CEUR-WS.org/Vol-3878/62_main_long.pdf

# Leveraging Advanced Prompting Strategies in Llama-8b for Enhanced Hyperpartisan News Detection

Michele Joshua Maggini[1,*], Erik Bran Marino[2] and Pablo Gamallo Otero[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes da USC, Spain, Galicia, Santiago de Compostela, 15782
[2]Universidade de Évora, Évora, Portugal

**Abstract**

This paper explores advanced prompting strategies for hyperpartisan news detection using the Llama3-8b-Instruct model, an open-source LLM developed by Meta AI. We evaluate zero-shot, few-shot, and Chain-of-Thought (CoT) techniques on two datasets: SemEval-2019 Task 4 and a headline-specific corpus. Collaborating with a political science expert, we incorporate domain-specific knowledge and structured reasoning steps into our prompts, particularly for the CoT approach. Our findings reveal that some prompting strategies work better than others, specifically on LLaMA, depending on the dataset and the task. This unexpected result challenges assumptions about ICL efficacy on classification tasks. We discuss the implications of these findings for In-Context Learning (ICL) in political text analysis and suggest directions for future research in leveraging large language models for nuanced content classification tasks.

**Keywords**

natural language processing, large language models, hyperpartisan detection, disinformation

## 1. Introduction

The proliferation of hyperpartisan news content in digital media has become a significant challenge for modern societies, potentially undermining democratic processes and social cohesion. Hypepartisan news consists of politically polarized content presented through the usage of rhetorical bias. In the media landscape, news outlets disseminate information using proprietary websites and social networks. Each news outlet frames the narratives of the facts based on their political leaning, influencing the content with rhetorical biases, emotional purposes, ideology, and reporting the facts while omitting parts [1, 2]. To improve the virality of the news, even mainstream journalists adopted click-bait practices like eye-catching titles [3]. Furthermore, the news not only stands for one opinion but could have an underlying political background that manifests through a specific vocabulary or assumptions against the opposite political leaning [4]. This type of news could radicalize the voters because of their emotional language [5]. When there is a massive usage of these techniques, we can consider news extremely partisan toward a particular political leaning. Although hyperpartisan news can share traits with misinformation and disinformation, it cannot be classified within these domains because the intent is not deceptive. For this reason, hyperpartisan news detection is closer to propaganda.

Recent advancements in large language models (LLMs) have opened new avenues for tackling complex NLP tasks, including detecting nuanced linguistic phenomena such as bias and partisanship. Among these models, LLama3 [6], developed by Meta AI.

This research makes use of the new LLM recently released by Meta AI, Llama3-8b-Instruct, fine-tuned and optimized for dialogue/chat use cases, to explore its application in the detection of both hyperpartisan news headlines and articles. LLMs can be prompted with instructions to perform classification tasks. Thus, we intend to use this open source model. In our case, by prompting the model with instructions and context, we are in the In-Context Learning (ICL) domain, a learning approach different from fine-tuning that does not require to update models' weights [7]. The study aims to investigate the efficiency and compare the performances of the following ICL techniques: 0-shot with a general prompt and a specific prompt, few-shot with a different number of examples and Multi-task Guided CoT. We investigate how carefully crafted prompts with the help of a political expert can guide the model to identify subtle indicators of extreme political bias in news articles, leveraging the model's deep understanding of language and context. Our approach aims to overcome the limitations of traditional machine learning methods, which often struggle with the complex and evolving nature of partisan language. Furthermore, we can include definitions of the political phenomena of our interest in the prompt to further define the task and narrow the application domain.

By focusing on ICL to provide context and background information, we seek to:

---

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

*Corresponding author.

michelejoshua.maggini@usc.es (M. J. Maggini); erik.marino@uevora.pt (E. B. Marino); pablo.gamallo@usc.es (P. G. Otero)
0009-0001-9230-9202 (M. J. Maggini); 0009-0008-4757-7540 (E. B. Marino); 0000-0002-5819-2469 (P. G. Otero)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

- Develop a flexible and adaptable system that can identify hyperpartisan content across various topics and writing styles without the need for extensive retraining;
- Reduce ambiguity and guide the model towards the desired outcome;
- Minimize the influence of biases in the training data, by incorporating diverse perspectives and examples. This research not only contributes to the field of automated content analysis but also aims to compare the efficiency of prompting techniques and to analyze if LLMs are valuable tools for classification task via ICL.

The structure of the paper is as follows. In section 2 we discuss the related literature; section 3 describes the experimental set-up we adopted and the methodology; section 4 covers the findings of our experiment comparing them based on the method used and highlights the limitation of our approach; section 5 reports the main findings and future research.

The main contributions of the paper are the following:

- We evaluated the state-of-the-art model Llama3-8b-Instruct on two benchmark datasets in political domain;
- We assessed how well the model performs under different inference approaches: zero-shot learning, few-shot learning, and Multi-task Guided Chain-of-Thought reasoning
- Introduction of external in-domain knowledge in the prompt and segmentation of reasoning steps in the CoT considering the difficulty of the micro-tasks.

# 2. Related Work

## 2.1. Hyperpartisan News and Political Leaning Detection

Hyperpartisan news detection has overlapped with similar tasks like fake news and political orientation detection. In this section, we report the main contributions in the field. Two main approaches were identified related to content analysis: topic- and stylistic-based [8, 2, 9]. Particularly, by comparing which of these features contributed the most to making news hyperpartisan or fake, Potthast et al. [2] found that stylistic traits differ between hyperpartisan and mainstream news and that both extreme left-wing and right-wing articles show similar writing styles. Along the same research line, Sánchez-Junquera et al. [9] applied masking techniques to distinguish the best methodology among these. They trained the model to focus separately on the writing style or topics within the articles. This confirmed the relevance of the topic-based approach in distinguishing between hyperpartisan left- and right-wing articles, aligned with the results of Potthast et al. [2]. Building on these works, we choose to focus on controversial topics because, by definition, they are polarizing and often characterized by extreme language [1]. We believe that by leveraging generative models, we can address effectively at the same time both the content and the style.

In the literature, researchers used different parts of the articles for the classification task: Lyu et al. [1] focused on the titles; quotes in the body were investigated by Pérez-Almendros et al. [10]; while others encompassed both titles and body content [5, 11]. Other works focused on meta-information, such as the political leaning of the journalist [12], or the hyperlinks between different media ecosystems [13]. In our study, we will focus on entire articles and headlines, to evaluate model performance across inputs of varying lengths.

## 2.2. In-Context Learning

Recently, generative models with billions of parameters have been released and perform not only generative tasks, but also more discriminative ones, such as named entity recognition, sentiment classification, or even unseen tasks [14]. Users directly interact with them using prompts, which are specific textual templates containing instructions written in natural language. Their structure varies depending on the model being used. Thus, by leveraging the instructions, even with different degrees of complexity, the model can perform a task without prior training on it [15]. While interacting with the model, we can distinguish between the following prompting techniques: zero-shot, few-shot, and guided CoT [16].

ICL has emerged as a crucial technique in natural language processing, particularly with the advent of recently decoder-only LLMs. This field builds upon earlier work in transfer learning and few-shot learning [17], but focuses specifically on optimizing input prompts to elicit desired behaviors from language models. Early work in ICL primarily focused on manual prompt design. Kojima et al. [18] demonstrated the effectiveness of CoT prompting, which encourages step-by-step reasoning in language models. Building on this, Wei et al. [16] introduced the concept of zero-shot CoT prompting, further improving model performance on complex reasoning tasks without task-specific examples. More recent research has explored automated methods for prompt optimization. AutoPrompt [19] introduced a gradient-based approach to automatically generate prompts, while Prefix-Tuning [20] proposed a method to learn task-specific continuous prompts. Lester et al. [21] further developed this idea with their work on prompt tuning, demonstrating that in some cases tuning only with soft prompts can be as effec-

**Table 1**
Overview of the datasets adopted for the experimentation.

| Dataset | Data | Language | Domain | Type | Partisan Data |
|---|---|---|---|---|---|
| Hyperpartisan news[1] | 2,200 | English | News | Headlines | 898 |
| SemEval-2019[2] | 1,273 | English | News | Articles | 552 |

tive as fine-tuning the entire model. Both Prefix-Tuning and prompt tuning are actually fine-tuning techniques, as they imply to retrain the model, even though only in a partial way. The development of zero-shot and few-shot prompting techniques has significantly expanded the capabilities of LLMs. Zero-shot prompting, as demonstrated by Brown et al. [17], allows models to perform tasks without any task-specific training examples, relying solely on the task description in the prompt. Few-shot prompting, on the other hand, provides a small number of examples in the prompt to guide the model's behavior. Raffel et al. [22] explored these approaches in their work on T5 model, showing how different prompting strategies can affect model performance across various tasks. Furthermore, Lu et al. [23] investigated the impact of prompt format and example selection in few-shot learning, highlighting the importance of careful prompt design in maximizing model performance. These aspects reflect the critical role that well-crafted prompts play in unlocking the potential of large language models for tasks with limited or no task-specific training data.

# 3. Experimental Setting

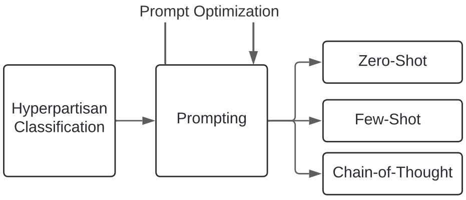

**Figure 1:** Pipeline of the experiment.

## 3.1. Datasets

For our experiment, we selected datasets tailored for binary classification. The datasets focus distinctly on headlines and the whole article. Specifically, we selected the SemEval-2019 by-article dataset [24] and the hyperpartisan news headlines dataset by Lyu et al. [1]. Both of these datasets are tailored for hyperpartisan classification. The former consists of 1,273 news articles collected by hyperpartisan and mainstream news outlets and manually labeled by 3 annotators. The latter is a collection of 2,200 news headlines manually labeled. The datasets are described in Table 1.

## 3.2. Model selection

We performed the classification as a text generation task, by inferencing the LLMs on the hyperpartisan dataset via ICL. We adopted a SOTA model: Llama3-8b-Instruct quantized in 4-bit with the QLoRA configuration [25]. The temperature of the model was fixed at 0.1 and max_tokens=1 to lower randomness in the outputs and maximizing the consistency. As a countereffect, the generated reasoning might become overly simplistic or stereotypical, lacking the nuance that slightly higher randomness could provid. Our computing infrastructure consisted of two Tesla P40 and one NVIDIA GeForce RTX 2080 Ti. Each experiment was run on a single GPU. With our approach, the class label predicted is modeled based on the previous tokens given as textual inputs through the prompts.

## 3.3. Prompt design

Earlier studies like Wei et al. [16], Jung et al. [26], Mishra et al. [14] have demonstrated the effectiveness of using task-specific prompts. Therefore, following Edwards and Camacho-Collados [27] and Labrak et al. [7], we constructed the prompts concatenating the following elements: 1) an instruction detailing the task and describing the label; 2) the input argument, supplying essential information for the task; 3) the constraints on the output space, namely inserting special symbols ” as place holders for the label, guiding the model during output generation. To improve the coherence, the specificity of the prompt and the fine-grained reasoning in CoT for the political domain, we collaborated with a Ph.D student in Political Science.

For this purpose, we designed the experimental pipeline depicted in Figure 1. We test different prompting strategies such as zero-shot, few-shot with $n$ numbers of examples (1, 2, 3, 5, 10), and a variant of guided CoT [28], namely Multi-task Guided CoT. We will compare the results given by prompting the models with instructions containing different levels of complexity: general instructions and specialized instructions with more context provided.

## 3.4. Method

To investigate the ability of LLMs on hyperpartisan detection, we audit Llama3-8b-Instruct by prompting it. In the *n*-shot configuration, we adopted the General Prompt along with examples and labels from the dataset. Examples of these prompts can be found in the Appendices.

### 0-Shot

- **0-shot General Prompt:** In this setting, we provided as context to the model the hyperpartisan article or the headline and asked the model to classify the text with the correct label. With this configuration, we leverage the internal knowledge of the model to predict the answer, aware that it can suffer from political bias [29].
- **0-shot Specific Prompt:** In this case, we provided as context to the model the article or the headline. In the instruction, we introduced a 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. With this, we insert external knowledge and introduce a political definition to narrow the task and improve the output.

**Few-shot:** In this circumstance, we evaluated the few-shot learning capabilities of LLMs across five k-shot settings and with the 0-shot General Prompt instruction: 1-shot, 2-shot, 3-shot, 5-shot, and 10-shot. In each setting, we sampled K examples from the dataset balancing the classes. Additionally, when an odd number of examples were provided, the hyperpartisan class was more represented.

**Multi-task Guided Chain-of-Thought:** In this approach, we prompted the model to break down its reasoning process step-by-step before arriving at a final classification [30]. Each step corrispond to a classification task. Previous works have treated hyperpartisan detection as a binary classification task [24, 12]. However, hyperpartisan detection can also be approached through methodologies that focus on distinct parts of the text [31]. Thus, while we frame the macro-task as binary classification, our goal is to investigate whether the model could benefit from incorporating reasoning steps into its process. These reasoning steps align with various NLP tasks that have been used to tackle hyperpartisan detection. The subtasks we focused on include sentiment analysis [32], rhetorical bias, framing bias [33], ideology detection [2], and political positioning.

By introducing complexity and dividing hyperpartisan detection into these related subtasks, we aim to enhance explainability, as the final output, namely the step-by-step generated explanation, is based on previously generated tokens. We provided the article or headline as context, along with instructions to analyze various aspects of the text—ranging from word-level features to meta-semantic reasoning—that could indicate partisan content. This method encourages the model to consider multiple factors and explicitly articulate its thought process, potentially leading to more robust and explainable classifications.

By guiding the model through a structured reasoning path, we aim to mitigate hasty judgments and foster a more nuanced analysis of the content. This technique allows us to observe how the model weighs different textual elements in its decision-making process, that is how it uses the existing internal knowledge [34], and it also provides the opportunity to identify any biases or limitations in the model's reasoning.

To develop the step-by-step chain-of-thought (CoT) reasoning and the specific prompt, we collaborated with a third-year Ph.D. student in Political Science. We preliminarily tested various prompts and configurations to craft the one used in this experiment, which led to the best results. Notably, the prompt optimization process was manual rather than automated.

# 4. Results and Discussions

The results shown in Table 2offer valuable insights into the performance of Llama3-8b-Instruct on the hyperpartisan classification task using various ICL techniques and few-shot learning approaches.

Table 2 compares the model's performance using 0-shot techniques with General (G), Specific (S) prompts, as well as Few-shot and guided CoT prompting. On the Hyperpartisan news dataset [1], 0-shot with general prompts slightly outperforms the other techniques, achieving an accuracy of 0.756 and an F1 score of 0.758. The 0-shot with Specific prompts follows closely, with an accuracy of 0.733 and an F1 score of 0.734. The CoT approach shows a slight decrease in performance, with an accuracy of 0.712 and an F1 score of 0.704. These findings suggest that for the Hyperpartisan news dataset, simpler prompting techniques may be more effective than more complex ones like CoT. This could indicate that the model already has a good grasp of the task without requiring additional reasoning steps.

With regards to the SemEval-2019 dataset [24], we observe low performance across all techniques, with the best results achieved by CoT (Acc: 0.647, F1: 0.696). This

**Table 2**

Llama3-8b-Instruct results on SemEval-2019 and Hyperpartisan news headline in 0-shot with General and Specific Prompts, Few-shot and CoT. The reported weighted Accuracy and weighted F1 scores are the averages obtained by running each model five times on the same dataset.

| Method | Hyperpartisan news |  | SemEval-2019 |  |
|---|---|---|---|---|
|  | Acc | F1 | Acc | F1 |
| 0-shot (G) | **0.756 ± .002** | **0.758 ± .010** | 0.600 ± .003 | 0.561 ± 0.036 |
| 0-shot (S) | 0.733 ± .008 | .734 ± 0.009 | .633 ± .008 | .603 ± .010 |
| CoT | .712 ± .013 | .704 ± .003 | **.647 ± .018** | **.696 ± .014** |
| Few-shot |  |  |  |  |
| 1-shot | **.752 ± .008** | **.742 ± .008** | **.639 ± .003** | **.614 ± .031** |
| 2-shot | .729 ± .012 | .717 ± .016 | .583 ± .017 | .540 ± .020 |
| 3-shot | .735 ± .018 | .737 ± .019 | .474 ± .018 | .351 ± .027 |
| 5-shot | .713 ± .011 | .712 ± .008 | .466 ± .002 | .340 ± .016 |
| 10-shot | .725 ± .018 | .725 ± .015 | .517 ± 0.008 | .437 ± .030 |

discrepancy between datasets highlights the importance of dataset characteristics in model performance.

Table 2 presents the results of few-shot learning experiments, ranging from 1-shot to 10-shot. For the Hyperpartisan news dataset, we observe an unstable performance as the number of shots increases, with the best results achieved at 1-shot (Acc: 0.752, F1: 0.742). The performance increase is not linear, with some fluctuations observed, such as a slight increase at 3-shot. For the SemEval-2019 dataset, we see a general trend of decreasing performance as the number of shots increases, with the best results at 1-shot (Acc: 0.639, F1: 0.614).

Taken this into account, with Hyperpartisan news dataset, the model not always benefit from additional examples, suggesting that it rarely can leverage this information to improve its understanding of the task. Furthermore, additional examples and context do not improve the performance with 0-shot (G) prompt configuration. Conversely, for SemEval-2019, the performance degradation with increased shots could indicate potential overfitting or confusion introduced by the additional examples.

We hypothesize that the ineffectiveness of introducing external knowledge and additional context stems from the Llama-3-8b-instruct model's optimization for dialogue and instruction-following tasks. This specialization enables the model to excel in zero-shot scenarios. Consequently, the few-shot setting may introduce complexity that exceeds the model's current capabilities, potentially interfering with its performance rather than enhancing it.

These findings underscore the complexity of ICL in the context of hyperpartisan classification. The results suggest that the optimal approach may vary depending on the specific dataset, the length of input-tokens, complexity of the instructions and task characteristics.

## 4.1. Limitations

**Outputs' inconsistency** We observed unexpected behaviors from the model despite providing clear instructions and a specific output template. The model generated extra text that wasn't requested in the instructions. We tackle this, by specifying a placeholder for the label. Additionally, it misspelled output labels, deviating from the format specified in the prompt. These issues highlight the challenges in controlling language model outputs, even with explicit guidelines. When the output did not correspond to our instructions, we considered this output as misclassified.

**Order of examples** During few-shot learning experiments, we noticed that the model performance was sensitive to examples' order [35, 23]. This fact raises concerns about the stability and reproducibility of few-shot learning techniques with LLMs. To quantify this effect, we conducted controlled experiments with systematically permuted example orders. Results revealed substantial fluctuations in performance metrics, with variations in accuracy and F1 scores exceeding 5-6% in some cases. This variability underscores the need for careful consideration of example selection and ordering in few-shot prompting strategies, highlighting a critical area for future research.

**Limited context window** Llama3-8b-Instruct has a context window of 8,200 tokens. This limitation prevented us from performing 10-shot learning with the SemEval-2019 dataset due to the length of the articles. The combined size of the articles and the necessary instructions exceeded the model's maximum context capacity.

**Quantizied model** In this study, we exclusively employed 4-bit quantized models to optimize computational efficiency. While this approach significantly reduced memory requirements and inference time, we acknowledge its potential impact on model performance. Quantization, particularly at the 4-bit level, can lead to a compression of the model's parameters, potentially resulting in a trade-off between efficiency and accuracy.

## 5. Conclusion

In this paper, we study the reliability of a SOTA model like Llama3-8b-Instruct for classification tasks in the political domain, namely to detect hyperpartisan articles and headlines comparing different prompting techniques. We cast the problem of the classification task using the generative capabilities of LLMs. Experiment results contradict the hypothesis that feeding the model with more context could lead to better performances [16]. Indeed, in our case, the 0-shot approach was the most efficient. An interesting future direction would be building a new dataset of instructions to improve models' capability in zero-shot [36] in identifying hyperpartisan news, inspired by datasets used for false information detection, such as Truthful-QA [37]. Indeed, this dataset could be used to fine-tune generative models to enhance their performance. Additionally, we plan to explore more sophisticated prompting techniques in zero-shot and few-shot settings like prompt tuning in the political domain [38]. Finally, we would like to investigate Retrieval-Augmented Generation (RAG) and implement neuro-symbolic strategies, to incorporate retrieved documents or knowledge bases into the process. By pursuing these research directions, we aim to develop more effective and reliable systems for detecting hyperpartisan news and promoting media literacy.

## Acknowledgments

This work is supported by the EUHORIZON2021 European Union's Horizon Europe research and innovation programme (https://cordis.europa.eu/project/id/101073351/es) the Marie Skłodowska-Curie Grant No.: 101073351. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. The authors have no relevant financial or non-financial interests to disclose.

## References

[1] H. Lyu, J. Pan, Z. Wang, J. Luo, Computational assessment of hyperpartisanship in news titles, 2023. doi:https://doi.org/10.48550/arXiv.2301.06270.

[2] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, B. Stein, A stylometric inquiry into hyperpartisan and fake news, in: I. Gurevych, Y. Miyao (Eds.), Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2018, pp. 231–240. doi:https://doi.org/10.18653/v1/P18-1022.

[3] F. Pierri, A. Artoni, S. Ceri, HoaxItaly: a collection of italian disinformation and fact-checking stories shared on twitter in 2019, 2020. doi:https://doi.org/10.48550/arXiv.2001.10926.

[4] G. K. W. Huang, J. C. Lee, Hyperpartisan news and articles detection using BERT and ELMo, in: 2019 International Conference on Computer and Drone Applications (IConDA), IEEE, 2019, pp. 29–32. doi:https://doi.org/10.1109/IConDA47345.2019.9034917.

[5] N. R. Naredla, F. F. Adedoyin, Detection of hyperpartisan news articles using natural language processing technique, International Journal of Information Management Data Insights 2 (2022) 100064. doi:https://doi.org/10.1016/j.jjimei.2022.100064.

[6] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, ArXiv abs/2302.13971 (2023). URL: https://api.semanticscholar.org/CorpusID:257219404.

[7] Y. Labrak, M. Rouvier, R. Dufour, A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2049–2066. URL: https://aclanthology.org/2024.lrec-main.185.

[8] Y. Liu, X. F. Zhang, D. Wegsman, N. Beauchamp, L. Wang, POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, 2022, pp. 1354–1374. doi:https://doi.org/10.18653/v1/2022.findings-naacl.101.

[9] J. Sánchez-Junquera, P. Rosso, M. Montes, S. Ponzetto, Masking and transformer-based models for hyperpartisanship detection in news, 2021, pp. 1244–1251. doi:10.26615/978-954-452-072-4_140.

[10] C. Pérez-Almendros, L. Espinosa-Anke, S. Schockaert, Cardiff university at SemEval-2019 task 4: Lin-

guistic features for hyperpartisan news detection, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 929–933. doi:`https://doi.org/10.18653/v1/S19-2158`.

[11] D.-V. Nguyen, T. Dang, N. Nguyen, NLP@UIT at SemEval-2019 task 4: The paparazzo hyperpartisan news detector, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, 2019, pp. 971–975. doi:`https://doi.org/10.18653/v1/S19-2167`.

[12] K. M. Alzhrani, Political ideology detection of news articles using deep neural networks, Intelligent Automation & Soft Computing 33 (2022) 483–500. doi:`https://doi.org/10.32604/iasc.2022.023914`.

[13] A. Hrckova, R. Moro, I. Srba, M. Bielikova, Quantitative and qualitative analysis of linking patterns of mainstream and partisan online news media in central europe, Online Information Review 46 (2021) 954–973. doi:`https://doi.org/10.1108/OIR-10-2020-0441`, publisher: Emerald Publishing Limited.

[14] S. Mishra, D. Khashabi, C. Baral, H. Hajishirzi, Cross-task generalization via natural language crowdsourcing instructions, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3470–3487. URL: https://aclanthology.org/2022.acl-long.244. doi:`10.18653/v1/2022.acl-long.244`.

[15] A. Efrat, O. Levy, The turking test: Can language models understand instructions?, ArXiv abs/2010.11982 (2020). URL: https://api.semanticscholar.org/CorpusID:225062157.

[16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. hsin Chi, F. Xia, Q. Le, D. Zhou, Chain of thought prompting elicits reasoning in large language models, ArXiv abs/2201.11903 (2022). URL: https://api.semanticscholar.org/CorpusID:246411621.

[17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, ArXiv abs/2005.14165 (2020). URL: https://api.semanticscholar.org/CorpusID:218971783.

[18] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, ArXiv abs/2205.11916 (2022). URL: https://api.semanticscholar.org/CorpusID:249017743.

[19] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4222–4235. URL: https://aclanthology.org/2020.emnlp-main.346. doi:`10.18653/v1/2020.emnlp-main.346`.

[20] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, Online, 2021, pp. 4582–4597. URL: https://aclanthology.org/2021.acl-long.353. doi:`10.18653/v1/2021.acl-long.353`.

[21] B. Lester, R. Al-Rfou, N. Constant, The power of scale for parameter-efficient prompt tuning, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 3045–3059. URL: https://aclanthology.org/2021.emnlp-main.243. doi:`10.18653/v1/2021.emnlp-main.243`.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

[23] Y. Lu, M. Bartolo, A. Moore, S. Riedel, P. Stenetorp, Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 8086–8098. URL: https://aclanthology.org/2022.acl-long.556. doi:`10.18653/v1/2022.acl-long.556`.

[24] J. Kiesel, M. Mestre, R. Shukla, E. Vincent, P. Adineh, D. Corney, B. Stein, M. Potthast, Semeval-2019 task 4: Hyperpartisan news detection, in: International Workshop on Semantic Evaluation, 2019. URL: https://api.semanticscholar.org/CorpusID:120224153.

[25] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettle-

moyer, Qlora: Efficient finetuning of quantized llms, Advances in Neural Information Processing Systems 36 (2024).

[26] J. Jung, L. Qin, S. Welleck, F. Brahman, C. Bhagavatula, R. Le Bras, Y. Choi, Maieutic prompting: Logically consistent reasoning with recursive explanations, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1266–1279. URL: https://aclanthology.org/2022.emnlp-main.82. doi:10.18653/v1/2022.emnlp-main.82.

[27] A. Edwards, J. Camacho-Collados, Language models for text classification: Is in-context learning enough?, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 10058–10072. URL: https://aclanthology.org/2024.lrec-main.879.

[28] J. Lee, F. Yang, T. Tran, Q. Hu, E. Barut, K.-W. Chang, Can small language models help large language models reason better?: LM-guided chain-of-thought, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 2835–2843. URL: https://aclanthology.org/2024.lrec-main.252.

[29] Y. Bang, D. Chen, N. Lee, P. 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.

[30] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, S.-Y. Yun, HARE: Explainable hate speech detection with step-by-step reasoning, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5490–5505. URL: https://aclanthology.org/2023.findings-emnlp.365. doi:10.18653/v1/2023.findings-emnlp.365.

[31] M. Michele Joshua, B. Davide, P. Paloma, D. Gaël, G. O. Pablo, A systematic review of hyperpartisan news detection: A comprehensive framework for definition, detection, and evaluation, 2024. doi:https://doi.org/10.21203/rs.3.rs-3893574/v1.

[32] M. Hitesh, V. Vaibhav, Y. A. Kalki, S. H. Kamtam, S. Kumari, Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model, in: 2019 2nd international conference on intelligent communication and computational techniques (ICCT), IEEE, 2019, pp. 146–151.

[33] S. Roy, D. Goldwasser, Weakly supervised learning of nuanced frames for analyzing polarization in news media, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 7698–7716. URL: https://aclanthology.org/2020.emnlp-main.620. doi:10.18653/v1/2020.emnlp-main.620.

[34] M. Ren, B. Cao, H. Lin, C. Liu, X. Han, K. Zeng, W. Guanglu, X. Cai, L. Sun, Learning or self-aligning? rethinking instruction fine-tuning, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 6090–6105. URL: https://aclanthology.org/2024.acl-long.330. doi:10.18653/v1/2024.acl-long.330.

[35] Z. Zhao, E. Wallace, S. Feng, D. Klein, S. Singh, Calibrate before use: Improving few-shot performance of language models, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 12697–12706. URL: https://proceedings.mlr.press/v139/zhao21c.html.

[36] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, ArXiv abs/2109.01652 (2021). URL: https://api.semanticscholar.org/CorpusID:237416585.

[37] O. Tafjord, B. Dalvi, P. Clark, Entailer: Answering questions with faithful and truthful chains of reasoning, in: Conference on Empirical Methods in Natural Language Processing, 2022. URL: https://api.semanticscholar.org/CorpusID:253097865.

[38] K.-M. Kim, M. Lee, H.-S. Won, M.-J. Kim, Y. Kim, S. Lee, Multi-stage prompt tuning for political perspective detection in low-resource settings, Applied Sciences 13 (2023) 6252. doi:https://doi.org/10.3390/app13106252.

## 6. Appendices

**Prompt templates** In this section we show the prompt used in the different tasks.

**General prompt:** **System message**: "role": "system", "content": "You have been provided with an instruction describing a task and it is combined with an input giving more context. Reply as indicated in the instruction."

**User query**: Instruction: Detect if an headline is hyperpartisan. The output's format will be only an integer, like 'integer'. If the article is hyperpartisan reply '1', else with '0'. Input: { }
Output: {model response}

**Specific prompt:** **System message**: "role": "system", "content": "You are an expert in Political Communication Science. You have been provided with an instruction describing a task and it is combined with an input giving more context. Reply as indicated in the instruction."

**User query**: Instruction: Determine if the headline is hyperpartisan whether not. Hyperpartisan articles contain biases, particularly ad hominem attack, loaded language and evidences of political ideology. Sometimes they rely on cherry-picking strategy. The output's format will be only an integer, like 'integer'. If the article is hyperpartisan reply '1', else with '0'. Input: { }
Output: {model response}

**Multi-task Guided CoT** **System message:** "role": "system", "content": "You are an expert in Political Communication Science. You have been provided with an instruction describing a task and it is combined with an input giving more context. Reply as indicated in the instruction."

**User query**: Instruction: The output's format will be only an integer, like 'integer'. If the article is hyperpartisan reply '1', else with '0'. Determine if the headline is hyperpartisan whether not. Reason step by step following this reasoning schema: 1. Sentiment analysis: Analyze the tone and the language to see if there are polarizing words, or emotional language. 2. Rhetorical bias: Rhetoric refers to speaking or writing that is designed to have a persuasive or impressive effect but is lacking in meaningful content. Analyze the presence of rhethorical biases like ad hominem attack. 3. Framing bias: involves presenting information to shape or influence people's perceptions of an issue or event by emphasizing certain aspects while downplaying others 4. Ideological bias: Analyze the presence of ideological bias, that is if by carefully reading the text specific moral values appear related to a specific ideology. 5. Intention: Analyze the intent of the article. Does it aim to persuade or just to inform? 6. Unilateral: Does the article provide only a point of view on the subject? Is the article unilateral in its coverage of the subject? 7. Left-wing hyperpartisan: Consider you are a left-wing reader. Would you consider this article as hyperpartisan to your political stance? 8. Right-wing hyperpartisan: Consider you are a right-wing reader. Would you consider this article as hyperpartisan to your political stance? Input: { }
Output: {model response}

**Output of CoT** "Based on the provided article, I will analyze it according to the given instruction.

**Sentiment analysis:** The tone of the article is critical and negative towards President Trump, using words like ""viciously attacked"", ""cruel"", and ""on their own"" to describe his response to the situation in Puerto Rico. The language used is emotional and polarizing.

**Rhetorical bias:** The article uses ad hominem attacks, criticizing Trump's response to the situation in Puerto Rico and implying that his actions are motivated by political considerations rather than a genuine concern for the well-being of the affected people.

**Framing bias:** The article frames Trump's response as a failure and a reflection of his character, emphasizing the negative consequences of his actions and downplaying any potential positive aspects.

**Ideological bias:** The article appears to be critical of Trump's ideology and his perceived lack of empathy for marginalized communities, such as the people of Puerto Rico.

**Intention:** The article's intention is to persuade the reader to view Trump's response to the situation in Puerto Rico as unacceptable and to criticize his character.

**Unilateral:** The article presents a one-sided view of the situation, criticizing Trump's response without providing a balanced perspective.

**Left-wing hyperpartisan:** As a left-wing reader, I would consider this article as not hyperpartisan, as it is critical of Trump's response and his perceived lack of empathy for marginalized communities, which aligns with my political stance.

**Right-wing hyperpartisan:** As a right-wing reader, I would consider this article as hyperpartisan, as it is critical of Trump's response and his perceived lack of empathy for marginalized communities, which may be seen as an attack on Trump's character and political ideology.

**Output:** 'label': '1' (hyperpartisan)"