Automated Media Assessment Using Large Language Models

Author 1[1], Author 2[1,3], and Author 3[2,3]

[1] ### Hidden for blind review ####
[2] ### Hidden for blind review ####
[3] ### Hidden for blind review ####

Abstract. Often, media producers exhibit certain types of bias that are passed through to news consumers, potentially shaping their opinions and perception of reality. This can have multiple negative effects, including polarized speech and general distrust on institutions and news producers themselves. The main goal of this research is to cover the European Portuguese news landscape and create an automated system that frames news pieces according to three axes: *political bias*, *reliability*, and *objectivity*. We adopt two methodologies: zero-shot prompting and a vector database technique, both working directly with a large language model and, more specifically, tested with EuroLLM 9B and LLaMA 3.3 70B. We report interesting results as the system reaches close to 90% accuracy on some cases. Furthermore, we observe that a medium-sized model works significantly better than a small-sized one.

Keywords: Media bias · Political bias · Reliability · Objectivity · Natural language processing · Large language models

1 Introduction

The media plays a crucial role in shaping public opinion and influencing societal discourse [6]. With this power comes the responsibility of adhering to journalistic principles of factuality, objectivity, and neutrality. However, media bias remains a persistent issue, often manifesting itself through framing, selective fact presentation, and sensationalism [13]. While some level of subjectivity is perhaps inevitable, unrecognized biases can distort public perception and erode trust in mainstream news sources.

Media outlets are often biased in some way [13],[19]. As shown by Rodrigo-Ginés et al. [12], media bias can take multiple forms, each with its own nuances and implications. In this research, we address three main issues with news media:

- **Political Bias:** Slant or favoritism in reporting that reflects a particular political perspective, often influencing how facts are presented and the tone or framing used.
- **Reliability:** Accuracy, consistency, and trustworthiness of the information presented, based on evidence, credible sources, and adherence to journalistic standards.

– **Objectivity:** Presentation of facts and events in a neutral, objective manner, without letting opinions influence the tone and overall reporting.

In order to create an automated classification system to frame news articles into these three axes, we adopt two methodologies based on leveraging open-weight large language models – EuroLLM 9B and LLaMA 3.3 70B. The first methodology simply involves prompting the model to classify the articles while also providing additional context and output requirements – also known as zero-shot prompting. The second method is based on vector database techniques, which provide a high level of explainability and flexibility.

This work addresses the lack of research and resources in media bias detection for the European Portuguese language. Unlike the wealth of studies focused on English-language media [13], particularly from the United States, Portuguese media have been largely overlooked. By employing large language models and their in-context learning capabilities, this research explores prompt-based methods for text classification tasks.

The rest of this paper is organized as follows. Section 2 reviews related work on media bias detection. Section 3 details the proposed methodology, while Section 4 describes the datasets used. Section 5 presents the experimental results and key findings. Finally, Section 6 offers concluding remarks.

2 Related Work

Media bias, as described by Spinde et al. [13], is an intentional or unintentional expression of tendency in favor of a perspective, ideology, or result that introduces systematic slants in how events are reported. Rooted in factors such as ownership, funding, and ideological influences [17], bias in media coverage can distort public perception, shaping the general public's opinions and decisions [5, 6, 13].

Looking into research working within the European Portuguese landscape, Moura et al. [10] proposed a system for automatic detection of fake news. The study uses traditional machine learning techniques with features such as n-grams, frequencies, metadata, and readability metrics. Tavares [15] used the output of BERT-based models such as Albertina, BERTimbau, and XLM-RoBERTa, in conjunction with other features extracted from the news articles to classify them according to their reliability, objectivity, and political bias. The author used the same dataset as Moura et al. [10] for reliability; for objectivity, BABE [14], an English dataset labeled for political stance and objectivity was translated into Portuguese using DeepL; and for political bias detection, the author used a proxy dataset by scraping Portuguese parliamentary minutes. Furthermore, Afonso and Rosas [1] worked on reliability classification, experimenting with traditional machine learning methods and also deep learning methods such as BERT-based methods and LSTM networks. The research experimented with a wide variety of possible training methods and different sets of features, and interestingly, traditional machine learning techniques were capable of keeping up in performance.

Finally, Caled et al. [3] developed the Mainstream and Independent News Text corpus (MINT), a complete dataset that groups news articles into five classes: hard news, soft news, opinion, satire, and conspiracy. The author's objective was to create a dataset that would be suitable for a more nuanced look at what fake news classification usually entails. Arguing that true/fake news is too strict in terms of labeling, the authors provide a more granular look distinguishing conspiracy from satire, and what should be considered as hard, factual news, from soft, gossip-related news.

Now focusing on prompt-based methods, Alghamdi et al. [2] proposed a prompt-based learning framework for fake news detection using GPT-2, focusing on zero-shot and few-shot learning. Regarding methodology, the authors employed prompt engineering, where templates and verbalizers were designed to guide the model's predictions. Additionally, domain information was incorporated into prompts to provide context. Sensitivity analyses revealed that slight changes in prompt design and wording can affect performance, highlighting the importance of careful prompt design. The results suggested that the proposed framework is effective for fake news detection, offering a scalable solution for scenarios with limited labeled data. Another study researching reliability detection in news content is that of Wen and Younes [18]. This paper evaluated ChatGPT's zero-shot capability to detect six types of bias-related characteristics, comparing its performance to fine-tuned models like BART, ConvBERT, and GPT-2. The prompts themselves were generated by ChatGPT on several iterations and then tested on a validation set to choose the best-performing prompt. To solve the task in a zero-shot manner, ChatGPT performs respectably on racial or context biases but underperforms fine-tuned models in fake news detection, probably due to the complexity and ambiguity of the task, according to the authors. The paper concludes by highlighting the potential for scalability when using zero-shot approaches, but underlines limitations in detecting nuanced biases without specific training. Finally, working on political stance detection, Lin et al. [8] presented IndiVec, a framework for political bias detection by adapting in-context learning methods. The methodology essentially relies on a vector database and bias indicators generated by the model across dimensions like tone, language, and citations. When analyzing new text, IndiVec matches descriptors derived from the input text to the stored indicators, assigning bias labels based on a majority voting system. The authors also fine-tuned BERT-based models for the same task and observed that although these would outperform the framework when working with in-domain data, IndiVec excelled in adaptability to different datasets, maintaining its performance while the fine-tuned models would fall apart.

3 Methodology

This section presents the details on the methodology we follow for bias classification in the context of Portuguese news articles. The core methods rely on prompting, meaning the employment of large language models (LLMs) and their in-context learning capabilities, without the need for parameter updates. Each

experiment was executed using two different open-weight models: EuroLLM 9B [9] and LLaMA 3.3 70B [16]. EuroLLM is a series of language models that support the 24 official European languages and a few additional ones. This choice is based on the fact that it was specifically trained to support European Portuguese. LLaMA is Meta's offering and one of the most used open-weight solutions, also providing multilingual support. Using the two models, we can assess how well they perform on this particular task and how the jump from a small to a medium-sized model is reflected in terms of final results.

3.1 Zero-shot Classification

Zero-shot learning is the ability of a model to perform tasks it has not been explicitly trained on, just by using general knowledge acquired during its pre-training stages and task instructions. This essentially means we prompt the LLM to classify a given article according to its political bias, reliability, and objectivity, separately. Given that the target of this research is Portuguese media, every article used is written in European Portuguese, and as such, so are the prompts. We illustrate a prompt that was used for zero-shot classification in the context of objectivity, translated to English:

```
You are a model trained to classify the objectivity of
    opinion and news articles. Your task is to classify the
    provided article as "objective" or "subjective", based on
     its language, structure, and transparency, without
    assessing the veracity of its contents.
An objective text tends to use neutral language, avoids
    expressing opinions, presents verifiable data, or
    maintains an impartial tone and clear structure.
A subjective text tends to use opinionated or emotionally-
    charged language, expresses value judgments, favors a
    point of view without presenting counterpoints, or makes
    assumptions not supported by verifiable data.

The output must always be a JSON in the following format:
{"class": "objective"} or {"class": "subjective"}

The article below is clearly classified with one of those
    options. Classify the objectivity of the following text
    and output the answer in the correct format only:
{ARTICLE BODY}
```

The prompt provides context into what is required, some indications into what we consider to be objective and subjective, and finally, a specific mention to how the output should be formatted. Formatting the output in JSON style is a simple and effective way of extracting information from the output of an LLM in a structured and programmatic way. Prompts for political bias and reliability are included in Appendix A. System prompts are omitted due to space constraints and are available, as well as the code behind the classification system, in `https://github.com/**hidden-for-blind-review**`

3.2 Vector Databases

Vector database techniques encompass a set of methods designed to efficiently store, index, and retrieve high-dimensional vector representations of data, in this case, text data. These embeddings capture the semantic features of the input data, enabling similarity-based operations. This concept is the backbone of this second type of employed method.

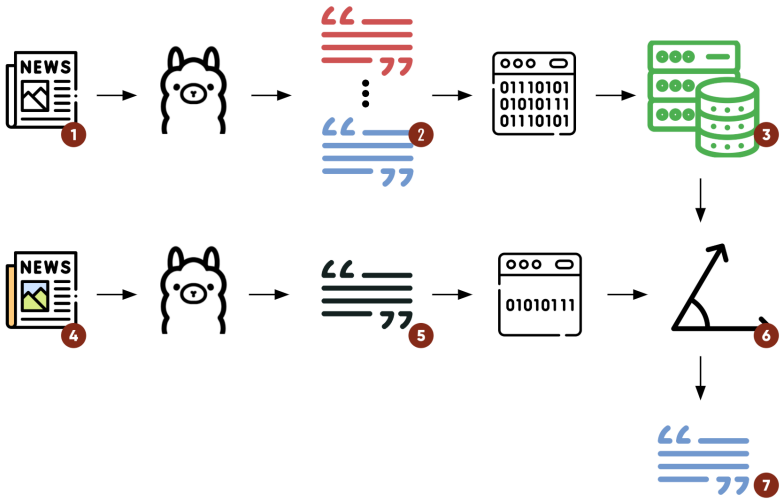

Fig. 1. Vector databases classification pipeline

Figure 1 visually represents the whole classification pipeline. The top row represents what could be considered the "training" of the system, although no training is involved, and the second row represents the testing/deployment phase. The process starts at step 1, where a number of labeled articles are provided to the LLM one by one through a series of prompts that also include the label and an instruction to generate an *indicator*, considering said label. The main idea is to guide the model to generate a small snippet of text that characterizes the article itself, based on the label. This label is attached to the generated indicator, as shown in step 2 of the schematic. To give an actual example, a right-leaning indicator of political bias generated by LLaMA 3.3 70B is "it refers to immigration as something negative for society and associates it with problems of security and integration". After this step, every indicator is converted into an embedding vector and stored in a vector database (step 3), as well as its corresponding label. The conversion is done using a Sentence Transformer [11], specifically `paraphrase-multilingual-mpnet-base-v2`, which performs well for similarity search tasks in a multilingual context. In step 4, the classification of a new instance of news is started. The process is initially very similar, and the LLM is

prompted to generate a *descriptor* (using a more neutral prompt that does not include the label). These differ from indicators in the sense that descriptors are not associated with any ground truth label since they come from new, unlabeled data (or a test set). Nevertheless, descriptors should look very similar to indicators in any other way, as step 5 illustrates. The newly generated descriptor is converted into a vector representation using the same embedding model and then matched against the whole vector database. Using cosine similarity for the search, as denoted in step 6, the top K most similar indicator vectors are chosen to classify the descriptor into one of the classes (step 7). This whole process is repeated for each of the three axes, creating three different vector databases.

Regarding the classification itself, for each of the three vector databases already built at this point, a score of -100 or 100 is also attached to each of the stored indicators, based on the label. For political bias, -100 was assigned to each left-wing indicator and 100 to each right-wing indicator; for reliability, -100 to indicators representing fake news and 100 otherwise; and for objectivity, -100 to subjectivity-related indicators and 100 to objectivity-related ones. As explained in the previous paragraph, the process involves a similarity search across the respective vector database that retrieves a cosine similarity value for each indicator. Subsequently, a weighted average score is computed, where the cosine similarities between the descriptor under classification and the indicators serve as the weighting factors, which is expressed in equation 1.

$$Axis\ Score = \frac{\sum_{i=1}^{n} similarity_i \times score_i}{\sum_{i=1}^{n} similarity_i} \tag{1}$$

Considering the objectivity classification task, an example of a translated prompt for indicator generation, specifically for objective articles[4], is:

```
An objectivity indicator is defined here as a descriptive
    label that represents the presence of elements in an
    article that contribute to its impartiality and rigor.
What you will detect are specifically linguistic, structural,
     or stylistic cues that reinforce the perceived
    objectivity of the article.
The output should always be a JSON with the following format:
{"objective": "identified indicator"}

Examples:
{"objective": "Avoids emotional language or subjective
    adjectives."}
{"objective": "Uses a neutral and descriptive tone, without
    expressing opinion."}
```

[4] For clarity, indicator generation prompts are adjusted for each class within an axis to enforce the label to the model.

```
{"objective": "Provides multiple perspectives on the topic,
    without favoring a specific position."}
{"objective": "Clear and logical structure, facilitating the
    understanding of the information."}

Assuming the following text as objective, extract only one
    indicator of objectivity from the text and format the
    output exactly as in the examples:
{ARTICLE BODY}
```

The user prompts for indicator generation of the 'right' and 'reliable' classes from the political bias and reliability axes, respectively, are shown in Appendix B. The remaining prompts – all system prompts, user prompts for indicator generation of the missing classes, and user prompts for descriptor generation – are made available in the same GitHub repository linked in the previous section.

The advantages of this methodology are clear. Firstly, it is adaptable and flexible since it is not fine-tuning a model into a specific dataset, avoiding domain overfitting. Secondly, it is an explainable method since indicators and descriptors are represented in natural language, allowing for direct human interpretation. And lastly, it doesn't require a significant amount of data in order to build a sufficiently big vector database that would allow for an effective system.

4 Data

There is a significant lack of available resources for media bias detection in European Portuguese. Due to this issue, some alternative routes had to be taken regarding the datasets chosen for assessing the effectiveness of our methodologies. To the best of our knowledge, there is no dataset that is labeled for all three axes simultaneously. For that reason, we consider different datasets for each axis. The final result is three datasets, each with two possible labels, making each axis a binary classification problem.

4.1 Political Bias

There are a few labeled datasets in the scope of political bias, mostly covering the American political landscape. Popular examples of this are BABE [14] and FlipBias [4]. As for the Portuguese landscape, to the best of our knowledge, there is no publicly available dataset. With this in mind, the only solution would be to build a dataset from scratch.

Of the three axes, political bias is possibly the most subjective one, making the labeling of articles a challenging task. This, and the fact that it is a costly and laborious process, led us to adopt an alternative approach for building the dataset that is both more reliable and possible to automate. We resort to opinion articles that are often published in the same outlets as news. In particular, we select those that have been written by personalities with specific and well-known political affiliations. Based on this information, we infer the political leaning of

the article based on authorship (similar to the "by publisher" dataset annotation in Kiesel et al. [7]). We posit that this approach provides a reliable proxy for framing the political bias of news articles, despite the differences in text genre.

The final dataset is composed of 6901 opinion articles: 3914 right-wing biased and 2987 left-wing biased. All articles were extracted from popular Portuguese media outlets such as Expresso, Público, Dinheiro Vivo, and Eco Sapo, and were written by 22 authors, each associated with a political party covering the whole spectrum. The final labels were set to either "right" or "left", depending on the party's positioning.

4.2 Reliability

For the reliability axis, the dataset used was built and made available by Afonso and Rosas [1]. It exclusively targets the problem of fake news detection within the European Portuguese context and is comprised of 31,716 credible news articles and 31,520 fake news articles, with both sides sourced from multiple news distributors and websites, adding variety to the domain. After some cleaning and source selection, as well as a random sampling process, we end up using 10,000 articles, evenly divided between the two classes.

4.3 Objectivity

MINT is a dataset that groups news articles into five classes: hard news, soft news, opinion, satire, and conspiracy. The most extreme version of what could be seen as a subjective news piece is an opinion-based article and, with this in mind, the data used to evaluate our methodologies' capability to frame an article into the objectivity axis is composed of the hard news and opinion examples from MINT, both representing each opposite of the one-dimensional spectrum between what is objective news and what is subjective news. This translates into 12,000 articles in total – 6,000 for each class.

5 Results and Discussion

5.1 Experimental Evaluation

For every axis, only 80% of the corresponding dataset is used to build each vector database, while the remaining 20% is used to test. In the case of the zero-shot methodology, we only consider the test sets. The splits were performed randomly, keeping class distribution in both sets, and are the same when using both models for a fair comparison. Table 1 shows the number of articles for each split.

Table 2 presents the results obtained in every experiment. We show results for the EuroLLM 9B model and for LLaMA 3.3 70B, each one of them first used with the zero-shot approach and then with the vector database methodology.

The first obvious remark is that, as expected, LLaMA outperforms EuroLLM from every possible perspective. This reflects the increased general capability of

Table 1. Dataset splits for each axis

	Train set	**Test set**
Political Bias	5520	1381
Reliability	8000	2000
Objectivity	9600	2400

Table 2. Performance comparison across models for political bias, reliability, and objectivity. "ZS" refers to the zero-shot methodology; "VD" refers to the vector database approach. The F1-score metric presented is the weighted-averaged version.

		EuroLLM$_{ZS}$	**EuroLLM$_{VD}$**	**LLaMA$_{ZS}$**	**LLaMA$_{VD}$**
Political Bias	Accuracy	0.56	0.59	0.67	**0.69**
	F1	0.55	0.59	0.66	**0.69**
Reliability	Accuracy	0.66	0.80	**0.87**	0.84
	F1	0.63	0.79	**0.87**	0.84
Objectivity	Accuracy	0.61	0.69	**0.89**	0.84
	F1	0.55	0.68	**0.89**	0.84

a medium-sized model when compared to a small one. When considering the same axis and same methodology, the increase in accuracy varies significantly, from as low as 5% to a maximum of around 46%, with the average being an increase of 23.5%.

Looking at the axes individually, we see that the performance on political bias detection is across the board worse than for the other two axes. This might highlight the subjectiveness and higher complexity of the problem of political stance detection. Between reliability and objectivity, the differences are, in general, not significant. In our experiments, EuroLLM has better performance when classifying reliability when compared to objectivity, whereas LLaMA has similar performance between the two axes, with a slight edge towards the objectivity side in the zero-shot approach.

Lastly, some interesting insights can be extracted from comparing the results obtained via zero-shot and the vector database methodology. When using the smallest model, the latter method is a strict upgrade in terms of performance for every axis. This is important because the performance of these smaller models can lag behind in a zero-shot fashion, and having a way of improving it reliably can be valuable. We want to highlight the significant boost in the reliability axis, where the reported 80% accuracy rivals that of the larger 70B parameter model. On the other hand, when working with the LLaMA model, a similar performance gain is not observed. When the model performed the worst – detecting political bias – the vector database approach brought a slight improvement, but for the other two axes, this is not the case. The model is able to achieve a surprisingly high performance when using a zero-shot approach, with 87% and 89% of ac-

curacy for reliability and objectivity, respectively, which is not topped by the vector database methodology.

In conclusion, both methods are effective choices for performing media bias detection when used with sufficiently capable models. However, for smaller models, the vector database approach appears particularly promising, as it enhances the model's performance compared to zero-shot scenarios. Furthermore, since our approach does not involve fine-tuning, we avoid the risk of domain overfitting, which can occur in fine-tuning-based models, as reported by Lin et al. [8]. We can say that the reported performance is truly generalizable.

5.2 Further Analysis

Axis Relationship In order to look deeper into the nature of the news framing axes, we put together a small dataset of 200 articles. These articles come in part from the MINT dataset, used to build the objectivity detection system, and from the dataset made available by Afonso and Rosas [1], used for reliability. All articles are classified using LLaMA 3.3 70B and the vector database methodology, since this allows us to get numerical axis values in $[-100, 100]$, as explained in Section 3.2. An interesting insight is the noticeable relationship between objectivity and reliability. This is expected since objective language in news articles should contribute to their reliability, while subjective content should contribute to the opposite. Figure 2 plots exactly this relationship, and with the addition of a simple linear regression, we observe an R^2 of 67.4%. When it comes to the political bias axis, we do not find any specific relation worth noting.

Indicator Analysis As explained in Section 4, the classification system is built using a different dataset for each axis. The size of each train set used for the construction of the respective vector databases is shown in Table 1. The logic behind the system makes it so that only one indicator is retrieved from each article, and there is no place for duplicates. This means that the final number of generated indicators by any LLM is determined by its ability to capture the nuance of each article and translate it into a small snippet of text that explains the provided label. Table 3 shows the number of unique indicators generated by each model for each axis. Looking at the values in this table, we can see a very interesting phenomenon: the more capable 70B model is very consistently generating a larger amount of distinct indicators than the 9B model, for exactly the same dataset. In fact, for reliability and objectivity in particular, the difference is massive, highlighting the ability of the larger model to capture small particularities of each article, while the smaller model tends to repeat indicators, taking too much inspiration from the examples provided in the prompt. Another interesting remark is that the difference between models is much smaller in the political bias axis, which can indicate that this axis promotes a more varied set of outputs, as the indicator generation becomes more of a subjective process than on the other two axes.

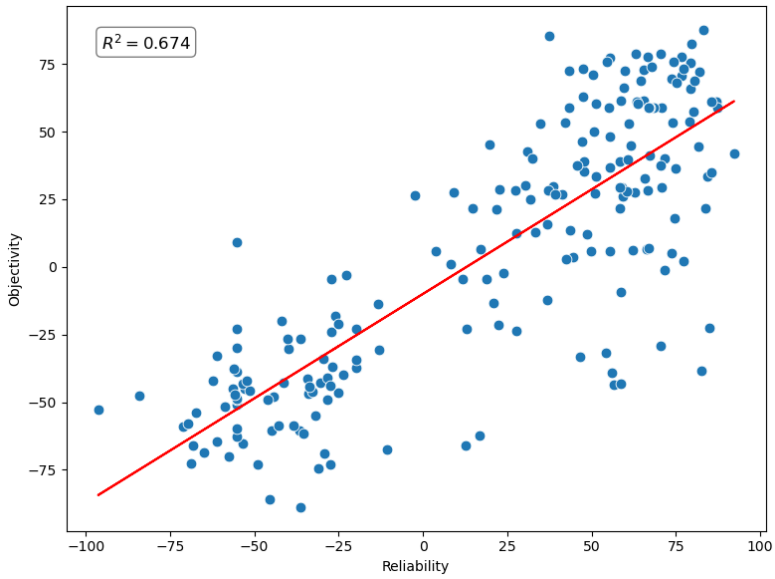

Fig. 2. Relation between Reliability and Objectivity

Table 3. Number of unique generated indicators in each vector database

	EuroLLM 9B	**LLaMA 3.3 70B**
Political Bias	2519	4630
Reliability	900	7505
Objectivity	522	9234

Indicator Vector Space Indicators are stored in their respective databases as vector embeddings to enable similarity search. These text embeddings, each with 768 dimensions, are generated for each indicator using a Sentence Transformer model. While this high dimensionality makes direct visualization impossible, there are effective techniques available to reduce dimensionality and enable meaningful plotting. Figure 3 shows the 2D representation of the indicators from each vector database, obtained by applying t-SNE for dimensionality reduction. The visualization yields very interesting observations that confirm previous conclusions. Firstly, the variety of outputs provided by the LLaMA model is clear by the increased cloud density of its respective vector databases, when compared to those of EuroLLM. Additionally, both plots relative to political bias are noticeably more muddled, with less segregation of classes, compared to the reliability and objectivity counterparts. This hints once more at the increased subjectiveness of the topic of political stance, which seems to have a complexity

not representable in two dimensions. The last and most remarkable observation is the very clear separation between classes in both reliability and objectivity indicators produced by LLaMA. The separation is so defined that a simple linear classifier, such as a logistic regression or a support vector machine, would achieve a high level of accuracy just by using these two dimensions, obtained from the t-SNE process, as features to predict the two classes.

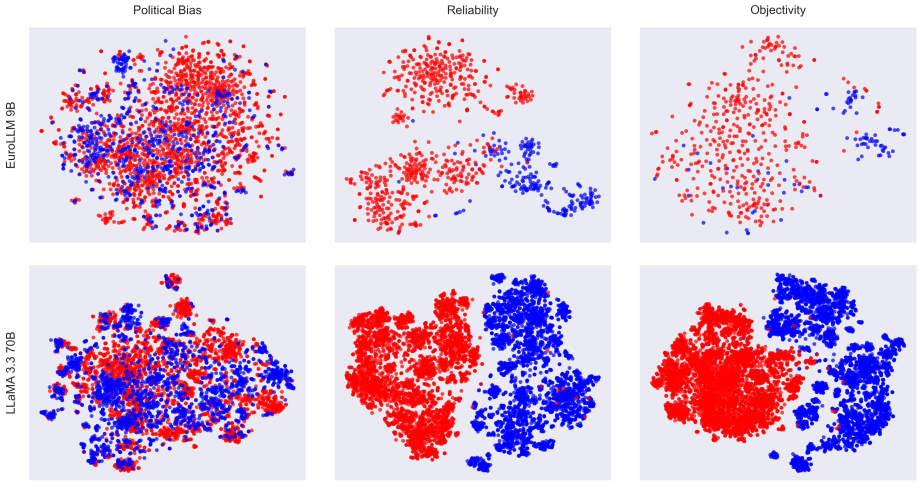

Fig. 3. Indicator embeddings in 2 dimensions. The blue color is assigned to the 'left', 'reliable', and 'subjective' classes, in the respective order of axes; red is assigned to the remaining classes.

6 Conclusion

In this paper, we examine media bias by framing news articles into three distinct axes: political bias, reliability, and objectivity. We propose two methodologies for building an automated classification system: one based on zero-shot prompting a large language model, and another that utilizes vector database techniques. Our study contributes to the broader field of media bias research by focusing on the European Portuguese context. Furthermore, the vector database approach we present not only enhances the performance of smaller models in this setting but also offers a high degree of explainability.

To conclude, we would set up future research by highlighting the still largely unexplored landscape covering the European Portuguese language. Although we deal with the lack of labeled data for political bias, a dataset with actual news articles expertly labeled would be of incredible usefulness. Additionally, there are still other axes to be explored from which news can be framed, such as

sensationalism and clickbait factor, for example. Finally, it would be interesting to add to this research the next level of model size with models such as GPT-4o or DeepSeek's offerings, and understand if the performance scales up further using similar methodologies.

References

1. Afonso, R., Rosas, J.: Development of a smartphone application and chrome extension to detect fake news in english and european portuguese. IEEE Latin America Transactions **22**(4), 294–303 (2024). https://doi.org/10.1109/TLA.2024.10472958, https://github.com/ro-afonso/fake-news-pt-eu
2. Alghamdi, J., Lin, Y., Luo, S.: Cross-domain fake news detection using a prompt-based approach. Future Internet **16**(8), 286 (2024), https://www.mdpi.com/1999-5903/16/8/286
3. Caled, D., Carvalho, P., Silva, M.J.: Mint - mainstream and independent news text corpus. In: Pinheiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., Pinto, H. (eds.) Computational Processing of the Portuguese Language. pp. 26–36. Springer International Publishing, Cham (2022)
4. Chen, W.F., Wachsmuth, H., Khatib, K.A., Stein, B.: Learning to flip the bias of news headlines. In: International Conference on Natural Language Generation (2018)
5. Dallmann, A., Lemmerich, F., Zoller, D., Hotho, A.: Media bias in german online newspapers (2015). https://doi.org/10.1145/2700171.2791057, https://doi.org/10.1145/2700171.2791057
6. Hamborg, F., Donnay, K., Gipp, B.: Automated identification of media bias in news articles: an interdisciplinary literature review. International Journal on Digital Libraries **20**(4), 391–415 (2019). https://doi.org/10.1007/s00799-018-0261-y, https://doi.org/10.1007/s00799-018-0261-y
7. Kiesel, J., Mestre, M., Shukla, R., Vincent, E., Adineh, P., Corney, D., Stein, B., Potthast, M.: SemEval-2019 task 4: Hyperpartisan news detection. In: May, J., Shutova, E., Herbelot, A., Zhu, X., Apidianaki, M., Mohammad, S.M. (eds.) Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 829–839. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019). https://doi.org/10.18653/v1/S19-2145, https://aclanthology.org/S19-2145/
8. Lin, L., Wang, L., Zhao, X., Li, J., Wong, K.F.: Indivec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. pp. 1038–1050. Findings of the Association for Computational Linguistics: EACL 2024, Association for Computational Linguistics (2024), https://aclanthology.org/2024.findings-eacl.70/
9. Martins, P.H., Fernandes, P., Alves, J., Guerreiro, N.M., Rei, R., Alves, D.M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J.G.C., Birch, A., Martins, A.F.T.: Eurollm: Multilingual language models for europe (September 01, 2024 2024). https://doi.org/10.48550/arXiv.2409.16235, https://ui.adsabs.harvard.edu/abs/2024arXiv240916235M
10. Moura, R., Sousa-Silva, R., Lopes Cardoso, H.: Automated fake news detection using computational forensic linguistics. In: Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20. pp. 788–800. Springer (2021)

11. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), http://arxiv.org/abs/1908.10084
12. Rodrigo-Ginés, F.J., Carrillo-de Albornoz, J., Plaza, L.: 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). https://doi.org/https://doi.org/10.1016/j.eswa.2023.121641, https://www.sciencedirect.com/science/article/pii/S0957417423021437
13. Spinde, T., Hinterreiter, S., Haak, F., Ruas, T., Giese, H., Meuschke, N., Gipp, B.: The media bias taxonomy: A systematic literature review on the forms and automated detection of media bias (December 01, 2023 2023). https://doi.org/10.48550/arXiv.2312.16148, https://ui.adsabs.harvard.edu/abs/2023arXiv231216148S
14. Spinde, T., Plank, M., Krieger, J.D., Ruas, T., Gipp, B., Aizawa, A.: Neural media bias detection using distant supervision with babe – bias annotations by experts (September 01, 2022 2022). https://doi.org/10.48550/arXiv.2209.14557, https://ui.adsabs.harvard.edu/abs/2022arXiv220914557S
15. Tavares, J.L.S.: Towards an Automated Media Chart: Framing News Articles with Natural Language Processing Techniques. Msc thesis, Universidade do Porto (2023)
16. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (February 01, 2023 2023). https://doi.org/10.48550/arXiv.2302.13971, https://ui.adsabs.harvard.edu/abs/2023arXiv230213971T
17. University of Michigan: News bias explored, https://websites.umich.edu/~newsbias/, [Accessed: 07-Feb-2025]
18. Wen, Z., Younes, R.: Chatgpt v.s. media bias: A comparative study of gpt-3.5 and fine-tuned language models (March 01, 2024 2024). https://doi.org/10.48550/arXiv.2403.20158, https://ui.adsabs.harvard.edu/abs/2024arXiv240320158W, 9 pages, 1 figure, published on Applied and Computational Engineering; ACE (2023) Vol. 21: 249-257.; doi:10.54254/2755-2721/21/20231153
19. Wolton, S.: Are biased media bad for democracy? American Journal of Political Science **63**(3), 548–562 (2019). https://doi.org/https://doi.org/10.1111/ajps.12424, https://doi.org/10.1111/ajps.12424

A Zero-shot User Prompts

Listing 1.1. Political Bias

```
You are a model trained to identify the political bias of
    opinion articles or news reports. Your task is to
    classify the provided article as having either a "right-
    wing" or "left-wing" bias.
A text with a right-wing bias tends to emphasize values such
    as free market, reduction of the role of the state,
    conservatism, nationalism, and criticism of progressive
    policies.
```

```
A text with a left-wing bias tends to value state
    intervention in the economy, social justice and workers'
    rights, wealth redistribution, and criticism of
    neoliberalism.

The output must always be a JSON in the following format:
{"class": "right"} or {"class": "left"}

The article below is clearly biased. Classify the bias of the
     following text and return only the correctly formatted
    response:
{ARTICLE BODY}
```

Listing 1.2. Reliability

```
You are a model trained to identify the reliability of
    opinion articles or news reports. Your task is to
    classify the provided article as either "reliable" or "
    unreliable", based on its language, structure, and
    transparency, without assessing the truthfulness of the
    content.
A reliable text tends to present verifiable sources, use
    neutral and objective language, have a clear and error-
    free structure, and avoid exaggeration and sensationalism
    .
An unreliable text tends to use sensationalist and
    emotionally charged language, contain grammatical or
    structural errors, and exaggerate or distort facts to
    support a point of view.

The output must always be a JSON in the following format:
{"class": "reliable"} or {"class": "unreliable"}

The article below is clearly classified as one of these
    options. Classify the reliability of the following text
    and return only the correctly formatted response:
{ARTICLE BODY}
```

B Vector Database User Prompts

Listing 1.3. Political Bias Indicators for 'Right' Class

```
A political bias indicator is here defined as a descriptive
    label that represents the presence of political bias in a
     text.
What you will detect is specifically right-wing political
    bias.
The output must always be a JSON in the following format:
```

```
{"bias": "description of the identified bias"}

Examples:
{"bias": "Refers to immigration as something negative for
    society."}
{"bias": "Mentions the advantages of low taxes for economic
    growth."}
{"bias": "Supports that the government should have little
    involvement in the economy and society."}
{"bias": "Devalues the welfare state and values individual
    merit."}
{"bias": "Criticizes the position of certain left-wing
    parties negatively."}

Extract only one new indicator of right-wing political bias
    from the following text and format the output exactly as
    in the examples:
{ARTICLE BODY}
```

Listing 1.4. Reliability Indicators for 'Reliable' Class

```
A reliability indicator is defined here as a descriptive
    label that represents the presence of elements in an
    article that may indicate greater credibility.
What you will detect are specifically linguistic, structural,
     or stylistic signs that reinforce the perceived
    reliability of the article.
The output should always be a JSON with the following format:
{"reliable": "description of the identified indicator"}

Examples:
{"reliable": "Presents verifiable sources for the information
     mentioned."}
{"reliable": "Uses neutral and objective language."}
{"reliable": "Avoids exaggerations or distortions when
    presenting the facts."}
{"reliable": "The text is well-structured and free of
    grammatical errors."}
{"reliable": "Provides multiple points of view on the topic
    addressed."}
{"reliable": "Presents statistics that support the argument
    ."}

Assuming the following text as reliable, extract only one
    indicator of reliability from the text and format the
    output exactly as in the examples:
{ARTICLE BODY}
```