignment between evidence and prediction [Zhang et al., 2024, Piot and Parapar, 2025]. An example of a rationale-enhanced instance is shown in Figure 3.7.

Text characteristics

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

Sociodemographic Data

Table 3.16 illustrates the sociodemographic data we collected for each annotator. All of those dimensions have been used to craft the Persona Prompts.

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

Table 3.15: Example of sociodemographic data. *COMPASS Economic* ranges from negative to positive values, where lower scores indicate a left-leaning economic stance, and higher scores indicate a right-leaning stance. *COMPASS Social* follows the same logic, with lower values reflecting a libertarian orientation and higher values an authoritarian one.

Annotator demographic

Table 3.16 presents the demographics of the annotators who participated in the curation of this dataset.

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

Table 3.16: Participant demographic overview

Evaluation metrics

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

Setup

To answer this research question, we measure mean pairwise Cohen's κ for human–human (H–H) and LLM–human pairs using two models: (*i*) `Llama 3.3--70B` fine-tuned with labels + rationales (L–H), and (*ii*) `gpt-5-mini` in inference mode (G–H). We chose the 70B model with rationales due to its robustness across tasks and include `gpt-5-mini` to study

whether a recent inference model narrows the human gap. The agreement results for all the annotation pairs (human pairs and LLM-human pairs) are summarized in Table 4.6[40].

3.4.3 Persona prompt

We build structured persona prompts representing each annotator's political and demographic background. To do so, each human annotator completed two anonymous instruments: the Political Compass test,[41] which places respondents on Left–Right and Authoritarian–Libertarian axes, and the PRISM quiz,[42] which captures social and economic orientations. We standardise test scores within each dimension and combine them with self-reported sociodemographics (e.g., age range, gender) to construct concise, non-identifying textual profiles. Personas are generated from predefined templates that map score ranges to qualitative descriptors (e.g., "economically left-leaning, socially moderate, libertarian-leaning"). Full sociodemographic summaries and persona templates are included in the Tables 3.15 and 3.18.

Tables 3.17 and Table 3.18 show the prompts used in the experiments.

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

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

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

Table 3.17: Zero-shot and few-shot prompt.

We implement persona prompting with two models: `Llama 3.1--70B` and `gpt5-mini`. For both, we extend the zero-shot classification prompt with the same user-persona context,

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

[41] `https://www.politicalcompass.org/`

[42] `https://prismquiz.github.io/`

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

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

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

Table 3.18: Zero-shot Persona prompts.

with all other prompt components remaining identical. The persona context states the annotator's traits (gender, age range, and economic/social/authority–liberty positions derived from Compass/PRISM) before the task instructions. For evaluation, we compare each persona-LLM to: (i) its matched human counterpart (the annotator whose profile it simulates) and (ii) non-matched annotators in the same language. We also contrast these results with the human–human agreement baseline (Table 4.6).

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

Design Principle

The primary objective of this study is to investigate the linguistic realization of hyperpartisan discourse in Italian and, more specifically, the rhetorical strategies employed to construct and convey polarized narratives in climate change reporting. Climate change represents a particularly relevant domain for this analysis, as it remains affected by forms of denialism and ideological polarization despite the overwhelming scientific consensus regarding its anthro-

pogenic origins and consequences.

To this end, we selected *nicolaporro.it* as the source of analysis. While focusing on a single news outlet inevitably raises concerns regarding source bias, the goal of this study is not to characterize the ideological position of a specific publisher. The selection of this source was further supported by the findings of Potthast et al. [2018b], who demonstrated that extremist outlets from opposite sides of the political spectrum tend to share common stylistic characteristics associated with hyperpartisan communication. Consequently, the study focuses on the linguistic realization of hyperpartisanship rather than on ideological orientation itself. However, the chore of this work is to examine the linguistic realization of hyperpartisan communication and to identify the textual mechanisms through which polarized meanings are constructed. This focus aligns with the broader objective of the thesis, namely, the development of linguistically informed and interpretable models of hyperpartisan language within a European context.

Following the definition proposed by Maggini et al. [2025e], hyperpartisan news detection is understood as the task of identifying news content that exhibits extreme one-sidedness through a pronounced use of bias. In this work, the task is formulated as a binary classification problem at the sentence level. This methodological choice is motivated by both theoretical and practical considerations.

From a theoretical perspective, the thesis adopts a progressive and scalable research design. The sentence-level dataset represents the first step toward the modelling of hyperpartisan language, focusing on its micro-structural realization. By analyzing individual sentences, we aim to isolate the linguistic and rhetorical features that contribute to the expression of hyperpartisan viewpoints, independently of broader discourse structures. This fine-grained perspective allows for a detailed examination of the linguistic mechanisms underlying hyperpartisan communication and provides the empirical foundation for the larger units of analysis explored in the subsequent studies presented in this thesis.

Furthermore, this preliminary investigation facilitated a deeper understanding of the stylistic and rhetorical patterns that characterize hyperpartisan discourse. Through the analysis of articles published by *nicolaporro.it*, we observed recurring linguistic strategies across topics that are particularly relevant to the HYBRIDS project, including climate change and immigration.

From a practical perspective, the sentence-level approach was also motivated by annotation constraints. The taxonomy adopted in this work includes a rich set of rhetorical and linguistic

categories, making the annotation process particularly demanding in terms of time, effort, and expert supervision. Restricting the unit of analysis to individual sentences allowed for a more reliable and consistent annotation procedure while preserving the level of granularity necessary to capture the targeted linguistic phenomena.

Finally, the sentence-level formulation enabled a quantitative analysis of the internal composition of news articles. Beyond article-level classification, it became possible to estimate the degree of hyperpartisan contamination within individual texts and to examine how hyperpartisan content is distributed throughout an article. This perspective provided valuable insights into the structural organization of hyperpartisan narratives and informed the design of the subsequent datasets developed in this thesis.

Article selection

We collected articles in Italian from `https://www.nicolaporro.it/articoli/ambiente-sostenibilita/`, an independent libertarian media outlet. The articles were manually collected from the page on "Green policies", assuring homogeneous topics. They were published in 2024, revolving around various globally discussed topics, including climate change, green policies and Euroscepticism. We considered this "alternative" media source that could potentially spread misinformation about climate change.

Annotation Process

The articles have been split into sentences, following the HTML <p> tags. Then, texts were annotated for hyperpartisanship and rhetorical bias using the taxonomy described above. Both hyperpartisanship and the persuasion techniques were annotated at the sentence level, ensuring a fine-grained analysis of manipulative strategies. The goal of this multi-layer annotation was to capture various complementary aspects of rhetorical manipulation and to investigate the correlation between hyperpartisan sentences and rhetorical biases.

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.

We divided the annotation process into three phases: **Training phase**: annotators studied detailed guidelines, participated in interactive training sessions, and completed structured online training with multiple-choice assessments, and performed pilot annotations; **Annotation**

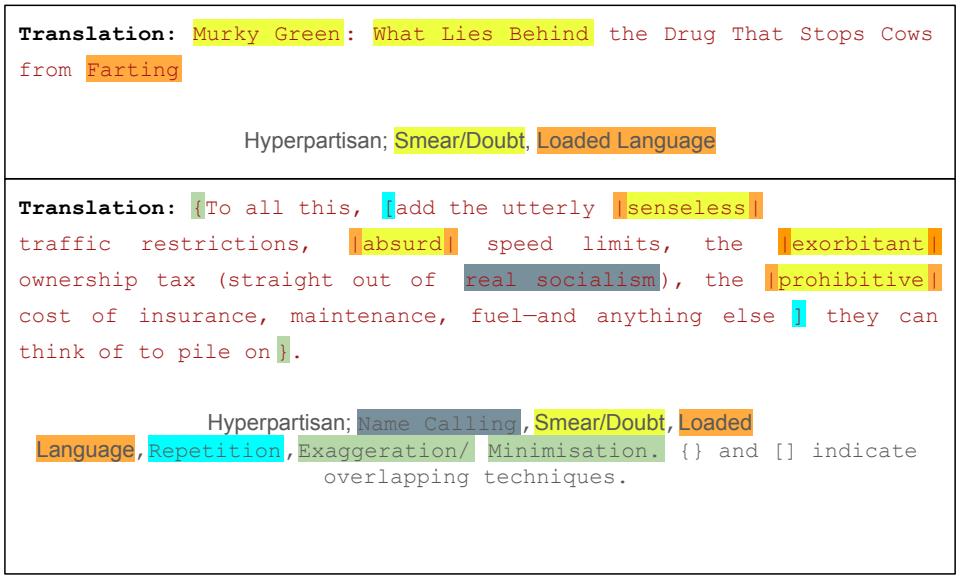


Figure 3.8: Example of the annotation scheme for one paragraph for the sentence-level dataset on climate change.

Phase: Each document was independtently annotated by both annotators; **Curation Phase**: Discrepancies between annotations were discussed and resolved with a senior annotator, who acted as an adjudicator to ensure final label consistency. An example of an annotated entry is shown in Fig. 3.8.

Rhetorical Biases

Given a news article, we identify the presence of specific rhetorical biases or logical fallacies used to influence readers' opinions. We employ a comprehensive taxonomy of 17 rhetorical techniques (see Table 3.19), which builds upon and extends previous work in propaganda detection and rhetorical analysis [Da San Martino et al., 2019, Piskorski et al., 2023c, Sprenkamp et al., 2023]. These techniques are characterized by specific linguistic patterns and argumentative strategies designed to manipulate readers' perceptions. Our rhetorical bias taxonomy encompasses the following categories:

Attack on reputation: The argument does not address the topic, but rather targets the participant (personality, experience, deeds) in order to question and/or to undermine their credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity. **Justification**: The argument is made of two parts, a statement and an explanation or an appeal, where the latter is used to justify and/or to support the statement. **Simplification**: The argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices. **Manipulative wording**: the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally. These six types are further subdivided into 17 fine-grained techniques, i.e., one less than in Da San Martino et al. [2019] and three more than Sprenkamp et al. [2023].

Dataset Annotation

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., 2025e]. We slightly modified the annotations from Da San Martino et al. [2019] and Piskorski et al. [2023c], to match the guidelines of this work, as hyperpartisan detection was not present. The dataset comprised 1,010 sentences from 48 articles, annotated for both hyperpartisan classification and rhetorical bias detection.

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 3.19: 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.

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

Design Principle

Building upon the sentence-level analysis introduced in the previous study, *IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyperpartisan Italian News* extends the investigation of hyperpartisan language to the discourse level. The primary objective of this work is twofold: first, to release a new Italian dataset for hyperpartisan news detection covering heterogeneous and socially relevant topics, such as the war in Ukraine and climate change, which are particularly susceptible to partisan framing [Falkenberg et al., 2022]; second, to

advance the modeling of hyperpartisanship by incorporating linguistic features grounded in pragmatics and discourse analysis.

This work reflects one of the central motivations of the thesis: moving beyond purely predictive approaches to hyperpartisan news detection and toward a linguistically informed and interpretable framework. While the sentence-level dataset allowed us to investigate the micro-structure of hyperpartisan language, many rhetorical phenomena require a broader contextual window to be adequately interpreted. Pragmatic mechanisms emerge through the interaction of multiple propositions and cannot always be captured within isolated sentences. For this reason, the unit of analysis was extended to short paragraphs composed of one to three sentences, enabling the examination of how hyperpartisan discourse develops within coherent segments of text while preserving a level of granularity suitable for detailed annotation.

To operationalize this perspective, the dataset introduces span-level annotations for six Linguistic Traits (LTs): *figurative speech*, *irony/sarcasm*, *epithet*, *hyperbole*, *loaded language*, and *argumentative fallacies*. These phenomena were selected because they represent explicit linguistic realizations through which hyperpartisan discourse reshapes the interpretation of political and social events. Rather than treating hyperpartisanship as a purely stylistic or lexical phenomenon, the proposed framework models it as a semantic-pragmatic process in which rhetorical devices actively influence the construction of meaning.

This interpretation is rooted in pragmatic theory and, in particular, in Grice's Cooperative Principle [Grice, 1975]. According to Grice, successful communication relies on the assumption that speakers cooperate by adhering to the conversational maxims of quantity, quality, relation, and manner. Hyperpartisan communication operates differently. Since hyperpartisan news often addresses audiences that already hold polarized views [Barkho, 2021], rhetorical devices are employed to reinforce existing beliefs through shared assumptions, emotionally charged language, and ideologically aligned interpretations [Waltinger, 2009]. As a result, discourse may shift from an informative function toward a persuasive one.

At the discourse level, the introduction of LTs can alter not only the communicative function of a paragraph, but also, with a downstream effect, the whole article. A news segment that begins with factual reporting but subsequently incorporates epithets, hyperbolic comparisons, or ironic framings undergoes a functional transformation, where rhetorical persuasion becomes more salient than information transmission. Such shifts can be interpreted as violations of Grice's maxim of relation [Grice, 1975], insofar as the inserted content serves persuasive objectives rather than contributing directly to the informative purpose of journalistic

communication.

At the semantic level, LTs contribute to the reconstruction of meaning itself. Traditional journalistic discourse generally relies on a relatively denotative relationship between linguistic expressions and their referents [Ogden and Richards, 1930]. Hyperpartisan rhetoric, however, frequently modifies this relationship through figurative and evaluative language. In particular, figurative speech and loaded language introduce conceptual associations that extend beyond the literal meaning of events. For example, describing a political statement as "dropping a bomb" (see Figure 3.9) invokes a conceptual metaphor [Lakoff and Johnson, 2008] that transfers notions of violence, disruption, and urgency onto an otherwise neutral event. Through these mechanisms, rhetorical devices influence how readers conceptualize political reality and interpret the information presented to them.

From this perspective, LTs are not peripheral characteristics of hyperpartisan discourse but constitute one of its primary linguistic mechanisms. This distinction is particularly important when differentiating hyperpartisan language from generic bias. Biased language may emerge from structural journalistic choices, including source selection, framing, or agenda-setting, and can occur without explicit persuasive intentions [Blodgett et al., 2020]. Hyperpartisan language, by contrast, strategically employs rhetorical and linguistic devices in politically polarized contexts with the explicit objective of reinforcing pre-existing beliefs and identities [Maggini et al., 2025e, Walker et al., 2025]. The intentional manipulation of semantic and pragmatic interpretations therefore represents a defining characteristic of hyperpartisan communication and motivates the use of LTs as computational proxies for its detection.

The paragraph-level design also serves a methodological purpose within the broader research trajectory of this thesis. By moving from sentences to short discourse units, it becomes possible to investigate the internal composition of news articles and to analyze how neutral and hyperpartisan content coexist and interact within the same text. At the same time, restricting the context to relatively short paragraphs preserves annotation reliability and remains compatible with the context-window limitations of the neural architectures employed in this study.

Finally, consistent with the thesis-wide emphasis on transparency and interpretability, this work combines computational linguistic resources with Explainable Artificial Intelligence (XAI) techniques. The annotated Linguistic Traits are incorporated into BERT-based classifiers through embedding-fusion strategies, and their contribution is evaluated through systematic ablation studies. Furthermore, SHAP and Integrated Gradients are employed to

examine whether the models rely on the annotated linguistic phenomena in a manner consistent with their measured predictive contribution. This approach allows us not only to assess model performance but also to investigate the extent to which computational systems capture the semantic-pragmatic mechanisms that characterize hyperpartisan discourse.

Source Selection

To investigate HP language in Italian, we combined data from different sources. We re-annotated Italian data points from SemEval 2023 Piskorski et al. [2023a]'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[43] to achieve sufficient coverage of HP instances for robust analysis. This yielded 192 articles from Piskorski et al. [2023a], 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.

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 paragraph tags.

Annotation Process

Our annotation framework addresses two core constructs: hyperpartisanship at the paragraph level, and LT at the span level (see Figure 3.9).

Hyperpartisanship is characterized as extremely one-sided content marked by strong bias [Maggini et al., 2025e]. We model it as a binary paragraph-level classification to track the distribution of HP content throughout paragraphs (see Table 3.14). LT are grounded in pragmatics theory, which is the study of language in context [Scott, 2022] and they potentially

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

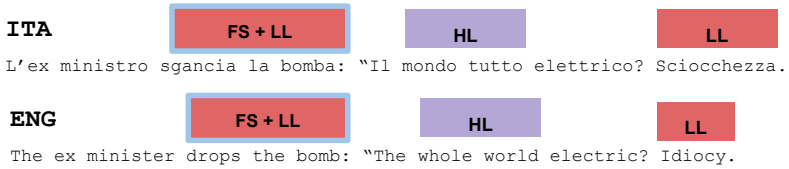


Figure 3.9: Example of the annotation scheme for one paragraph for the HIPP dataset. FS: Figurative Speech, LL: Loaded Language, HL: Hyperbolic Language.

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 3.20 presents our taxonomy and the LT's distribution.

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 3.20: Taxonomy of LT and HP language detection used in our annotation scheme.

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.

3.6.1 Experimental Setup

Models

For our experiments, we selected: 1) monolingual and multilingual encoders with different sizes and training strategies: **bert-base-italian-uncased** (ita-uncased), **bert-base-italian-xxl-uncased** (ita-xxl-uncased), **sentence-bert-base-italian-uncased** (sent-ita) and **sentence-bert-base-italian-xxl-uncased** (sent-ita-xxl), and **bert-base-multilingual-uncased** (multi-uncased).

Fine-tuning and Ablation study

Firstly, we fine-tuned (FT) the selected models for HP detection. Then, we vectorized the linguistic labels using one-hot encoding and concatenated the resulting embeddings with the text (FT+Embeddings). To demonstrate the effectiveness of introducing LT, we FT the BERT models for HP classification using the enriched representations.

To achieve this, we customized the model architecture to process text and linguistic traits as distinct segments. We encoded both the text and the linguistic traits tokens using the tokenizer, separating them with `token_type_ids` (segment 0 for text and segment 1 for traits). The custom model extracts hidden states from the encoder and applies average pooling separately over the text segment and traits segment, explicitly masking out special tokens like CLS, SEP, and PAD.

After extracting the text and traits features independently via pooling, they are concatenated and passed through a fusion linear layer followed by a sigmoid activation to produce a gating weight. This gating mechanism allows the model to dynamically blend the representations: the fused output is computed as a weighted sum of the text features and the traits features, controlled by the gate. The fused representation is then fed into a classification head, consisting of a linear layer (hidden size of 256), a ReLU activation, 20% dropout, and a final linear layer for binary hyperpartisan classification. Lastly, we performed an ablation study to investigate the most contributive features and compared this setting against FT+Embeddings.

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 integrating gradients along this path from a zero-embedding baseline to the input embedding. We respectively calculated the attribution ratio of LT and non-trait tokens as:

$$\text{attribution ratio} = \frac{\text{median}\Big(\{\hat{a}_i \mid i \in \text{span}\}\Big)}{\text{median}\Big(\{\hat{a}_j \mid j \in \text{all valid tokens}\}\Big)} \tag{3.3}$$

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, both individually and in combination, push the prediction from the baseline value. Both SHAP and Integrated Gradients, however, directly explain the original model's behavior without the need for an intermediate surrogate. SHAP uses game theory to divide the contribution, while Integrated Gradients uses the model's own gradients. We first mapped token span positions, removing special tokens and computed Integrated Gradients and SHAP for tokens annotated with linguistic traits. 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.

3.7 Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization

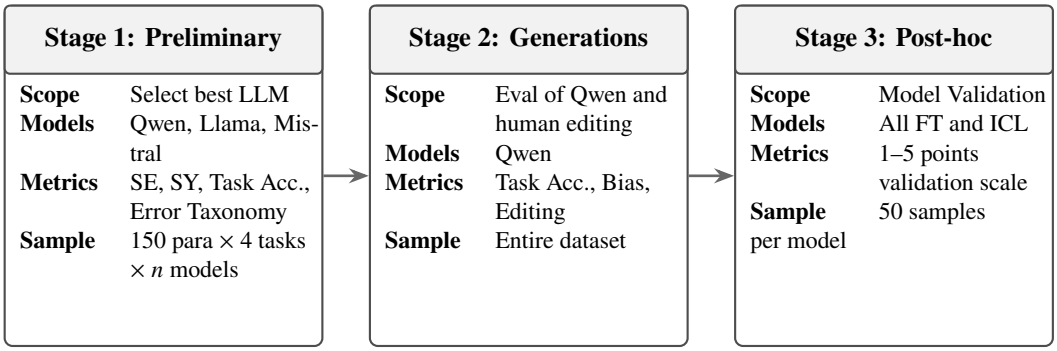


Figure 3.10: The three-stage annotation and evaluation pipeline of Par-ITA.

3.7.1 Design Principle

Par-ITA: Benchmarking Seq2Seq and LLMs on a Human-Supervised Parallel Corpus for Italian Hyperpartisan Neutralization represents the final stage of the research trajectory developed throughout this thesis. While the previous studies focused on identifying and characterizing hyperpartisan language through increasingly complex linguistic units—from sentences to paragraphs—this work addresses a complementary challenge: the transformation of hyperpartisan content into neutral journalistic discourse. The objective is not merely to detect hyperpartisan language, but to investigate whether computational models can learn to remove or mitigate the rhetorical and pragmatic mechanisms that contribute to polarization while preserving the informational content of the original text.

This shift from detection to neutralization is motivated by one of the broader goals of the HYBRIDS project and of this thesis: developing computational tools that not only identify disinformation-related phenomena but also contribute to mitigating their effects. In this context, neutralization provides a controlled framework for studying hyperpartisan language, as it requires the explicit identification and modification of the linguistic elements responsible for partisan framing. Consequently, the task offers a complementary perspective to classification, enabling a deeper understanding of the relationship between linguistic form, rhetorical

strategies, and perceived neutrality.

To support this investigation, we introduce PAR-ITA, the first parallel corpus for Italian hyperpartisan neutralization, consisting of 2.475 paragraph pairs manually curated through a human-supervised annotation process. The choice of paragraphs as the unit of analysis follows directly from the findings of the previous studies. As demonstrated in *IHPP*, many of the rhetorical and pragmatic phenomena that characterize hyperpartisan discourse emerge through interactions between multiple sentences and therefore require a discourse-level context to be properly interpreted. At the same time, paragraph-sized units remain manageable for annotation and revision, ensuring high-quality human supervision throughout the dataset creation process.

The corpus was developed through a rigorous three-stage pipeline (Figure 3.10). The first stage, **Preliminary Selection**, evaluated several open-source Large Language Models (LLMs) with respect to both syntactic preservation and semantic fidelity in order to identify the most reliable generator for the subsequent annotation phases. The second stage, **Supervised Generation**, employed the selected model to produce neutralized and hyperpartisan variants, which were then carefully revised and corrected by human annotators. This human-in-the-loop strategy was adopted to ensure that the resulting transformations accurately reflected the intended neutralization process while maintaining the informational content of the source paragraphs. The third stage, **Post-hoc Validation**, consisted of expert evaluation aimed at assessing the effectiveness and reliability of the generated neutralizations. Across all stages, we additionally investigated the potential role of GPT-4o-mini as an automated judge capable of supporting annotation and evaluation procedures.

From a methodological perspective, the design of PAR-ITA reflects the thesis-wide emphasis on transparency, interpretability, and resource quality. Rather than relying exclusively on automatically generated parallel data, the corpus was constructed through extensive human supervision, ensuring that the linguistic transformations remained consistent with journalistic neutrality and with the semantic content of the original texts. This decision is particularly important in the context of hyperpartisan language, where subtle rhetorical devices, pragmatic implications, and ideological framing may not be adequately captured through fully automatic generation processes.

The resulting corpus enables a systematic evaluation of different approaches to hyperpartisan neutralization. Specifically, we benchmark both sequence-to-sequence architectures, such as IT5 and BART, and contemporary LLMs under multiple adaptation paradigms, including

standard Fine-Tuning (FT), Direct Preference Optimization (DPO) [Rafailov et al., 2024], and In-Context Learning (ICL). In addition to automatic metrics, we employ a human evaluation protocol that assesses linguistic quality, semantic preservation, and perceived neutrality, while also examining the degree of alignment between automated evaluation measures and expert judgments.

The main contributions of this work are therefore threefold. First, it introduces PAR-ITA, the first parallel corpus specifically designed for Italian hyperpartisan neutralization. Second, it proposes a comprehensive annotation methodology that integrates model selection, human-supervised generation, and expert validation into a unified framework. Third, it provides a systematic benchmark of contemporary training and adaptation strategies, offering insights into the strengths and limitations of different architectures for the neutralization of hyperpartisan discourse. More broadly, the dataset extends the research agenda of this thesis from the analysis and detection of hyperpartisan language toward the development of practical methods for reducing its influence, thereby contributing to the creation of more transparent and trustworthy information ecosystems.

3.7.2 Dataset Construction

Data Collection and Preparation

In order to build a hyperpartisan parallel corpus, we selected SemEval 2023 task 3 dataset [Piskorski et al., 2023b] tailored for detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup.

Type	Task	Text ITA	Text ENG
HP	O	L'Italia si barcamena con Di Maio. . .	*Italy is muddling through with Di Maio. . .*
HP	Rep	l'Italia cammina sul filo del rasoio con Di Maio. . .	*Italy is walking on the razor's edge with Di Maio...*
HP	Per	l'Italia naviga a vista con Di Maio. . .	*Italy is navigating by sight with Di Maio. . .*
N	Rew	L'Italia fatica a gestire la situazione con Di Maio.	*Italy is struggling to manage the situation with Di Maio.*

Table 3.21: Examples of model-generated and human-supervised neutral rewrites for hyperpartisan input phrases. **O**: original text, **Rep**: Rephrased, **Per**: Perturbed, **Rew**: Rewritten.

Annotation Protocol

After having filtered Italian data from the corpus (2.5k entries), two native speakers—a Ph.D. student in Disinformation and an expert journalist—annotated paragraphs as hyperpartisan or neutral. Hyperpartisanship indicates one-sided or ideologically extreme views expressed through irony, sarcasm, framing bias, or cherry-picking [de León et al., 2024]. The annotation process involved training on guidelines, pilot annotations, and interactive sessions to refine edge cases. Independent annotation achieved 0.85 Cohen's κ, with disagreements resolved through discussion and third-party mediation, yielding perfect final agreement on 825 hyperpartisan paragraphs. The entire process required 300 hours across three custom web interfaces.

3.7.3 Three-Stage Annotation Pipeline

Our three-stage approach involves: (1) preliminary LLM selection, (2) full dataset generation with human supervision, and (3) post-hoc model validation.

Stage 1: Preliminary LLM Selection

To select the optimal model for data augmentation, we evaluated three LLMs: Mistral-Nemo-2407 [Mistral, 2024], Llama-3.1-8b-Instruct [Touvron et al., 2023a], and Qwen2.5-14B-Instruct [Yang et al., 2025], chosen for their open availability and balanced efficiency. These models generated three variants per paragraph via backtranslation (ita $\rightarrow$ eng $\rightarrow$ ita): one neutralized (*Rewritten*) and two hyperpartisan versions (*Rephrased* with synonym substitution, *Perturbed* focusing on changing only loaded language and rhetoric).

Evaluation Protocol

Annotators independently assessed 150 sampled paragraphs on 3-point scales for **SY** (1: Major Errors, 2: Minor Errors, 3: Correct) and **SE** (1: Poor Preservation, 2: Partial Preservation, 3: Full Preservation), plus binary neutrality classification. We applied a multi-dimensional error taxonomy categorizing failures in neutralization, factuality, fluency, and structure. Additionally, GPT-4o-mini served as an automated judge, being prompted with the guidelines and the taxonomy.

Justification for LLM-Based Data Generation

While comprehensive benchmarking reveals seq2seq models achieve superior performance, we selected Qwen-2.5-14B-Instruct for data generation based on Stage 1 evaluation comparing only LLMs. This decision reflects three considerations: (1) Stage 1 established Qwen as best among LLMs for generation quality and human alignment; (2) High editing rates mean final quality depends on expert supervision rather than raw model output—Levenshtein distances of 25-32% confirm substantial human refinement; (3) At experimentation time, we prioritized selecting a versatile model capable of both neutralization and bias amplification tasks.

Stage 2: Full Dataset Generation with Human Supervision

Based on Stage 1 results, we selected Qwen-2.5-14B-Instruct to generate variants for all 825 hyperpartisan paragraphs. At the time of experimentation, we wanted to assess the capabilities of the selected model to accomplish the task and use it as a tool to speed up the writing process for the data augmentation. Annotators evaluated each output using a 3-level task accomplishment scale (No/Partially/Yes) and performed necessary editing. Failed generations were collaboratively rewritten to ensure corpus quality.

Final Corpus Specifications

The final corpus comprises 2,475 training pairs and 489 test pairs (see Table 3.22), ensuring high-quality parallel data for neutralization tasks.

Metric	Train	Test	Total
# Pairs	1,986	489	2,475
# Unique Originals	662	163	825
Avg. input chars	365.7	377.5	368.0
Avg. input words	55.3	57.2	55.6
Avg. output chars	352.9	402.0	362.6
Avg. output words	51.3	59.2	52.9
Vocabulary size	15,639	6,760	18,032

Table 3.22: Statistics of the human-curated Par-ITA corpus.

Stage 3: Post-hoc Model Validation

Following model training, we conducted human evaluation to assess generation quality beyond automated metrics. Two expert annotators independently rated 50 randomly sampled outputs per model on a 5-point scale, evaluating both neutralization success and semantic preservation. GPT-4o-mini provided parallel judgments while being prompted with the task and the taxonomy.

GPT-4o-mini as a Judge Prompts

In this section, we report the prompts used for data augmentation with LLMs. The prompt instructs the model to perform three parallel tasks on each hyperpartisan paragraph: surface-level rephrasing (preserving bias), unperturbed back-translation (preserving placeholders), and perturbed back-translation (amplifying bias by replacing placeholders with synonymous loaded expressions). The structured JSON output facilitates automatic extraction and subsequent human supervision.

3.7.4 Benchmark Experiments: Setup

Models

Our experimental setup combines various seq2seq models in both base and large sizes: BART [Lewis et al., 2020], IT5 [Sarti and Nissim, 2024], FLAN-T5, [Chung et al., 2024], and mT5 [Xue et al., 2021]; and instructed LLMs loaded through `unsloth` library: Gemma-3-4b-it [Team et al., 2025], Phi-3.5-mini-instruct [Abdin et al., 2024], Qwen2.5-3b-instruct, Mistral-7b-it-v0.3, Llama-3.2-3b-instruct/-3.1-8b-instruct, and with `vllm`: Apertus-70B-2509 [Apertus et al., 2025], Gemma-3-27b-Instruct/-2.5-72B-Instruct, Qwen-3-30B-Instruct.

Guided SFT Classifier

In this approach, we augment standard maximum likelihood estimation (MLE) with an auxiliary loss derived from a fine-tuned binary neutrality classifier for italian hyperpartisan paragraph detection on Maggini et al. [2025b]'s dataset, consisting of 1,010 paragraphs. During training, the seq2seq model generates outputs via standard teacher forcing, which are then periodically (every n steps) decoded to text and scored by the frozen classifier. The classifier

produces neutrality probabilities p_{neutral} for each generated sequence, which we convert into a guidance loss term:

$$\mathcal{L}_{\text{guide}} = \lambda \cdot \max(0, \tau - p_{\text{neutral}}) \tag{3.4}$$

where τ is a target neutrality rate and λ is a weighting coefficient. The total training objective then becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MLE}} + \mathcal{L}_{\text{guide}}. \tag{3.5}$$

DPO

With our current setup, we crafted a DPO dataset. Indeed, for each task we had a preferred and optimal variant, the direct and valid output of Qwen, or the human-rewritten version, and an underoptimal one, namely the failed generation or a corrupted version of it. We prioritized the human-edited texts as the preferred version. While when we kept the Qwen's generated text as the preferred option, we corrupted it to generate the suboptimal choice with the following methods: (*i*) swapping the order of the words, (*ii*) inserting loaded adverbs like "ovviamente", "incredibilmente", (*iii*) random truncation, (*iv*) word repetition, (*v*) random masking and (*vi*) random word deletion.

Component	Prompt Content

User Prompt	You are an expert Italian rewriter and back-translator for hyper-partisan text.
	
	- CONTEXT -
	1. Original Italian paragraph:
	{original}
	2. Masked Italian paragraph:
	{masked}
	3. Span mapping:
	{span_table}
	
	The placeholders have this structure: `[linguistic_trait_n]`, where the linguistic traits are: `LOADED_LANGUAGE`, `EPITHET`, `IRONY/SARCASM`, `HYPERBOLIC_LANGUAGE`, `FIGURATIVE_SPEECH`, `NEOLOGISM` and the *n* indicates the occurrence of each linguistic trait.
	
	- TASK 1: REPHRASE -
	Rephrase the **original Italian paragraph** in natural, fluent Italian while **preserving the exact meaning**.
	Avoid changing tone or bias. Output should be different in wording but identical in semantics.
	
	- TASK 2: BACK-TRANSLATION (UNPERTURBED) -
	• Translate masked Italian to English (keep placeholders)
	• Translate that English back to Italian (keep placeholders)
	• The placeholders **MUST** remain in unperturbed outputs both in Italian and English outputs. Keep them in bold text as in the original text.
	
	This is an example of the output for TASK 2: BACK-TRANSLATION (UNPERTURBED):
	"Text Text Text [LOADED_LANGUAGE_n] Text Text [EPITHET_n] Text"
	
	- TASK 3: BACK-TRANSLATION (PERTURBED) -
	• Replace each placeholder with a **synonymous, bias-amplifying Italian expression**
	• Translate the full replaced Italian to English
	• Translate that English back to Italian
	• **NO placeholder may remain** in perturbed outputs
	This is an example of the output for TASK 3: BACK-TRANSLATION (PERTURBED):
	"Text Text Text Text Text Text Text Text"
	Return **ONLY** this JSON (no markdown, no extra text):
	`''rephrased'': ''rephrased Italian'',`
	`''it_unperturbed'': ''Italian with placeholders'',`
	`''it_perturbed'': ''Italian with replaced expressions''`

Component	Prompt Content

Judge Prompt (Italian)	Sei un esperto di analisi del linguaggio giornalistico italiano. Esegui il seguente compito seguendo la definizione di hyperpartisan adottata nelle guidelines: 1-Etichetta il testo come hyperpartisan o neutrale.
	
	{guidelines}
	
	Testo da valutare:
	{text}
	
	Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo):
	{
	"label": <hyperpartisan, neutrale>,
	}

Table 3.23: Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini.

Component	Prompt Content
System Instruction	Sei un giornalista esperto. Riscrivi il seguente paragrafo in modo completamente neutrale, rimuovendo ogni tono hyperpartisan, sarcastico, caricaturale. Mantieni tutti i fatti e il significato originale, usando un linguaggio oggettivo e imparziale. Ti verranno forniti degli esempi di neutralizzazione di paragrafi hyperpartisan nel seguente formato: ESEMPIO: Hyperpartisan:"text", Neutrale: "text".
Few-Shot Examples	Esempio 1: Hyperpartisan: "{text}" Esempio 2: Hyperpartisan: "{text}" Esempio 3: Hyperpartisan: "{text}"
User Instruction	Questo è il paragrafo che devi neutralizzare: Paragrafo: {input_text} Rispondi esclusivamente con la versione neutralizzata, senza aggiunte, spiegazioni o commenti.

Table 3.24: Prompt used for zero-shot neutralization with big size LLMs (> 70B parameters). The system instruction defines the role and task, three high-quality few-shot examples illustrate the expected neutralization style (preserving facts while removing hyperpartisan tone, sarcasm, and loaded language), and the user instruction provides the target paragraph with a strict requirement to output only the neutralized version.

Component	Prompt Content

Judge Prompt (Italian)	Sei un esperto di analisi del linguaggio giornalistico italiano. Esegui i seguenti giudizi: 1-Etichetta il testo come hyperpartisan o neutrale. 2-Assegna un valore (int) in una scala da 1 a 3 per la semantica del testo prodotto e per la correttezza sintattica seguendo le guidelines.
	
	{guidelines}
	
	Testo da valutare:
	{text}
	
	Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo):
	{
	"label": <hyperpartisan, neutrale>,
	"semantic_score": [1,2,3],
	"syntax_score": [1,2,3],
	}

Table 3.25:

tab:llm-judge-LLMs$_{p}reliminary - generation - prompt$

Component	Prompt Content
Judge Prompt (Italian)	JUDGE_PROMPT = Sei un esperto di analisi del linguaggio giornalistico italiano. Valuta quanto il seguente testo sia neutrale e imparziale. Scala di valutazione (1-5): 1 = Molto di parte, linguaggio fortemente emotivo o carico. Oppure generazione fallita: presenza di errori gravi, testo incomprensibile, output ripetuto. 2 = Parzialmente di parte, diverse espressioni cariche. Generazione con errori ed incoerenze semantiche come l'aggiunta di contesto inesistente o troncamento del contesto fornito. 3 = Moderatamente neutrale, ma alcune tracce di bias. Generazione generalmente corretta ma con alcune imprecisioni grammaticali e semantiche. Il testo può contenere lievi allucinazioni e del contesto aggiunto con correttezza referenziale. 4 = Principalmente neutrale, linguaggio equilibrato. Generazione sostanzialmente corretta con lievi imperfezioni grammaticale (concordanza, piccoli refusi). Non viene aggiunto alcun tipo di contesto, tutte le informazioni nel testo originale vengono preservate, ma neutralizzate. 5 = Completamente neutrale, oggettivo e imparziale. Linguaggio chiaro e professionale. Generazione priva di errori grammaticali o semantici. Testo originale NON da valutare: {input} Testo generato da un LLM: {generated} Rispondi SOLO con un JSON in questo formato (nessun testo aggiuntivo): neutrality_score: [1-5]

Table 3.26: Prompt used for LLM-as-a-judge evaluation with GPT-4o-mini. The prompt instructs the model to rate the neutrality of a given text on a 1–5 Likert scale. The strict JSON-only output format ensures structured, parseable responses for automated analysis and comparison with human judgments.

3.8 Summary of Methodology

This section presented the methodologies underlying the studies included in this thesis, organized around the selected articles for discussion. The step-by-step description of each experimental pipeline reflects the rigor applied in designing, organizing, and conducting this work, from the mathematical formulation of the algorithms to the specification of prompts and annotation protocols. The next section presents the results in the same order, covering:

- LLMs for hyperpartisan and related tasks using fine-tuning and in-context learning;
- new resources for hyperpartisan detection in low-resource languages;
- persona-prompting strategies;
- hyperpartisanship, PRCT, and stance on multilingual headlines;
- hyperpartisanship through rhetorical bias;
- explainable AI techniques to pragmatic linguistic features in hyperpartisan detection;
- fine-tuning Transformers using embeddings of pragmatic linguistic features;
- LLMs as judges;
- LLMs as annotators;
- synthetic data with LLMs;
- neutralizing hyperpartisan language.

CHAPTER 4

RESULTS AND DISCUSSION

This section presents the results of the doctoral research, organized around two complementary pillars: (i) a systematic review of the state-of-the-art in automated hyperpartisan news detection, and (ii) the original contributions developed during the doctoral studies. The systematic review provides essential context on the research landscape, including publication trends, annotation methodologies, available resources and methodological approaches. Subsequently, we present five research contributions that progressively address challenges in hyperpartisan detection, dataset creation across multiple languages and domains, and explainability techniques applied to the pragmatical characteristics of hyperpartisan languages as well as neutralization techniques. Each subsection details the key findings and implications of the respective contribution.

4.1 Are LLMs Enough for Hyperpartisan, Fake, Polarized and Harmful Content Detection? Evaluating In-Context Learning vs. Fine-Tuning

Content moderation at scale requires robust detection systems for diverse harmful content categories spanning multiple languages. The emergence of LLMs with ICL capabilities has prompted widespread speculation that gradient-free inference could replace traditional fine-tuning pipelines. This paper systematically benchmarks encoder-only and decoder-only transformer architectures across 10 diverse datasets covering hyperpartisan content, fake news, polarized speech, and hateful language detection in English, Spanish, Portuguese, Arabic, and

Bulgarian. We evaluate both fine-tuning and multiple in-context learning strategies (zero-shot, few-shot with curriculum learning, chain-of-thought prompting, and codebook-based prompting) to provide empirical guidance for practitioners selecting approaches for production moderation systems.

Model		HV	SH	Macro Avg. HP	FNN	SFN	FB	Macro Avg. FN	C1A	C1B	C1E	Macro Avg. HF	QB	C3A	Macro Avg. PL
RoBERTa-base	Acc	.822	**.865**	.843	.879	–	–	–	–	–	.919	–	.622	.659	.640
	F1	.818	**.865**	.841	.880	–	–	–	–	–	.915	–	.604	.660	.632
RoBERTa-large	Acc	**.852**	**.865**	**.858**	.893	–	–	–	–	–	**.925**	–	.683	.663	.673
	F1	**.850**	**.865**	**.857**	.893	–	–	–	–	–	**.923**	–	.674	.660	.667
XLM-RoBERTa	Acc	.827	.801	.814	.842	.623	.957	.807	**.917**	.917	.917	**.917**	.582	.632	.607
	F1	.825	.798	.811	.844	.577	.957	.793	**.883**	.884	.885	**.884**	.553	.629	.591
POLITICS	Acc	.831	.854	.842	.867	–	–	–	–	–	.916	–	.682	.679	.680
	F1	.826	.854	.840	.868	–	–	–	–	–	.911	–	.673	.678	.675
mDeBERTaV3	Acc	.819	.776	.797	.893	–	–	–	–	–	.915	–	.539	.590	.564
	F1	.816	.772	.794	.841	–	–	–	–	–	.874	–	.534	.570	.552
ModernBERT-large	Acc	.829	.854	.839	.858	**.863**	.941	.883	.815	.965	.835	.872	.658	.654	.653
	F1	.824	.853	.839	.846	**.863**	.941	.815	.815	.949	.803	–	.649	.657	.667
ModernBERT-base	Acc	.764	.780	.767	.852	.782	.942	.854	.830	**.966**	.830	.875	.584	.617	.592
	F1	.755	.779	.767	.840	.781	.942	.854	.792	**.966**	.774	.844	.571	.612	.592
Llama-3.1-8b	Acc	.830	.810	.820	**.945**	.812	.975	**.911**	.864	.869	.921	.884	.788	.762	.775
	F1	.830	.801	.815	**.945**	.801	.975	**.907**	.858	.832	.920	.870	.786	.763	.774
Llama-3.1-8b-Instruct	Acc	.784	.820	.801	.875	.823	**.976**	.890	.878	.915	.862	.885	.786	.796	.791
	F1	.782	.820	.801	.869	.823	**.976**	.889	.867	.928	.829	.875	.781	**.802**	.792
Mistral-Nemo-Instruct-2407	Acc	.834	.736	.783	.867	.725	**.976**	.851	.850	.943	.838	.877	**.790**	**.846**	**.819**
	F1	.833	.733	.783	.855	.722	**.976**	.851	.859	.946	.825	.877	**.787**	.789	**.849**
Qwen2.5-7B-Instruct	Acc	.819	.686	.745	.864	.685	.974	.835	.856	.913	.824	.864	.735	.698	.711
	F1	.812	.677	.745	.855	.675	.974	.835	.863	.928	.806	.866	.728	.693	.711

Table 4.1: Performance of models in the FT setting. The reported weighted Accuracy and weighted F1 scores are the averages obtained by running each model five times on the same dataset, reporting standard deviation.

4.1.1 Fine-Tuning

Table 4.1 presents the results for fine-tuning. On average across datasets, decoder-based models tend to outperform encoder-based models on tasks that require factual world knowledge, such as fake news detection and political bias identification. For instance, in fake news detection (Macro Avg. FN F1 score), the Llama-3.1-8b achieves .907, while the best-performing encoder with a directly comparable macro average, ModernBERT-base, scores .854. Similarly, for political leaning detection (Macro Avg. PL F1 score), the Mistral-Nemo-Instruct reaches .849, significantly surpassing the top encoder, POLITICS, which scores .675. Conversely, encoders achieve better results on linguistically oriented tasks, specifically harmful tweet detection and hyperpartisan language identification. RoBERTa-large (encoder) records an F1 score of .850 on HV, compared to the best decoder, Mistral-Nemo-Instruct, at .833. For SH, RoBERTa-large again leads with an F1 score of .865, while the best-performing decoder on this task, Llama-3.1-8b-Instruct, achieves .820. We hypothesize that this difference arises because the bidirectional attention mechanism of encoders may be better at capturing nuanced linguistic features, whereas decoders might excel at tasks more reliant on content or semantic

understanding. Surprisingly, continuous pretraining of RoBERTa-base aimed at adapting it to either the political domain (POLITICS) or multilingual contexts (XLM-RoBERTa) has, in several instances, not led to improved performance over the original RoBERTa-base model and sometimes resulted in a reduction. For example, on the Macro Avg. HP (Hyperpartisan) task, RoBERTa-base achieves an F1 score of .841, whereas POLITICS scores .840 and XLM-RoBERTa scores .811. Regarding decoder models, we observe that a larger parameter count or expanded training corpus does not necessarily equate to superior results. This is demonstrated by LlaMA 3.1-8B outperforming Mistral Nemo-Instruct-2407 in hyperpartisan detection (SH F1 score of .801 for LlaMA 3.1-8b versus .733 for Mistral Nemo). Meanwhile, the decoder models exhibit relatively comparable high performance in harmful text detection (Macro Avg. HF F1 scores): Llama-3.1-8b (.870), Llama-3.1-8b-Instruct (.875), and Mistral-Nemo-Instruct-2407 (.877). During the FT experiment, we reached the SOTA in HV, FNN, SFN, and FB.

4.1.2 In-Context Learning

For the results discussed in this section, please refer to Figures 4.1 and 4.2 for the zero-shot configurations and CoT; and 4.3 for FS. Detailed results are reported in Table 4.2

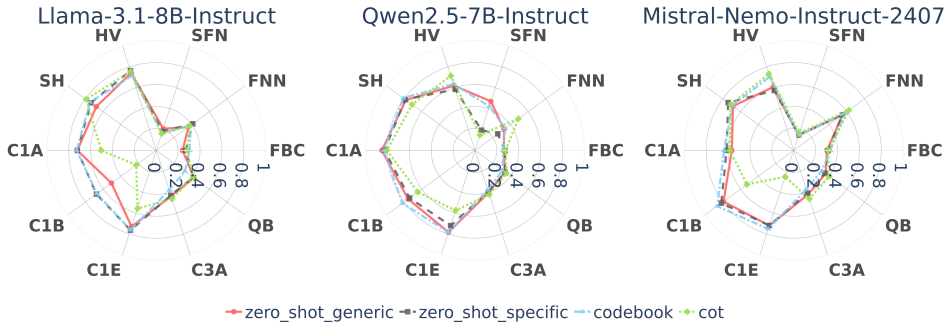


Figure 4.1: Results for zero-shot and CoT grouped by models.

HP

zero-shot-general vs zero-shot-specific In the hyperpartisan (HP) task, moving from generic to specific zero-shot prompting results in moderate improvements across all three models,

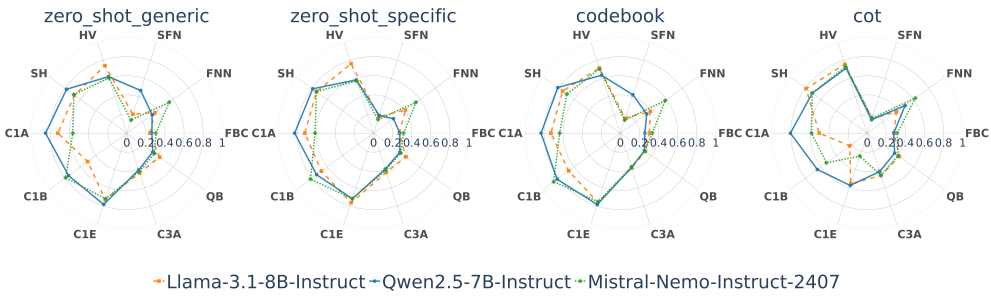


Figure 4.2: Results zero-shot and CoT grouped by configuration.

especially for LLaMA 3.1-8B Instruct and Mistral-Nemo-Instruct, whose F1 scores increase from 0.678 and 0.686 to 0.738 and 0.740, respectively. This performance gap indicates that these models have learned robust representations of hyperpartisan content during pre-training, and that providing more detailed task descriptions helps further refine their predictions. Lastly, SH predictions are generally stronger, largely because this dataset includes full articles rather than just headlines, as in HV. The richer contextual information in SH provides models with more linguistic and semantic cues, enabling a deeper understanding of the content and improving their ability to detect hyperpartisan narratives. In contrast, the limited context in headlines offers fewer signals for accurate classification.

Codebook The codebook approach yields improvements for SH, where Llama and Qwen demonstrate the most significant gains, particularly on the SH dataset, where their performance approaches F1 .810 and .748 respectively. This suggests that providing explicit criteria for identifying partisan language can help address edge cases where the models depend solely on their internal task representations. The performance gap between models narrows with codebook prompting, indicating that structured guidance can help equalize performance differences stemming from models' architecture, which may rely on different definitions of the phenomenon investigated.

FN

zero-shot-general vs zero-shot-specific Fake news detection tasks show variable performance under zero-shot conditions. The FNN task proves more challenging for the multilingual datasets: SFN and FBC. The transition from generic to specific prompting yields minimal

gains, with Mistral-Nemo-Instruct-2407 maintaining the highest performance on 2 out of 3 fake news (FN) datasets. We observed a 0.275-point drop in F1 score for Qwen on the Spanish dataset, which may be attributed to a misalignment between the fake news definitions—specifically, the model's internally assumed definition in the zero-shot generic prompt versus the expert-crafted definition used in the zero-shot specific prompt. This limited improvement suggests that, beyond clear definitions, models need additional contextual or world knowledge to effectively differentiate factual from fabricated content.

Codebook The codebook approach yields minimal improvements for FNN and SFN. indicates that providing explicit criteria for evaluating factual claims, source credibility markers, and stylistic indicators of fabricated content may help models overcome the inherent complexity of fact verification. The codebook's little effectiveness in this domain suggests that FN does not completely benefit from structured evaluation frameworks. Regarding FN, with the different prompt strategies tested in ICL, we found out that the small LLMs are not effectively capable of detecting this kind of disinformation because they can not rely properly on the ontological structures encoded in their world-knowledge. Thus, employing these models as out-of-the-box tools with an ICL setup proves to be inefficient for this task.

PL

zero-shot-general vs zero-shot-specific Political leaning (PL) classification tasks exhibit low performance across all models under zero-shot conditions. It is important to note that this is a multiclass classification task, which adds complexity. Generally, specific prompting produces a slight decrease for this task. Llama maintains consistently higher performance (on average F1 .416) than the other models. The limited effectiveness with a specific definition of the political wings suggests that PL requires more than definitional refinement to overcome the inherent subjectivity involved.

Codebook Providing more detailed and specific knowledge through the codebook generally resulted in decreased performance across all models. This highlights the complexity of the task and suggests that, despite offering explicit rules to interpret the U.S. political context and agendas, the cultural nuances and linguistic factors involved are insufficient for effectively addressing the task. This result reveals the complexity underlying political bias detection.

HF

zero-shot-general vs zero-shot-specific HF covered three languages and Qwen reached the best results in C1A with zero-shot-generic prompts (F1 .851). However, when prompted with zero-shot-specific prompts, its performance decreased. On the other hand, Llama particularly benefited from the introduction of the specific knowledge for C1B (from F1 .507 to .676) and C1E (from F1 .730 to .764).

Codebook With the introduction of specific dimensions to frame the task, all the models across the HF datasets (except for Qwen in C1A) improved their performance. Specifically, Mistral reached F1 .864 in C1B. This marked improvement highlights the value of explicit harm taxonomies and classification criteria for this sensitive domain. The codebook's effectiveness for harmful content detection suggests that these tasks require clearly articulated boundaries and examples to overcome potential ambiguities in what constitutes harmful material.

4.1.3 Summary for Zero-shot configurations

Across all classification domains, the transition from generic to specific zero-shot prompting led only to marginal improvements. This pattern suggests that merely elaborating on task definitions provides insufficient guidance for these models to significantly alter their classification behavior. The codebook approach demonstrates modest effectiveness across all task categories, with particular improvements for HF and FN classification. This pattern indicates that providing structured classification criteria helps models overcome the inherent complexities of these judgment tasks. The codebook's effectiveness stems from its ability to bridge the gap between abstract classification concepts and concrete textual indicators, providing models with clearer decision boundaries for ambiguous cases. Lastly, we noticed that more subjective and nuanced tasks like HF and PL, which require extensive knowledge of the facts and their truthfulness, show improvement with advanced prompting strategies than more stylistic-based tasks (e.g. HP detection), suggesting that prompt optimization benefits may correlate with task complexity. Indeed, the rule-based approach is the best ICL configuration for 3 out of 10 datasets: SH, C1B, and FBC.

4.1.4 FS: DPP-selected vs Random examples

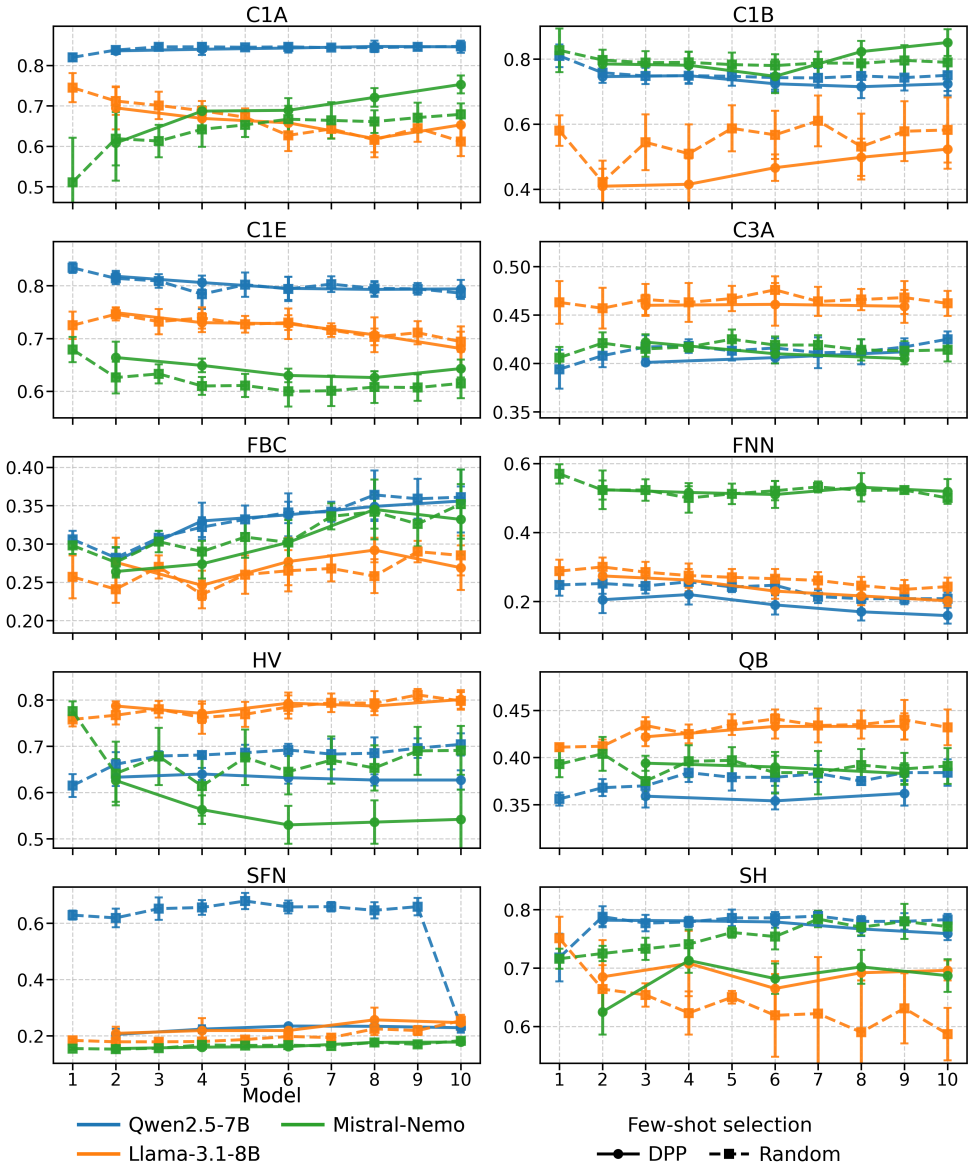


Figure 4.3: FS DPP vs Random results.

To test the Few-Shot capacity of the model, we decided to compare the performance using random data points against a representative set of examples, maintaining the dataset diversity using DPP. Table 4.3 reports the results for this comparison. Random selection risks subsets that lack diversity or fail to represent edge cases, while DPP ensures prompt stability by challenging the model with dissimilar patterns. In Few-Shot Learning, where models rely on limited examples to generalize, diverse subsets prevent overfitting to specific features and improve generalization. DPP-selected examples enhance prompt informativeness by showcasing varied cases, enabling the model to better understand nuanced relationships. This diversity reduces classification errors and improves accuracy by covering a broader range of inputs. A key observation across both FS Random and FS DPP is that increasing the number of shots does not consistently or monotonically improve performance for all models and datasets. Generally, performance often peaks at an intermediate number of shots (e.g., 1-shot, 5-shot, 6-shot) and can then plateau, fluctuate, or even decline as more shots are added (e.g., in the range 7-10 shots). This implies that simply providing more examples is not always the best strategy, regardless of the representativeness of the selected examples. There is no ideal threshold for the n-shots, since the performances vary across models and datasets. For instance, while using FS Random examples in Hyperpartisan Detection on the SH dataset, Llama-3.1-8b-Instruct's F1 score varies from .751 (1-shot) down to .623 (4-shot) and then to .587 (10-shot). However, the results do not indicate a clear best method across all scenarios. Peak performance for a given model and dataset can be achieved by either method, often at different n-shot values. For example, with Llama-3.1-8B-Instruct on the HV dataset, Random 9-shot yields an F1 of .811, while DPP 10-shot gives .801. Conversely, Mistral-Nemo-Instruct-2407 on C1B achieves its highest few-shot F1 of .851 with DPP 10-shot. FS Random can achieve high scores, possibly when the random selection happens to include particularly effective examples. However, its performance can be inherently more variable. FS DPP, by design, selects for diversity, which might be expected to lead to more robust or consistent improvements. While it achieves strong results in some cases, it also exhibits fluctuations and doesn't always outperform Random FS.

4.1.5 Chain of Thought

This setting provided reasoning steps with increasing levels of abstraction modeled as successive sub-task steps. Llama-3.1-8B-Instruct shows good performance with CoT on HP tasks, achieving the best F1/Accuracy in its section for both SH (F1 .792, Acc .795) and HV (F1 .757, Acc .764) datasets. It also performs well on PL Detection for C3A (F1 .465, Acc .459)

and QB (F1 .416, Acc .405), again leading in its section. Mistral with CoT achieves the overall best scores for FN Detection across all models and configurations on the FNN dataset (F1 .623, Acc .680) and across sections in the FBC dataset (F1 .315, Acc .397), and SFN (F1 .167, Acc .168). This suggests CoT might be particularly beneficial for this model on these specific complex tasks. Regarding Qwen, CoT helps it in achieving strong results on C1A (F1 .804, Acc .783), but only F1 .646 in C1B and F1 .575 in C1E, leading this section for these datasets. Nevertheless, in most cases, it proved to be suboptimal. The CoT prompting proved largely suboptimal across our experiments, showing significant improvement only for the FNN task. Our analysis suggests this underperformance stems primarily from language representation issues in the model's training data. When prompted in underrepresented languages with insufficient training tokens, the models struggled to process the unfamiliar linguistic patterns. Rather than aiding reasoning, these novel tokens appeared to confuse the models, disrupting their inference capabilities. Additionally, we observed that even in zero-shot-specific, codebook, and few-shot configurations, the models sometimes generated explanations unprompted, suggesting they were trained to occasionally provide reasoning alongside their answers. This built-in explanatory behavior likely accounts for why explicit CoT prompting offered minimal additional benefits despite its theoretical advantages and added complexity.

4.1.6 Insights from Fine-Tuning vs. In-Context Learning in LLMs

Across all the configurations tested in our experiments, FT emerged as the most effective method to apply models to the political domain, and, in particular, the disinformation subdomain. Specifically, for LLMs, the FT configuration demonstrated its efficacy in 28 out of 33 cases. Notably, Qwen particularly benefited from ICL, achieving its best performance with the following configurations: few-shot random for C1E (F1: 0.833) and SFN (F1: 0.678), and zero-shot codebook for SH (F1: 0.810). These results suggest that updating model parameters through FT is generally the most reliable way to optimize performance for downstream tasks in this domain. However, ICL remains a valid and convenient strategy for probing a model's task-specific knowledge without parameter updates. Despite our efforts to optimize prompts—by incorporating external domain-specific knowledge, employing rule-based approaches, and eliciting reasoning capabilities—ICL configurations still showed more limited effectiveness compared to FT. Lastly, both model architectures benefited from FT, with encoder-based models achieving superior performance on 6 out of 10 datasets, and smaller LLMs performing better on the remaining 4—particularly in tasks such as fake news and

political leaning detection, which require deeper world knowledge. It is important to note, however, that FT —especially when applied to LLMs—demands significant computational resources, making it a considerably resource-intensive approach.

4.1.7 Final Remarks

This work provides a comprehensive empirical benchmark comparing fine-tuning (FT) and in-context learning (ICL) across encoder-only and decoder-only transformer architectures on 10 diverse misinformation detection tasks spanning five languages (English, Spanish, Portuguese, Arabic, Bulgarian). Our extensive evaluation across 9 models and multiple prompting strategies reveals critical insights for practitioners selecting approaches for content moderation systems.

First, fine-tuning consistently outperforms in-context learning across nearly all settings. Despite ICL's appeal for rapid deployment without gradient updates, the performance gap is substantial: few-shot prompting with up to 10 examples still underperforms FT by significant margins on most datasets. Even chain-of-thought reasoning with decoders (CoT) fails to close this gap, challenging recent claims about LLMs' reasoning capabilities for classification. This finding justifies continued investment in FT infrastructure for production systems handling diverse moderation tasks.

Second, architectural choices interact meaningfully with task types. In FT, decoders (Llama-3.1-8B, Mistral-Nemo, Qwen2.5-7B) substantially outperform encoders on polarization and fake news detection (PB, FN), achieving F1 scores 15–20% higher. Conversely, encoders (RoBERTa, XLM-RoBERTa) dominate on hyperpartisan and hateful content detection (HP, HF), suggesting that encoder inductive biases better capture stylistic markers of extremism. This encoder-decoder task affinity should inform model selection based on deployment objectives.

Third, our analysis of ICL techniques reveals that prompt engineering rarely recovers FT-level performance. Determinantal point processes (DPP) for shot selection provide marginal improvements over random sampling but inconsistently—effectiveness depends on both the model and the task. Codebook-based prompting (listing classes and descriptions) emerges as the most reliable ICL strategy for classification, outperforming few-shot and chain-of-thought except on fake news tasks. However, even optimal ICL configurations achieve F1 scores 8–18% below FT baselines, reinforcing that ICL should be reserved for scenarios with strict latency requirements or where annotation budget prohibits FT.

		Hyperpartisan Det.				Political Leaning Det.				Harmful Text Det.						Fake News Det.			
		SH		HV		C3A		QB		C1E		C1A		C1B		FNN		FBC		SFN	
		F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc
Zero-Shot	**Generic**																				
	Llama-3.1-8B-Instruct	0.678	0.701	0.742	0.739	0.437	0.425	0.426	0.425	0.730	0.689	0.722	0.679	0.507	0.372	0.364	0.329	0.240	0.256	0.208	0.210
	Mistral-Nemo-Instruct-2407	0.686	0.707	0.607	0.623	0.422	0.421	0.360	0.360	0.720	0.677	0.565	0.516	0.793	0.695	0.551	0.542	0.304	0.327	0.144	0.145
	Qwen2.5-7B-Instruct	0.780	0.782	0.621	0.678	0.404	0.399	0.336	0.335	0.783	0.761	**0.851**	**0.857**	0.755	0.643	0.328	0.321	0.263	0.265	0.469	0.335
	Specific																				
	Llama-3.1-8B-Instruct	0.738	0.748	0.764	0.764	0.431	0.419	0.419	0.417	0.764	0.737	0.721	0.678	0.676	0.548	0.412	0.371	0.269	0.326	0.182	0.182
	Mistral-Nemo-Instruct-2407	0.740	0.750	0.574	0.593	0.409	0.408	0.352	0.356	0.719	0.677	0.613	0.562	0.818	0.732	0.549	0.540	0.315	0.354	0.150	0.150
	Qwen2.5-7B-Instruct	0.789	0.790	0.585	0.668	0.395	0.391	0.340	0.335	0.721	0.685	0.827	0.818	0.738	0.622	0.257	0.278	0.275	0.276	0.194	0.205
	Codebook																				
	Llama-3.1-8B-Instruct	0.748	0.753	0.715	0.722	0.382	0.407	0.324	0.344	0.762	0.729	0.728	0.686	0.670	0.538	0.388	0.349	0.304	0.405	0.165	0.166
	Mistral-Nemo-Instruct-2407	0.694	0.713	0.706	0.704	0.372	0.385	0.318	0.340	0.753	0.717	0.636	0.585	**0.864**	**0.806**	0.580	0.595	0.335	**0.440**	0.144	0.145
	Qwen2.5-7B-Instruct	**0.810**	**0.811**	0.636	0.688	0.377	0.395	0.311	0.323	0.785	0.769	0.832	0.838	0.820	0.735	0.341	0.320	0.256	0.256	0.422	0.295
Few-Shot	**Random 1-shot**																				
	Llama-3.1-8B-Instruct	0.751 ±0.037	0.755 ±0.034	0.758 ±0.015	0.756 ±0.015	0.463 ±0.022	0.457 ±0.024	0.411 ±0.003	0.411 ±0.006	0.725 ±0.026	0.684 ±0.032	0.745 ±0.036	0.708 ±0.044	0.580 ±0.047	0.446 ±0.047	0.288 ±0.033	0.270 ±0.023	0.257 ±0.028	0.305 ±0.036	0.184 ±0.006	0.185 ±0.006
	Mistral-Nemo-Instruct-2407	0.716 ±0.017	0.728 ±0.014	0.776 ±0.021	0.774 ±0.021	0.406 ±0.011	0.390 ±0.013	0.393 ±0.014	0.395 ±0.009	0.679 ±0.024	0.629 ±0.028	0.511 ±0.110	0.473 ±0.091	0.827 ±0.067	0.752 ±0.096	0.570 ±0.028	0.580 ±0.049	0.298 ±0.011	0.334 ±0.022	0.155 ±0.005	0.156 ±0.006
	Qwen2.5-7B-Instruct	0.718 ±0.041	0.730 ±0.035	0.615 ±0.025	0.683 ±0.013	0.394 ±0.020	0.402 ±0.014	0.356 ±0.007	0.355 ±0.006	**0.834 ±0.010**	**0.838 ±0.015**	0.820 ±0.009	0.854 ±0.005	0.810 ±0.035	0.723 ±0.051	0.248 ±0.031	0.273 ±0.011	0.306 ±0.011	0.332 ±0.021	0.629 ±0.016	0.640 ±0.012
	Random 2-shot																				
	Llama-3.1-8B-Instruct	0.664 ±0.041	0.685 ±0.031	0.767 ±0.020	0.765 ±0.019	0.457 ±0.021	0.451 ±0.018	0.412 ±0.007	0.416 ±0.008	0.746 ±0.012	0.708 ±0.015	0.712 ±0.034	0.668 ±0.039	0.422 ±0.066	0.302 ±0.052	0.300 ±0.027	0.278 ±0.020	0.241 ±0.018	0.269 ±0.019	0.179 ±0.011	0.179 ±0.011
	Mistral-Nemo-Instruct-2407	0.725 ±0.013	0.735 ±0.009	0.641 ±0.069	0.653 ±0.057	0.421 ±0.011	0.407 ±0.011	0.404 ±0.018	0.400 ±0.017	0.626 ±0.030	0.574 ±0.032	0.619 ±0.066	0.571 ±0.066	0.797 ±0.032	0.703 ±0.045	0.523 ±0.028	0.505 ±0.041	0.276 ±0.020	0.296 ±0.018	0.153 ±0.007	0.155 ±0.007
	Qwen2.5-7B-Instruct	0.788 ±0.018	0.789 ±0.017	0.661 ±0.026	0.711 ±0.014	0.408 ±0.012	0.411 ±0.010	0.368 ±0.009	0.361 ±0.009	0.814 ±0.011	0.804 ±0.013	0.838 ±0.007	0.855 ±0.003	0.758 ±0.012	0.648 ±0.016	0.252 ±0.030	0.277 ±0.013	0.282 ±0.004	0.299 ±0.013	0.619 ±0.033	0.640 ±0.022
	Random 3-shot																				
	Llama-3.1-8B-Instruct	0.654 ±0.020	0.678 ±0.017	0.780 ±0.018	0.778 ±0.019	0.466 ±0.016	0.461 ±0.019	0.434 ±0.009	0.428 ±0.011	0.732 ±0.024	0.692 ±0.029	0.701 ±0.034	0.655 ±0.039	0.544 ±0.086	0.412 ±0.078	0.285 ±0.030	0.268 ±0.021	0.270 ±0.015	0.331 ±0.028	0.179 ±0.005	0.180 ±0.005
	Mistral-Nemo-Instruct-2407	0.733 ±0.019	0.740 ±0.016	0.678 ±0.062	0.684 ±0.056	0.415 ±0.008	0.404 ±0.009	0.375 ±0.019	0.374 ±0.015	0.633 ±0.018	0.581 ±0.019	0.613 ±0.040	0.563 ±0.039	0.789 ±0.035	0.693 ±0.050	0.523 ±0.031	0.507 ±0.043	0.303 ±0.014	0.309 ±0.017	0.156 ±0.005	0.157 ±0.005
	Qwen2.5-7B-Instruct	0.777 ±0.014	0.780 ±0.013	0.679 ±0.010	0.720 ±0.007	0.417 ±0.012	0.421 ±0.005	0.370 ±0.006	0.359 ±0.005	0.809 ±0.013	0.799 ±0.018	0.846 ±0.007	0.852 ±0.004	0.748 ±0.025	0.634 ±0.033	0.245 ±0.022	0.271 ±0.010	0.308 ±0.009	0.323 ±0.013	0.652 ±0.040	0.663 ±0.032
	Random 4-shot																				
	Llama-3.1-8B-Instruct	0.623 ±0.037	0.656 ±0.028	0.762 ±0.035	0.760 ±0.035	0.463 ±0.020	0.456 ±0.022	0.425 ±0.010	0.421 ±0.007	0.739 ±0.020	0.700 ±0.024	0.688 ±0.024	0.640 ±0.027	0.509 ±0.090	0.380 ±0.081	0.275 ±0.035	0.264 ±0.023	0.234 ±0.018	0.267 ±0.047	0.180 ±0.012	0.183 ±0.013
	Mistral-Nemo-Instruct-2407	0.741 ±0.025	0.747 ±0.020	0.614 ±0.064	0.631 ±0.052	0.417 ±0.005	0.404 ±0.0007	0.396 ±0.017	0.392 ±0.013	0.610 ±0.017	0.558 ±0.017	0.642 ±0.043	0.593 ±0.044	0.790 ±0.053	0.694 ±0.069	0.500 ±0.043	0.475 ±0.059	0.290 ±0.014	0.291 ±0.014	0.168 ±0.003	0.172 ±0.004
	Qwen2.5-7B-Instruct	0.779 ±0.009	0.780 ±0.009	0.681 ±0.006	0.721 ±0.002	0.418 ±0.007	0.421 ±0.003	0.384 ±0.010	0.372 ±0.010	0.784 ±0.028	0.763 ±0.039	0.846 ±0.009	0.851 ±0.005	0.749 ±0.023	0.636 ±0.030	0.257 ±0.017	0.276 ±0.008	0.322 ±0.013	0.335 ±0.013	0.656 ±0.026	0.667 ±0.020
	Random 5-shot																				
	Llama-3.1-8B-Instruct	0.650 ±0.011	0.676 ±0.008	0.769 ±0.027	0.767 ±0.027	0.467 ±0.013	0.460 ±0.016	0.435 ±0.011	0.432 ±0.011	0.727 ±0.016	0.686 ±0.019	0.672 ±0.022	0.623 ±0.024	0.587 ±0.071	0.455 ±0.076	0.270 ±0.024	0.261 ±0.016	0.260 ±0.025	0.327 ±0.051	0.187 ±0.007	0.192 ±0.008
	Mistral-Nemo-Instruct-2407	0.761 ±0.009	0.762 ±0.009	0.676 ±0.060	0.682 ±0.053	0.425 ±0.010	0.411 ±0.011	0.397 ±0.014	0.392 ±0.013	0.611 ±0.022	0.559 ±0.021	0.653 ±0.030	0.604 ±0.032	0.783 ±0.037	0.684 ±0.052	0.512 ±0.030	0.490 ±0.042	0.309 ±0.027	0.314 ±0.031	0.166 ±0.005	0.169 ±0.006
	Qwen2.5-7B-Instruct	0.786 ±0.014	0.788 ±0.013	0.686 ±0.017	0.723 ±0.011	0.413 ±0.010	0.417 ±0.0006	0.379 ±0.014	0.369 ±0.016	0.802 ±0.023	0.786 ±0.032	0.845 ±0.004	0.842 ±0.008	0.747 ±0.029	0.634 ±0.038	0.242 ±0.015	0.270 ±0.006	0.332 ±0.018	0.339 ±0.019	**0.679 ±0.029**	**0.684 ±0.023**
	Random 6-shot																				
	Llama-3.1-8B-Instruct	0.619 ±0.071	0.657 ±0.047	0.785 ±0.025	0.783 ±0.026	**0.476 ±0.014**	**0.471 ±0.016**	**0.441 ±0.010**	0.438 ±0.010	0.730 ±0.014	0.689 ±0.017	0.627 ±0.039	0.576 ±0.040	0.567 ±0.074	0.434 ±0.076	0.266 ±0.028	0.259 ±0.017	0.265 ±0.027	0.327 ±0.059	0.198 ±0.010	0.202 ±0.011
	Mistral-Nemo-Instruct-2407	0.754 ±0.022	0.757 ±0.020	0.645 ±0.049	0.655 ±0.043	0.419 ±0.009	0.408 ±0.012	0.384 ±0.022	0.382 ±0.017	0.600 ±0.029	0.548 ±0.029	0.667 ±0.033	0.618 ±0.036	0.780 ±0.035	0.679 ±0.050	0.521 ±0.028	0.500 ±0.038	0.302 ±0.030	0.305 ±0.032	0.167 ±0.006	0.170 ±0.007
	Qwen2.5-7B-Instruct	0.786 ±0.010	0.786 ±0.009	0.692 ±0.013	0.727 ±0.007	0.416 ±0.010	0.421 ±0.004	0.379 ±0.009	0.369 ±0.009	0.794 ±0.023	0.775 ±0.034	0.846 ±0.007	0.843 ±0.009	0.743 ±0.037	0.630 ±0.048	0.247 ±0.026	0.273 ±0.011	0.341 ±0.014	0.348 ±0.011	0.658 ±0.023	0.667 ±0.017
	Random 7-shot																				
	Llama-3.1-8B-Instruct	0.622 ±0.097	0.661 ±0.066	0.794 ±0.020	0.792 ±0.020	0.464 ±0.015	0.457 ±0.018	0.434 ±0.018	0.432 ±0.017	0.716 ±0.013	0.672 ±0.015	0.643 ±0.023	0.592 ±0.024	0.610 ±0.078	0.479 ±0.082	0.261 ±0.024	0.257 ±0.013	0.268 ±0.017	0.346 ±0.048	0.194 ±0.012	0.199 ±0.013
	Mistral-Nemo-Instruct-2407	0.784 ±0.014	0.785 ±0.014	0.670 ±0.051	0.676 ±0.047	0.419 ±0.010	0.407 ±0.011	0.384 ±0.023	0.383 ±0.017	0.601 ±0.029	0.549 ±0.028	0.664 ±0.045	0.616 ±0.048	0.788 ±0.035	0.691 ±0.051	0.532 ±0.016	0.516 ±0.022	0.336 ±0.017	0.345 ±0.023	0.164 ±0.004	0.168 ±0.004
	Qwen2.5-7B-Instruct	0.789 ±0.009	0.790 ±0.009	0.683 ±0.033	0.724 ±0.018	0.411 ±0.016	0.417 ±0.008	0.383 ±0.009	0.373 ±0.010	0.803 ±0.015	0.788 ±0.023	0.844 ±0.006	0.838 ±0.010	0.742 ±0.030	0.627 ±0.039	0.214 ±0.019	0.259 ±0.008	0.342 ±0.013	0.344 ±0.013	0.659 ±0.017	0.667 ±0.013
	Random 8-shot																				
	Llama-3.1-8B-Instruct	0.590 ±0.089	0.639 ±0.053	0.793 ±0.026	0.791 ±0.026	0.466 ±0.011	0.459 ±0.014	0.435 ±0.015	0.433 ±0.013	0.703 ±0.016	0.657 ±0.018	0.616 ±0.043	0.565 ±0.044	0.533 ±0.101	0.402 ±0.098	0.246 ±0.025	0.248 ±0.014	0.258 ±0.022	0.321 ±0.050	0.225 ±0.022	0.230 ±0.026
	Mistral-Nemo-Instruct-2407	0.770 ±0.014	0.771 ±0.013	0.653 ±0.049	0.661 ±0.043	0.414 ±0.011	0.400 ±0.012	0.392 ±0.017	0.389 ±0.014	0.608 ±0.030	0.556 ±0.030	0.661 ±0.028	0.612 ±0.030	0.787 ±0.041	0.690 ±0.059	0.522 ±0.014	0.500 ±0.019	0.342 ±0.022	0.355 ±0.025	0.177 ±0.006	0.182 ±0.006
	Qwen2.5-7B-Instruct	0.780 ±0.009	0.780 ±0.009	0.685 ±0.034	0.724 ±0.022	0.412 ±0.013	0.416 ±0.007	0.375 ±0.004	0.364 ±0.004	0.795 ±0.014	0.775 ±0.018	0.844 ±0.005	0.837 ±0.009	0.748 ±0.030	0.635 ±0.040	0.208 ±0.031	0.259 ±0.013	**0.364 ±0.032**	0.366 ±0.032	0.646 ±0.029	0.656 ±0.024
	Random 9-shot																				
	Llama-3.1-8B-Instruct	0.631 ±0.060	0.665 ±0.041	**0.811 ±0.013**	**0.809 ±0.013**	0.468 ±0.017	0.462 ±0.022	0.440 ±0.021	**0.439 ±0.021**	0.711 ±0.022	0.667 ±0.025	0.645 ±0.034	0.595 ±0.036	0.578 ±0.092	0.447 ±0.097	0.235 ±0.027	0.242 ±0.014	0.290 ±0.014	0.384 ±0.019	0.219 ±0.016	0.229 ±0.024
	Mistral-Nemo-Instruct-2407	0.780 ±0.030	0.781 ±0.028	0.690 ±0.052	0.694 ±0.046	0.413 ±0.010	0.400 ±0.011	0.388 ±0.017	0.385 ±0.014	0.607 ±0.025	0.555 ±0.024	0.671 ±0.037	0.623 ±0.040	0.796 ±0.047	0.704 ±0.067	0.523 ±0.009	0.501 ±0.013	0.326 ±0.029	0.341 ±0.026	0.170 ±0.008	0.174 ±0.008
	Qwen2.5-7B-Instruct	0.780 ±0.014	0.781 ±0.013	0.696 ±0.021	0.732 ±0.014	0.417 ±0.009	0.416 ±0.009	0.384 ±0.008	0.373 ±0.008	0.794 ±0.011	0.773 ±0.016	0.846 ±0.008	0.840 ±0.014	0.743 ±0.040	0.630 ±0.051	0.208 ±0.016	0.256 ±0.008	0.359 ±0.026	0.362 ±0.027	0.659 ±0.051	0.668 ±0.025
	Random 10-shot																				
	Llama-3.1-8B-Instruct	0.587 ±0.045	0.635 ±0.029	0.798 ±0.019	0.796 ±0.019	0.462 ±0.013	0.457 ±0.017	0.432 ±0.019	0.431 ±0.019	0.694 ±0.019	0.647 ±0.021	0.612 ±0.036	0.561 ±0.037	0.582 ±0.100	0.453 ±0.103	0.243 ±0.026	0.245 ±0.012	0.285 ±0.026	0.367 ±0.048	0.254 ±0.012	0.273 ±0.019
	Mistral-Nemo-Instruct-2407	0.771 ±0.016	0.772 ±0.015	0.691 ±0.053	0.695 ±0.049	0.414 ±0.012	0.401 ±0.012	0.391 ±0.019	0.388 ±0.016	0.615 ±0.028	0.563 ±0.028	0.679 ±0.027	0.631 ±0.030	0.790 ±0.060	0.697 ±0.084	0.500 ±0.007	0.470 ±0.009	0.352 ±0.045	0.364 ±0.050	0.183 ±0.005	0.190 ±0.008
	Qwen2.5-7B-Instruct	0.783 ±0.010	0.783 ±0.010	0.704 ±0.024	0.736 ±0.016	0.425 ±0.008	0.421 ±0.007	0.384 ±0.014	0.373 ±0.014	0.786 ±0.011	0.760 ±0.016	0.847 ±0.006	0.841 ±0.010	0.750 ±0.048	0.640 ±0.063	0.208 ±0.018	0.256 ±0.008	0.361 ±0.014	0.362 ±0.013	0.236 ±0.020	0.237 ±0.021
	DPP 2-shot																				
	Llama-3.1-8B-Instruct	0.685 ±0.063	0.704 ±0.048	0.787 ±0.010	0.784 ±0.010	-	-	-	-	0.749 ±0.010	0.712 ±0.012	0.695 ±0.053	0.650 ±0.057	0.409 ±0.054	0.292 ±0.042	0.274 ±0.025	0.261 ±0.017	0.276 ±0.032	0.309 ±0.035	0.209 ±0.024	0.226 ±0.041
	Mistral-Nemo-Instruct-2407	0.625 ±0.039	0.660 ±0.027	0.626 ±0.046	0.640 ±0.038	-	-	-	-	0.664 ±0.030	0.614 ±0.032	0.608 ±0.093	0.562 ±0.090	0.785 ±0.022	0.686 ±0.031	0.524 ±0.056	0.511 ±0.074	0.264 ±0.006	0.276 ±0.017	0.155 ±0.009	0.158 ±0.011
	Qwen2.5-7B-Instruct	0.782 ±0.010	0.784 ±0.009	0.633 ±0.019	0.694 ±0.010	-	-	-	-	0.818 ±0.010	0.814 ±0.017	0.836 ±0.007	0.848 ±0.010	0.746 ±0.019	0.633 ±0.025	0.205 ±0.039	0.258 ±0.012	0.278 ±0.017	0.298 ±0.011	0.205 ±0.020	0.213 ±0.025
	DPP 3-shot																				
	Llama-3.1-8B-Instruct	-	-	-	-	0.460 ±0.011	0.453 ±0.013	0.422 ±0.010	0.416 ±0.011	-	-	-	-	-	-	-	-	-	-	-	-
	Mistral-Nemo-Instruct-2407	-	-	-	-	0.422 ±0.008	0.417 ±0.010	0.394 ±0.008	0.386 ±0.006	-	-	-	-	-	-	-	-	-	-	-	-
	Qwen2.5-7B-Instruct	-	-	-	-	0.401 ±0.003	0.407 ±0.012	0.359 ±0.012	0.350 ±0.010	-	-	-	-	-	-	-	-	-	-	-	-
	DPP 4-shot																				
	Llama-3.1-8B-Instruct	0.708 ±0.056	0.721 ±0.043	0.771 ±0.021	0.768 ±0.021	-	-	-	-	0.730 ±0.018	0.690 ±0.022	0.669 ±0.022	0.619 ±0.024	0.415 ±0.079	0.298 ±0.061	0.262 ±0.032	0.253 ±0.021	0.246 ±0.019	0.264 ±0.025	0.219 ±0.044	0.238 ±0.075
	Mistral-Nemo-Instruct-2407	0.713 ±0.021	0.724 ±0.015	0.563 ±0.031	0.590 ±0.022	-	-	-	-	0.649 ±0.013	0.598 ±0.014	0.687 ±0.012	0.639 ±0.013	0.781 ±0.041	0.681 ±0.057	0.516 ±0.018	0.496 ±0.026	0.274 ±0.019	0.279 ±0.019	0.160 ±0.008	0.164 ±0.008
	Qwen2.5-7B-Instruct	0.781 ±0.005	0.782 ±0.006	0.640 ±0.013	0.695 ±0.008	-	-	-	-	0.806 ±0.013	0.793 ±0.019	0.840 ±0.014	0.841 ±0.023	0.749 ±0.025	0.636 ±0.033	0.220 ±0.029	0.259 ±0.010	0.330 ±0.024	0.343 ±0.017	0.224 ±0.014	0.231 ±0.017
	DPP 6-shot																				
	Llama-3.1-8B-Instruct	0.665 ±0.047	0.686 ±0.032	0.793 ±0.023	0.791 ±0.023	0.461 ±0.022	0.455 ±0.021	0.433 ±0.013	0.429 ±0.017	0.728 ±0.029	0.688 ±0.034	0.658 ±0.035	0.608 ±0.037	0.466 ±0.041	0.337 ±0.035	0.230 ±0.022	0.235 ±0.013	0.277 ±0.031	0.291 ±0.024	0.219 ±0.017	0.226 ±0.025
	Mistral-Nemo-Instruct-2407	0.682 ±0.026	0.698 ±0.021	0.530 ±0.041	0.566 ±0.028	0.410 ±0.010	0.410 ±0.015	0.390 ±0.012	0.381 ±0.011	0.630 ±0.013	0.578 ±0.015	0.689 ±0.030	0.643 ±0.032	0.747 ±0.052	0.636 ±0.070	0.510 ±0.039	0.490 ±0.056	0.302 ±0.029	0.304 ±0.028	0.162 ±0.007	0.170 ±0.007
	Qwen2.5-7B-Instruct	0.779 ±0.010	0.780 ±0.010	0.632 ±0.013	0.691 ±0.008	0.406 ±0.006	0.413 ±0.009	0.354 ±0.009	0.345 ±0.008	0.795 ±0.022	0.777 ±0.033	0.843 ±0.011	0.840 ±0.017	0.724 ±0.024	0.604 ±0.030	0.190 ±0.028	0.249 ±0.009	0.338 ±0.028	0.352 ±0.024	0.235 ±0.003	0.242 ±0.006
	DPP 8-shot																				
	Llama-3.1-8B-Instruct	0.692 ±0.013	0.703 ±0.011	0.787 ±0.011	0.786 ±0.012	-	-	-	-	0.707 ±0.033	0.663 ±0.036	0.619 ±0.034	0.568 ±0.035	0.498 ±0.057	0.367 ±0.052	0.216 ±0.027	0.227 ±0.012	0.292 ±0.016	0.303 ±0.020	0.257 ±0.043	0.287 ±0.079
	Mistral-Nemo-Instruct-2407	0.702 ±0.029	0.710 ±0.025	0.536 ±0.047	0.571 ±0.033	-	-	-	-	0.626 ±0.012	0.575 ±0.013	0.721 ±0.023	0.678 ±0.026	0.823 ±0.033	0.742 ±0.050	0.531 ±0.041	0.521 ±0.063	0.345 ±0.039	0.352 ±0.041	0.177 ±0.007	0.187 ±0.006
	Qwen2.5-7B-Instruct	0.767 ±0.012	0.767 ±0.012	0.627 ±0.012	0.688 ±0.007	-	-	-	-	0.793 ±0.014	0.772 ±0.021	0.847 ±0.014	0.841 ±0.020	0.715 ±0.035	0.593 ±0.045	0.170 ±0.025	0.243 ±0.008	0.349 ±0.019	0.362 ±0.028	0.234 ±0.015	0.236 ±0.016
	DPP 9-shot																				
	Llama-3.1-8B-Instruct	-	-	-	-	0.459 ±0.017	0.454 ±0.019	0.433 ±0.014	0.428 ±0.014	-	-	-	-	-	-	-	-	-	-	-	-
	Mistral-Nemo-Instruct-2407	-	-	-	-	0.405 ±0.006	0.407 ±0.009	0.383 ±0.011	0.376 ±0.011	-	-	-	-	-	-	-	-	-	-	-	-
	Qwen2.5-7B-Instruct	-	-	-	-	0.412 ±0.007	0.416 ±0.009	0.362 ±0.013	0.353 ±0.011	-	-	-	-	-	-	-	-	-	-	-	-
	DPP 10-shot																				
	Llama-3.1-8B-Instruct	0.696 ±0.017	0.708 ±0.013	0.801 ±0.020	0.800 ±0.021	-	-	-	-	0.681 ±0.042	0.634 ±0.044	0.653 ±0.036	0.603 ±0.037	0.523 ±0.060	0.390 ±0.059	0.202 ±0.016	0.220 ±0.005	0.269 ±0.029	0.280 ±0.034	0.247 ±0.028	0.267 ±0.042
	Mistral-Nemo-Instruct-2407	0.687 ±0.028	0.694 ±0.020	0.542 ±0.064	0.576 ±0.046	-	-	-	-	0.643 ±0.017	0.592 ±0.018	0.753 ±0.022	0.715 ±0.026	0.851 ±0.041	0.788 ±0.066	0.519 ±0.036	0.502 ±0.052	0.332 ±0.046	0.346 ±0.058	0.178 ±0.008	0.186 ±0.008
	Qwen2.5-7B-Instruct	0.759 ±0.011	0.760 ±0.011	0.627 ±0.021	0.688 ±0.012	-	-	-	-	0.794 ±0.017	0.774 ±0.024	0.846 ±0.015	0.841 ±0.022	0.724 ±0.036	0.604 ±0.044	0.159 ±0.023	0.239 ±0.005	0.356 ±0.041	0.368 ±0.050	0.229 ±0.018	0.230 ±0.019
CoT	**CoT**																				
	Llama-3.1-8B-Instruct	0.792	0.795	0.757	0.764	0.465	0.459	0.416	0.405	0.557	0.502	0.503	0.454	0.226	0.160	0.377	0.340	0.280	0.358	0.161	0.161
	Mistral-Nemo-Instruct-2407	0.713	0.728	0.732	0.729	0.464	0.465	0.401	0.386	0.252	0.275	0.584	0.530	0.528	0.394	**0.623**	**0.680**	0.315	0.397	0.167	0.168
	Qwen2.5-7B-Instruct	0.711	0.725	0.714	0.729	0.422	0.428	0.355	0.349	0.575	0.522	0.804	0.783	0.646	0.511	0.491	0.456	0.277	0.340	0.147	0.149

Table 4.2: Comparison of F1 and accuracy across all models, configurations, and datasets. Each cell shows the mean F1 and accuracy (± one standard deviation for Few-Shot only, since Zero-Shot and CoT use greedy decoding and have no standard deviation). **Boldface** = best overall; underlined = best within each section.

Fourth, multilingual generalization shows both structured patterns and task variance. Models fine-tuned on English transfer surprisingly well to Arabic and Bulgarian (F1 drop $\leq 8\%$), but Romance languages (Spanish, Portuguese) exhibit greater variance depending on the task. This suggests that linguistic distance alone does not determine cross-lingual transfer; task-specific linguistic phenomena (e.g., sarcasm, cultural references) significantly impact generalization.

Finally, our work underscores that the rapid evolution of LLMs should not obscure the fundamentals: for production content moderation systems, FT remains the gold standard. While ICL offers benefits for development and exploration, practitioners should recognize its inherent performance ceiling and invest in curated training data and FT pipelines for domains where moderation accuracy directly impacts user safety and platform trust.

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

While the previous section benchmarked fine-tuning and in-context learning approaches across various architectures and datasets, we now apply these findings to PartisanLens, a novel multilingual dataset designed specifically for detecting hyperpartisan and conspiratorial narratives in immigration-related European media. This dataset addresses an important gap in resources for underrepresented Romance languages (Spanish, Italian, Portuguese) and enables us to evaluate how our selected models perform on domain-specific content with real annotator disagreement and explicit rhetorical reasoning.

4.2.1 Annotation Process Results and Dataset Statistics

The IAA results are presented in Table 4.3 and indicate substantial to almost perfect agreement across all languages [Landis and Koch, 1977]. Hyperpartisan labels were most consistent in Portuguese and Spanish, while PRCT achieved slightly higher agreement in Italian. Stance annotations were somewhat lower than the other tasks but still showed substantial reliability.

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

Table 4.3: Inter-Annotator Agreement (Fleiss' κ) across annotation tasks. Agreement ranges from substantial to almost perfect across all languages.

Stance-conditioned patterns. Figure 4.4 shows how PRCT and hyperpartisan content correlate with stance. The strongest signal comes from anti-immigration headlines: 81% contain PRCT claims and 56% are hyperpartisan. Neutral headlines are minor (14% PRCT, 27% hyperpartisan), while pro-immigration headlines are the least polemical (3% PRCT, 12% hyperpartisan). These results confirm that conspiratorial narratives cluster much more around negative stances than hyperpartisan style, which is more broadly distributed.

Language-level variation Figure 4.5 highlights how the two labels distribute across the languages' headlines. Italian outlets display the strongest conspiratorial signal: 27% of their immigration headlines contain PRCT language, roughly three times the Spanish rate (8%) and nine times the Portuguese rate (3%). Hyperpartisan style shows a different trend: it peaks in Portuguese (70%) and remains consistently high in Spanish and Italian (≈60%). Taken together, these findings suggest that while hyperpartisan framing is common across languages, conspiratorial narratives concentrate more on Italian media.

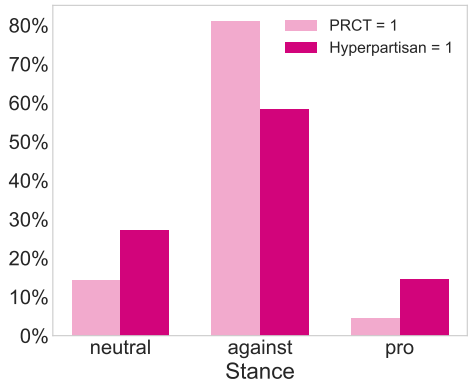


Figure 4.4: Hyperpartisan and PRCT stance proportions.

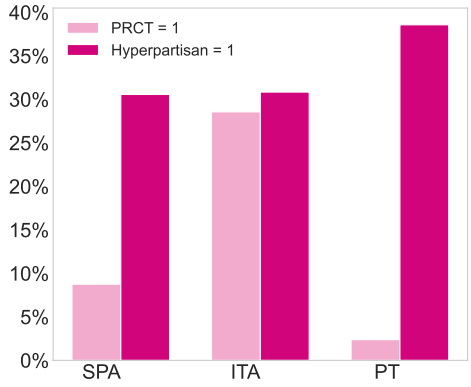


Figure 4.5: Hyperpartisan and PRCT content by language.

4.2.2 Classification Results

Table 4.4 reports macro-averaged precision (P), recall (R), and F1 across the three tasks and all our proposed models.

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

Table 4.4: Macro-averaged precision (P), recall (R), and F1 across tasks and all our considered models. Best metrics per task are marked in **bold**.

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

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

Stance. Stance is the hardest task (best F1=0.61), with Llama 70B (fine-tuned / ratio-

nales) and Nemo (few-shot) tying at the top. Gains mostly come from modest recall (0.61), while smaller models stay below (F1=0.58). The pronounced difficulty likely reflects both class distributions (e.g., very rare pro stance in Spanish) and the need for contextual inferences beyond surface style, since rationale inputs can even hurt (with Nemo obtaining a F1=0.42). The overall difficulty suggests that stance requires finer-grained semantic understanding than hyperpartisan or PRCT, and remains an open challenge even for very large models.

Three main insights emerge from these results. (*i*) Supervision matters: few-shot and fine-tuning typically add 4–7 F1 points over zero-shot. (*ii*) Scale helps but is not enough: Llama 70B is most consistent, yet Llama 8B few-shot matches top hyperpartisan scores. (*iii*) Task complexity varies: *PRCT* > *hyperpartisan* > *stance*. PRCT benefits from distinctive lexical/conceptual signatures, hyperpartisan is stylistic and diffuse, and stance seems to demand finer pragmatic cues. A detailed per-language analysis, including $F1_{MACRO}$, P/R, and error profiles, is provided in subsection 4.2.2.

Per-language Analysis of Classification Results

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

Table 4.5: Fine-tuning with rationales results by language (macro-averaged).

Table 4.5 presents the performance by language across Spanish (SPA), Italian (ITA), and Portuguese (PT) under the fine-tuning with rationales setting. Looking at the results, hyperpartisan is robust in all three languages, with the best scores in Italian ($F1_{MACRO}$= 0.77, Llama 3 70B) and a tie at the top in Spanish ($F1_{MACRO}$= 0.75, Llama 3 8B and Nemo).

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

Table 4.6: Mean Cohen's κ by language and label. The agreement pairs are: **H–H** (human vs. human), **L–H** (`Llama 3.3--70B` vs. human), **G–H** (`gpt-5-mini` vs. human).

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

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

4.2.3 LLMs as Potential Data Annotators

To evaluate whether LLMs can serve as reliable annotators, we compare their agreement against the reference level of human agreement across languages. The core question is: *does an LLM approach human agreement levels on our three targets (Hyperpartisan, PRCT, Stance) in the three languages?* We measure mean pairwise Cohen's κ for human–human (H–H) and LLM–human pairs using two models: (*i*) `Llama 3.3--70B` fine-tuned with labels and rationales (L–H), and (*ii*) `gpt-5-mini` in inference mode (G–H), comparing their performance across the three target languages.[1]

Results

Human agreements (H-H)

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

LLM vs. human agreements (L–H / G–H)

Models remain below H–H across languages, but `gpt-5-mini` (G–H) is closer to human agreement than `Llama 3.3--70B` (L–H). In Spanish, `gpt-5-mini` improves over `Llama 3.3--70B` on all tasks, especially on the Stance category (0.633 vs. 0.364). In Italian, however, `Llama 3.3--70B` obtains better results in Hyperpartisan and PRCT than `gpt-5-mini`, but `gpt-5-mini` still obtains way better performance on Stance, being close to human agreement (0.567 vs. 0.598). In Portuguese, `gpt-5-mini` yields the large PRCT difference in comparison to `Llama 3.3--70B` (0.662 vs. L–H 0.124), but, in this case, for the three labels, the LLMs are still way below human agreement. In sum, although we evaluated two large and competitive LLMs, their limited agreement with experts shows that fully unsupervised annotation remains out of reach for these nuanced categories. Human supervision is therefore still indispensable, making resources in this domain crucial for advancing robust detection methods.

4.2.4

User-Persona LLMs

Persona-based prompting was used to condition LLM outputs to reflect specific personas, enabling simulation of human annotators' socio-demographic and ideological perspectives.

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

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

Table 4.7: Cohen's κ for each user-persona vs. its corresponding human annotator, comparing `Llama 3.1-70B` (L70B) and `gpt5-mini` (gpt5-m) inference.

We examined whether persona-conditioned LLMs can better emulate patterns of human annotation behaviour when grounded in realistic user profiles—that is, whether an LLM simulating a particular profile aligns more closely with annotators sharing that same profile. Personas were constructed from real socio-demographic and political data (Political Compass and PRISM quiz responses) combined with annotator demographics.
footnoteFurther details on persona construction and experimental setup are provided in the Methodology section.

4.2.5 Results

Agreement user-persona vs. corresponding human

Table 4.7 reports Cohen's κ between each persona-conditioned LLM and the human it simulates, using `Llama--70B` and `gpt5-mini`. If we compare the overall results to the generic L–H baselines in Table 4.6, LLM persona conditioning helps in several settings, and `gpt5-mini` often amplifies these gains, especially for PRCT and Stance. Regarding Spanish annotators, `Llama--70B` personas improves PRCT over the generic baseline (0.599 vs. 0.483) and slightly Hyperpartisan (0.459 vs. 0.424), but lower Stance (0.252 vs. 0.364). This suggests that persona conditioning may help capture content focus and priorities rather

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

Table 4.8: Selected headlines with human and user-personas disagreements.

than polarity judgments. With `gpt5-mini`, agreement rises across all three: Hyperpartisan (0.598 vs. 0.635), PRCT (0.711 vs. 0.754), and Stance (0.663 vs. 0.767).

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

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

Looking at the overall results, persona conditioning increases LLM–human PRCT agreement in all languages (largest gains in ITA and PT). Using `gpt5-mini`, it clearly improves Stance in SPA/ITA. Hyperpartisan performance remains mixed (SPA gains, ITA near parity, and PT drops). Overall levels are still way down compared to the H–H reference (Table 4.6), but grounding in real profiles recovers meaningful human-like variability for PRCT and, with stronger inference, also for stance.

4.2.6 Error analysis

We explore LLMs' deficiencies and attribute them to several factors. One important issue is the limited understanding of sarcastic and metaphorical sentences. In Table 4.8, we can see some examples of headlines with disagreements among annotators and user personas. A second limitation concerns the model's insufficient understanding of semantics and context. For example, the Portuguese term *Imigrantes* can denote both "immigrants" and the name of a Brazilian highway. The model struggled to disambiguate the meaning based on context. Similarly, the word "invading" carries different connotations across Portuguese, Italian, and Spanish, and the LLM's rigid interpretation suggests a lack of sensitivity to cultural context. Another illustrative shortcoming is the model's failure to interpret irony, sarcasm, or implicit critique. Such use of rhetorical devices is a subtle but powerful tool in political discourse, where it frequently conveys unstated or indirect opinions. As a matter of fact, LLMs struggle when elaborating on data that is scarcely represented in their training set, given the limited token distribution for languages like Spanish, Italian, and, especially, Portuguese [Roxas, 2024]. Aware of the hurdles that underrepresented languages pose, it is necessary to take action to develop tailored resources. Thus, in the next chapter, we are going to discuss the creation of ad-hoc solutions (i.e. datasets) focusing on the Italian Language.

4.2.7 Classification Errors

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

to counteract default negative bias.

4.2.8 Final Remarks

PartisanLens addresses a critical gap in multilingual resources for detecting hyperpartisanship and conspiratorial narratives in immigration-related European media, covering three underrepresented Romance languages (Spanish, Italian, Portuguese). Our dataset and comprehensive evaluation reveal how LLMs and fine-tuned architectures perform on nuanced multilingual content with annotator disagreement and explicit rhetorical reasoning.

First, our inter-annotator agreement analysis reveals robust dataset quality: hyperpartisan and PRCT labels achieved substantial-to-almost-perfect agreement across languages (Fleiss' κ 0.706–0.880), validating that our taxonomy captures consistent linguistic phenomena. The strong stance-conditioned signal (81% PRCT in anti-immigration headlines) confirms that conspiratorial narratives cluster around negative political positions, while hyperpartisan style distributes more broadly across stance dimensions.

Second, language-level analysis uncovers distinct rhetorical patterns. Italian media display the strongest conspiratorial signal (27% PRCT vs. 8% Spanish, 3% Portuguese), suggesting cultural and institutional factors shape how conspiracy narratives manifest in different political environments. Conversely, all languages maintain high hyperpartisan rates (60%–70%), indicating that extreme framing transcends linguistic boundaries. These findings underscore that computational approaches must account for language-specific discourse patterns rather than assuming monolithic global misinformation phenomena.

Third, our extensive LLM evaluation across zero-shot, few-shot, fine-tuning, and rationale-enhanced prompting strategies reveals that LLMs struggle with understudied languages and rhetorical subtlety. Llama 70B with fine-tuning and rationales achieves the best performance on Italian (F1 0.77 hyperpartisan, 0.86 PRCT), yet fine-tuning remains essential—zero-shot LLMs heavily underperform despite near-perfect neutrality classifications, exposed through human evaluation. This dependency on expensive annotation contradicts the zero-shot LLM narrative and justifies continued dataset creation for low-resource and underrepresented language communities.

Fourth, our error analysis identifies systematic LLM limitations in Romance languages: (1) sarcasm and metaphorical reasoning elude current models; (2) context-dependent polysemy (e.g., Portuguese *Imigrantes* as a highway name) confuses token-based reasoning; (3) cultural specificity of rhetorical devices requires language-community expertise that general-purpose

models lack. These findings suggest that future advances require language-specific fine-tuning and culturally grounded annotation frameworks, not just larger multilingual models trained on web-scale data.

Finally, PartisanLens establishes a foundation for investigating how hyperpartisan framing and conspiratorial narratives localize across political systems and media landscapes. By combining high-quality human annotation, extensive LLM evaluation, and granular multilingual analysis, we demonstrate that bias detection is fundamentally contextual and language-specific. Our work advocates for targeted resource creation in underrepresented languages and cultural contexts, recognizing that one-size-fits-all misinformation detection approaches inevitably fail underrepresented communities.

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

The contributions of this paper are threefold. Firstly, we developed resources for a low-represented language like Italian in the landscape of misinformation detection, covering the most recent and polarizing events like climate change, green policies and Euroscepticism. For this reason, we employed a complex fine-grained taxonomy of 17 rhetorical biases, mutuated and adapted from Piskorski et al. [2023c]. The final dataset contains 48 articles and 1,010 sentences manually annotated. Secondly, the annotation methodology employed allowed us to study the composition of the articles from the selected libertarian-right media source, providing informative insights on how hyperpartisanship spreads and evolves across the news' structure as well as how it is systematized through the usage of rhetorical devices. Lastly, the crafted dataset was used for benchmarking and evaluating a small set of models (GPT-4-mini and BERTbase), architectures and learning strategies (ICL, FT).

4.3.1 Dataset

The key general statistics are represented in Table 4.9. On average, each article is composed of 21.12 sentences of 40.26 words and contains 2.12 instances of rhetorical bias. Thus, the document with which we are working are short texts that are slightly unbalanced: hyperpartisan (304) and neutral (706). Considering the ratio of rhetorical biases, this distribution stresses the high frequency with which these dialogic techniques are adopted by the selected bias source.

Topic Distribution

As mentioned before, the topics are related to climate change and Euroscepticism, which were selected for their polarizing attributes. To inspect further their granularity providing a better understanding on how they fork, we applied the BERTopic framework [Grootendorst, 2022] with parameters optimized to preserve local structure[2]. Thanks to human inspection, strict keywords were selected to detect three main topics: science, institutions and *Other* and their further subdivision into subtopics like *institutions.Italy*, *science.cars*, etc. Table 4.10 presents the topic distribution. It emerges that the data covers adjacent actual topics like electric cars, specific countries (China, Italy) as main actors, and politically correct, enriching the variety of the content hypothesized at the beginning.

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 4.9: Dataset Statistics

4.3.2 Hyperpartisan-Rhetorical Bias Relation

Although theoretically the relationship between rhetorical biases and hyperpartisanship is taken for granted in the literature [Potthast et al., 2018b, Kiesel et al., 2019c], by calculating the correlation patterns amongst them, it is possible to determine their granular and intertwined properties (see Figure 4.6). By looking at this correlation matrix, evident patterns emerge suggesting that there is a correlation between the usage of rhetorical biases and hyperpartisan sentences. In other words, these devices appear to polarize the meaning. Particularly, Appeal to Values/Flag Waving is correlated with Appeal to Fear (.40), implying that patriotic feelings or moral values justifying positions or actions are relevant while stimulating responses in the audience towards specific viewpoints or actions. Additionally, Consequential Oversimplifica-

[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)

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 4.10: Topic Distribution.

tion is a good strategy to manipulate the readers' fears. In other words, the oversimplification of events serves as a means to carry and stimulate anxiety, touching only sensitive topics, triggering certain behaviours that are picked by supporting an argument superficially and with a frame to select the most emotional information to be framed in a way to create a semantic and logical short circuit, eliciting (illogical) fears promoted by a given ideology. On top of that, in hyperpartisan sentences, Loaded Language and Exaggeration/Minimisation (hyperbolic language) appear together (.34 correlation), maximizing their overstating power. Furthermore, Reductio ad Hitlerum, a technique aiming at discrediting someone, usually a political opponent, by attributing nazist traits, is combined with Name Calling (.25 correlation). This result is quite expected, since it highlights the common ground between these two devices, resulting in addressing someone. In addition to this, Reductio ad Hitlerum shows a .27 correlation with Appeal to Values, too. This points out how, after presenting an actor in a bad light, the authors put this character in contrast with the promotion of politicized values. Lastly, it is important to notice that the fearing rhetoric is stressed with hyperbolic language (Exaggeration/Minimisation) co-occurring together (.21).

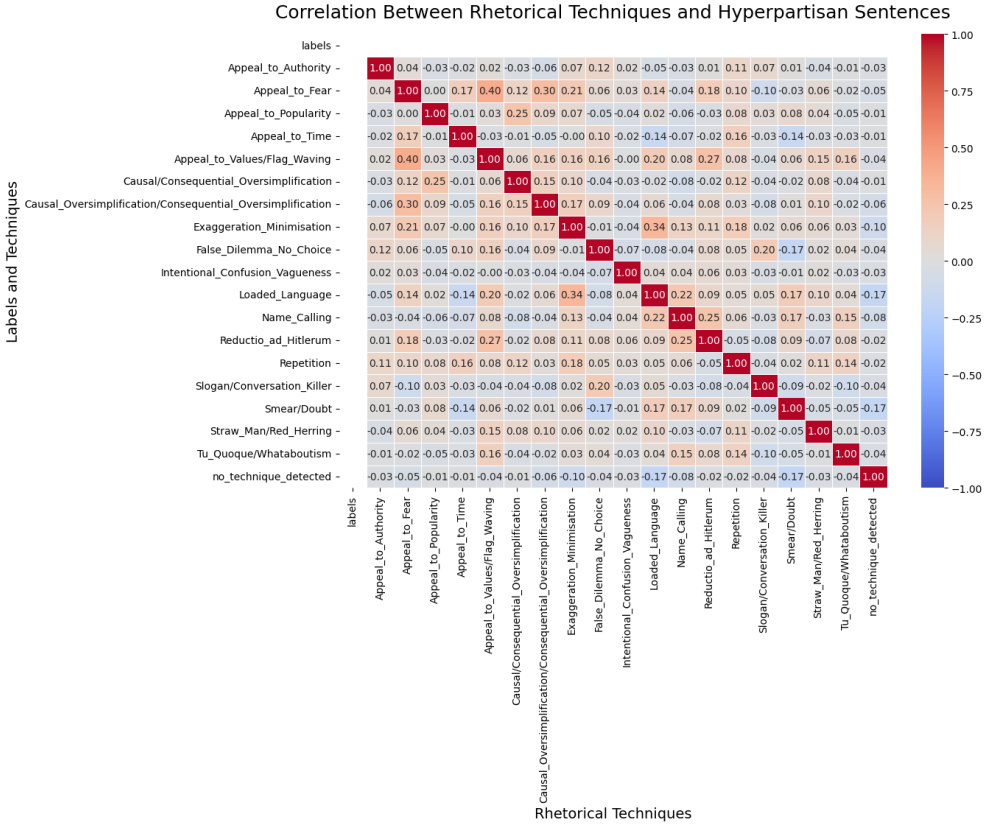


Figure 4.6: Correlation between Hyperpartisan sentences and techniques.

After that, the measure of Log Ratio with regard to the label gives information on the discriminative magnitude of the rhetorical bias. 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 negative values indicate a higher frequency in the neutral group. As a result, Figure 4.7 depicts which rhetorical biases are most determinant in distinguishing between hyperpartisan and neutral sentences, showing how these two distinct types of utterances differ in the involvement of rhetorical biases. Thus, neutral sentences are usually characterized by no rhetorical biases (”no_technique_detected”), whereas ”Re-

duction_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. [2025e]'s definition of hyperpartisanship as well as the previous definitions used in the literature by Kiesel et al. [2019b], Lyu et al. [2023].

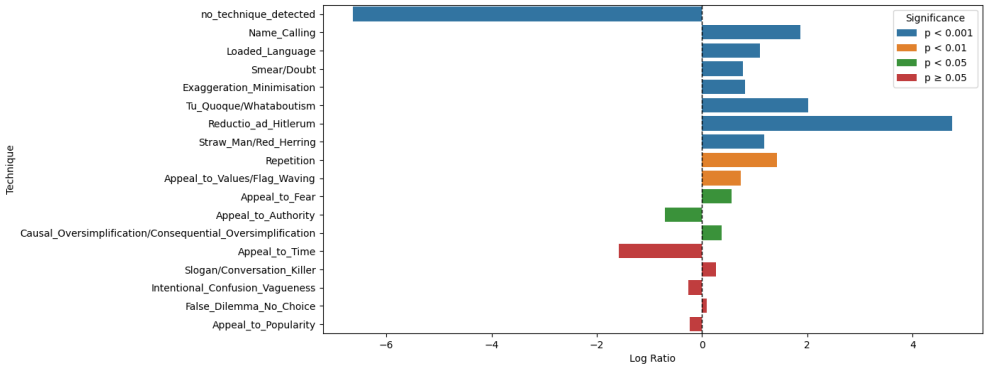


Figure 4.7: Log Ratio of the Rhetorical Techniques with regards to Hyperpartisan/Neutral Label.

Lastly, the χ^2 test comparing the frequency of each rhetorical technique across the binary labels demonstrates that in most of the cases a relationhip between independent variables exist with statistical evidence. This evidence is described in Table 4.11.

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 4.11: 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.

4.3.3 Topological Distribution of Rhetorical Biases

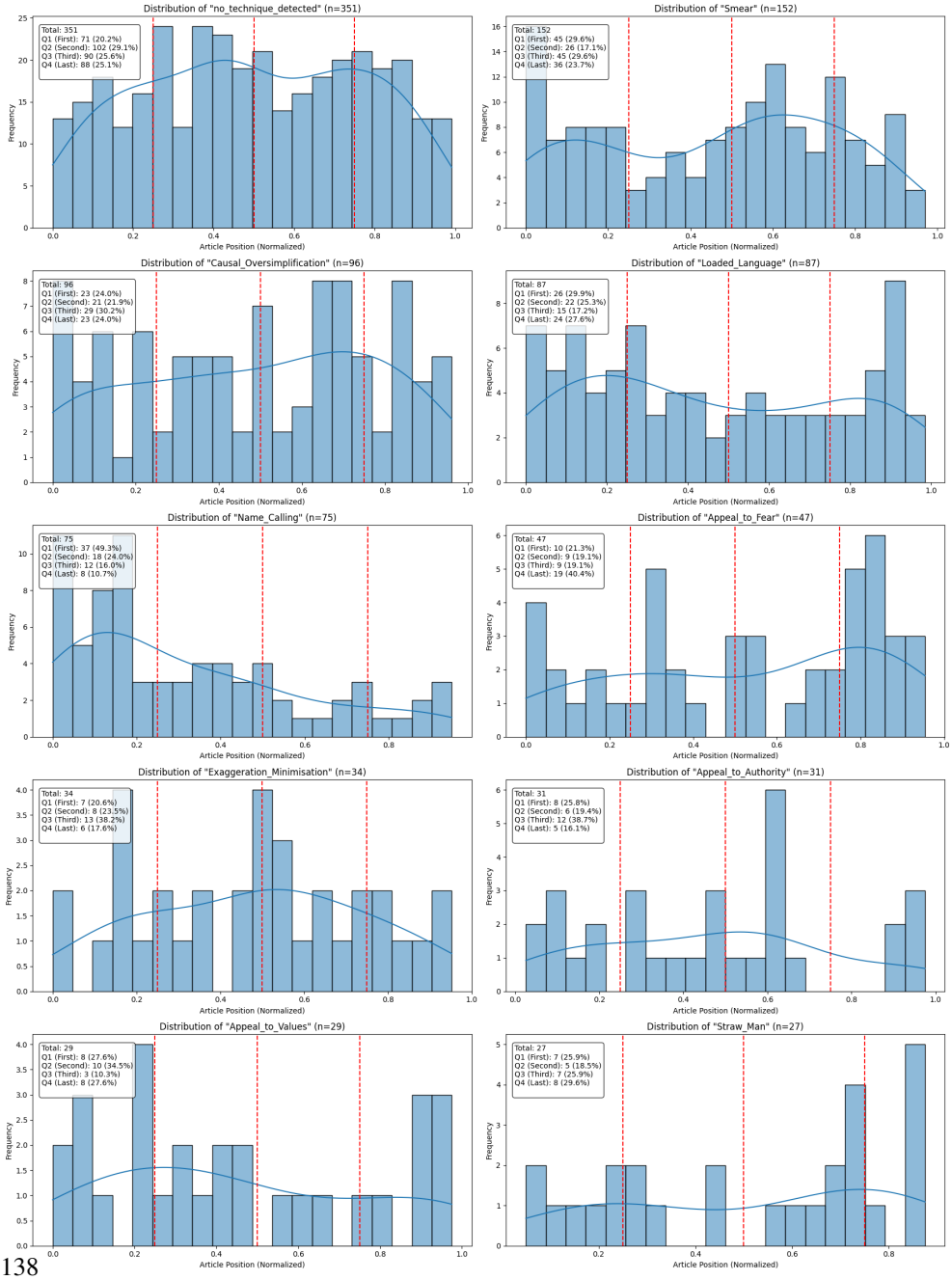


Figure 4.8: Distribution of the Rhetorical Techniques across article quantiles.

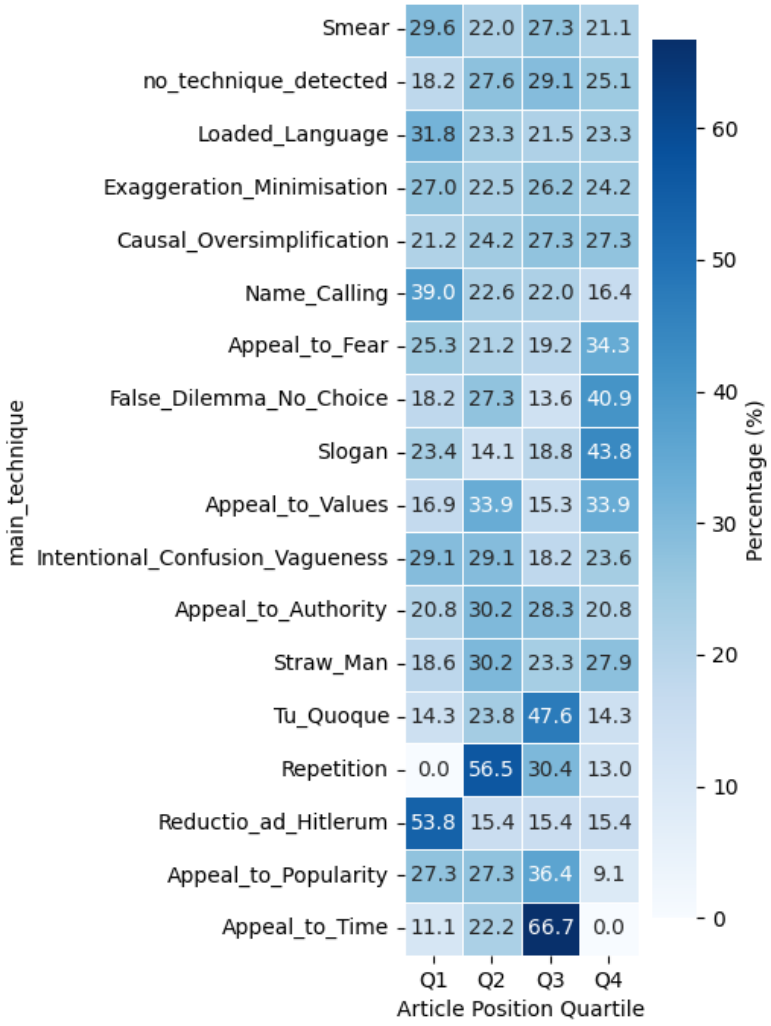


Figure 4.9: Distribution of Rhetorical Techniques across article quartiles

Fig. 4.10 shows the average of hyperpartisan sentences across the articles' structure, while Fig. 4.9 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 position bins. Lastly, we computed the average hyperpartisan score for each position bin. Fig. 4.10 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 how rhetorical techniques are adopted to convey and shape the message (Fig. 4.9 and Fig. 4.11). Firstly, 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.

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 mostly 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

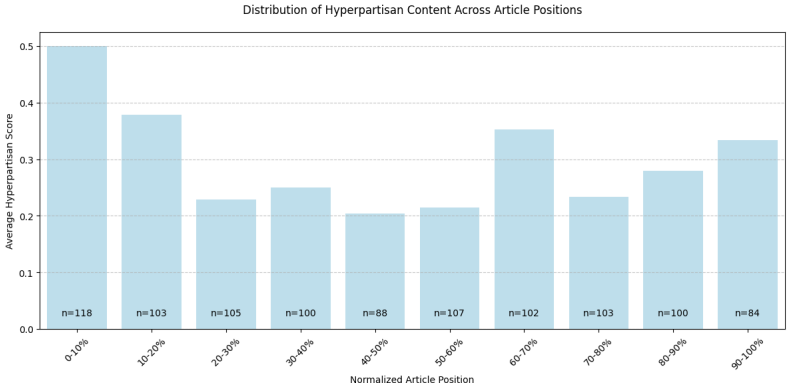


Figure 4.10: Hyperpartisan Contamination Level per sentence position (deciles). *n* represents the number of sentences. Because of articles have dissimilar number of sentences, we normalize their length.

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 4.12: Comparison of Hyperpartisan and Rhetorical Bias Classification Models

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 clickbait and hyperpartisan content.

4.3.4 Computational Baselines

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

The hyperparameters employed and the results of the evaluation on the detection of hy-

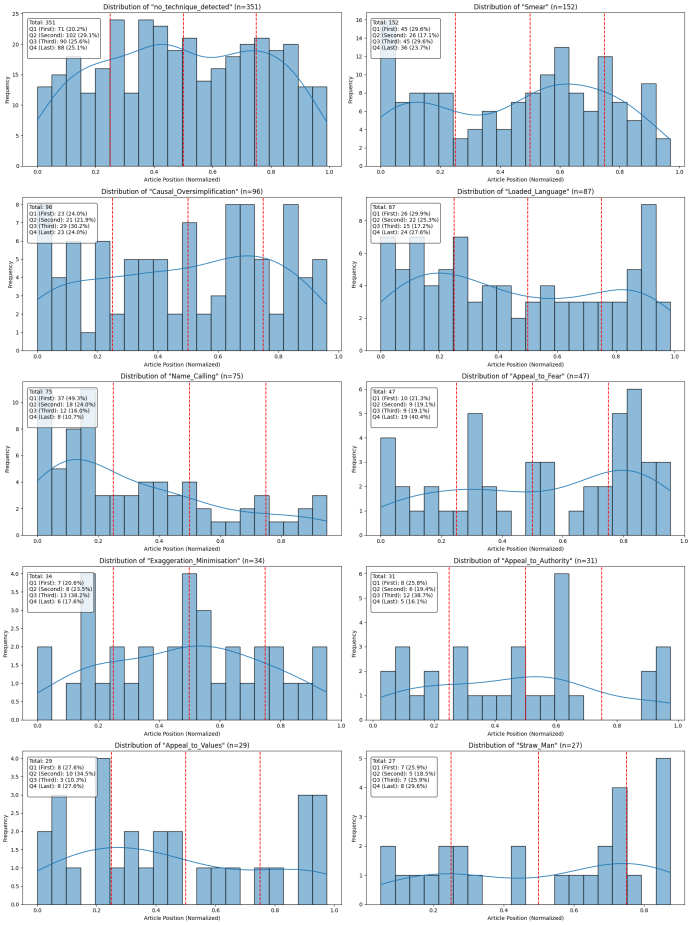


Figure 4.11: Distribution of Rhetorical Techniques across articles' bodies normalized.

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

perpartisan and rhetorical bias are shown in Tables 4.13 and 4.12. The results demonstrate significant variability in the metric scores between different models and methodologies.

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 increases 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 highly 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 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.

4.3.5 Final Remarks

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.

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

With the previous work, we have validated the intertwined logical and linguistic relationship between the political construct of hyperpartisanship and a taxonomy of rhetorical biases, investigating how a specific Italian source utilizes these devices to shape the communication process. With *IHPP: A Paragraph-Level Dataset for Investigating the Pragmatics of Hyper-*

partisan Italian News the intent is to expand that previous work with the introduction of more Italian data points and testing the assumption that HP language breaks Gricean's maxims, exploiting non-literal language at the semantic and pragmatic levels. Namely, by leveraging some rhetorical figures (e.g. Figurative Speech, Hyperbolic Language) hyperpartisan content creates vivid imagery emphasizing points, or describing something in a charming way, departing from standard definitions to convey deeper meaning. Thus, we have shifted the focus from rhetorical biases to non-literal figures. The chapter is structured as follows, presenting the results for the dataset we have created in 4.4.1 and the experiments on the explainability in 4.4.3.

4.4.1 Dataset

In the Chapter 3.6 we already provided superficial information about the Annotation Process. Here we are going to discuss it in more detail.

Annotation Framework

Our annotation framework addresses three core constructs: hyperpartisanship at the paragraph level and LT, which has been applied at the span level. For hyperpartisanship, we adopted the definition from Maggini et al. [2025e], characterizing HP news as extremely one-sided content marked by strong bias. We implemented this as a binary paragraph-level classification to track the distribution of HP content throughout paragraphs (see Figure 4.12). LT are grounded in pragmatics theory, which is the study of language in context [Scott, 2022] and in the hyperpartisan language, they could be used to undermine the Gricean's maxims of quality, relevance, quantity, and manner. 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, Sarcasm $\rightarrow$ Irony/Sarcasm, Metaphors and Similarities $\rightarrow$ Figurative Speech), leading to a total of 6 LT types 3.20.

Source selection

To investigate hyperpartisan language in Italian, we combined data from different sources. We re-annotated the Italian data points from SemEval 2023 [Piskorski et al., 2023a] and De Mattei et al. [2020]'s datasets because they cover different topics, such as the Ukraine war, Italian and European politics (see Table 4.14) from mainstream and "alternative" media sources that

could potentially spread mis/disinformation. Given this scale, we prompted `Llama 3.1 8B` with our annotation guidelines to perform an initial classification along the hyperpartisan dimension. Guided by this filtering, we selected a batch of articles whose prediction score was above 0.7. Then we inspected, read and manually selected 192 articles from Piskorski et al. [2023a] and 45 from De Mattei et al. [2020]. This human-in-the-loop approach ensured coverage of low-frequency classes. Subsequently, to cover different topics like climate change, we selected 119 articles from an independent outlet, "Green Policies" section: nicolaporro.it[3].

Data Preparation

The data from SemEval 2023 required no additional processing, as they were already divided at the paragraph level. In contrast, the dataset from De Mattei et al. [2020] needed to be split into paragraphs. Some of these textual segments consisted only of partial sentences or verbless clauses, undermining natural discourse units and the logical coherence of the discourse. Consequently, we merged consecutive segments containing fewer than 20 tokens. This procedure provided a broader analytical context and produced more coherent paragraphs, since very short segments were unlikely to contain LT. Regarding the data collected from Porro's news outlet, duplicate articles were first removed. The remaining texts were then segmented into paragraphs using the HTML *paragraph* tags.

Annotation Process

Annotations were performed by two native Italian Ph.D. students with expertise in NLP and disinformation analysis. After a training phase involving guideline review and pilot annotations, each paragraph was independently annotated for hyperpartisanship, and LT, using a custom web interface that masked outlet names to prevent source-related bias. Annotations were performed at the paragraph level for the HP label and at the word/span level for LT. Although annotators had access to the full article, the dataset was annotated paragraph by paragraph in consecutive order. This choice aimed to preserve the local narrative flow and discursive coherence within each article. We then conducted a thorough curation phase to resolve disagreements through collaborative review that consisted of analyzing this type of data while discussing the previous annotations. With this method, all the annotations were solved.

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

Training Phase: Annotators studied the guidelines, conducted pilot annotations, and participated in interactive sessions to clarify edge cases and resolve initial disagreements. **Annotation Phase**: Each sentence was independently annotated by both annotators for hyper-partisanship and LT. **Curation Phase**: Disagreements were reviewed and resolved collaboratively to finalize the dataset. All annotators gave informed consent, were fully briefed on the study's objectives, and retained the right to withdraw at any time. Inter-annotator reliability was high, with both Cohen's and Krippendorf'sα of 0.937 for HP detection, and an average Jaccard Similarity of 0.814 for LT, indicating high agreement across all dimensions.

The final dataset's label distribution is shown in Table 3.20.

Statistic	Train	Test
Articles in COVID	253	68
Articles in EU_institutions	801	214
Articles in climate change	1254	318
Articles in immigration	503	123
Articles in war.UR	1077	250
Avg words per paragraph	47.54	48.0
Avg words per article	520.62	154.65
Total words	184819	46705

Table 4.14: IHPP topic-wise statistics for train and test splits.

IAA Error Analysis

Analysis of the Jaccard's similarity for LT reveals a clear hierarchy of annotation reliability. Neologism exhibits near-perfect agreement with a score of 0.949, annotators successfully detected newly coined words belonging to an in-group dictionary. Similarly, Epithet (0.88), Irony/Sarcasm (0.847), and Figurative Language (0.801) show high agreement on more abstract aspects of pragmatics. For example, they adopt metaphors, similarities, adjectives used critically to express someone's quality and the paradoxical alteration of reality via Irony and Sarcasm. Only Loaded and Hyperbolic Language scored a consistent agreement of 0.709 and 0.701 due to their similar confounding nature.

Data statistics and analysis

Italian **H**yper**P**artisan and **P**ragmatics in News is, to our knowledge, the largest dataset on Italian HP news detection focused on the following main topics: Ukraine War (16.7%), climate

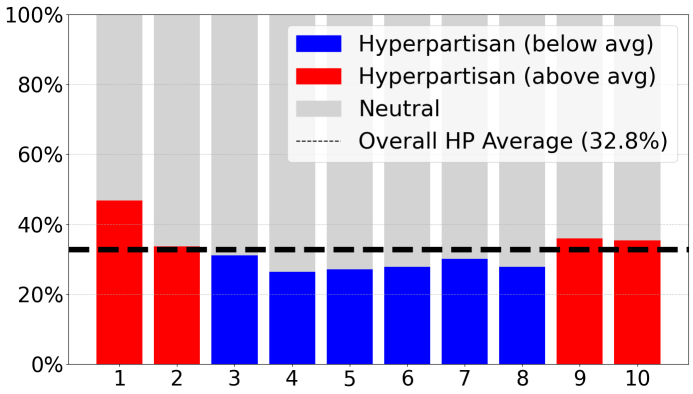


Figure 4.12: Distribution of HP paragraphs across article deciles.

change (13.6%), Immigration (12.9%) and European Institution (11.4%) and is annotated considering pragmatic layers. The remaining topics are subtopics of the corresponding macro-families. 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 them. We applied a stratified split to maintain consistent proportions of HP and neutral paragraphs across training and test sets. The dataset contains paragraphs with the following labels: hyperpartisanship and LT. This enables multifaceted analysis of how partisan language operates in polarizing topics discourse across mainstream and independent media. Table 3.20 presents our taxonomy of LT together with their distributions in the data set. 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 strongly biased articles must include neutral and functional paragraphs to ensure clarity and coherence in communication, rather than consisting solely of hyperpartisan content. By looking at Figure 4.12, HP paragraphs are concentrated in the first decile at 46.8 %. In the second decile, they occur about one-third of the time—a proportion that is also observed at the end of the articles. In all other positions, which correspond to the main body, the presence consistently remains below the average of HP paragraphs (32.8%) in the total dataset. The analysis of LT reveals a dual strategy in HP content. All LT categories peak sharply in the articles' first decile, which often corresponds to titles and leads. This initial surge directly connects our findings to the use of clickbait and persuasive language to influence reader engagement and shape initial narratives [Chakraborty et al., 2016], a tactic common in sensationalist

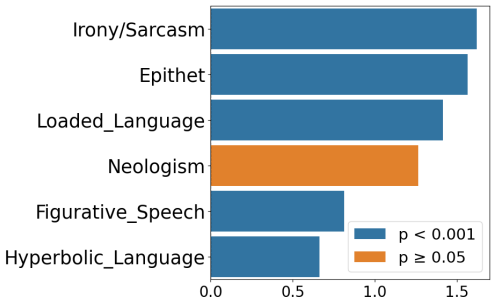


Figure 4.13: We measured the effect sizes using Log Ratio (LR), that is calculated as the logarithm base 2 of the ratio of the frequencies between the two groups. LR = 0: equal frequency in both groups; LR>0: higher frequency in the hyperpartisan group, and LR¡0: higher frequency in the neutral group.

media [Munger, 2020, Blom and Hansen, 2015]. However, while the overall frequency of HP paragraphs decreases after the opening, the core persuasive function is sustained internally. Specifically, Hyperbolic Language, Irony/Sarcasm, and Figurative Speech maintain a stable frequency across the remainder of the article. This sustained presence suggests that these traits are crucial not just for attracting the reader but for continuously reinforcing bias and pushing particular agendas within the main body of the text. Quantitatively, all identified LT categories demonstrate their value in distinguishing HP content, as evidenced by their positive Log Ratio (LR) (Fig. 4.13). This positive LR indicates that every LT is more frequent in the HP group compared to the neutral group. We confirmed the statistical significance of this distinction using a chi-squared test; however, only Neologism shows a non-significant p-value, likely due to its comparatively low frequency within the corpus.

4.4.2 Fine-Tuning and Ablation Study

Fine-Tuning: For FT, we selected the optimal hyperparameters for each model using Optuna[4] running the models for 5 runs each and 2 epochs to avoid overfitting, given the limited size of the dataset. The FT models achieved F1 scores ranging from .673 (multi-uncased) to .729 (sent-ita-xxl). In this setting, the models pretrained on NLI datasets and with a larger vocabulary outperformed those pre-trained on full-text corpora (bert-base-ita and its xxl variant). This difference stems from their objectives: sentence-based models like sentence-

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

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 4.15: Performance comparison of Italian language models on HP detection task. Best scores within the FT configuration are underlined, while overall best scores (all from FT+Embeddings) are bolded.

BERT optimize for sentence-level semantics, whereas traditional masked language models focus on word-level predictions. In particular, sent-ita-xxl achieved the highest F1 (.867), precision (.866) and accuracy (.867). These results suggest sentence-BERT's architecture is particularly suited for document classification tasks like HP detection. The XXL variants, with larger vocabularies and more parameters, offer better precision-recall trade-offs. For both model families, upgrading to XXL versions increased precision at a slight recall cost, ultimately improving F1 scores. Thus, their predictions are more confident, but miss more true positives. 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.

Error Analysis: Ita-xxl-uncased predicted on the test set 220 true positives and 605 true negatives, one less than sent-ita-xxl, which, in contrast, predicted 591 true negatives. The multilingual-uncased labeled 126 false-positive cases. Across all the models, the paragraphs containing Loaded Language were the most frequently classified as false negatives and false positives, but they were also more frequent than the others. 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 F1 performance 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

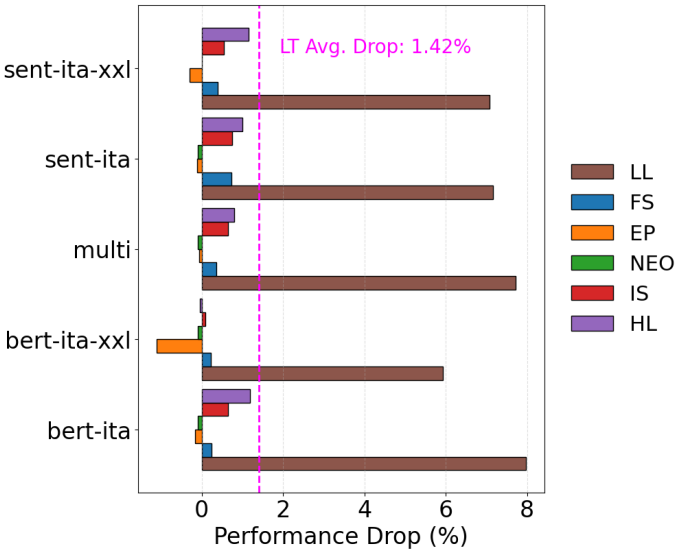


Figure 4.14: LT average drop by model in ablation.

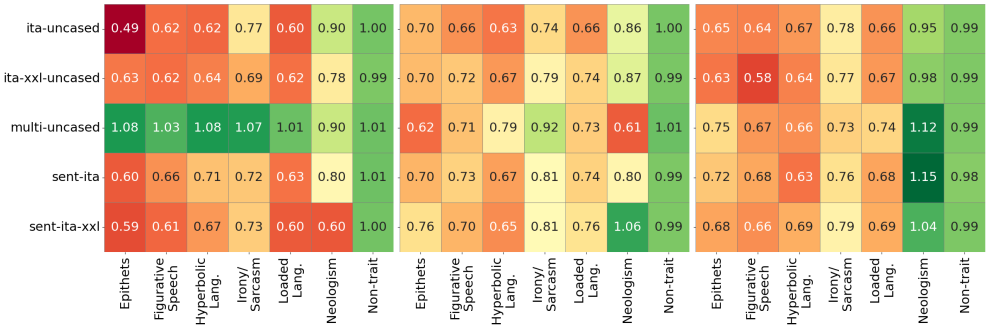


Figure 4.15: Attribution Ratios: Models × LT × Variants

embeddings. **Error Analysis**: This setting confirms that the trend for the misclassification is the same as before, with a declared hierarchy. There is the exception of NEO, which now appears only in paragraphs annotated as false positives.

Ablation Study: Ablating LT categories revealed different impacts (Figure 4.14) and an average F1 drop 1.42%. Removing LL caused the largest drop (7.17%), while FS, IS,

and HL had minimal negative effects (0.39%, 0.54%, 0.82%). Paragraphs containing LL are the most frequent amongst the ones containing LT and account for 94% of all errors, confirming the difficulty of the task and the positive contribution of embedding this specific LT. Instead, removing EP and NEO slightly improved F1 by 0.35% and 0.08%. Despite their negative ablation impact, they do not introduce confusion since their error rates are low. This behavior suggests they may suffer from redundancy with co-occurring LT like LL, and they are insufficient as standalone predictive features in current model architectures. To investigate this redundancy, we analyzed co-occurrence patterns between LT categories (Table 4.16). The high conditional probabilities reveal that most LT tokens co-occur with LL: 82.5% of EP instances and 76.5% of NEO instances appear alongside LL. However, their Pearson correlations with LL differ substantially (0.41 for EP vs. 0.11 for NEO). This discrepancy reflects their different distribution patterns: EP appears frequently throughout HP paragraphs (18.8% coverage) with moderate correlation to LL, while NEO is rare (1.8% coverage) but highly clustered in LL-heavy contexts when it does appear. The combination of (1) proportional error rates, (2) high P(LL—LT), and (3) negative ablation impact suggests EP and NEO function as LL proxies rather than independent discriminative features. When LL is present (in their typical context), these features provide redundant information; when ablated, models rely more heavily on LL's robust signal, improving performance. Only in the 23% of cases where EP/NEO appear without LL do they provide a weak independent signal—insufficient to justify their inclusion given the overfitting risk on rare patterns. The ablation impact was not clearly tied to vocabulary size or language, as IS affected all models negatively, but bert-ita-xxl reduced its impact compared to bert-ita, while sentence models showed the opposite trend. Learning polarized vocabulary (e.g., NEO expressing slurs, violent expressions) via FT+EMB aided HP detection, and its ablation increased false positives for this LT, underscoring LTs' role as hyperpartisanship proxies.

LT	Occur.	Errors	Err%	FP	FP%	FN	FN%
LL	225	165	73.3	101	44.9	129	57.3
FS	195	113	58.0	88	45.1	58	29.7
HL	180	89	49.4	66	36.7	58	32.2
IS	165	72	43.6	52	31.5	34	20.6
EP	165	72	43.6	70	42.4	17	10.3
NEO	70	19	27.1	19	27.1	0	0.0
No LT	270	40	14.8	0	0.0	40	14.8

Table 4.16: LT FT+EMB error statistics. Coverage: % of corpus containing LT. Error Share: % of total errors involving the specific LT. Error Rate: % of LT-containing paragraphs misclassified.

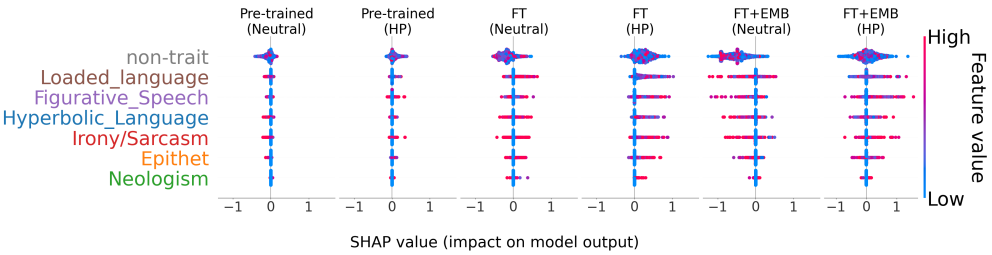


Figure 4.16: Comparison of SHAP values across learning strategies ordered by magnitude. The plot is split by predicted label. Samples predicted as HP SHAP contributions toward label HP, viceversa for the labelled ones as Neutral.

4.4.3 Integrated Gradient and SHAP Analysis

Integrated Gradient Figure 4.15 shows IG attribution ratios for LTs, with non-trait tokens averaging 1 as a baseline (per Formula 3.3). Comparing pre-trained, FT, and FT+EMB models, LT embeddings generally reduced attribution importance, with FT outperforming in 17 of 35 model-variant-LT combinations. A consistent hierarchy of attribution was observed across the LT categories: FS led in FT except for multi-uncased, followed by IS, LL, and EP, indicating model-agnostic prioritization of certain LT. Pre-trained multi-uncased highlighted HL (ratio 1.08), but ignored NEO, which rose to 1.12 in FT+EMB, possibly due to Italian vocabulary and task optimization. NEO gains were: FT vs. pre-trained +5.1%, FT+EMB vs. pre-trained +31.2%, FT+EMB vs. FT +24.9%. Except for multi-uncased, LTs were underscored, but NEO exceeded non-trait averages in FT+EMB (e.g., 1.04 XXL, 1.15 non-XXL, 0.98 ita-xxl-uncased), suggesting a hyperpartisan cue. However, ablation data tempers this: NEO removal improved F1 by 0.08%, EP by 0.35%, versus LL's 7.17% drop, hinting at NEO's over-optimization on rare instances (96/4681 paragraphs) and LL co-occurrence (correlation 0.11 vs. 0.41 for EP-LL). This consistently higher attribution for NEO relative to non-trait tokens suggests its potential role as a hyperpartisan cue. While IG scores highlight NEO as a primary contributor, the ablation study provided further context. To resolve the IG-ablation discrepancy and disentangle the possible effects of co-occurring LT, we conducted conditional attribution analysis comparing NEO and EP against LL, the most important. We defined the boosting factor as:

$$\frac{\text{median attribution with LL} - \text{median attribution isolated}}{|\text{median attribution isolated}|} \tag{4.1}$$

Conditional IG scores in (computed across isolated, without LL, and with LL conditions) showed NEO's context dependency boosting factors ranging from +81.4% to +100% in ita-xxl-uncased and sent-ita-xxl, and -41.2% to -100% in others, yet δ F1 $\leq$ 0.08% (p > 0.05) across models. Models with larger vocabularies (ita-xxl, sent-ita-xxl) may have dedicated embeddings for NEO terms, amplifying their attribution when LL co-occurs. In contrast, smaller vocabulary models may merge NEO into subword tokens, reducing their distinct signal. However, the consistent near-zero F1 impact (δF1 $\leq$ 0.08%, p>0.05) across all architectures confirms that this variability reflects architectural artifacts rather than genuine predictive value. Sent-ita-uncased's NEO hit 1.0 attribution when isolated, but δF1. EP showed significant δF1 (+1.1%, p < 0.05) only in ita-xxl, likely due to its larger vocabulary disentangling LL (P(LL—EP) = 0.825, P(LL—NEO) = 0.765). Other models' EP boosting (67.7%–97.1%) lacked impact, indicating LL dependency. Given the high attribution on NEO, the model currently relies on it when making predictions, but its negative ablation effect forces the model to use more reliable features. The model might have overweighted rare, co-occurring features during training, similar to shortcut learning phenomena [Geirhos et al., 2020]. Those results provide cues about the limited discriminative power of the NEO and EP. IG reveals what models do learn and not what they should learn. Thus, high attribution scores indicate model reliance, but not feature utility. Lastly, two attribution mechanisms emerged: independent (high isolated IG, minimal LL boost, significant δ F1, e.g., EP in ita-xxl) and correlation-driven (low δ F1 < 0.01, p > 0.05, LL-sensitive, e.g., NEO, most EP cases), reflecting redundancy with LL.

SHAP analysis Figure 4.16 represents SHAP values toward HP predictions across different models and learning strategies. Each dot represents a paragraph, with red indicating frequent trait presence and blue indicating infrequent occurrence. The pre-trained model shows SHAP values concentrated near zero across both neutral and HP predictions, indicating it does not rely on linguistic traits 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 regardless 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. LL, FS, and IS 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, the FT+EMB model exhibits substantially increased SHAP value variability, with LL and FS 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., LL 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 weigh trait importance based on broader context, leading to more nuanced but less linearly separable feature contributions. The non-trait feature exhibits this contextual sophistication, showing minimal spread in pre-trained 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 LTs prove equally discriminative. Indeed, Neologism and Epithet consistently show weaker SHAP magnitudes and tighter clustering across all models. For NEO, this stems from its rarity (96 instances, 2%) combined with high LL co-occurrence (76.5%). For EP, despite moderate frequency (1,242 instances), its high correlation with LL (82.5% co-occurrence) limits independent discriminative power.

4.4.4 Final Remarks

The chore of IHPP was to release a paragraph-level Italian corpus annotated for HP detection and enriched with span-level LT across six categories. Secondly, we incorporated LT into BERT-based classifiers via embedding fusion and conducted an ablation study to quantify each trait's discriminative contribution. Lastly, we apply Integrated Gradients and SHAP to examine whether models rely on LT tokens in ways consistent with their actual predictive value. With these experiments, our contribution pushes a little further the understanding of internal representations of LT with regard to hyperpartisan detection for a low-resource language like Italian. Furthermore, with this work, we also confirmed that hyperpartisan language relies on not only rhetorical biases but also on breaking Gricean maxims and on other figurative devices shaping the semantics of the utterances.

4.5 Par-ITA: Seq2seq vs LLMs on neutralizing capabilities

Par-ITA is the last work we are presenting in this thesis and aims to provide insights on the seq2seq capabilities for neutralizing a hyperpartisan text, as well as to test the support of LLMs as a judge in the subjective annotation task. Additionally, the final scope is creating a synthetic human-curated dataset for the neutralization task on the Italian language.

4.5.1 Data Generation and Pipeline Evaluation

Stage 1: Preliminary LLM Selection Results

Human annotators achieved high agreement on model evaluations (Table 4.17): Krippendorff's α of 0.890 (Syntax), 0.863 (Semantics), and 0.815 (Neutrality). Model performance varied significantly. Llama-3.1-8b-Instruct excelled in Syntax for paraphrasing tasks (2.88 Rephrased, 2.54 Perturbed) but frequently generated English text, yielding poor semantic scores. Mistral-Nemo-Instruct-2407 struggled with neutralization, retaining hyperpartisan cues in 70% of outputs. Qwen-2.5-14B-Instruct demonstrated consistent quality across all tasks (Rewritten: 2.90 Syntax, 2.73 Semantics) with only 13% neutralization failures, achieving the following human alignment (Cohen's κ = 0.801, MCC = 0.806).

	α	κ_w	r	ρ	κ_c	MCC
Syntax	.890	.890	.891	.902	-	-
Semantics	.863	.863	.865	.871	-	-
Neutrality	.815	.815	.815	.812	-	-
Best Model	-	-	-	-	.871	-
Llama Rew.	-	-	-	-	.834	.840
Mistral Rew.	-	-	-	-	.762	.764
Qwen Rew.	-	-	-	-	.801	.806

Table 4.17: Inter-annotator agreement for human experts and model-human alignment on neutrality classification in the preliminary LLM selection phase. α: Krippendorff's α, κ_w: weighted Cohen's κ, r: Pearson correlation, ρ: Spearman correlation, κ_c: Cohen's κ, MCC: Matthews Correlation Coefficient.

GPT-4o-mini as Judge in Data Generation GPT-4o-mini showed moderate correlation with human consensus on quality scores (Figure 4.17), particularly diverging on Rewritten texts where humans assigned higher ratings. Considering the agreement between humans and GPT on the neutralization task (Figure 4.18), it ranged from poor to fair (Cohen's κ: -0.039 to 0.454), with systematic over-classification of hyperpartisanship on sensitive topics

(COVID, immigration, climate change). This aligns with documented ”safety over-refusal” in RLHF-tuned models, which rely on lexical triggers rather than nuanced framing analysis.

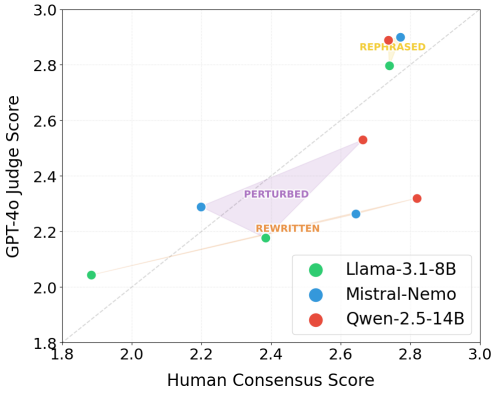


Figure 4.17: Correlation between human consensus and GPT-4o-mini scores (averaged Syntax and Semantics) during data generation.

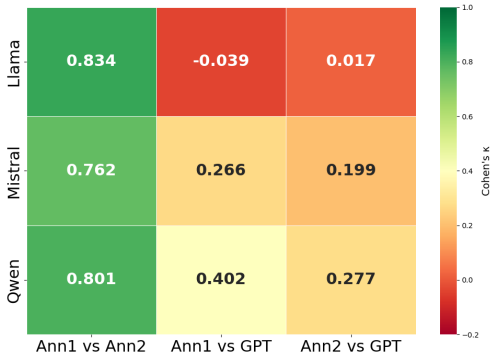


Figure 4.18: Cohen's κ between human experts and GPT-4o-mini on binary label for the Rewritten task across models.

Stage 2: Full Dataset Generation Statistics

For the full dataset generation via Qwen-2.5-14B-Instruct, the model successfully generated hyperpartisan variants in most cases (Table 4.18), but achieved only 44.4% complete success on Rewritten. Consequently, annotators edited 32.8% of Rephrased texts, 41.9% of Perturbed

texts, and 67.8% of Rewritten texts, with Avg. Levenshtein distances of 29%, 25%, and 32% respectively.

Task	Edited	Lev.	N	P	Y
Rephrased	32.8%	29%	12.4%	11.6%	76.0%
Perturbed	41.9%	25%	17.5%	8.1%	74.4%
Rewritten	67.8%	32%	27.9%	27.8%	44.4%

Table 4.18: Generation quality. Lev.: Avg. Levenshtein distance from original text. N: No, P: Partially, Y: Yes.

Human-GPT Agreement on Qwen's Rewritten Task For Rewritten texts, human annotators achieved substantial agreement (Cohen's $\kappa = 0.694$), while GPT-4o-mini agreement remained fair ($\kappa = 0.302$–0.454), confirming its tendency to over-classify contentious topics as hyperpartisan regardless of actual framing bias.

Common Error Patterns *Rephrased*: Difficulty replacing neologisms and managing noisy elements (calls-to-action, malformed punctuation). *Perturbed*: While successfully amplifying bias markers, the model occasionally introduced grammatical errors, nonsense neologisms, or output placeholder text instead of content. *Rewritten*: Most failures involved incomplete neutralization (residual loaded terms, preserved biased framing) or difficulty detecting subtle irony. To preserve semantic meaning, we avoided complete context removal except when sentences contained only insults or slurs, prioritizing reduction of loaded language over topic elimination.

Stage 3: Post-hoc Generation Metrics Validation

Inter-Annotator Agreement on LLMs Generations Post-Hoc Human annotators achieved substantial-to-strong agreement across models (Figure 4.19), with Cohen's κ ranging from 0.459 (gemma-3-27b) to 0.763 (llama-3.2-3b). Higher agreement on models like llama-3.2-3b and qwen-25-72b suggests these models produce more consistent, interpretable outputs. Lower agreement on larger models indicates greater output variability.

Human vs. GPT-4o-mini Rating Alignment GPT-4o-mini consistently assigned higher scores (concentrated in ratings 3-5), while human annotators distributed ratings more evenly, particularly using lower scores (1-2) to penalize subtle neutralization failures, semantic drift, or hallucinations. This pattern confirms GPT-4o-mini's tendency toward leniency and insufficient sensitivity to residual bias or factual inconsistencies that humans readily detect.

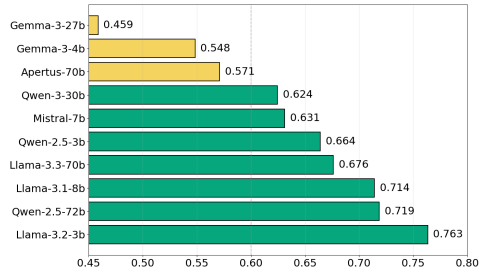


Figure 4.19: Inter-annotator agreement (Cohen's κ) between humans across evaluated LLMs. Higher values indicate more consistent human judgments.

Implications for Automated Evaluation These findings reinforce that while GPT-4o-mini provides a useful signal for high-level quality assessment, it cannot replace human judgment for nuanced neutralization tasks. The divergence between high automated semantic similarity scores and lower human ratings highlights that lexical/embedding similarity does not capture subtle hyperpartisan cues, framing bias, or hallucinated content.

4.5.2 Seq2seq Model Results

We evaluated four seq2seq model families (IT5, mT5, FLAN-T5, BART-IT) across supervised fine-tuning (SFT), classifier-guided fine-tuning (SFT-CLASS), and Direct Preference Optimization (DPO) configurations.

Training Configuration Analysis. **SFT** provides balanced performance with high semantic preservation ($\text{SBERT}_{\text{src-gen}} > 0.83$) but moderate BLEU scores (14–19). **SFT-CLASS** marginally improves neutrality (+2–5%) at a slight cost to fluency (–1–3% BLEU), confirming that classifier-guided decoding offers modest but consistent neutrality gains. **DPO** achieves the highest BLEU scores (+23% average improvement) but with variable neutrality outcomes.

DPO Performance. DPO significantly improves generation quality for seq2seq models. BART-base DPO achieves 29.1 BLEU—the highest across all configurations—while FLAN-T5-large DPO maintains both high quality (24.1 BLEU) and acceptable neutrality (0.73). However, DPO shows higher variance in neutrality outcomes: while some models maintain high neutrality (IT5-large DPO: 0.73, mT5-base DPO: 0.72), others exhibit significant degradation (BART-base DPO: 0.13).

Model Family Comparison. Among model families, FLAN-T5 demonstrates the most consistent performance across all configurations, benefiting from its instruction-tuning pre-

Model	Config	BLEU	ROUGE-L	BERTScore	Neutrality	$\text{SBERT}_{\text{src-gen}}$	$\text{SBERT}_{\text{gen-ref}}$
IT5-base	SFT	11.9	0.32	**0.74**	0.52	0.83	0.69
	SFT-CLASS	12.2	0.32	**0.74**	0.53	0.83	0.69
	DPO	14.0	0.39	0.35	0.57	0.62	0.64
IT5-large	SFT	13.7	0.33	0.35	0.59	0.83	0.70
	SFT-CLASS	13.6	0.34	0.35	0.62	0.84	0.70
	DPO	4.3	0.15	-0.01	0.73	0.55	0.43
mT5-base	SFT	14.3	0.34	0.39	0.72	0.85	0.70
	SFT-CLASS	14.5	0.34	0.39	0.71	0.86	0.70
	DPO	23.0	0.39	0.41	0.72	0.72	0.68
mT5-large	SFT	14.9	0.35	0.41	0.67	0.88	0.72
	SFT-CLASS	15.8	0.36	0.41	0.69	0.88	0.72
	DPO	21.1	0.37	0.36	0.57	0.75	0.69
FLAN-T5-base	SFT	18.4	0.38	0.42	0.83	0.91	0.73
	SFT-CLASS	17.5	0.44	0.26	**0.85**	0.79	0.63
	DPO	22.7	0.38	0.41	0.71	0.68	0.65
FLAN-T5-large	SFT	19.3	0.38	0.43	0.81	0.91	0.73
	SFT-CLASS	18.8	0.38	0.43	0.79	0.90	0.74
	DPO	24.1	0.43	0.47	0.73	0.69	0.73
BART-base	SFT	18.1	0.37	0.73	0.33	0.80	0.62
	SFT-CLASS	16.4	0.36	0.20	0.73	0.81	0.62
	DPO	**29.1**	**0.44**	0.44	0.13	0.75	0.73
BART-large	SFT	18.1	0.37	0.40	0.75	0.91	0.72
	SFT-CLASS	18.0	0.38	0.44	**0.85**	**0.92**	**0.75**
	DPO	24.0	0.39	0.45	0.36	0.62	0.70

Table 4.19: Seq2seq model performance on Par-ITA test set. Best values per column are **bolded**. $\text{SBERT}_{\text{src-gen}}$ measures semantic preservation from source; $\text{SBERT}_{\text{gen-ref}}$ measures alignment with neutral reference. Neutrality is measured by our fine-tuned BERT classifier.

training. BART-IT achieves the highest peak BLEU but with unstable neutrality. mT5 provides a balanced middle ground. IT5 shows the weakest overall performance despite being Italian-native: IT5-base achieves only 11.9–12.2 BLEU with lower neutrality (0.52–0.53). This suggests that domain-specific pretraining does not guarantee task-specific performance, and cross-lingual models (mT5, FLAN-T5) may better generalize to the neutralization task.

Semantic Preservation. The $\text{SBERT}_{\text{src-gen}}$ metric reveals an important trade-off: SFT models preserve source semantics best (0.83–0.91), while DPO models show moderate semantic drift (0.62–0.75). This suggests that preference optimization encourages more extensive rewriting, which increases BLEU but may alter the original meaning beyond neutralization requirements.

4.5.3 LLM Results

We evaluate 11 LLMs across two configurations: zero-shot prompting for large models (27B–72B parameters) and supervised fine-tuning (FT) using LoRA for smaller models (3B–8B). Table 4.20 summarizes performance across both automatic metrics and human evaluation.

Model	Config	BLEU	ROUGE-L	BERTScore	Neutr.	$\text{SBERT}_{\text{src}}$	Human	Rank
Large Models (0-shot prompting)								
Gemma3-27B	0-shot	8.1	0.25	0.34	**1.00**	0.62	**3.96±1.56**	**1**
Qwen3-30B	0-shot	9.4	0.26	0.33	0.98	0.63	3.42±1.69	2
Qwen2.5-72B	0-shot	**12.8**	**0.31**	**0.34**	0.94	**0.68**	3.10±1.61	6
Llama3.3-70B	0-shot	6.8	0.18	0.23	0.98	0.60	2.52±1.73	9
Apertus-70B	0-shot	1.5	0.06	0.01	**0.99**	0.51	1.64±1.21	10
Small Models (LoRA Fine-tuned)								
Mistral-7B	FT	16.3	0.37	0.48	0.91	0.76	3.24±1.59	3
Llama3.1-8B	FT	**16.6**	**0.38**	**0.48**	0.92	**0.76**	3.20±1.54	4
Llama3.2-3B	FT	13.3	0.33	0.42	0.91	0.72	3.10±1.56	5
Gemma3-4B	FT	11.0	0.26	−0.16	0.89	0.56	3.00±1.60	7
Qwen2.5-3B	FT	10.8	0.29	0.39	0.92	0.70	2.84±1.43	8
Phi3.5-Mini	FT	0.3	0.03	−0.33	0.03	0.25	1.00±0.00	11

Table 4.20: LLM performance on Par-ITA test set. **Human** column shows mean±std ratings (1–5 scale, n=50). **Rank** is determined by human evaluation.

Zero-shot Performance LLMs demonstrate zero-shot neutralization capabilities, achieving near-perfect neutrality rates (0.94–1.00) with simple prompting. Qwen2.5-72B leads in automatic metrics with 12.8 BLEU. However, human evaluation reveals a different picture: Gemma3-27B ranks highest (3.96/5) despite lower BLEU, suggesting qualitative factors beyond surface-level similarity matter for perceived neutralization quality.

Fine-tuning Fine-tuning smaller models substantially improves alignment with reference neutralizations. Llama3.1-8B FT achieves the highest BLEU (16.6) and maintains excellent neutrality (0.92). Mistral-7B FT performs comparably and achieves the best human rating among FT models (3.24/5).

Notable Failure Case Phi3.5-Mini FT represents a complete failure case, producing incoherent outputs with near-zero BLEU (0.3). All 50 human-annotated samples received the lowest possible rating (1/5).

4.5.4 Human Evaluation and Metric Alignment

To validate automatic metrics and gain deeper insight into model behavior, we conducted human evaluation on a stratified sample of 50 examples across all 11 LLM configurations.

Metric	**Pearson** r	p	**Spearman** ρ	p
$\text{SBERT}_{\text{gen-ref}}$	+0.84	.001	+0.61	.048
ROUGE-L	+0.82	.002	+0.53	.091
$\text{SBERT}_{\text{src-gen}}$	+0.79	.004	+0.59	.055
BLEU	+0.75	.008	+0.54	.085
BERTScore	+0.75	.009	+0.57	.065
Neutrality	+0.70	.016	+0.35	.299

Table 4.21: Correlation between mean human ratings and automatic metrics across LLM configurations.

Table 4.21 shows that semantic similarity best predicts human judgment: $\text{SBERT}_{\text{gen-ref}}$ achieves the strongest correlation ($r = 0.84$). Overlap metrics like ROUGE-L ($r = 0.82$) and BLEU ($r = 0.75$) show strong linear correlation with human ratings. Conversely, the neutrality rate is weakly aligned. Often, zero-shot models produce over-neutralized or generic phrasing that the classifier assesses as neutral, but humans penalize.

Implications for Metric Selection

Our analysis suggests the following hierarchy for metric reliability in predicting human preferences: **Best predictors**: $\text{SBERT}_{\text{gen-ref}}$, ROUGE-L (strong linear and rank correlation); **Good predictors**: BLEU, BERTScore, $\text{SBERT}_{\text{src-gen}}$; **Weak predictor**: Neutrality rate alone. For practical evaluation, we recommend using **BLEU + Neutrality jointly**, as high performance on both metrics is required for human acceptance.

4.5.5 Final Remarks

Par-ITA contributes to the neutralization task in Italian by creating a human-curated synthetic dataset and benchmarking both seq2seq and LLM approaches for text neutralization. Our findings reveal critical insights into the role of training strategies and model scale in neutralization quality.

First, our three-stage pipeline demonstrates that LLM-assisted data generation combined with human curation is feasible for creating high-quality neutralization datasets. Qwen-2.5-

14B-Instruct emerged as the most reliable generator with strong inter-annotator agreement (Cohen's κ = 0.801 on neutrality), though even the best model required substantial human correction (67.8% on Rewritten texts). This underscores that while LLMs provide valuable scaffolding, human judgment remains indispensable for nuanced neutralization tasks where subtle framing and semantic preservation matter.

Second, our seq2seq benchmarking reveals that training configuration significantly impacts the neutralization-quality trade-off. SFT provides balanced semantic preservation ($\text{SBERT}_{\text{src-gen}} > 0.83$) but moderate BLEU scores. DPO substantially boosts generation quality (BART-base achieves 29.1 BLEU), yet it introduces high variance in neutrality outcomes, suggesting that preference optimization can degrade task-specific constraints without explicit neutrality-aware reward design. Among families, FLAN-T5 demonstrates superior robustness across configurations, benefiting from instruction-tuning pretraining, while domain-specific models like IT5 underperform, challenging assumptions about language-specific pretraining advantages.

Third, our extensive LLM evaluation (11 models, both zero-shot and fine-tuned) shows that larger zero-shot models achieve near-perfect neutrality (0.94–1.00) but produce generic outputs ranked lower by humans. Conversely, fine-tuned smaller models (Llama3.1-8B: BLEU 16.6, ranked 4th) generate more natural neutralizations with stronger human preference. This validates that fine-tuning on task-specific data outweighs raw model scale for neutralization quality.

Finally, our human evaluation and metric correlation analysis reveals that SBERT semantic similarity best predicts human judgments ($r = 0.84$), while neutrality classifiers alone are insufficient ($r = 0.70$). This highlights a critical limitation of automated evaluation: lexical/embedding similarity metrics cannot capture subtle hyperpartisan cues, semantic drift, or hallucinations that humans readily detect. Our recommended dual-metric approach (BLEU + Neutrality jointly) provides practical guidance for practitioners balancing generation quality and task fidelity.

These findings position Par-ITA as a foundational resource for Italian text neutralization research, advancing our understanding of how to bridge the gap between LLM-aided generation, human curation, and rigorous evaluation for bias mitigation in low-resource languages.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Detecting hyperpartisan news is challenging because the phenomenon changes over time (diachronic linguistic variation) and varies across countries, where different media outlets reflect distinct linguistic styles.

This thesis proposes several methodologies that address the full news production and dissemination process. It analyzes country-specific linguistic factors in Europe using a representative sample of native-speaking annotators, with attention to gender balance. In addition, it develops and evaluates specialized resources for both monolingual and multilingual approaches.

The work also explores practical methods for both detecting and neutralizing biased content. All proposed tools and resources have been rigorously tested and successfully validated.

5.1 Contributions

This work involves the development, deployment, and testing of models, as well as the creation of new tools and resources. The study focuses on hyperpartisan news in Europe, specifically in Italian, Spanish, and Portuguese contexts, as these countries are considered at higher risk of political polarization.

Given this geographical scope, we conducted several experiments to advance empirical understanding of how machine learning models perform in low- and mid-resource language settings, particularly in tasks related to harmful political content.

The main contributions of this work include not only the design and implementation of methods for automatically detecting hyperpartisan news, but also the development of new

linguistic resources. These resources are made available to the research community to support and encourage further studies in this domain.

1. **Literature Review**: The main contribution of this work is systematizing the literature about hyperpartisan news detection. The process gathers the methods, architectures and resources scientists have been using, highlighting the preference for Encoder-only models due to the time range of our study and the limited attempts in adopting LLMs as the main tool. Firstly, after comparing the different definitions of hyperpartisan news and their implications in terms of feature engineering, we reframed it properly, embodying both linguistic, social and political aspects. While doing this, we clarify the scope and the domain of this phenomenon. Then, for the first time, the literature on this topic has been scientifically systematized following the PRISMA methodology. This allowed us not only to collect the documents across different online databases but also to discuss the wide range of approaches researchers crafted to tackle the detection of hyperpartisan news. Of course, the results reported in the literature have been collected in proper tables, analyzed and critically discussed along with the embedding approaches and the feature-engineering. Additionally, the datasets were tracked, gathered, and inspected, so that this investigation offers insights about their availability (e.g. open vs restricted access), their shape, the languages taken into account and the rate of reproducibility of these works. Lastly, the literature review served to delineate prevailing research gaps and challenges in hyperpartisan news detection.

2. **Benchmarking** Our focus is on testing different architectures, training and learning methods on hyperpartisan news detection and similar tasks spanning from harmful tweets to fake news detection. Thus, we conducted several evaluations in both mono- and multi-lingual settings, covering European and Semitic languages while testing both a wide range of models and LLM families. Particularly, these experiments covered the most popular ICL techniques: 0-shot, Few-shot, Chain-of-Thought, and eliciting rule-based methods like codebooks, or employing them as human-simulating tools via persona-prompting. Considering the Few-shot tests, the experiments involved the usage of random data points and their selection through accurate algorithms like DPP, ensuring that the most representative points across the datasets could be used as guiding examples for the models. Besides, across the evaluations, we also fine-tuned the LLMs for the detection and neutralization tasks employing Direct Preference Optimization, and in

other cases, distillation techniques. In comparison, we assessed the capabilities of transformers, both encoder-only and seq2seq, customizing their architectures (e.g. gate fusion mechanism), ablating the features used at the classification time, and through explainability techniques (e.g. SHAPS, IG), we were able to assess how they leverage the linguistic knowledge we annotated. As a result, the work provides a rigorous comparison of adaptation paradigms, evaluation and learning strategies, optimization of example selection and a framework for interpretable classification, covering most of the recently developed methodologies.

3. **Resource Creation** Given the scarcity of the resources available with regard to European languages covering divisive and urgent topics that are shaping our democracies, we concentrated our effort on crafting several resources for Italian, Spanish and Portuguese. In total, 4 datasets have been released, of which 3 are for Italian, and 1 is multilingual (Italian, Spanish, and Portuguese). These resources covered the textual detection of hyperpartisan news, but we always modeled the construct according to the definition provided in our systematic review, trying to incorporate several dimensions. For instance, *A sentence level...* and the *PartisanLens* model HP according to rhetorical biases mutuated from Piskorski et al. [2023c]; *IHPP* inspects how HP works at the pragmatic and semantic level thanks to the introduction of linguistic categories like Metaphors, Neologisms, etc.; lastly, **Par-ITA** is the first parallel corpus for neutralizing Italian hyperpartisan paragraphs. Under these circumstances, the creation of the dataset involved data augmentation. To this end, we leveraged LLMs to augment the pre-existing data supervised by a human-in-the-loop approach. During the validation phases and the annotation process, GPT-4o has been used as a judge, so that we were able to evaluate its sensitivity to biased outputs with regard to the human benchmark. Actually, in terms of task, we provided meaningful assets to the NLP community, ensuring high-quality datasets with high IAA covering both detection and rephrasing.

4. **Neutralization** The thesis contributes to neutralization as a complementary task to detection, moving from binary classification to constructive rewriting of hyperpartisan language. Through **Par-ITA**, we released the first Italian parallel resource for paragraph-level hyperpartisan neutralization, with source-target pairs curated through a three-stage pipeline combining LLM generation and human validation. The experimental campaign showed that the neutralization problem cannot be reduced to lexical substitution: high automatic neutrality scores are not sufficient when semantic faithfulness and discourse

quality degrade. In particular, seq2seq experiments highlighted the trade-off between content preservation and style correction, while LLM experiments showed that smaller models, when properly fine-tuned, can outperform larger zero-shot models in human preference. This result is relevant for practical deployment because it indicates a feasible path for low-resource languages: targeted data curation plus supervised adaptation can provide robust neutralization systems without relying exclusively on frontier-scale models.

5. **Explainability** A further contribution concerns interpretability and model diagnostics. In *IHPP*, explainability techniques were not treated as post-hoc visualization only, but as a methodological component to validate whether models rely on linguistically meaningful evidence. By combining feature ablation, embedding fusion, Integrated Gradients, and SHAP analyses, we quantified the contribution of pragmatic traits and rhetorical devices to hyperpartisan prediction. The analysis showed that linguistic traits are not uniformly informative: some categories consistently improve discrimination, while others introduce noise depending on topic and context. This evidence supports a principled use of linguistic knowledge in model design and demonstrates that explainability can guide architecture choices, annotation refinement, and error analysis in a cycle of continuous dataset-model co-improvement.

5.2 Conclusion

This thesis investigated hyperpartisan news as a multilingual, multi-level, and evolving phenomenon, where political framing, rhetorical style, and pragmatic devices interact differently across languages and media ecosystems. Starting from the systematic review, we established that the field has grown rapidly but remains methodologically fragmented, with strong dependence on English-centric resources and frequent simplifications of the hyperpartisan construct. This motivated our empirical agenda: benchmark existing architectures and adaptation paradigms, then build new resources tailored to underrepresented European languages. Across the experimental studies, a coherent pattern emerged. Fine-tuning remains the most reliable adaptation strategy for complex political classification tasks, while in-context learning is useful but generally less stable in low-/mid-resource settings. At the same time, no single architecture is universally superior: encoder-based systems retain competitive advantages in several detection tasks, whereas decoder-based and instruction-tuned models offer flexibility

for generation and reformulation. The evidence from *PartisanLens*, climate-focused sentence-level data, and *IHPP* further showed that hyperpartisanship is better modeled when rhetorical and pragmatic dimensions are explicitly represented rather than collapsed into purely topical signals. Another central outcome is that resource creation is not a side product but a scientific contribution in itself. The datasets introduced in this work provide new benchmarks, richer annotation layers, and stronger grounding for reproducible experimentation in Italian, Spanish, and Portuguese. They also expose language-specific and culturally situated patterns that are often hidden in English-dominant evaluations. In this sense, the thesis contributes both technically and epistemologically: it broadens what is measurable and, therefore, what can be modeled. Finally, by integrating neutralization and explainability, the thesis moves beyond detection-only pipelines. Neutralization experiments demonstrate that effective rewriting requires balancing semantic preservation, stylistic correction, and human acceptability; explainability analyses show how linguistic traits can be traced, validated, and operationalized in model development. Together, these dimensions support a more actionable and responsible NLP framework for political-content analysis: not only identifying problematic discourse, but also understanding model decisions and enabling controlled mitigation strategies.

Overall, the work provides a cohesive contribution spanning theory, benchmarking, resource design, interpretability, and generation, with a specific focus on underrepresented European languages. The resulting framework offers practical guidance for future research and deployment in multilingual settings where data scarcity, political context, and linguistic nuance are central constraints.

5.3 Future Work

The results of this thesis open several directions that are both scientifically relevant and practically necessary for robust multilingual hyperpartisan analysis.

First, future work should expand cross-lingual and cross-cultural evaluation beyond the current language set. Although Italian, Spanish, and Portuguese already reveal meaningful divergence in rhetorical and pragmatic patterns, broader coverage is needed to test the portability of our annotation frameworks and adaptation strategies. A natural continuation is to design controlled transfer experiments where models are trained in one linguistic context and evaluated in another, with explicit analysis of which features transfer and which remain language-bound.

Second, the detection pipeline should evolve toward richer prediction targets. Our findings suggest that binary hyperpartisan labels can hide important internal structure, especially when rhetorical devices and pragmatic traits co-occur. Future datasets and models should therefore move toward multi-label and hierarchical formulations, integrating stance, rhetorical devices, and pragmatic markers in a unified framework. This would better reflect the complexity observed in real political discourse and improve model usefulness for downstream analysis.

Third, neutralization requires stronger control mechanisms. The experiments in Par-ITA showed that high automatic scores do not guarantee acceptable rewriting from a human perspective. A promising direction is to develop controllable generation pipelines that explicitly optimize three constraints jointly: neutrality, semantic fidelity, and fluency. This can be pursued through preference-aware objectives, constraint-based decoding, and larger human-in-the-loop evaluation protocols. In parallel, extending neutralization resources to multilingual settings would allow testing whether rewriting strategies generalize across political and linguistic contexts.

Fourth, explainability should be integrated into training and data design, not only into post-hoc evaluation. The evidence from IHPP indicates that interpretability tools can identify both informative and noisy linguistic traits. Future research can operationalize this signal by creating iterative loops where attribution analyses drive guideline refinement, data re-annotation, and architecture updates. This direction would strengthen both model transparency and annotation quality, especially in sensitive domains where accountability is crucial.

Fifth, future work should investigate temporal robustness. Hyperpartisan language changes over time with political events, media trends, and platform-specific styles. Building longitudinal benchmarks and evaluating models under temporal shift would clarify how quickly performance degrades and which adaptation strategies are most reliable. Continual learning and lightweight adaptation methods could then be tested to maintain model validity without complete retraining.

Sixth, in this field, hybrid systems can present advantages over the current architectures. Data-driven and knowledge-driven systems can complement each other to boost their strengths and alleviate weaknesses. A hybrid architecture could address this gap by combining the generalization capacity of neural sequence-to-sequence and decoder-only models with an explicit, knowledge-driven layer that encodes the rhetorical and pragmatic taxonomy developed across *PartisanLens*, the climate-focused Italian dataset, and *IHPP*. Concretely, three components could be integrated: a symbolic detection layer, a knowledge graph for cross-lingual general-

ization, and a constrained generation layer for neutralization. Indeed, rather than relying solely on a classifier's binary or scalar output, an ontology of rhetorical biases and pragmatic markers (loaded language, hyperbole, irony, epithets, neologisms, ad hominem attacks, appeals to fear) could be used to tag spans of text with the specific technique they instantiate. Because the same rhetorical strategies recur across languages even when their lexical realization differs, encoding the ontology as a knowledge graph would allow language-specific surface patterns (e.g., Italian epithets, Spanish hyperbolic constructions) to be linked to language-agnostic functional categories. This addresses one of the central limitations identified throughout this thesis: that fine-tuned models trained on one language's annotated data do not transfer cleanly to another. A knowledge graph layer would not replace the need for multilingual training data, but it would provide a structured prior that constrains and explains model behavior when operating on a new language or domain, functioning as a form of distant supervision or as a feature source that can be relations of other parts of speech, e.g. modifiers, intensifiers, and their typical rhetorical function within the graph.

Finally, from an application perspective, the next step is to move from isolated model benchmarking to integrated decision-support systems. Such systems should combine detection, explanation, and neutralization modules, while preserving human oversight for high-impact decisions. In this setting, the resources and methodologies developed in this thesis can serve as the backbone for practical tools that are linguistically grounded, culturally aware, and methodologically transparent.

Bibliography

Toril Aalberg, Frank Esser, Carsten Reinemann, Jesper Strömbäck, Claes Holger Vreese, et al. Populist political communication in europe. 2017.

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly

capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219.

Nazanin Afsarmanesh, Jussi Karlgren, Peter Sumbler, and Nina Viereckel. Team harry friberg at SemEval-2019 task 4: Identifying hyperpartisan news through editorially defined metatopics. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1004–1006. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2174.

Rodrigo Agerri. Doris martin at SemEval-2019 task 4: Hyperpartisan news detection with generic semi-supervised features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 944–948. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2161.

Ali Aghababaei, Jan Nikadon, Magdalena Formanowicz, Maria Laura Bettinsoli, Carmen Cervone, Caterina Suitner, and Tomaso Erseghe. Application of integrated gradients explainability to sociopsychological semantic markers, 2025. URL https://arxiv.org/abs/2503.04989.

Usman Ahmed, Jerry Chun-Wei Lin, and Gautam Srivastava. Semisupervised federated learning for temporal news hyperpatism detection. *IEEE Transactions on Computational Social Systems*, 10(4):1758–1769, 2023a. ISSN 2329-924X, 2373-7476. doi: https://doi.org/10.1109/TCSS.2023.3247602.

Usman Ahmed, Jerry Chun-Wei Lin, and Gautam Srivastava. Temporal positional lexicon expansion for federated learning based on hyperpatism detection. *Expert Systems*, 40(5): e13183, 2023b. ISSN 1468-0394. doi: https://doi.org/10.1111/exsy.13183.

Dmitrii Aksenov, Peter Bourgonje, Karolina Zaczynska, Malte Ostendorff, Julian Moreno-Schneider, and Georg Rehm. Fine-grained classification of political bias in german news: A data set and initial experiments. In Aida Mostafazadeh Davani, Douwe Kiela, Mathias Lambert, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem, editors, *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 121–131. Association for Computational Linguistics, 2021. doi: https://doi.org/10.18653/v1/2021.woah-1.13.

Amal Alabdulkarim and Tariq Alhindi. Spider-jerusalem at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Seman-*

tic Evaluation, pages 985–989. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2170.

Laura Alonso-Muñoz and Andreu Casero-Ripollés. Populism against europe in social media: The eurosceptic discourse on twitter in spain, italy, france, and united kingdom during the campaign of the 2019 european parliament election. *Frontiers in communication*, 5:54, 2020.

Khudran Alzhrani. Ideology detection of personalized political news coverage: A new dataset. In *Proceedings of the 2020 the 4th International Conference on Compute and Data Analysis*, pages 10–15. ACM, 2020. ISBN 978-1-4503-7644-0. doi: https://doi.org/10.1145/3388142.3388149.

Mussa Aman. Large language model based fake news detection. *Procedia Computer Science*, 231:740–745, 2024.

Evan Amason, Jake Palanker, Mary Clare Shen, and Julie Medero. Harvey mudd college at SemEval-2019 task 4: The d.x. beaumont hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 967–970. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2166.

Bharat Anand, Rafael Di Tella, and Alexander Galetovic. Information or opinion? media bias as product differentiation. *Journal of Economics & Management Strategy*, 16(3):635–682, 2007. doi: https://doi.org/10.1111/j.1530-9134.2007.00153.x.

Talita Anthonio and Lennart Kloppenburg. Team kermit-the-frog at SemEval-2019 task 4: Bias detection through sentiment analysis and simple linguistic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1016–1020. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2177.

Project Apertus, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, et al. Apertus: Democratizing open and compliant llms for global language environments. *arXiv preprint arXiv:2509.14233*, 2025.

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

Dennis Assenmacher, Paloma Piot, Katarina Laken, David Jurgens, and Claudia Wagner. Beyond the explicit: A bilingual dataset for dehumanization detection in social media, 2025. URL https://arxiv.org/abs/2510.18582.

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

Chris Atton. Alternative media. 2001.

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. APPDIA: A discourse-aware transformer-based style transfer model for offensive social media conversations. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.530/.

Dilshod Azizov and Preslav Nakov. Overview of the CLEF-2023 CheckThat! Lab Task 3 on Political Bias of News Articles and News Media. 3497:250–259, 2023.

Dilshod Azizov, Preslav Nakov, and Shangsong Liang. Frank at checkthat! 2023: Detecting the political bias of news articles and news media. In *Conference and Labs of the Evaluation Forum*, 2023.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. Predicting factuality of reporting and bias of news media sources. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/D18-1389.

Ramy Baly, Georgi Karadzhov, Abdelrhman Saleh, James Glass, and Preslav Nakov. Multi-task ordinal regression for jointly predicting the trustworthiness and the leading political ideology of news media. In Jill Burstein, Christy Doran, and Thamar Solorio, editors,

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 2109–2116. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/N19-1216.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. We can detect your bias: Predicting the political ideology of news articles. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991. Association for Computational Linguistics, 2020. doi: https://doi.org/10.18653/v1/2020.emnlp-main.404.

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. *ArXiv*, abs/2403.18932, 2024. URL `https://api.semanticscholar.org/CorpusID:268732713`.

Micaela Bangerter, Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Alberto Volpe, Carmen De Maio, and Claudio Stanzione. 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, 2023a.

Micaela Bangerter, Giuseppe Fenza, Mariacristina Gallo, Vincenzo Loia, Alberto Volpe, Carmen De Maio, and Claudio Stanzione. Unisa at SemEval-2023 task 3: A SHAP-based method for propaganda detection. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 885–891, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.122. URL `https://aclanthology.org/2023.semeval-1.122/`.

Rémi Bardenet and Adrien Hardy. Monte carlo with determinantal point processes, 2019. URL `https://arxiv.org/abs/1605.00361`.

Leon Barkho. 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, 2021.

Matthew Barnidge and Cynthia Peacock. A third wave of selective exposure research? the challenges posed by hyperpartisan news on social media. *Media and Communication*, 7(3): 4–7, 2019. ISSN 2183-2439. doi: https://doi.org/10.17645/mac.v7i3.2257.

Davide Bassi, Søren Fomsgaard, and Martín Pereira-Fariña. Decoding persuasion: a survey on ML and NLP methods for the study of online persuasion. *Frontiers in Communication*, 9, October 2024. ISSN 2297-900X. doi: 10.3389/fcomm.2024.1457433. URL https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2024.1457433/full. Publisher: Frontiers.

Marco Bastos, Shawn Walker, and Michael Simeone. The imped model: Detecting low-quality information in social media. *American Behavioral Scientist*, 65(6):863–883, 2021. doi: 10.1177/0002764221989776. URL https://doi.org/10.1177/0002764221989776.

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

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, editors, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado, May-June 2015. Association for Computational Linguistics. doi: 10.3115/v1/N15-1171.

Tilman Beck, Hendrik Schuff, Anne Lauscher, and Iryna Gurevych. Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2589–2615, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.159. URL https://aclanthology.org/2024.eacl-long.159/.

Márton Bene, Melanie Magin, Jörg Haßler, Uta Russmann, Darren Lilleker, Simon Kruschinski, Daniel Jackson, Vicente Fenoll, Xénia Farkas, Pawe l Baranowski, et al. Populism in context: A cross-country investigation of the facebook usage of populist appeals during

the 2019 european parliament elections. *The International Journal of Press/Politics*, 30(1): 100–121, 2025.

Yochai Benkler, Robert Faris, and Hal Roberts. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press, 2018.

Yves Bestgen. Tintin at SemEval-2019 task 4: Detecting hyperpartisan news article with only simple tokens. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1062–1066. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2186.

Su Lin Blodgett, Solon Barocas, Hal Daumé Iii, and Hanna Wallach. 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, 2020.

Nico Blokker, André Blessing, Erenay Dayanik, Jonas Kuhn, Sebastian Padó, and Gabriella Lapesa. 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, 2023.

Jonas Nygaard Blom and Kenneth Reinecke Hansen. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics*, 76:87–100, 2015. ISSN 0378-2166. doi: https://doi.org/10.1016/j.pragma.2014.11.010. URL `https://www.sciencedirect.com/science/article/pii/S0378216614002410`.

Jay Blumler and Michael Gurevitch. *The crisis of public communication*. Routledge, 2002.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In Octavian Popescu and Carlo Strapparava, editors, *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89. Association for Computational Linguistics, 2017. doi: https://doi.org/10.18653/v1/W17-4215.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: http://dx.doi.org/10.1023/A:1010933404324.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,

Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020. URL https://api.semanticscholar.org/CorpusID:218971783.

Michael Brüggemann, Sven Engesser, Florin Büchel, Edda Humprecht, and Laia Castro. Hallin and mancini revisited: Four empirical types of western media systems. *Journal of communication*, 64(6):1037–1065, 2014.

Axel Bruns, Stephen Harrington, and Edward Hurcombe. Corona? 5g? or both?': the dynamics of covid-19/5g conspiracy theories on facebook. *Media International Australia*, 177(1):12–29, 2020.

Beatriz Buarque. The politics of fear: the shameless normalization of far-right discourses: by ruth wodak, london, sage publications, 2020, 337 pp.,£ 59.52 (paperback), isbn 9781526499219, 2023.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018. doi: https://doi.org/10.48550/arXiv.1803.11175.

Andrew Chadwick. *The hybrid media system: Politics and power*. Oxford University Press, 2017.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. FLUTE: Figurative language understanding through textual explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.481. URL https://aclanthology.org/2022.emnlp-main.481/.

Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 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, 2016. doi: 10.1109/ASONAM.2016.7752207.

Nikhil Chakravartula, Vijayasaradhi Indurthi, and Bakhtiyar Syed. Fermi at SemEval-2019 task 4: The sarah-jane-smith hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 954–956. Association for Computational Linguistics, 2019. URL https://aclanthology.org/S19-2163.

Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *AI Magazine*, 45(3):354–368, 2024.

Celena Chen, Celine Park, Jason Dwyer, and Julie Medero. Harvey mudd college at SemEval-2019 task 4: The carl kolchak hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 957–961. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2164.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. pages 785–794, 08 2016. doi: https://doi.org/10.1145/2939672.2939785.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *J. Mach. Learn. Res.*, 25(1), January 2024. ISSN 1532-4435.

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

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

European Commission. European commission: The strengthened code of practice on disinformation 2022, 2022.

European Commission. Commission opens formal proceedings against facebook and instagram under the digital services act, 2024a.

European Commission. Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act). OJ L2024/1689, 2024b.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Rebekah Cramerus and Tatjana Scheffler. Team kit kittredge at SemEval-2019 task 4: LSTM voting system. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1021–1025. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2178.

André Cruz, Gil Rocha, Rui Sousa-Silva, and Henrique Lopes Cardoso. Team fernando-pessa at SemEval-2019 task 4: Back to basics in hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 999–1003. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2173.

André Ferreira Cruz, Gil Rocha, and Henrique Lopes Cardoso. On document representations for detection of biased news articles. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, pages 892–899. ACM, 2020. ISBN 978-1-4503-6866-7. doi: 10.1145/3341105.3374025.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-Grained Analysis of Propaganda in News Articles, October 2019. URL `http://arxiv.org/abs/1910.02517`. arXiv:1910.02517 [cs].

Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova, editors, *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1377–1414, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.186. URL `https://aclanthology.org/2020.semeval-1.186/`.

Samuel Caetano Da Silva and Ivandre Paraboni. Politically-oriented information inference from text. *JUCS - Journal of Universal Computer Science*, 29(6):569–594, 2023. ISSN 0948-6968, 0948-695X. doi: https://doi.org/10.3897/jucs.96652.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. Text detoxification using large pre-trained neural models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7979–7996, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.629. URL https://aclanthology.org/2021.emnlp-main.629/.

Russell J. Dalton. Modeling ideological polarization in democratic party systems. *Electoral Studies*, 72:102346, 2021. ISSN 02613794. doi: https://doi.org/10.1016/j.electstud.2021.102346.

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

Lorenzo De Mattei, Michele Cafagana, Felice Dell'Orletta, Malvina Nissim, and Albert Gatt. 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), 2020.

Viola De Ruvo, Arianna Muti, Daryna Dementieva, and Debora Nozza. Detoxify-IT: An Italian parallel dataset for text detoxification. In Agostina Calabrese, Christine de Kock, Debora Nozza, Flor Miriam Plaza-del Arco, Zeerak Talat, and Francielle Vargas, editors, *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 267–275, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-105-6. URL https://aclanthology.org/2025.woah-1.24/.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. Multiparadetox: Extending text detoxification with parallel data to new languages. *arXiv preprint arXiv:2404.02037*, 2024a.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, et al. Overview of the multilingual text detoxification task at pan 2024. *Working Notes of CLEF*, 2024b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. doi: https://doi.org/10.48550/arXiv.1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL `https://arxiv.org/abs/1810.04805`.

Liviu P. Dinu, Ioan-Bogdan Iordache, Ana Sabina Uban, and Marcos Zampieri. A computational exploration of pejorative language in social media. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3493–3498, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.296. URL `https://aclanthology.org/2021.findings-emnlp.296/`.

Anna Veronika Dorogush, Andrey Gulin, Gleb Gusev, Nikita Kazeev, Liudmila Ostroumova Prokhorenkova, and Aleksandr Vorobev. Fighting biases with dynamic boosting. *CoRR*, abs/1706.09516, 2017. doi: https://doi.org/10.48550/arXiv.1706.09516.

Mehdi Drissi, Pedro Sandoval Segura, Vivaswat Ojha, and Julie Medero. Harvey mudd college at SemEval-2019 task 4: The clint buchanan hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 962–966. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2165.

Jinhao Duan, Shiqi Wang, James Diffenderfer, Lichao Sun, Tianlong Chen, Bhavya Kailkhura, and Kaidi Xu. ReTA: Recursively thinking ahead to improve the strategic reasoning of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(Volume 1: Long Papers), pages 2232–2246, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.123. URL https://aclanthology.org/2024.naacl-long.123.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, ..., and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Vlad Cristian Dumitru and Traian Rebedea. Fake and hyper-partisan news identification. In Alin Moldoveanu and Alan J. Dix, editors, *16th International Conference on Human-Computer Interaction, RoCHI 2019, Bucharest, Romania, October 17-18, 2019*, pages 60–67. Matrix Rom, 2019.

Ries T. E. Edelman trust barometer 2021, report, 2021.

Aleksandra Edwards and Jose Camacho-Collados. Language models for text classification: Is in-context learning enough? In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.879.

Avia Efrat and Omer Levy. The turking test: Can language models understand instructions? *ArXiv*, abs/2010.11982, 2020. URL https://api.semanticscholar.org/CorpusID:225062157.

Sven Engesser, Nicole Ernst, Frank Esser, and Florin Büchel. Populism and social media: How politicians spread a fragmented ideology. *Information, communication & society*, 20 (8):1109–1126, 2017.

Soufiane Essahli, Oussama Sarsar, Ahmed Bentajer, Anas Motii, and Imane Fouad. Fakezero: Real-time, privacy-preserving misinformation detection for facebook and x, 2025. URL https://arxiv.org/abs/2510.25932.

European Parliament and Council. Proposal for a regulation of the european parliament and of the council on a single market for digital services (digital services act) and amending directive 2000/31/ec, 2020.

Max Falkenberg, Alessandro Galeazzi, Maddalena Torricelli, Niccolò Di Marco, Francesca Larosa, Madalina Sas, Amin Mekacher, Warren Pearce, Fabiana Zollo, Walter Quattrociocchi, et al. Growing polarization around climate change on social media. *Nature Climate Change*, 12(12):1114–1121, 2022. doi: https://doi.org/10.1038/s41558-022-01527-x.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. In plain sight: Media bias through the lens of factual reporting. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6343–6349, Hong Kong, China, nov 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/D19-1664.

Miguel Fernández, Maximiliano Ojeda, Lilly Guevara, Diego Varela, Marcelo Mendoza, and Alberto Barrón-Cedeño. Victor vectors@ dipromats 2024: Propaganda detection with llm paraphrasing and machine translation. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2024) co-located with the conference of the Spanish society for natural language processing (SEPLN 2024), Valladolid, Spain*, page 3756, 2024.

Tine Ustad Figenschou and Karoline Andrea Ihlebæk. Challenging journalistic authority: Media criticism in far-right alternative media. *Journalism studies*, 20(9):1221–1237, 2019.

Richard Fletcher, Alessio Cornia, Lucas Graves, and Rasmus Kleis Nielsen. Measuring the reach of" fake news" and online disinformation in europe. *Australasian Policing*, 10(2): 25–33, 2018.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746, 2025.

Leon Fröhling, Gianluca Demartini, and Dennis Assenmacher. Personas with attitudes: Controlling LLMs for diverse data annotation. In Agostina Calabrese, Christine de Kock, Debora Nozza, Flor Miriam Plaza-del Arco, Zeerak Talat, and Francielle Vargas, editors, *Proceedings of the The 9th Workshop on Online Abuse and Harms (WOAH)*, pages 468–481, Vienna, Austria, August 2025. Association for Computational Linguistics. ISBN 979-8-89176-105-6. URL https://aclanthology.org/2025.woah-1.43/.

Dawid Aristotelis Fusiek, Angeliki Elli Stougiannou, and Theoharris William Efthymiou-Egleton. Digital democracy and disinformation: The european approach to disinformation on social media in the case of 2019 european parliament elections. *Journal of Politics and Ethics in New Technologies and AI*, 1(1):e31215–e31215, 2022.

Michael Färber, Agon Qurdina, and Lule Ahmedi. Team peter brinkmann at SemEval-2019 task 4: Detecting biased news articles using convolutional neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1032–1036. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2180.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. Detecting political bias in news articles using headline attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 77–84. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/W19-4809.

Sonal Garg and Dilip Kumar Sharma. Role of ELMo embedding in detecting fake news on social media. In *2022 11th International Conference on System Modeling & Advancement in Research Trends (SMART)*, pages 57–60. IEEE, 2022. ISBN 978-1-66548-732-0 978-1-66548-734-4. doi: https://doi.org/10.1109/SMART55829.2022.10046789.

Guillaume Gautier, Guillermo Polito, Rémi Bardenet, and Michal Valko. Dppy: Dpp sampling with python. *Journal of Machine Learning Research*, 20(180):1–7, 2019. URL `http://jmlr.org/papers/v20/19-179.html`.

Eleni Gavriil and Georgios Pavlidis. The fog of information: The eu ai act and legal strategies against ai-fuelled disinformation. *Available at SSRN 5635130*, 2025.

Lukas Gebhard and Felix Hamborg. The polusa dataset: 0.9m political news articles balanced by time and outlet popularity. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, pages 467–468, 2020. doi: https://doi.org/10.1145/3383583.3398567. Conference Name: JCDL '20: The ACM/IEEE Joint Conference on Digital Libraries in 2020 ISBN: 9781450375856 Place: Virtual Event China Publisher: ACM.

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

Huang Gerald Ki Wei and Lee Jun Choi. Hyperpartisan news classification with ELMo and bias feature. *Journal of Information Science and Engineering*, 37(5), 2021. doi: https://doi.org/10.6688/JISE.202109_37(5).0013.

Paolo Gerbaudo. *The digital party: Political organisation and online democracy*. Pluto books, 2018.

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

Erving Goffman. *Frame analysis: An essay on the organization of experience.* 1974.

Jennifer Golbeck, Matthew Mauriello, Brooke Auxier, Keval H. Bhanushali, Christopher Bonk, Mohamed Amine Bouzaghrane, Cody Buntain, Riya Chanduka, Paul Cheakalos, Jennine B. Everett, Waleed Falak, Carl Gieringer, Jack Graney, Kelly M. Hoffman, Lindsay Huth, Zhenya Ma, Mayanka Jha, Misbah Khan, Varsha Kori, Elo Lewis, George Mirano, William T. Mohn IV, Sean Mussenden, Tammie M. Nelson, Sean Mcwillie, Akshat Pant, Priya Shetye, Rusha Shrestha, Alexandra Steinheimer, Aditya Subramanian, and Gina Visnansky. Fake news vs satire: A dataset and analysis. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 17–21, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450355636. doi: 10.1145/3201064.3201100. URL https://doi.org/10.1145/3201064.3201100.

Helena Gómez-Adorno, Juan Pablo Posadas-Durán, Gemma Bel Enguix, and Claudia Porto Capetillo. Overview of fakedes at iberlef 2021: Fake news detection in spanish shared task. *Procesamiento del lenguaje natural*, 67:223–231, 2021. doi: https://doi.org/10.26342/2021-67-19.

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

Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure, 2022. URL https://arxiv.org/abs/2203.05794.

Viresh Gupta, Baani Leen Kaur Jolly, Ramneek Kaur, and Tanmoy Chakraborty. Clark kent at SemEval-2019 task 4: Stylometric insights into hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 934–938. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2159.

Michael Gusenbauer and Neal R Haddaway. Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources. *Research Synthesis Methods*, 11:181 – 217, 2020. doi: 10.1002/jrsm.1378.

Fabian Haak and Philipp Schaer. Qbias - a dataset on media bias in search queries and query suggestions. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 239–244, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700897. doi: 10.1145/3578503.3583628. URL https://doi.org/10.1145/3578503.3583628.

Daniel C Hallin and Paolo Mancini. Americanization, globalization and secularization. *Comparing political communication*, pages 25–44, 2004a.

Daniel C Hallin and Paolo Mancini. *Comparing media systems: Three models of media and politics*. Cambridge university press, 2004b.

Daniel C Hallin and Paolo Mancini. *Comparing media systems beyond the Western world*. Cambridge University Press, 2011.

Andrew Halterman and Katherine A. Keith. Codebook llms: Evaluating llms as measurement tools for political science concepts, 2025. URL https://arxiv.org/abs/2407.10747.

Michael Hameleers and Andreas C Goldberg. Europe against the people: does eurosceptic news exposure relate to populist attitudes? evidence from a linkage study across nine european countries. *Political Research Exchange*, 4(1):2126792, 2022.

Michael Hameleers and Desirée Schmuck. It's us against them: A comparative experiment on the effects of populist messages communicated via social media. *Information, Communication & Society*, 20(9):1425–1444, 2017.

Michael Hameleers, Linda Bos, and Claes H De Vreese. "they did it": The effects of emotionalized blame attribution in populist communication. *Communication Research*, 44 (6):870–900, 2017.

Kazuaki Hanawa, Shota Sasaki, Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. The sally smedley hyperpartisan news detector at SemEval-2019 task 4. In *Proceedings of the 13th Interna-*

tional Workshop on Semantic Evaluation, pages 1057–1061. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2185.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. Can GPT-4 identify propaganda? annotation and detection of propaganda spans in news articles. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2724–2744, Torino, Italia, May 2024a. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.244/.

Maram Hasanain, Fatema Ahmad, and Firoj Alam. Large language models for propaganda span annotation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14522–14532, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.850. URL https://aclanthology.org/2024.findings-emnlp.850/.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, 1998. doi: https://doi.org/10.1109/5254.708428.

Annett Heft, Eva Mayerhöffer, Susanne Reinhardt, and Curd Knüpfer. Beyond breitbart: Comparing right-wing digital news infrastructures in six western democracies. *Policy & internet*, 12(1):20–45, 2020.

Edward S. Hermann and Noam Chomsky. *Manufacturing consent. The political economy of the mass media*. Vintage, London, 1994.

Michael Heseltine and Bernhard Clemm von Hohenberg. Large language models as a substitute for human experts in annotating political text. *Research & Politics*, 11(1): 20531680241236239, 2024. doi: 10.1177/20531680241236239. URL https://doi.org/10.1177/20531680241236239.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9: 1735–80, 12 1997. doi: https://doi.org/10.1162/neco.1997.9.8.1735.

Kristoffer Holt. *Right-wing alternative media*. Routledge, 2019.

Kristoffer Holt, Tine Ustad Figenschou, and Lena Frischlich. Key dimensions of alternative news media. *Digital Journalism*, 7(7):860–869, 2019.

Nathan Honeycutt and Lee Jussim. Political bias in the social sciences: A critical, theoretical, and empirical review. *Ideological and Political Bias in Psychology: Nature, Scope, and Solutions*, pages 97–146, 2023. doi: https://doi.org/10.1007/978-3-031-29148-7_5.

Benjamin D. Horne and Sibel Adali. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news, 2017. URL `https://arxiv.org/abs/1703.09398`.

Benjamin D. Horne, William Dron, Sara Khedr, and Sibel Adali. Sampling the news producers: A large news and feature data set for the study of the complex media landscape. *CoRR*, abs/1803.10124, 2018. doi: https://doi.org/10.48550/arXiv.1803.10124.

Tomáš Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. The promises and pitfalls of LLM annotations in dataset labeling: a case study on media bias detection. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1370–1386, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.75. URL `https://aclanthology.org/2025.findings-naacl.75/`.

Md Zobaer Hossain, Md Ashraful Rahman, Md Saiful Islam, and Sudipta Kar. BanFakeNews: A dataset for detecting fake news in Bangla. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2862–2871, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.349/`.

Andrea Hrckova, Robert Moro, Ivan Srba, and Maria Bielikova. Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in central europe. *Online Information Review*, 46(5):954–973, 2021. ISSN 1468-4527. doi: https://doi.org/10.1108/OIR-10-2020-0441. Publisher: Emerald Publishing Limited.

Yibo Hu, Erick Skorupa Parolin, Latifur Khan, Patrick Brandt, Javier Osorio, and Vito D'Orazio. Leveraging codebook knowledge with NLI and ChatGPT for zero-shot political relation classification. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 583–603, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.35. URL https://aclanthology.org/2024.acl-long.35/.

Gerald Ki Wei Huang and Jun Choi Lee. Hyperpartisan news and articles detection using BERT and ELMo. In *2019 International Conference on Computer and Drone Applications (IConDA)*, pages 29–32. IEEE, 2019. ISBN 978-1-72816-592-9 978-1-72816-593-6. doi: https://doi.org/10.1109/IConDA47345.2019.9034917.

Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. Development of a codebook of online anti-vaccination rhetoric to manage covid-19 vaccine misinformation. *International Journal of Environmental Research and Public Health*, 18(14), 2021. ISSN 1660-4601. doi: 10.3390/ijerph18147556. URL https://www.mdpi.com/1660-4601/18/14/7556.

Edda Humprecht, Frank Esser, and Peter Van Aelst. Resilience to online disinformation: A framework for cross-national comparative research. *The international journal of press/politics*, 25(3):493–516, 2020.

Tim Isbister and Fredrik Johansson. Dick-preston and morbo at SemEval-2019 task 4: Transfer learning for hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 939–943. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2160.

So jeong Lim, Adam Jatowt, and Masatoshi Yoshikawa. Understanding characteristics of biased sentences in news articles. In *CIKM Workshops*, 2018.

Ye Jiang, Johann Petrak, Xingyi Song, Kalina Bontcheva, and Diana Maynard. Team bertha von suttner at SemEval-2019 task 4: Hyperpartisan news detection using ELMo sentence representation convolutional network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 840–844. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2146.

Youngjun Joo and Inchon Hwang. Steve martin at SemEval-2019 task 4: Ensemble learning model for detecting hyperpartisan news. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 990–994. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2171.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.82. URL `https://aclanthology.org/2022.emnlp-main.82`.

Lucas Raniére Juvino Santos, Leandro Balby Marinho, Claudio Elizio Calazans Campelo, Filippo Menczer, and Alessandro Flammini. Can large language models effectively mitigate polarization in social media text? In *Proceedings of the 17th ACM Web Science Conference 2025*, Websci '25, page 348–357, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714832. doi: 10.1145/3717867.3717904. URL `https://doi.org/10.1145/3717867.3717904`.

Eleni Kapantai, Androniki Christopoulou, Christos Berberidis, and Vassilios Peristeras. A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, 23(5):1301–1326, 2021. doi: https://doi.org/10.1177/1461444820959296.

Giorgos Katsambekis. Constructing 'the people'of populism: A critique of the ideational approach from a discursive perspective. *Journal of Political Ideologies*, 27(1):53–74, 2022.

Staffs Keele et al. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse, 2007.

Nina Khairova, Bogdan Ivasiuk, Fabrizio Lo Scudo, Carmela Comito, and Andrea Galassi. A first attempt to detect misinformation in Russia-Ukraine war news through text similarity. In Sara Carvalho, Anas Fahad Khan, Ana Ostroški Anić, Blerina Spahiu, Jorge Gracia, John P. McCrae, Dagmar Gromann, Barbara Heinisch, and Ana Salgado, editors, *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 559–564, Vienna, Austria, September 2023. NOVA CLUNL, Portugal. URL https://aclanthology.org/2023.ldk-1.60/.

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

Majid KhosraviNik and Eleonora Esposito. Online hate, digital discourse and critique: Exploring digitally-mediated discursive practices of gender-based hostility. *Lodz papers in pragmatics*, 14(1):45–68, 2018.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, D. Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *International Workshop on Semantic Evaluation*, 2019a. URL https://api.semanticscholar.org/CorpusID:120224153.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news detection. In Jonathan May, Ekaterina Shutova, Aurelie Herbelot, Xiaodan Zhu, Marianna Apidianaki, and Saif M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA, June 2019b. Association for Computational Linguistics. doi: 10.18653/v1/S19-2145. URL https://aclanthology.org/S19-2145/.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. Semeval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 829–839, 2019c.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. SemEval-2019 task 4: Hyperpartisan news

detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839. Association for Computational Linguistics, 2019d. doi: https://doi.org/10.18653/v1/S19-2145.

Jongin Kim, Byeo Rhee Bak, Aditya Agrawal, Jiaxi Wu, Veronika Wirtz, Traci Hong, and Derry Wijaya. COVID-19 vaccine misinformation in middle income countries. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3903–3915, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.237. URL https://aclanthology.org/2023.emnlp-main.237/.

Kang-Min Kim, Mingyu Lee, Hyun-Sik Won, Min-Ji Kim, Yeachan Kim, and SangKeun Lee. Multi-stage prompt tuning for political perspective detection in low-resource settings. *Applied Sciences*, 13(10):6252, 2023b. ISSN 2076-3417. doi: https://doi.org/10.3390/app13106252.

Michelle YoungJin Kim and Kristen Marie Johnson. CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2780–2793. International Committee on Computational Linguistics, 2022.

Jürgen Knauth. Orwellian-times at SemEval-2019 task 4: A stylistic and content-based classifier. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 976–980. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2168.

Yunyong Ko, Seongeun Ryu, Soeun Han, Youngseung Jeon, Jaehoon Kim, Sohyun Park, Kyungsik Han, Hanghang Tong, and Sang-Wook Kim. KHAN: Knowledge-aware hierarchical attention networks for accurate political stance prediction. *Proceedings of the ACM Web Conference 2023*, pages 1572–1583, 2023. doi: https://doi.org/10.1145/3543507.3583300. Conference Name: WWW '23: The ACM Web Conference 2023 ISBN: 9781450394161 Place: Austin TX USA Publisher: ACM.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *ArXiv*, abs/2205.11916, 2022. URL https://api.semanticscholar.org/CorpusID:249017743.

Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–12, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206. doi: 10.1145/3173574.3174012.

Micha l Krzyżanowski. Normalization and the discursive construction of "new" norms and "new" normality: Discourse in the paradoxes of populism and neoliberalism, 2020.

Dilek Küçük and Fazli Can. A tutorial on stance detection. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1626–1628, 2022.

Alex Kulesza. Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286, 2012. ISSN 1935-8245. doi: 10.1561/2200000044. URL http://dx.doi.org/10.1561/2200000044.

Vivek Kulkarni, Junting Ye, Steve Skiena, and William Yang Wang. Multi-view models for political ideology detection of news articles. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3518–3527. Association for Computational Linguistics, 2018. doi: https://doi.org/10.18653/v1/D18-1388.

Murathan Kurfali, Shorouq Zahra, Joakim Nivre, and Gabriele Messori. ClimateEval: A comprehensive benchmark for NLP tasks related to climate change. In Kalyan Dutia, Peter Henderson, Markus Leippold, Christoper Manning, Gaku Morio, Veruska Muccione, Jingwei Ni, Tobias Schimanski, Dominik Stammbach, Alok Singh, Alba (Ruiran) Su, and Saeid A. Vaghefi, editors, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 194–207, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-259-6. doi: 10.18653/v1/2025.climatenlp-1.13. URL https://aclanthology.org/2025.climatenlp-1.13/.

Yanis Labrak, Mickael Rouvier, and Richard Dufour. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani

Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.185.

Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. 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, 2018.

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

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

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159, March 1977. ISSN 0006-341X. doi: 10.2307/2529310. URL http://dx.doi.org/10.2307/2529310.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*, 6(2):443–484, 2023.

Quoc V. Le and Tomás Mikolov. Distributed representations of sentences and documents. *CoRR*, abs/1405.4053, 2014. doi: https://doi.org/10.48550/arXiv.1405.4053.

Jooyoung Lee, Fan Yang, Thanh Tran, Qian Hu, Emre Barut, and Kai-Wei Chang. Can small language models help large language models reason better?: LM-guided chain-of-thought. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2835–2843, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.252.

Nayeon Lee, Zihan Liu, and Pascale Fung. Team yeon-zi at SemEval-2019 task 4: Hyperpartisan news detection by de-noising weakly-labeled data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1052–1056. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2184.

Peter T Leeson and Christopher J Coyne. *Manipulating the media*. SSRN, 2011.

João A. Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. Euvsdisinfo: A dataset for multilingual detection of pro-kremlin disinformation in news articles. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 5380–5384, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704369. doi: 10.1145/3627673.3679167. URL https://doi.org/10.1145/3627673.3679167.

Matthew Levendusky. *How partisan media polarize America*. University of Chicago Press, 2024.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

Chang Li and Dan Goldwasser. Encoding social information with graph convolutional networks forPolitical perspective detection in news media. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2594–2604, Florence, Italy, July 2019. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/P19-1247.

Chang Li and Dan Goldwasser. Using social and linguistic information to adapt pretrained representations for political perspective identification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4569–4579. Association for Computational Linguistics, 2021. doi: https://doi.org/10.18653/v1/2021.findings-acl.401.

Sora Lim, Adam Jatowt, and Masatoshi Yoshikawa. Creating a dataset for fine-grained bias detection in news articles. volume https://db-event.jpn.org/deim2020/post/proceedings/, 2020.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021a.

Ruibo Liu, Lili Wang, Chenyan Jia, and Soroush Vosoughi. Political depolarization of news articles using attribute-aware word embeddings. *CoRR*, abs/2101.01391, 2021b. URL https://arxiv.org/abs/2101.01391.

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

Y Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. doi: https://doi.org/10.48550/arXiv.1907.11692.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374. Association for Computational Linguistics, 2022a. doi: https://doi.org/10.18653/v1/2022.findings-naacl.101.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States, July 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.101. URL https://aclanthology.org/2022.findings-naacl.101.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. ParaDetox: Detoxification

with parallel data. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.469. URL https://aclanthology.org/2022.acl-long.469/.

Sheng Lu, Hendrik Schuff, and Iryna Gurevych. How are prompts different in terms of sensitivity? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.325. URL https://aclanthology.org/2024.naacl-long.325.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. URL https://arxiv.org/abs/1705.07874.

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

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

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723, June 2024b. doi: 10.1162/coli_a_00511. URL https://aclanthology.org/2024.cl-2.6/.

Khudran M. Alzhrani. Political ideology detection of news articles using deep neural networks. *Intelligent Automation & Soft Computing*, 33(1):483–500, 2022. ISSN 1079-8587. doi: https://doi.org/10.32604/iasc.2022.023914.

Michele Maggini and Pablo Gamallo Otero. Leveraging advanced prompting strategies in LLaMA3-8B for enhanced hyperpartisan news detection. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli, editors, *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 531–539, Pisa, Italy, December 2024. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.63/.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In Kalyan Dutia, Peter Henderson, Markus Leippold, Christoper Manning, Gaku Morio, Veruska Muccione, Jingwei Ni, Tobias Schimanski, Dominik Stammbach, Alok Singh, Alba (Ruiran) Su, and Saeid A. Vaghefi, editors, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 168–187, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-259-6. doi: 10.18653/v1/2025.climatenlp-1.11. URL https://aclanthology.org/2025.climatenlp-1.11/.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. Detecting hyperpartisanship and rhetorical bias in climate journalism: A sentence-level Italian dataset. In Kalyan Dutia, Peter Henderson, Markus Leippold, Christoper Manning, Gaku Morio, Veruska Muccione, Jingwei Ni, Tobias Schimanski, Dominik Stammbach, Alok Singh, Alba (Ruiran) Su, and Saeid A. Vaghefi, editors, *Proceedings of the 2nd Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2025)*, pages 168–187, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-259-6. doi: 10.18653/v1/2025.climatenlp-1.11. URL https://aclanthology.org/2025.climatenlp-1.11/.

Michele Joshua Maggini, Davide Bassi, and Pablo Gamallo. 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, 2025c.

Michele Joshua Maggini, Davide Bassi, Paloma Piot, Gaël Dias, and Pablo Gamallo Otero. A systematic review of automated hyperpartisan news detection. *PloS one*, 20(2):e0316989, 2025d.

Michele Joshua Maggini, Davide Bassi, Paloma Piot, Gaël Dias, and Pablo Gamallo Otero. A systematic review of automated hyperpartisan news detection. *PLOS ONE*, 20(2):1–39, 02 2025e. doi: 10.1371/journal.pone.0316989. URL https://doi.org/10.1371/journal.pone.0316989.

Michele Joshua Maggini, Paloma Piot, Anxo Pérez, Erik Bran Marino, Lúa Santamaría Montesinos, Ana Lisboa, Marta Vázquez Abuín, Javier Parapar, and Pablo Gamallo. Partisanlens: A multilingual dataset of hyperpartisan and conspiratorial immigration narratives in european media, 2026. URL https://arxiv.org/abs/2601.03860.

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

Adrien Mazoyer, Jean-François Coeurjolly, and Pierre-Olivier Amblard. Projections of determinantal point processes, 2020. URL https://arxiv.org/abs/1901.02099.

Cory Merow, Matthew Smith, and John Silander. A practical guide to maxent for modeling species' distributions: What it does, and why inputs and settings matter. *Ecography*, 36, 10 2013. doi: https://doi.org/10.1111/j.1600-0587.2013.07872.x.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. Automated stance detection in complex topics and small languages: The challenging case of immigration in polarizing news media. *PLOS ONE*, 19(4):1–16, 04 2024. doi: 10.1371/journal.pone.0302380. URL https://doi.org/10.1371/journal.pone.0302380.

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

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.244. URL https://aclanthology.org/2022.acl-long.244.

Mistral. Mistral nemo: Collaborative innovation with nvidia, 2024.

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

Benjamin Moffitt. *The global rise of populism: Performance, political style, and representation*. Stanford University Press, 2016.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In Steven Bethard, Marine Carpuat, Daniel Cer, David Jurgens, Preslav Nakov, and Torsten Zesch, editors, *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/S16-1003. URL https://aclanthology.org/S16-1003/.

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

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

David Moher, Douglas G. Altman, Alesandro Liberati, and Jennifer Tetzlaff. Prisma statement. *Epidemiology*, 22(1):128, 2011. doi: 10.1097/EDE.0b013e3181fe7825. Cited by: 466; All Open Access, Bronze Open Access.

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

Rafael A. Monteiro, Roney L. S. Santos, Thiago A. S. Pardo, Tiago A. de Almeida, Evandro E. S. Ruiz, and Oto A. Vale. Contributions to the study of fake news in portuguese: New corpus and automatic detection results. In *Computational Processing of the Portuguese Language*, pages 324–334. Springer International Publishing, 2018. ISBN 978-3-319-99722-3.

Jose G. Moreno, Yoann Pitarch, Karen Pinel-Sauvagnat, and Gilles Hubert. Rouletabille at SemEval-2019 task 4: Neural network baseline for identification of hyperpartisan publishers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 981–984. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2169.

Lorenzo Mosca and Mario Quaranta. News diets, social media use and non-institutional participation in three communication ecologies: Comparing germany, italy and the uk. *Information, Communication & Society*, 19(3):325–345, 2016.

Rachel R. Mourão and Craig T. Robertson. Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies*, 20(14):2077–2095, 2019. doi: https://doi.org/10.1080/1461670X.2019.1566871.

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

Sagnik Mukherjee, Muhammad Farid Adilazuarda, Sunayana Sitaram, Kalika Bali, Alham Fikri Aji, and Monojit Choudhury. Cultural conditioning or placebo? on the effectiveness of socio-demographic prompting. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15811–15837, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.884. URL https://aclanthology.org/2024.emnlp-main.884/.

Kevin Munger. All the news that's fit to click: The economics of clickbait media. *Political Communication*, 37(3):376–397, 2020. doi: 10.1080/10584609.2019.1687626. URL https://doi.org/10.1080/10584609.2019.1687626.

Arianna Muti, Federico Ruggeri, Cagri Toraman, Lorenzo Musetti, Samuel Algherini, Silvia Ronchi, Gianmarco Saretto, Caterina Zapparoli, and Alberto Barrón-Cedeño. Pejorativity:

Disambiguating pejorative epithets to improve misogyny detection in italian tweets. *arXiv preprint arXiv:2404.02681*, 2024.

Osman Mutlu, Ozan Arkan Can, and Erenay Dayanik. Team howard beale at SemEval-2019 task 4: Hyperpartisan news detection with BERT. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1007–1011. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2175.

EPRS Naja Bentzen. Information manipulation in the age of generative artificial intelligence, 2025.

Preslav Nakov, Alberto Barrón-Cedeño, Giovanni Da San Martino, Firoj Alam, Julia Maria Struß, Thomas Mandl, Rubén Míguez, Tommaso Caselli, Mucahid Kutlu, Wajdi Zaghouani, Chengkai Li, Shaden Shaar, Gautam Kishore Shahi, Hamdy Mubarak, Alex Nikolov, Nikolay Babulkov, Yavuz Selim Kartal, and Javier Beltrán. The clef-2022 checkthat! lab on fighting the covid-19 infodemic and fake news detection. In Matthias Hagen, Suzan Verberne, Craig Macdonald, Christin Seifert, Krisztian Balog, Kjetil Nørvåg, and Vinay Setty, editors, *Advances in Information Retrieval*, pages 416–428, Cham, 2022. Springer International Publishing. ISBN 978-3-030-99739-7.

Luca Nannini, Eleonora Bonel, Davide Bassi, and Michele Joshua Maggini. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, 5(2): 1241–1269, 2025.

Navakanth Reddy Naredla and Festus Fatai Adedoyin. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1):100064, 2022. ISSN 26670968. doi: https://doi.org/10.1016/j.jjimei.2022.100064.

Nic Newman, Richard Fletcher, Anne Schulz, Simge Andi, Craig T. Robertson, and Rasmus Kleis Nielsen. Reuters Institute Digital News Report 2021, June 2021. URL `https://papers.ssrn.com/abstract=3873260`.

Duc-Vu Nguyen, Thin Dang, and Ngan Nguyen. NLP@UIT at SemEval-2019 task 4: The paparazzo hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 971–975. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2167.

Dan S. Nielsen and Ryan McConville. Mumin: A large-scale multilingual multimodal fact-checked misinformation social network dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 3141–3153, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531744. URL https://doi.org/10.1145/3477495.3531744.

Zhiyuan Ning, Yuanzhen Lin, and Ruichao Zhong. Team peter-parker at SemEval-2019 task 4: BERT-based method in hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1037–1040. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2181.

Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali. NELA-GT-2018: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR*, abs/1904.01546, 2019. doi: https://doi.org/10.48550/arXiv.1904.01546.

Silje Nygaard. Boundary work: Intermedia agenda-setting between right-wing alternative media and professional journalism. *Journalism Studies*, 21(6):766–782, 2020.

Ronan Ó Fathaigh, Doris Buijs, and Joris van Hoboken. The regulation of disinformation under the digital services act. *Media and Communication*, 13, 2025.

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

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

Sahar Omidi Shayegan, Isar Nejadgholi, Kellin Pelrine, Hao Yu, Sacha Levy, Zachary Yang, Jean-François Godbout, and Reihaneh Rabbany. An evaluation of language models for hyperpartisan ideology detection in Persian Twitter. In Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu, and John P. McCrae, editors, *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, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.eurali-1.8.

Silviu Oprea and Walid Magdy. iSarcasm: A dataset of intended sarcasm. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1279–1289, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.118. URL https://aclanthology.org/2020.acl-main.118/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Niko Palić, Juraj Vladika, Dominik Čubelić, Ivan Lovrenčić, Maja Buljan, and Jan Šnajder. TakeLab at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 995–998. Association for Computational Linguistics, 2019. doi: 10.18653/v1/S19-2172.

Olga Papadopoulou, Giorgos Kordopatis-Zilos, Markos Zampoglou, Symeon Papadopoulos, and Yiannis Kompatsiaris. Brenda starr at SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 924–928. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2157.

Anish Patankar, Joy Bose, and Harshit Khanna. A bias aware news recommendation system. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 232–238. IEEE, 2019. ISBN 978-1-5386-6783-5. doi: https://doi.org/10.1109/ICOSC.2019.8665610.

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

Maja Pavlovic and Massimo Poesio. The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation. In Gavin Abercrombie, Valerio Basile, Davide Bernadi, Shiran Dudy, Simona Frenda, Lucy Havens, and Sara Tonelli,

editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 100–110, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.nlperspectives-1.11/.

Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1162.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. doi: https://doi.org/10.48550/arXiv.1802.05365. URL http://arxiv.org/abs/1802.05365.

Francesco Pierri, Alessandro Artoni, and Stefano Ceri. HoaxItaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019, 2020.

Paloma Piot and Javier Parapar. *Towards Efficient and Explainable Hate Speech Detection via Model Distillation*, page 376–392. Springer Nature Switzerland, 2025. ISBN 9783031887116. doi: 10.1007/978-3-031-88711-6_24. URL http://dx.doi.org/10.1007/978-3-031-88711-6_24.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.317. URL https://aclanthology.org/2023.semeval-1.317/.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In Atul Kr. Ojha, A. Seza Doğruöz, Giovanni Da San Martino, Harish Tayyar Madabushi, Ritesh Kumar, and Elisa Sartori, editors, *Proceedings of the*

17th International Workshop on Semantic Evaluation (SemEval-2023), pages 2343–2361, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.317. URL https://aclanthology.org/2023.semeval-1.317/.

Jakub Piskorski, Nicolas Stefanovitch, Nikolaos Nikolaidis, Giovanni Da San Martino, and Preslav Nakov. Multilingual multifaceted understanding of online news in terms of genre, framing, and persuasion techniques. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3001–3022, Toronto, Canada, July 2023c. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.169. URL https://aclanthology.org/2023.acl-long.169/.

Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.731. URL https://aclanthology.org/2022.emnlp-main.731/.

Konstantin Pogorelov, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, and Johannes Langguth. Wico text: A labeled dataset of conspiracy theory and 5g-corona misinformation tweets. In *Proceedings of the 2021 Workshop on Open Challenges in Online Social Networks*, OASIS '21, page 21–25, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450386326. doi: 10.1145/3472720.3483617. URL https://doi.org/10.1145/3472720.3483617.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. 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, 2018a.

Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computa-*

tional Linguistics (Volume 1: Long Papers), pages 231–240. Association for Computational Linguistics, 2018b. doi: https://doi.org/10.18653/v1/P18-1022.

Amrit Poudel, Maria Milkowski, and Tim Weninger. The power of framing: How news headlines guide search behavior. *Findings of the Association for Computational Linguistics: EMNLP*, pages 889–900, 2025.

Markus Prior. *Post-broadcast democracy: How media choice increases inequality in political involvement and polarizes elections*. Cambridge University Press, 2007.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. Automatically neutralizing subjective bias in text. In *Proceedings of the aaai conference on artificial intelligence*, volume 34, pages 480–489, 2020.

Carla Pérez-Almendros, Luis Espinosa-Anke, and Steven Schockaert. 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, 2019. doi: https://doi.org/10.18653/v1/S19-2158.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL `https://arxiv.org/abs/2412.15115`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL `https://arxiv.org/abs/2305.18290`.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL `https://aclanthology.org/D19-1410/`.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641, 2024. ISSN 09574174. doi: https://doi.org/10.1016/j.eswa.2023.121641.

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

Robert M. Ross, David G. Rand, and Gordon Pennycook. Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. *Judgment and Decision Making*, 16(2):484–504, 2021. ISSN 1930-2975. doi: http://dx.doi.org/10.1017/S1930297500008640. Publisher: Cambridge University Press.

Rachel Edita Roxas. Large language models and natural language processing on minority languages: A systematic review. In Nathaniel Oco, Shirley N. Dita, Ariane Macalinga Borlongan, and Jong-Bok Kim, editors, *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 1–8, Tokyo, Japan, December 2024. Tokyo University of Foreign Studies. URL https://aclanthology.org/2024.paclic-1.1/.

Shamik Roy and Dan Goldwasser. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716. Association for Computational Linguistics, 2020. doi: https://doi.org/10.18653/v1/2020.emnlp-main.620.

Qin Ruan, Brian Mac Namee, and Ruihai Dong. Bias bubbles: Using semi-supervised learning to measure how many biased news articles are around us. In *AICS*, pages 153–164, 2021.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning internal representations by error propagation. 1986. URL https://api.semanticscholar.org/CorpusID:62245742.

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

10776–10787, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.722. URL https://aclanthology.org/2023.findings-emnlp.722/.

Abdelrhman Saleh, Ramy Baly, Alberto Barrón-Cedeño, Giovanni Da San Martino, Mitra Mohtarami, Preslav Nakov, and James Glass. Team QCRI-MIT at SemEval-2019 task 4: Propaganda analysis meets hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1041–1046. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2182.

Gabriele Sarti and Malvina Nissim. IT5: Text-to-text pretraining for Italian language understanding and generation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9422–9433, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.823/.

Olufunke O. Sarumi, Charles Welch, Daniel Braun, and Jörg Schlötterer. The impact of annotator personas on LLM behavior across the perspectivism spectrum. In Mourad Abbas, Tariq Yousef, and Lukas Galke, editors, *Proceedings of the 8th International Conference on Natural Language and Speech Processing (ICNLSP-2025)*, pages 121–136, Southern Denmark University, Odense, Denmark, August 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.icnlsp-1.14/.

Johannes Schäfer, Aidan Combs, Christopher Bagdon, Jiahui Li, Nadine Probol, Lynn Greschner, Sean Papay, Yarik Menchaca Resendiz, Aswathy Velutharambath, Amelie Wuehrl, Sabine Weber, and Roman Klinger. Which demographics do LLMs default to during annotation? In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17331–17348, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.848. URL https://aclanthology.org/2025.acl-long.848/.

Ralph Schroeder. Digital media and the entrenchment of right-wing populist agendas. *Social Media+ Society*, 5(4):2056305119885328, 2019.

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

Kjetil Selvik and Jacob Høigilt. Journalism under instrumentalized political parallelism. *Journalism Studies*, 22(5):653–669, 2021.

Saptarshi Sengupta and Ted Pedersen. Duluth at SemEval-2019 task 4: The pioquinto manterola hyperpartisan news detector. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 949–953. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2162.

Kasidit Sermsri and Teerapong Panboonyuen. Debiasing large language models in thai political stance detection via counterfactual calibration. In *Proceedings of the 9th Widening NLP Workshop*, pages 56–64, 2025.

Colin Seymour-Ure. *The political impact of mass media*. London, Constable; Beverly Hills, Calif., Sage Publications, 1974.

Daniel Shaprin, Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. Team jack ryder at SemEval-2019 task 4: Using BERT representations for detecting hyperpartisan news. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1012–1015. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2176.

Ankur Sharma, Navreet Kaur, Anirban Sen, and Aaditeshwar Seth. Ideology detection in the indian mass media. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 627–634. IEEE, 2020. ISBN 978-1-72811-056-1. doi: https://doi.org/10.1109/ASONAM49781.2020.9381344.

Anu Shrestha and Francesca Spezzano. 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, 2021. Springer-Verlag. ISBN 978-3-030-72239-5. doi: 10.1007/978-3-030-72240-1_9. URL https://doi.org/10.1007/978-3-030-72240-1_9.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media: A data mining perspective. *SIGKDD Explor. Newsl.*, 19(1):22–36, sep 2017. ISSN

1931-0145. doi: 10.1145/3137597.3137600. URL https://doi.org/10.1145/3137597.3137600.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. Black holes and white rabbits: Metaphor identification with visual features. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1020. URL https://aclanthology.org/N16-1020/.

Craig Silverman, Lauren Strapagiel, Hamza Shaban, Ellie Hall, and Jeremy Singer-Vine. Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. *BuzzFeed News*, 2016.

Răzvan-Alexandru Smădu, Sebastian-Vasile Echim, Dumitru-Clementin Cercel, Iuliana Marin, and Florin Pop. From fake to hyperpartisan news detection using domain adaptation, 2023.

Zhivar Sourati, Vishnu Priya Prasanna Venkatesh, Darshan Deshpande, Himanshu Rawlani, Filip Ilievski, Hông Ân Sandlin, and Alain Mermoud. Robust and explainable identification of logical fallacies in natural language arguments. *Knowledge-Based Systems*, 266:110418, 2023. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2023.110418. URL https://www.sciencedirect.com/science/article/pii/S0950705123001685.

Rui Sousa-Silva. Fighting the fake: A forensic linguistic analysis to fake news detection. *International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique*, 35, 04 2022. doi: https://doi.org/10.1007/s11196-022-09901-w.

Francesca Spezzano, Anu Shrestha, Jerry Alan Fails, and Brian W. Stone. That's fake news! reliability of news when provided title, image, source bias & full article. *Proceedings of the ACM on Human-Computer Interaction*, 5:1–19, 2021. ISSN 2573-0142. doi: https://doi.org/10.1145/3449183.

Kilian Sprenkamp, Daniel Gordon Jones, and Liudmila Zavolokina. Large Language Models for Propaganda Detection, November 2023. URL http://arxiv.org/abs/2310.06422. arXiv:2310.06422 [cs].

Agastya Sridharan and Natarajan S. An automated news bias classifier using caenorhabditis elegans inspired recursive feedback network architecture. *CoRR*, abs/2207.12724, 2022. doi: https://doi.org/10.48550/ARXIV.2207.12724.

Vertika Srivastava, Ankita Gupta, Divya Prakash, Sudeep Kumar Sahoo, Rohit R.R, and Yeon Hyang Kim. Vernon-fenwick at SemEval-2019 task 4: Hyperpartisan news detection using lexical and semantic features. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1078–1082. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2189.

Bozhidar Stevanoski and Sonja Gievska. Team ned leeds at SemEval-2019 task 4: Exploring language indicators of hyperpartisan reporting. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1026–1031. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2179.

Jennifer Stromer-Galley, Brian McKernan, Saklain Zaman, Chinmay Maganur, and Sampada Regmi. The efficacy of large language models and crowd annotation for accurate content analysis of political social media messages. *Social Science Computer Review*, 0(0):08944393251334977, 2025. doi: 10.1177/08944393251334977. URL `https://doi.org/10.1177/08944393251334977`.

Natalie Jomini Stroud. *Niche news: The politics of news choice*. Oxford University Press, 2011.

Jiuding Sun, Chantal Shaib, and Byron C Wallace. Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv:2306.11270*, 2023.

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

Cass R Sunstein. Republic: Divided democracy in the age of social media. 2018.

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

Joanna Szwoch, Mateusz Staszkow, Rafal Rzepka, and Kenji Araki. Creation of polish online news corpus for political polarization studies. In Haithem Afli, Mehwish Alam, Houda Bouamor, Cristina Blasi Casagran, Colleen Boland, and Sahar Ghannay, editors,

Proceedings of the LREC 2022 workshop on Natural Language Processing for Political Sciences, pages 86–90. European Language Resources Association, 2022.

Javier Sánchez-Junquera. On the detection of political and social bias. 2021.

Javier Sánchez-Junquera, Paolo Rosso, Manuel Montes, and Simone Ponzetto. Masking and transformer-based models for hyperpartisanship detection in news. pages 1244–1251, 01 2021. doi: 10.26615/978-954-452-072-4_140.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan,

Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL `https://arxiv.org/abs/2503.19786`.

Martina Toshevska and Sonja Gievska. Llm-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721, 2025. doi: 10.1109/ACCESS.2025.3548967.

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. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971, 2023a. URL `https://api.semanticscholar.org/CorpusID:257219404`.

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. Llama: Open and efficient foundation language models, 2023b. URL `https://arxiv.org/abs/2302.13971`.

Minh Tran. How biased are american media outlets? a framework for presentation bias regression. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 4359–4364. IEEE, 2020. ISBN 978-1-72816-251-5. doi: https://doi.org/10.1109/BigData50022.2020.9377987.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 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, 2014.

Joshua A Tucker, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal, and Brendan Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

Petter Törnberg. Large language models outperform expert coders and supervised classifiers at annotating political social media messages. *Social Science Computer Review*, 0(2024): 08944393241286471, 2024. doi: 10.1177/08944393241286471. URL https://doi.org/10.1177/08944393241286471.

University Politehnica of Bucharest, Vlad Cristian Dumitru, Traian Rebedea, and University Politehnica of Bucharest. Topic-based models with fact checking for fake news identification. In *RoCHI - International Conference on Human-Computer Interaction*, pages 182–190. MATRIX ROM, 2021. doi: https://doi.org/10.37789/rochi.2021.1.1.28.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. Towards fine-grained classification of climate change related social media text. In Samuel Louvan, Andrea Madotto, and Brielen Madureira, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-srw.35. URL https://aclanthology.org/2022.acl-srw.35/.

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

Emmanuel Vincent and Maria Mestre. Crowdsourced measure of news articles bias: Assessing contributors' reliability. In *1st Workshop on Subjectivity, Ambiguity and Disagreement in Crowdsourcing, and Short Paper Proceedings of the 1st Workshop on Disentangling the Relation Between Crowdsourcing and Bias Management (SAD 2018 and CrowdBias 2018)*, 2018.

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

2019, Proceedings, pages 288–295, 2019. doi: 10.1007/978-3-030-30760-8_25. URL https://doi.org/10.1007/978-3-030-30760-8_25.

Silvio Waisbord. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies*, 19(13):1866–1878, 2018.

Robert Walecki, Ognjen Rudovic, Vladimir Pavlovic, and Maja Pantic. Copula ordinal regression for joint estimation of facial action unit intensity. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4902–4910, 2016. doi: https://doi.org/10.1109/CVPR.2016.530.

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

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

Douglas Walton. *Ad Hominem Arguments*. University Alabama Press, 1998.

Peiqi Wang, Yikang Shen, Zhen Guo, Matthew Stallone, Yoon Kim, Polina Golland, and Rameswar Panda. Diversity measurement and subset selection for instruction tuning datasets, 2024a. URL https://arxiv.org/abs/2402.02318.

Peng Wang, Xiaobin Wang, Chao Lou, Shengyu Mao, Pengjun Xie, and Yong Jiang. Effective demonstration annotation for in-context learning via language model-based determinantal point process, 2024b. URL https://arxiv.org/abs/2408.02103.

Claire Wardle and Hossein Derakhshan. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg, 2017.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference, 2024. URL https://arxiv.org/abs/2412.13663.

Azmine Toushik Wasi, Wahid Faisal, Taj Ahmad, Abdur Rahman, and Mst Rafia Islam. Dhoroni: Exploring bengali climate change and environmental views with a

multi-perspective news dataset and natural language processing, 2024. URL https://arxiv.org/abs/2410.17225.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903, 2022. URL https://api.semanticscholar.org/CorpusID:246411621.

Weiai Wayne Xu, Yoonmo Sang, and Christopher Kim. What drives hyper-partisan news sharing: Exploring the role of source, style, and content. *Digital Journalism*, 8(4):486–505, 2020. ISSN 2167-0811, 2167-082X. doi: https://doi.org/10.1080/21670811.2020.1761264.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yingxuan Yang, Huayi Wang, Muning Wen, Xiaoyun Mo, Qiuying Peng, Jun Wang, and Weinan Zhang. P3: A policy-driven, pace-adaptive, and diversity-promoted framework for data pruning in llm training, 2024. URL https://arxiv.org/abs/2408.05541.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. HARE: Explainable hate speech detection with step-by-step reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational*

Linguistics: EMNLP 2023, pages 5490–5505, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.365. URL https://aclanthology.org/2023.findings-emnlp.365.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. Representative demonstration selection for in-context learning with two-stage determinantal point process. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.331. URL https://aclanthology.org/2023.emnlp-main.331/.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California, jun 2016. Association for Computational Linguistics. doi: https://doi.org/10.18653/v1/n16-1174.

Chia-Lun Yeh, Babak Loni, and Anne Schuth. Tom jumbo-grumbo at SemEval-2019 task 4: Hyperpartisan news detection with GloVe vectors and SVM. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1067–1071. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2187.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *J. Data and Information Quality*, 11(3), may 2019. ISSN 1936-1955. doi: https://doi.org/10.1145/3309699.

Albin Zehe, Lena Hettinger, Stefan Ernst, Christian Hauptmann, and Andreas Hotho. Team xenophilius lovegood at SemEval-2019 task 4: Hyperpartisanship classification using convolutional neural networks. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 1047–1051. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2183.

Chiyu Zhang, Arun Rajendran, and Muhammad Abdul-Mageed. UBC-NLP at SemEval-2019 task 4: Hyperpartisan news detection with attention-based bi-LSTMs. In *Proceedings of*

the 13th International Workshop on Semantic Evaluation, pages 1072–1077. Association for Computational Linguistics, 2019. doi: https://doi.org/10.18653/v1/S19-2188.

Jiang Zhang, Qiong Wu, Yiming Xu, Cheng Cao, Zheng Du, and Konstantinos Psounis. Efficient toxic content detection by bootstrapping and distilling large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(19):21779–21787, Mar. 2024. doi: 10.1609/aaai.v38i19.30178. URL https://ojs.aaai.org/index.php/AAAI/article/view/30178.

Yiming Zhu, Ehsan-Ul Haq, Gareth Tyson, Lik-Hang Lee, Yuyang Wang, and Pan Hui. 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, 2024.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, March 2024. doi: 10.1162/coli_a_00502. URL https://aclanthology.org/2024.cl-1.8/.