

Generalizability of Media Frames: Corpus creation and analysis across countries

Agnese Daffara¹, Sourabh Dattawad¹, Sebastian Padó¹, Tanise Ceron²

¹Institute for Natural Language Processing, University of Stuttgart, Germany

²Bocconi University, Italy

{agnese.daffara,pado}@ims.uni-stuttgart.de

tanise.ceron@unibocconi.it

Abstract

Frames capture aspects of an issue that are emphasized in a debate by interlocutors and can help us understand how political language conveys different perspectives and ultimately shapes people’s opinions. The Media Frames Corpus (MFC) is the most commonly used framework with categories and detailed guidelines for operationalizing frames. It is, however, focused on a few salient U.S. news issues, making it unclear how well these frames can capture news issues in other cultural contexts. To explore this, we introduce FrameNews-PT, a dataset of Brazilian Portuguese news articles covering political and economic news and annotate it within the MFC framework. Through several annotation rounds, we evaluate the extent to which MFC frames generalize to the Brazilian debate issues. We further evaluate how fine-tuned and zero-shot models perform on out-of-domain data. Results show that the 15 MFC frames remain broadly applicable with minor revisions of the guidelines. However, some MFC frames are rarely used, and novel news issues are analyzed using general ‘fall-back’ frames. We conclude that cross-cultural frame use requires careful consideration.

1 Introduction

Frames are interpretation schemes used to organize reality by isolating and highlighting salient aspects of it (Otmakhova et al., 2024). They combine three levels: cognitive (concerning the mental representations of the world); semantic (concerning the linguistic structures involved); and communicative (their usage and impact on the audience).

Frames are finding increasing application in NLP as a framework to analyze social media and news coverage. In particular the communicative dimension of frames can be seen as a bridge between belief and action (Snow and Benford, 2005). Following Entman (1993), frames have four objectives: i) promote a particular definition of a problem;

ii) promote an interpretation of it; iii) propose a moral evaluation; and iv) recommend a treatment or a solution. This makes them a valuable tool for identifying, measuring, and quantifying particular worldviews and communication strategies (Card et al., 2015; Mendelsohn et al., 2021).

Frames are primarily a methodological framework, and not a fixed set of categories. However, a widely used set is based on the Media Frames Corpus (MFC), 15 frames (see Table 1) with annotation guidelines, developed to cover a limited number of issues concerning the U.S. context, e.g., gun control and climate change (Card et al., 2015).

As in other NLP tasks, however, *generalization* is an issue: coarse-grained frames risk failing to capture important aspects of a debate, notably ideological bias, while fine-grained frames may not generalize across debates. For example, Mendelsohn et al. (2021) demonstrate that in news about migration, the generic frame *cultural identity* is correlated with liberal texts; introducing two related sub-frames *hero: cultural diversity* and *threat: national cohesion* leads to a clear separation by political ideology – liberal in the former and conservative the latter case – but these subframes are presumably irrelevant for other debates. Similar questions arise in cross-lingual contexts: the MFC frames were applied by SemEval 2023 Task 3 to 9 languages (Piskorski et al., 2023b), but the applicability of the frames was not assessed, nor was inter-annotator agreement reported.

Indeed, full cross-linguistic generalization of frames is far from clear, given previous work on the cross-lingual generalization of semantic inventories (Peters et al., 1998; Reddy et al., 2017; Gilardi and Baker, 2018). For full generalization, frames would show high *agreement* – their definitions and annotation guidelines would allow each markable in a new language to be assigned to a unique frame – and *completeness* – for each instance in the new language, there would be a suitable frame. However,

we do not have empirical evidence so far.

Thus, we set out to answer these questions:

1. How high are agreement and completeness for MFC frames when annotating a representative sample of news articles despite having been originally developed on a small number of issues?
2. How high are agreement and completeness of the MFC frames when applied to news reporting from other countries?
3. Even if the frames generalize well, is it possible to transfer computational models of frame prediction between languages?

Our study proceeds by creating FrameNews-PT, a corpus of 300 news articles of ‘hard news’ (Reinmann et al., 2012) in Brazilian Portuguese.¹ FrameNews-PT is based on the News Portal Recommendation dataset (Lucas et al., 2023). It contains Brazilian news about, e.g., new regulations, taxes, and finance. We manually annotate the articles with MFC frames using a *perspectivist* approach (Cabitza et al., 2023). We carry out several rounds of annotation to a) adapt the MFC guidelines while maintaining their level of detail, b) value the individual perspectives of annotators, and c) examine generalizability and overlap of frames (RQ1+2).

After corpus creation, we conduct modeling experiments (RQ3). We use two transformer-based models fine-tuned on the MFC and 9 chat-instructed models in a zero-shot setting, and analyze how well models perform out of domain on FrameNews-PT in comparison with the MFC.

Our results are generally promising: (Most) MFC frames are annotated with high agreement in Brazilian hard news, with little confusion among them. However, some frames appear rarely, being bound to U.S.-specific issues. Also, completeness is not perfect: Brazilian issues outside the scope of the original MFC are often annotated with relatively general frames as a fallback (e.g., *Economic*), leading to some loss of information. On the modeling side, we find that zero-shot models perform better at frame prediction on the Brazilian data than transferred English classifiers, indicating that the data distributions are sufficiently dissimilar. Overall, use of MFC frames in new languages and issues calls for a reflected process.

¹The corpus is available for the community at this link: <https://github.com/tceron/FrameNews-PT>

2 Related Work

Previous work in NLP has investigated how framing political language influences society and public opinion (Lecheler et al., 2015), and how it shapes the spread of information (Gilardi et al., 2021) across communication contexts, such as agenda-setting campaigns (Tsur et al., 2015; Field et al., 2018) and news media (Card et al., 2015). Frames can help analyze argumentation, and frame detection can be seen as argument assessment (Lauscher et al., 2022). Given their connection to political ideologies (Alashri et al., 2015), frames are also useful for analyzing and diversifying perspectives, e.g. in news recommendation (Mulder et al., 2021; Dattawad et al., 2025).

The task of frame detection has often been approached through topic modeling with unsupervised methods (Nguyen et al., 2013; Roberts et al., 2014; Tsur et al., 2015; Ajour et al., 2019), which, however, tend to be highly corpus-specific. Other studies employ supervised classification techniques, including logistic regression (Card et al., 2016), neural networks (Naderi and Hirst, 2017; Liu et al., 2019), and fine-tuning of pre-trained models (Kwak et al., 2020), all of which require robust gold standard annotations.

The Media Framing Corpus (MFC) (Card et al., 2016), building on the foundational work of Boydston et al. (2014), introduced a novel annotation framework that has become a reference for frame identification (cf. Section 3). Several studies either use parts of the MFC or applied its guidelines to related datasets, e.g., Kwak et al. (2020) conduct a systematic analysis of the 15 frames in New York Times articles across 17 years.

Khanehazar et al. (2019) are the first to test the generalizability of MFC frames to other contexts. They apply the framework to Australian parliamentary speeches on same-sex marriage and migration, and fine-tuned three BERT-based classifiers. Their findings show a drop in accuracy when the models are applied across different contexts. It is unclear, however, if this is due to the failure of the frames to *generalize*, or only to a shift in frame *distribution*.

Other corpora and labeling schemes have been proposed, though they often focus on specific debate issues. Examples include the Ballistic Missile Defense corpus (Morstatter et al., 2018), the Gun Violence Corpus (Liu et al., 2019), and the migration corpus by Mendelsohn et al. (2021).²

²For a comprehensive overview of framing detection and

SemEval-2023 Task 3 introduced a sub-task on framing detection in online news within a multilingual setting (Piskorski et al., 2023b). The dataset spans multiple topics and languages. It is annotated with an adaptation of the MFC framework (Piskorski et al., 2023a) and the proposed models, based on Transformer architectures, achieved strong performance. The focus of the study was on modeling, however, not on assessing frame generalizability.

3 Data

The Policy Frames Codebook A key contribution to the study of frames is the Policy Frames Codebook by Boydstun et al. (2014), which introduced 15 framing dimensions applied to three debate issues: immigration, tobacco, and same-sex marriage. The guidelines for frame annotation in this codebook have been dynamically updated over time (Boydstun and Gross, 2014). This work led to the creation of the Media Frames Corpus (Card et al., 2015), later expanded to include three additional issues: gun control, death penalty, and climate (MFC v4.0). The MFC was explicitly developed for three debate issues and the U.S. context, which is a potential limitation. Additionally, as noted by Ali and Hassan (2022), it may not distinguish sufficiently between issues and frames by introducing overly broad categories, such as *Economic*.

For our study, we adopt the 15 framing dimensions as labeled in Card et al. (2015), cf. full list in Table 1. Our guidelines are a version of the Policy Frames Codebook, shortened from 45 to 9 pages to preserve the original information while making it more accessible. We remove specific edge cases and marginal notes, but our annotators are encouraged to consult the original guidelines in cases of doubt. Our version of the guidelines, in English, includes a description for each frame and the cues to recognize it (cf. Appendix A.6). In addition, they include: i) a contextualization of the project, presenting the tool used for annotation (Google Sheets with drop-down menus), and ii) for each frame, an example drawn from MFC headlines on migration where annotators agreed on the frame. In contrast to the version used in SemEval 2023 (Piskorski et al., 2023a), our guidelines retain all relevant content from the original codebook. The guidelines are made available in the repository.

analysis, see Ali and Hassan (2022) and Otmakhova et al. (2024), offering a multidisciplinary perspective on the field.

Frame	Example
Capacity and Resources	<i>Immigration debate: Illegals take jobs from Americans</i>
Crime and Punishment	<i>Two charged in deaths of illegal immigrants in truck</i>
Cultural Identity	<i>Ethnic shift: Immigration—an Irish enclave learns a new language; Mexican immigrants boost a growing Latino population</i>
Economic	<i>Society makes no-interest loans to New York’s immigrants</i>
Fairness and Equality	<i>Strict immigration law unfairly targets Hispanics</i>
External Regulation and Reputation	<i>‘International village’ gets hostile reception</i>
Health and Safety	<i>Colombian drug violence leads to exodus</i>
Legality, Constitutionality and Jurisprudence	<i>House approves bill to abolish INS; The Senate will begin work next week on its own measure dealing with the immigration agency</i>
Morality	<i>County’s undocumented workers say they aren’t here to ‘steal’</i>
Other	<i>U.S. under pressure to carry bigger load</i>
Policy Prescription and Evaluation	<i>President Donald Trump stalls on promise to eliminate J-1 visa program</i>
Political	<i>Following Trump voter fraud allegations, claim that 5.7 million non-citizens voted is wrong</i>
Public Opinion	<i>Immigration: Political evangelicals feel push to take sides</i>
Quality of Life	<i>Big money, cheap labor</i>
Security and Defense	<i>Decision on refugees overdue; U.S. officials must loosen immigration restrictions</i>

Table 1: The 15 frames and their examples extracted from the MFC headlines on migration.

Data collection Our dataset consists of 300 news articles randomly extracted from the News Portal Recommendation Corpus (NPR), a collection of 148,099 news articles in Portuguese from a Brazilian news portal called G1 with user click history (Lucas et al., 2023). The inclusion of click data makes it suitable for potential future applications in multi-perspective, frame-based news recommendation. The dataset was selected not only for its relevance to recommendation scenarios, but also because it covers a broad range of issues from a cultural and linguistic context different from the U.S. and the Global North generally. This allows us to investigate our research questions on the generalizability of framing annotations and model performance across news issues.

We select articles belonging to the economics and politics sections of the newspaper. The articles



		 Annotator 1	 Annotator 2
Heading	Ministro do Trabalho quer mesa de negociação entre trabalhadores e controladores das Americanas	Economic	Legality
Body	O ministro do Trabalho, Luiz Marinho, afirmou nesta terça-feira (31) que quer estabelecer uma mesa redonda entre os trabalhadores das lojas Americanas e os controladores da empresa. A declaração acontece um dia após as maiores centrais sindicais do país terem divulgado comunicado conjunto defendendo a participação direta do Ministério do Trabalho no processo de recuperação judicial da Americanas, ocorrido após a revelação no início do mês de pelo menos R\$ 20 bilhões em problemas contábeis. As centrais querem preservar empregos na varejista, que têm mais de 44 mil trabalhadores, além de "centenas de milhares" na rede de fornecedores. "Vamos providenciar mesa redonda das Americanas com representantes dos trabalhadores para buscar compreender o que aconteceu. Se houve fraude, vamos acionar os órgãos responsáveis pelas fraudes. Não posso afirmar que houve fraude, mas que tem cheiro tem cheiro", declarou o ministro. [...]	Economic Crime Morality	Economic Crime Fairness
Primary (global)		Morality	Legality

Figure 1: Example of an annotation in FramesNews-PT (full article and translation in English are provided in Appendix A.2). Each annotator assigns, in this order: (i) one frame to the article **heading**, (ii) multiple frames to the article **body**, and (iii) a **primary** frame to the article as a whole.

in the NPR dataset often belong to more than one category, so they also span across other news categories as well. In this way, we ensure that articles i) consist of hard news, typically including topics such as foreign and domestic politics, economy and finance (Reinemann et al., 2012), and ii) cover a broad domain, spanning multiple debate issues. A topic modeling analysis with BerTopic (Groendorst, 2022) on NPR and FrameNews-PT shows that our selection of topics is representative of the whole NPR corpus (see Appendix A.3). To comply with model input length limitations, we limit our selection to articles shorter than 300 words and remove boilerplate (banners, ads).

Data annotation We hired two annotators, both Master’s students: one Brazilian and one Portuguese. We expect their cultural backgrounds to influence the interpretation and application of frames during the task. In order to acknowledge individual perspectives and subjectivity, we keep the annotations disaggregated and we evaluate the models on each annotator, adopting a perspectivist approach (Cabitza et al., 2023).

Like in the MFC, our annotators label in the following order: (i) the heading frame (single label), (ii) the body frames (multiple labels), and (iii) the primary frame, i.e., the most prominent frame throughout the article (single label). Following the MFC guidelines, the primary frame must be selected from among the annotated frames; in case of doubt, it defaults to the heading frame. Unlike the MFC, however, our annotations are performed at

the article level. While span-level annotation may be preferable for certain applications, article-level annotation can still supports various tasks, such as news recommendation with frame diversification. Figure 1 presents an example annotation.

In the MFC, the annotators also decide whether an article is relevant and do not carry out annotations on irrelevant articles³. We skip this annotation phase, and therefore annotate all the articles, for two main reasons: i) models are presented with both relevant and potentially irrelevant articles for classification, and we want the annotation process to align with the modeling setup; and ii) our dataset includes only hard news and has already been cleaned. We aggregate irrelevant articles and those that do not match any frame’s guidelines under a single label, namely *Other*. This frequently includes news on a personal rather than societal level, such as advice on personal finance, new commercial products, or celebrity news. An example of this category is the following headline, which describes a private decision by President Lula that does not match any frame dimension:

- (1) Lula decide não se mudar para a Granja do Torto antes da posse.
Lula decides not to move to Granja do Torto before the inauguration.

³A relevant article: i) concerns the policy issue in question, ii) takes place in or explicitly regards the United States, iii) has more than four lines, and iv) is a proper article, not a correction of a previous error or an obituary.

The annotation process was organized into four weekly rounds, with an increasing number of articles, based on the assumption that annotators become more efficient and familiar with the task over time: the first two rounds comprised 50 articles each, and the last two rounds contained 100 articles each. After each round, the annotators participated in a discussion session to address ambiguous cases and refined the application of the guidelines. Inter-annotator agreement (IAA) is measured using Krippendorff’s α^4 after each round. We hypothesize that i) the IAA increases at each round because of the discussions, and ii) that the score before and after discussion becomes closer at each round.

Annotators’ perspectives In the discussion rounds, we observe the influence of the annotators’ cultural backgrounds on the usage of frames. The Portuguese annotator struggled to understand certain contexts due to her different cultural perspective and she would sometimes need to look up acronyms, such as STF (*Supremo Tribunal Federal*). The Brazilian annotator, familiar with the local context, often had a deeper understanding of the reasons and potential consequences of news articles. Consider the following heading sentence, referring to scholarships for students:

- (2) Reajuste de bolsas da Capes e do CNPq deve ser anunciado ainda em janeiro, diz Camilo Santana.
Adjustment of Capes and CNPq scholarships should be announced in January, says Camilo Santana.

The Brazilian annotator, aware that access to scholarships is a major issue in Brazil, considered the introduction of such a policy to be significant and labeled the article as *Policy Prescription and Evaluation*. By contrast, the Portuguese annotator, having associated the term *bolsas* with the stock exchange (*bolsa de valores*), labeled it as *Economic*. These differences show that a perspectivist approach is helpful in annotation and modeling. We keep the annotations disaggregated in the evaluation (see Section 5.1 and Table 4 for more details).

4 Analysis of Frame Generalization

We can now investigate the generalization of the MFC frames to the Brazilian context.

⁴Krippendorff’s α was adopted following the MFC. It calculates agreement considering the whole pool of labels rather than individual frequency of use (Card et al., 2015).

Inter-annotator agreement As Figure 2 shows, the agreement score on the primary frame kept improving generally over the four rounds, with the exception of the third round, which contained more complex cases. The discussion was useful in distinguishing difficult frames. After the final revision, the global agreement after discussion on the dataset was $\alpha=0.78$. It is substantially higher than the one reported on the MFC by (Card et al., 2015), which never exceeds 0.70. This can be attributed to four main reasons: i) in the MFC, new annotators with no previous knowledge were introduced in each round, potentially lowering the overall score; ii) discussing the guidelines with the same person, as we did, could in the worst case lead to overfitting (we alleviate the effects in modeling by keeping annotations disaggregated); iii) annotators usually fell back to general frames rather than specific ones⁵; iv) they used *Other* to annotate articles about personal stories that did not match other frames in the guidelines, consequently raising the agreement.

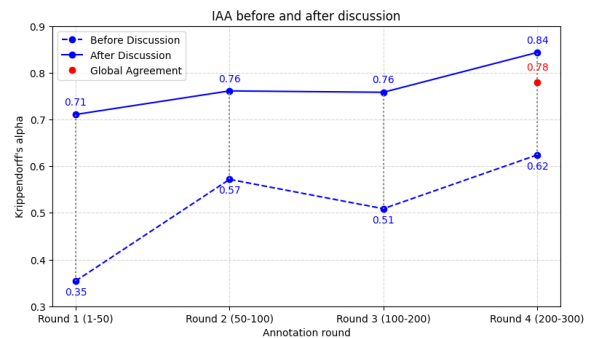


Figure 2: Inter-annotator agreement on the primary frame over four rounds, before and after discussion.

IAA on single frames We also compute agreement on single frames (Figure 3). The results confirm highest agreement for *Other*, arguably due to its particular function in our annotations (see Section 3). Other frames with high scores include *External Regulation and Reputation*, suggesting that it is particularly well defined, *Political*, and *Economic*, which is not surprising given that these are the dominant categories in the dataset. On the other hand, the frames with the lowest agreement are *Capacity and Resources* and *Quality of Life*.

⁵We use ‘general frames’ to refer to high-frequency frames — such as *Political*, *Economic*, and *Legality, Constitutionality and Jurisprudence* — that frequently overshadow other frames during annotation. By contrast, ‘specific frames’ denote low-frequency frames that are more easily obscured by broader categories and potentially tied to specific issues, for example *Capacity and Resources* (see Figures 4, 5, and 6).

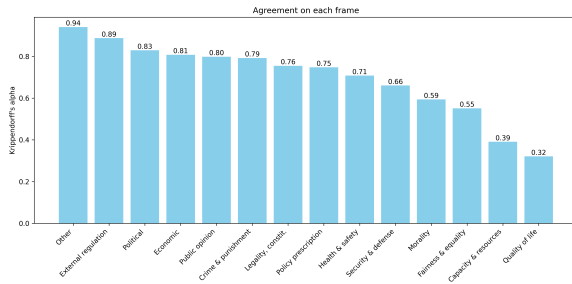


Figure 3: Agreement on each frame (primary).

Frame overlap Figure 4 shows the overlap between the two annotators on primary frames, indicating which frame pairs struggle with high agreement. These include: i) *Crime and Punishment* vs. *Legality, Constitutionality and Jurisprudence* ii) *Fairness and Equality* vs. *Quality of Life*, iii) *Policy Prescription and Evaluation* vs. *Economic*.

Quality of Life and *Fairness and Equality* frequently co-occur when discussing social issues, as in the following:

- (3) Ministério diz que novo Bolsa Família terá R\$ 18 bilhões para crianças de até 6 anos. *The Ministry says that the new Bolsa Família will have R\$ 18 billion for children up to 6 years old.*

Crime and Punishment and *Legality, Constitutionality and Jurisprudence* overlap when talking about legal measures to punish someone, while *Policy Prescription* and *Economic* occur together when speaking of policies that affect economy.

We would argue that this overlap is not due to failure of a frame to generalize. Rather, it is a case of a complex situation which is described as having an import on multiple aspects. Such multi-class cases are intrinsic to framing.

Completeness: Frame frequency While a low frequency does not imply that a frame does not generalize, high frequencies for *Other* or general frames may indicate lack of completeness. We therefore compare frame frequencies between MFC and FrameNews-PT (Figures 5 and 6).

Some frames, like *Economic*, *Fairness and Equality*, and *External Regulation and Reputation* appeared more frequently in FrameNews-PT, suggesting they transfer well to Brazilian debate coverage. In contrast, six frames were used fewer than 20 times: *Health and Safety*, *Security and Defense*, *Quality of Life*, *Capacity and Resources*, *Morality*,

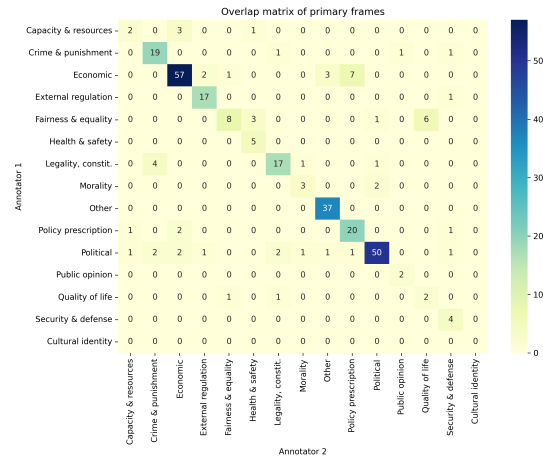


Figure 4: Primary frames overlap between annotators.

and *Public Opinion*. We hypothesize that these frames are not used frequently because they are tied to specific issues in the MFC; for example, *Capacity and Resources* is strongly associated to climate change. *Cultural Identity* was never used by our annotators, which appears due to its specific use in the U.S. issues migration and gun control.

The proportion of *Other* in our corpus is almost the same as in the MFC (13%), although it is used more broadly in our annotations (see Section 3). This means that, even though we sample from a much broader range of issues, the 15 frames still suffice to classify the news. However, annotators in our case tended to fall back to more general frames, indicating a lack of completeness.

Annotator survey To validate our interpretations from above, we asked the annotators to fill a survey about the annotation process, and discussed the results in a meeting. The feedback from the survey confirmed that annotators found our summarized guidelines easier to read and understand than the original one by [Boydston et al. \(2014\)](#). They identified two main limitations. First, some frame descriptions are overly specific to the U.S. context. For example, *Security and Defense* included immigration issues at the U.S. border (cf. Appendix A.6), a topic not prominent in Brazil. Second, some frames are not described sufficiently distinctly by the guidelines, leading to overlap.

Regarding the frequency of frames, the annotators attributed the low usage of certain frames to the topics covered, confirming our hypothesis that some frames are specific to particular issues. However, when asked whether they would add or eliminate any category, they agreed that the same

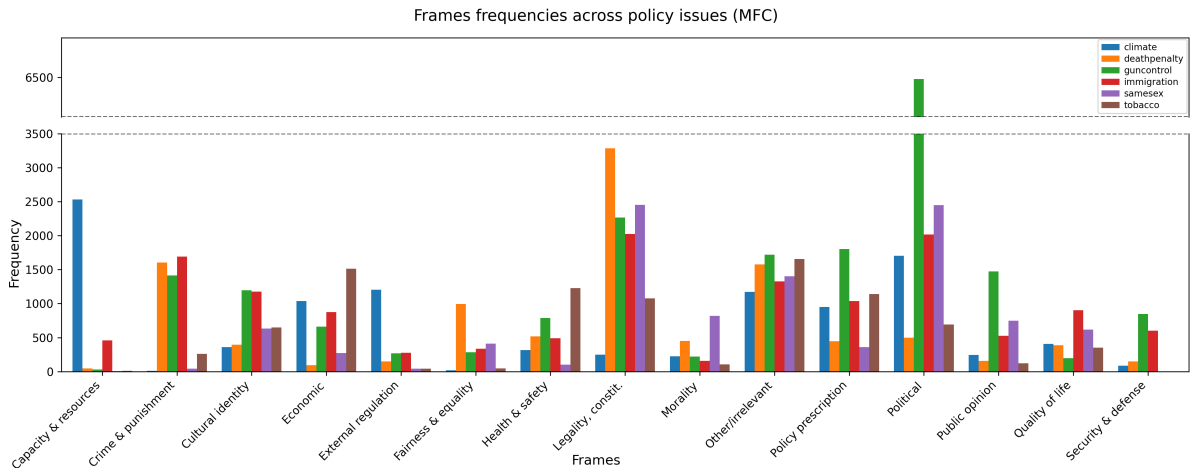


Figure 5: Frames frequencies across policy issues in the MFC.

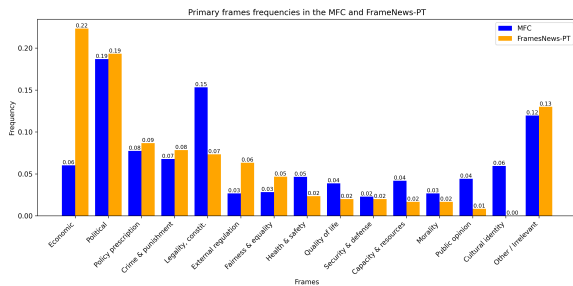


Figure 6: Primary frames frequency in the MFC compared to FrameNews-PT (normalized).

tagset could be used for future tasks, provided that some adjustments are made, making the frames more context-aware and contrastive.

The articles tagged as *Other* mainly focused on private issues rather than on societal-level concerns, which is why they did not match any frame. While these could be covered by new frames (e.g., *Personal Finance*, *Advertisement*), adding such fine-grained categories would be misaligned with the goals of framing analysis for hard news.

Conclusions In sum, we find that the 15 frames generalize well enough that they can be employed in future comparable projects, with the following revisions to the guidelines: i) replace U.S.-specific examples with locally relevant counterparts (e.g., substitute *Medicare* with *Bolsa Família*); ii) clarify contrasts between overlapping frames based on the local context (e.g., *Quality of Life* vs. *Health and Safety*); iii) better define the scope of *Other*. The addition of culturally specific frames does not appear necessary, as the existing inventory already provides adequate coverage; rather, adapting the

guidelines ensures better fit to the local context.

5 Frame Prediction on FrameNews-PT

Even very good frame generalizability does not imply that frame prediction models transfer well across languages, and at least for some frames we see substantial shifts in frequency (cf. Figure 6). Therefore, a) we run two classifiers trained on the MFC corpus, testing them both on MFC and on FrameNews-PT; b), we classify FrameNews-PT in a zero-shot setting with 8 chat-instructed LLMs ranging from 1B to 12B parameters as well as ChatGPT-4o (exact parameter size unknown). All chosen models are multilingual since FrameNews-PT is in Portuguese.

5.1 Experimental Setup

We split the Media Frames Corpus (MFC) into 70% training (21,751 examples), 10% validation (3,107), and 20% test (6,214). The test set is held out and only used for final evaluation. All splits are fixed across runs to ensure consistency. The FrameNews-PT corpus is always used as a test set, in order to understand how classifiers perform on a novel set of news issues from a different country.

Supervised classifiers We fine-tune two pre-trained multilingual language models: XLM-RoBERTa-base (Conneau et al., 2019) and Multilingual-E5-base (Wang et al., 2024) for single-label classification over 15 frames. Inputs are tokenized using the model tokenizers with a maximum sequence length of 512, and models are trained using cross-entropy loss. Each model is trained for up to 10 epochs with early stopping based on vali-

Model		Dataset	Accuracy	Global F ₁	F ₁ (ann 1)	F ₁ (ann 2)	Cohen’s κ
Majority class baseline		MFC	0.21	-	-	-	-
		FrameNews-PT	0.23	-	-	-	-
Fine-tuned	XLM-RoBERTa	MFC	0.67 ± 0.00	0.68 ± 0.00	-	-	-
		FrameNews-PT	0.53 ± 0.00	0.48 ± 0.01	0.49 ± 0.00	0.47 ± 0.01	-
Fine-tuned	Multilingual-E5	MFC	0.66 ± 0.00	0.67 ± 0.01	-	-	-
		FrameNews-PT	0.56 ± 0.01	0.50 ± 0.01	0.51 ± 0.01	0.49 ± 0.01	-
Zero-shot	Qwen2.5-7B-Instruct	MFC	0.40 ± 0.10	0.38 ± 0.01	-	-	0.50 ± 0.00
		FrameNews-PT	0.53 ± 0.02	0.44 ± 0.02	0.45 ± 0.03	0.42 ± 0.03	0.49 ± 0.18
Zero-shot	gpt-4o-2024-08-06	MFC	0.46 ± 0.00	0.46 ± 0.00	-	-	0.69 ± 0.00
		FrameNews-PT	0.59 ± 0.01	0.50 ± 0.03	0.51 ± 0.04	0.49 ± 0.03	0.75 ± 0.06

Table 2: Results of the models on the two datasets. For XLM-RoBERTa and Multi-E5, standard deviations are across random seeds. For Qwen and ChatGPT, they are across templates. Accuracy considers the output as correct if it matches at least one annotation. Global F₁ is calculated on the Gold Standard for the MFC, while it is averaged on both annotators for FrameNews-PT. Cohen’s κ measures reliability across prompt templates. Best results for encoder-based and generative models are **in bold**.

dition loss, using a patience of 3. We use a batch size of 64 for both training and evaluation. The learning rate is 5e-5. All experiments are repeated with five different random seeds.

Zero-shot classifiers For zero-shot classification, we prompt multilingual chat-instructed models from the Qwen (Qwen Team et al., 2024), Google Gemma (Kamath et al., 2025), Llama (Grattafiori et al., 2024), and GPT (Achiam et al., 2023) families. Our prompts include the article to be labeled, the shortened annotation guidelines, and a task description. We evaluate the models on 3 prompt templates (see examples in A.5) and temperature 0.

Evaluation metrics We evaluate the performance of the models on accuracy and F1-score. Accuracy is measured taking the perspectivist approach into account. That is, we consider the answer of the model correct when it matches the primary labels of one of the annotators. The F1-score for the MFC corpus is based on the gold labels while we report the F1-scores for each annotator in FrameNews-PT separately. We also calculate Cohen’s κ for checking the reliability of the models answers across the evaluated prompt templates.

5.2 Results

We report results for the two supervised models and the two best-performing zero-shot models in Table 2; for the remaining results see the Appendix.

The best model among the supervised learning classifiers for MFC is RoBERTa-XLM (acc=67%) while for FrameNews-PT it is Multi-E5 (acc=56%). RoBERTa-XLM clearly outperforms the zero-shot

classifiers on the MFC. On the other hand, ChatGPT-4o is the best model among the zero-shot classifiers for MFC (acc=46%) and FrameNews-PT (acc=59%) – not surprisingly, given its parameter count. Note that when comparing ChatGPT-4o and RoBERTa-XLM, ChatGPT-4o yields the best result on FrameNews-PT, while RoBERTa-XLM outperforms it by 8 points accuracy on the MFC. The F1-score results follow similar patterns.

These findings suggest a trade-off: supervised models trained on annotated data achieve strong in-domain performance but are less effective in transfer to new corpora, whereas large zero-shot models offer more robust cross-domain performance but fall short in accuracy when compared to supervised models in an in-domain setup.

As for reliability across prompt templates, larger models are more reliable (Ceron et al., 2024): ChatGPT-4o achieves 0.75 and 0.69 Cohen’s κ on the MFC and FrameNews-PT respectively, whereas Qwen2.5-7B reaches only 0.46 and 0.49.

Figure 7 shows model performance and inter-annotator agreement across frames. The agreement between human annotators is always higher than model-human agreement, except for *Quality of Life*. The models struggle the most with the category *Other* – for the supervised models, this is presumably because this category is very infrequent in the MFC; for the zero-shot models, we hypothesize that this is due to its status as a ‘catch-all’ category which is hard to learn.

While supervised models perform relatively well across all frames on the MFC (acc=66%-67%), they struggle to generalize to FrameNews-PT (acc=53%-

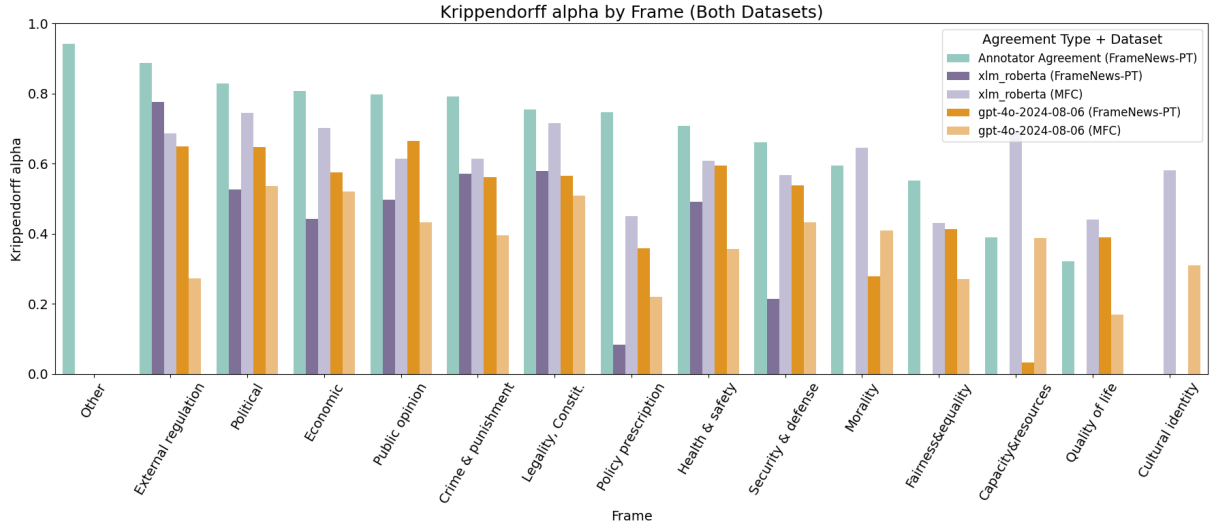


Figure 7: Agreement between models’ predictions and annotator 1 on FrameNews–PT and MFC, and agreement between annotators on FrameNews–PT.

56%), particularly for *Policy Prescription and Evaluation*, *Morality*, *Fairness and Equality*, *Capacity and Resources*, and *Quality of Life*. The latter three also have the lowest inter-annotator agreement. Our best zero-shot model, GPT-4o, performs worst for *Policy Prescription and Evaluation* and *Quality of Life* on the MFC, arguably due to blurred boundaries of these frames (cf. Section 4). On FrameNews–PT, it struggles the most with *Morality* and *Capacity and Resources*, likely due to the shift in topics.

These observations give rise to a final question: is the ‘limiting factor’ in the models’ performance on FrameNews–PT data quality on the MFC, on FrameNews–PT, or the shift in distribution? To understand this, we ran a linear regression to predict GPT-4o’s accuracy on the MFC at the frame level (cf. Figure 7). Using three predictors: IAA on MFC, IAA on FrameNews–PT, the regression explains 53% of the variance, with only the IAA on the MFC being a significant predictor at $p < 0.05$. This is surprising, since GPT-4o is not even trained on the MFC. Our take on this result is that the MFC inventory includes frames with vague or otherwise deficient definitions, which leads to problems both in terms of annotation (low IAA) and in terms of prediction (low performance).

6 Conclusion

This study introduces FrameNews–PT, the first dataset of Brazilian Portuguese news with both users’ click history (sampled from the NPR corpus) and annotation with the 15 Media Frames Corpus

frames. It offers a valuable resource for both the NLP and recommender systems communities. By including frame annotations, it also enables new research directions, such as investigating the relationship between frames and topics.

Our analysis focuses on the generalizability of the MFC frames to the Brazilian context. We find generally high agreement among annotators, which we interpret as a promising sign; however, the MFC frames are not complete. In the absence of specific frames, annotators tend to fall back to general ones, with reduced specificity as consequence. This trend is largely due to the original frames and annotation guidelines being developed for the U.S. context on a limited number of policy issues.

Our modeling experiments show that zero-shot generative models achieve robust results on our dataset. However, the very good in-domain results for supervised MFC models indicate that annotated data is still pivotal to reach strong performance; both model types suffer from shortcomings in frame definition and data creation.

We conclude that the 15 MFC frames remain broadly applicable and continue to capture essential dimensions of political discourse, even across countries and languages. To improve generalization, we recommend refining the annotation guidelines to offer clearer distinctions between overlapping categories and incorporating locally relevant examples (e.g., *Bolsa Família* instead of *Medicare*). Such refinements would enhance the framework’s adaptability to diverse cultural contexts and evolving policy debates.

Limitations

The main limitation of our study is the size of our dataset. We had to limit the number of articles to 300, which is a small sample. The main reason is the expense of running this type of study. Annotating frames in news articles is costly because the guidelines are extensive to learn and the articles are long. Annotators also need time to identify all body frames and then the primary frame for each article. For the same reason, we hired only two annotators and chose articles from a single Brazilian news portal, which may limit the generalizability of our results. In modeling, we focused on the primary frame, assuming it represents the article, but considering multiple frames could yield more accurate analysis. Moreover, including few-shot examples could help assess whether they improve performance.

Ethics Statement

The data used in this study is publicly available and was collected from the News Portal Recommendation Corpus (NPR) (Lucas et al., 2023), which contains news articles from Brazilian outlets. The annotation was conducted by two Master’s students who were compensated for their work. The study follows ethical guidelines for research involving human participants, ensuring that annotators were informed about the purpose of the study and their rights. The research does not involve sensitive or personal data, and all data is anonymized.

Acknowledgements

We gratefully acknowledge funding from the German Federal Ministry of Research, Technology and Space (BMFT) under the grant 01IS23072 for the Software Campus project MULTIVIEW.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yamen Ajjour, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. Modeling frames in argumentation. In *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 2922–2932.

Saud Alashri, Sultan Alzahrani, Lenka Bustikova, David Siroky, and Hasan Davulcu. 2015. What animates political debates? analyzing ideological perspectives in online debates between opposing parties. In *Proceedings of the ASE/IEEE International Conference on Social Computing (SocialCom-15)*, Stanford, CA. Academy of Science and Engineering.

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348.

Amber E Boydston, Dallas Card, Justin H Gross, Philip Resnik, and Noah A Smith. 2014. Tracking the development of media frames within and across policy issues. In *APSA 2014 annual meeting paper*.

Amber E Boydston and Justin Gross. 2014. Policy frames codebook. *Work in progress*.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Dallas Card, Amber Boydston, Justin H Gross, Philip Resnik, and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444.

Dallas Card, Justin Gross, Amber Boydston, and Noah A. Smith. 2016. *Analyzing framing through the casts of characters in the news*. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. *Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in llms*. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

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

Sourabh Dattawad, Agnese Daffara, and Tanise Ceron. 2025. Leveraging media frames to improve normative diversity in news recommendations. *arXiv preprint arXiv:2509.02266*.

Robert M Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of communication*, 43(4):51–58.

- Anjalie Field, Doron Kliger, Shuly Wintner, Jennifer Pan, Dan Jurafsky, and Yulia Tsvetkov. 2018. Framing and agenda-setting in russian news: a computational analysis of intricate political strategies. *arXiv preprint arXiv:1808.09386*.
- Fabrizio Gilardi, Charles R Shipan, and Bruno Wüest. 2021. Policy diffusion: The issue-definition stage. *American Journal of Political Science*, 65(1):21–35.
- Luca Gilardi and Collin Baker. 2018. Learning to align across languages: Toward multilingual framenet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Shima Khanehzar, Andrew Turpin, and Gosia Mikołajczak. 2019. Modeling political framing across policy issues and contexts. In *Proceedings of The 17th Annual Workshop of the Australasian Language Technology Association*, pages 61–66.
- Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th acm conference on web science*, pages 305–314.
- Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.
- Sophie Lecheler, Mario Keer, Andreas RT Schuck, and Regula Hänggli. 2015. The effects of repetitive news framing on political opinions over time. *Communication Monographs*, 82(3):339–358.
- Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding us gun violence. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pages 504–514.
- Joel Pinho Lucas, João Felipe Guedes da Silva, and Leticia Freire de Figueiredo. 2023. Npr: a news portal recommendations dataset. In *NORMALize@RecSys*.
- Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. **Modeling framing in immigration discourse on social media**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.
- Fred Morstatter, Liang Wu, Uraz Yavanoglu, Stephen R Corman, and Huan Liu. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing*, 1(2):1–18.
- Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing framing to support multiperspective recommendations of opinion pieces. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 478–488.
- Nona Naderi and Graeme Hirst. 2017. Classifying frames at the sentence level in news articles. *Pol-icy*, 9:4–233.
- Viet-An Nguyen, Jordan L Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. *Advances in neural information processing systems*, 26.
- Julia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428.
- Wim Peters, Piek Vossen, Pedro Díez-Orzas, and Geert Andriaens. 1998. **Cross-linguistic alignment of wordnets with an inter-lingual-index**. *Computers and the Humanities*, 32(2):221–251.
- Jakub Piskorski, Nicolas Stefanovitch, Valerie-Anne Bausier, Nicolo Faggiani, Jens Linge, Sopho Kharazi, Nikolaos Nikolaidis, Giulia Teodori, Bertrand De Longueville, Brian Doherty, et al. 2023a. News categorization, framing and persuasion techniques: Annotation guidelines. *European Commission, Ispra, JRC132862*.
- Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023b. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361.
- An Yang Qwen Team, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Siva Reddy, Oscar Täckström, Slav Petrov, Mark Steedman, and Mirella Lapata. 2017. **Universal semantic parsing**. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 89–101, Copenhagen, Denmark. Association for Computational Linguistics.

- Carsten Reinemann, James Stanyer, Sebastian Scherr, and Guido Legnante. 2012. [Hard and soft news: A review of concepts, operationalizations and key findings](#). *Journalism*, 13(2):221–239.
- Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4):1064–1082.
- David A Snow and Robert D Benford. 2005. Clarifying the relationship between framing and ideology. *Frames of protest: Social movements and the framing perspective*, 205:209.
- Oren Tsur, Dan Calacci, and David Lazer. 2015. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638.
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

A Appendix

A.1 Frames labels

We adopt the labels from [Card et al. \(2015\)](#). Note that some names from [Boydston et al. \(2014\)](#) were changed. For example, *Constitutionality and Jurisprudence* was modified to *Legality, Constitutionality and Jurisprudence*, while *Law and Order, Crime and Justice* was changed into *Crime and Punishment*. In the updated Codebook ([Boydston and Gross, 2014](#)), the names of some labels were adapted again (e.g., *Morality* is called *Morality and Ethics*).

A.2 Annotation example

Here we report the complete text of the example news introduced in [Figure 1](#) in Brazilian Portuguese (PT) and its translation in English (EN).

PT Headline: Ministro do Trabalho quer mesa de negociação entre trabalhadores e controladores das Americanas.

Body: O ministro do Trabalho, Luiz Marinho, afirmou nesta terça-feira (31) que quer estabelecer uma mesa redonda entre os trabalhadores das lojas Americanas e os controladores da empresa. A declaração acontece um dia após as maiores centrais sindicais do país terem divulgado comunicado conjunto defendendo a participação direta do Ministério do Trabalho no processo de recuperação judicial da Americanas, ocorrido após a revelação no início do mês de pelo menos R\$ 20 bilhões em problemas contábeis. As centrais querem preservar empregos na varejista, que têm mais de 44 mil trabalhadores, além de "centenas de milhares" na rede de fornecedores. "Vamos providenciar mesa redonda das Americanas com representantes dos trabalhadores para buscar compreender o que aconteceu. Se houve fraude, vamos acionar os órgãos responsáveis pelas fraudes. Não posso afirmar que houve fraude, mas que tem cheiro tem cheiro", declarou o ministro. Acionistas de referência da Americanas divulgam nota pública A Americanas tem entre os principais acionistas os bilionários Jorge Paulo Lemann, Carlos Sicupira e Marcel Telles. Luiz Marinho questionou nesta terça-feira onde estavam os "compliances" das lojas Americanas, ou seja, as regras de boa gestão e de governança da empresa. "Então é o momento de a gente colocar em dúvida alguns valores do chamado mercado, que é o senhor da razão. Dita regra para lá e para cá, e a regra sagrada não foi observada, que é cuidar da saúde de uma empresa que tem 1.790 lojas espalhadas pelo Brasil e tem 44 mil trabalhadores e trabalhadoras", disse ele. O ministro do Trabalho avaliou, ainda, que a Americanas "pode dar calote no Estado brasileiro, pode dar calote nos credores, nos que estavam de boa fé. E acrescentou que os "espertalhões provavelmente não têm problema". "Quem estava no comando, provavelmente

já se livrou", disse. "Os acionistas provavelmente vão poder se locupletar, é possível que tenham se locupletado. Os controladores né. Os acionistas minoritários, coitados, estão na mesma situação dos trabalhadores", concluiu.

EN Headline: Minister of Labor wants negotiation table between Americanas' workers and controllers.

Body: The Minister of Labor, Luiz Marinho, stated this Tuesday (31) that he wants to establish a roundtable between the workers of the Americanas stores and the company's controllers. The statement comes one day after the country's largest union federations released a joint communiqué defending the direct participation of the Ministry of Labor in Americanas' bankruptcy protection process, which followed the revelation earlier this month of at least R\$ 20 billion in accounting irregularities. The unions want to preserve jobs at the retailer, which employs more than 44,000 workers, in addition to "hundreds of thousands" in its supplier network. "We are going to set up a roundtable between Americanas and worker representatives to try to understand what happened. If there was fraud, we will trigger the authorities responsible for fraud cases. I cannot say there was fraud, but it certainly smells like it," said the minister. Major shareholders of Americanas release public note Americanas' main shareholders include billionaires Jorge Paulo Lemann, Carlos Sicupira, and Marcel Telles. On Tuesday, Luiz Marinho questioned where the "compliances" of Americanas were, meaning the company's good management and governance practices. "So this is the moment for us to question certain values of the so-called market, which claims to be the voice of reason. It sets rules here and there, yet the sacred rule was not observed: taking care of the health of a company with 1,790 stores across Brazil and 44,000 workers," he said. The Minister of Labor also argued that Americanas "may default on the Brazilian state, may default on creditors, on those who acted in good faith. And he added that the 'smart operators probably have no problems.' 'Those in charge probably already got away,' he said. "The shareholders will probably be able to enrich themselves, it is possible that they already have. The controllers, right. The minority shareholders, poor them, are in the same situation as the workers," he concluded.

A.3 FrameNews-PT and NPR analysis

For the NPR dataset, the average number of words in article titles is 14.9 ± 3.6 , while the average number of words in article bodies is 471.7 ± 391.9 ([Lucas et al., 2023](#)). [Table 3](#) shows the most frequent topics in NPR and FrameNews-PT.

A.4 Expenses

We spent 96 dollars for evaluating ChatGPT-4o with the OpenAI API (3 prompt templates x 6,214

id	FrameNews-PT	NPR
0	presidente, lula, bolsonaro, governo, ex, feira, nesta, petrobras, conselho, pt	lula, presidente, governo, ex, petrobras, conselho, feira, ministro, cargo, nesta
1	governo, haddad, sobre, fiscal, tributária, imposto, vai, ministro, câmara, texto	bolsonaro, presidente, federal, atos, stf, tribunal, ex, ministro, república, justiça
2	federal, atos, stf, tribunal, prisão, justiça, torres, ministro, polícia, pf	governo, haddad, fiscal, sobre, tributária, proposta, câmara, regra, ministro, imposto
3	mínimo, bilhões, inflação, salário, 2023, pib, governo, valor, ano, tesouro	juros, taxa, brasil, tesouro, dívidas, bancos, milhões, bilhões, inss, estados
4	dívidas, bancos, tesouro, milhões, bilhões, empréstimos, governo, juros, estados, taxa	café, rural, vídeos, bacalhau, globo, produtos, soja, onde, vem, produção
5	empresa, petrobras, bahia, companhia, vendas, marca, varejo, ford, refinaria, unilever	empresa, petrobras, bahia, companhia, vendas, marca, ford, varejo, refinaria, bilhões
6	rural, bacalhau, globo, vídeos, soja, produtos, saiba, água, assistidos, café	mínimo, inflação, salário, bilhões, 2023, pib, governo, ano, reajuste, família
7	brasil, auxílio, divórcios, mulheres, milhões, número, julho, ativos, datafolha, analisa	microsoft, internet, telemarketing, explorer, uso, bloqueadores, partir, clienti, codice, applicazione
8	microsoft, telemarketing, internet, explorer, ataque, clientes, secretaria, código, aplicativo, partir	crédito, cartão, americanas, causas, trabalho, bancos, linha, trabalhadores, diagnóstico, câmbio

Table 3: Comparison between top topics in FramesNews-PT and NPR, detected with BERTopic (Grootendorst, 2022). Topics: ■ politics, ■ economics, ■ justice, ■ production, ■ companies, ■ technology, ■ social support.

= 18,642 prompts). We paid 445 euros for each annotator to annotate the 300 examples in 27 hours of work.

A.5 Prompt templates

```

### PROMPT:
"{content}"

### TASK:
Classify the PROMPT above into exactly ONE of
these frame categories:

"1" Economic
"2" Capacity and resources
"3" Morality
"4" Fairness and equality
"5" Legality, Constitutionality, Jurisdiction
"6" Crime and punishment
"7" Security and defense
"8" Health and safety
"9" Quality of life
"10" Cultural identity
"11" Public opinion
"12" Political
"13" Policy prescription and evaluation
"14" External regulation and reputation
"15" Other

### CLASSIFICATION GUIDELINES:

{guidelines}

Base your answer only on the PROMPT and the
guidelines provided above. Answer as a
single number ("1", ..., "15") corresponding
to the most appropriate category.

### ANSWER:

```

Listing 1: Prompt template without guidelines: zero1

```

### CLASSIFICATION GUIDELINES:

{guidelines}

### PROMPT:
"{content}"

### TASK:
Classify the PROMPT above into exactly ONE of
the categories below.
"1" Economic
"2" Capacity and resources
"3" Morality
"4" Fairness and equality
"5" Legality, Constitutionality, Jurisdiction
"6" Crime and punishment
"7" Security and defense
"8" Health and safety
"9" Quality of life
"10" Cultural identity
"11" Public opinion
"12" Political
"13" Policy prescription and evaluation
"14" External regulation and reputation
"15" Other

Answer as a single number ("1", ..., "15")
corresponding to the most appropriate
category.

### ANSWER:

```

Listing 2: Prompt template with guidelines: The guidelines can be found in the github link.

A.6 Description of frames

Here we provide the descriptions of frames in our guidelines. The full guidelines are provided in the repository.

1. Economic: The costs, benefits, or any monetary/financial implications of the issue (to an individual, family, organization, community or to the economy as a whole). Can

include the effect of policy issues on trade, markets, wages, employment or unemployment, viability of specific industries or businesses, implications of taxes or tax breaks, financial incentives, etc.

2. Capacity and resources: The lack or availability of resources (time, physical, geographical, space, human, and financial resources). The capacity of existing systems and resources to carry out policy goals. The easiest way to think about it is in terms of there being "not enough" or "enough" of something. The capacity or resources may be an impediment to solving a problem or adequately addressing an issue.

3. Morality: Any perspective that is compelled by religious doctrine or interpretation, duty, honor, righteousness or any other sense of ethics or social or personal responsibility. It is sometimes presented from a religious perspective (i.e. "eye for an eye"), but non-religious frames can also be used. For example, the moral imperatives to help others can be used to justify military intervention or foreign aid, social programs such as Medicare, welfare, and food stamps. Appeals that a policy move "is just the right thing to do" or "would indicate a recognition of our shared humanity" may reflect humanist morality. The commitment aspect of marriage would evoke feelings of morality. Environmental arguments that focus on responsible stewardship or "leaving something for our children" are based in a sense of responsibility or morality. Lawbreakers, including illegal immigrants, can be presented as fundamentally immoral, conversely breaking a law that is bad or unjust can be presented as moral (e.g., Rosa Parks). Enacting protective legislation, such as laws that protect children from pedophiles, guns, violence, poverty, or failure to do so can also be presented using moral frames.

4. Fairness and equality: The fairness, equality or inequality with which laws, punishment, rewards, and resources are applied or distributed among individuals or groups. Also the balance between the rights or interests of one individual or group compared to another individual or group. Fairness and Equality frame signals often focus on whether society and its laws are equally distributed and enforced across regions, race, gender, economic class, etc. Many gender and race issues, in particular, include equal pay, access to resources such as education, healthcare and housing. Another example could be fairness considerations about whether punishments are proportional to crimes committed. The frame is also used when discussing social justice, discrimination and talk of an inmate's innocence or exonerated.

5. Legality, Constitutionality, Jurisdiction: The legal, constitutional, or jurisdictional aspects of an issue. Legal aspects include existing laws, reasoning on fundamental rights and court cases; constitutional aspects include all discussion of constitutional interpretation and/or potential revisions; jurisdiction includes any discussion of which government body should be in charge of a policy decision and/or the appropriate scope of a body's policy reach. This frame deals specifically with the authority of government to regulate, and the authority of individuals/corporations to act independently of government. Of special note are constraints imposed on freedoms granted to individuals, government, and corporations via the Constitution, Bill of Rights and other amendments. Some frequent arguments and issues are: i) the right to bear arms; ii) equal protection; iii) free speech and expression; iv) the constitutionality of restricting individual freedoms and imposing taxes; v) conflicts between state, local or federal regulation and authority, or between different branches of government; vi) legal documentation (green card, visas, passports, driver licenses, marriage license, etc.).

6. Crime and punishment: The violation of policies and its consequences. It includes enforcement and interpre-

tation of civil and criminal laws, sentencing and punishment with retribution or sanctions. This frame includes: i) deportation when an individual does not have the necessary documents that grant legal standing; ii) increases or reductions in crime; iii) punishment and execution; iv) resources analysis like DNA analysis. Usually found together with other frames, such as Economic, Legality, constitutionality and jurisdiction, Morality, and Capacity and resources. The primary frame should be chosen according to where the emphasis is.

7. Security and defense: Any threat to a person, group, or nation, or any defense that needs to be taken to avoid that threat. Security and Defense frames differ from Health and Safety frames in that Security and Defense frames address a preemptive action to stop a threat from occurring, whereas Health and Safety frames address steps taken to ensure safety in the event that something happens. It can include efforts to build a border fence or "secure the borders," issues of national security including resource security, efforts of individuals to secure homes, neighborhoods or schools, and efforts such as guards and metal detectors that would defend children from a possible threat. Discussion regarding terrorist activity should be coded as Security and Defense (e.g. arrests of terrorists, immigrants linked to terrorism activity, increased border security to prevent terrorism). Arrests at the border will receive both a Crime and Punishment and Security and Defense frame but the primary frame would be Security and Defense since the action is taking place on the border. All terrorist attacks are coded as Security and Defense, but attention should be paid to potential criminal, legal, or any other aspects and double coded accordingly.

8. Health and safety: The potential health and safety outcomes of any policy issue (e.g. health care access and effectiveness, illness, disease, sanitation, carnage, obesity, mental health infrastructure and building safety). Also policies taken to ensure safety in case of a tragedy would fit under this (e.g. emergency preparedness kits, lock down training in schools, disaster awareness classes for teachers). It includes any discussion of the various capital punishment methods and procedures and any mentions of refugees. Often used in conjunction with Quality of Life.

9. Quality of life: The benefits and costs of any policy on quality of life. The effects of a policy on people's wealth, mobility, access to resources, happiness, social structures, ease of day-to-day routines, quality of community life, etc. It includes any mention of people receiving generic "benefits", adoptions, and weddings. Often used in conjunction with Health and Safety.

10. Cultural identity: The social norms, trends, values and customs constituting any culture(s). It includes: i) language issues and language learning; ii) patriotism and national traditions, the history of an issue or the significance of an issue within a group or subculture; iii) census and demographics; iv) cultural shifts in a group or society; v) cultural norms of ethnic and political groups. May also include stereotypes or assumed preferences and reactions of a group (e.g., an affinity for Republicans to wear cowboy hats); vi) references and quotations of famous people like politicians, leaders or representatives of a subculture.

11. Public opinion: The opinion of the general public. It includes references to general social attitudes, protests, polling and demographic information, as well as any public passage of a proposition or law (i.e. "California voters passed Prop 8"). All the opinions that represent the sentiment of a group will be coded as Public opinion. However, a group of experts in a particular domain gets coded according to their domain (e.g. police officers in Crime and Punishment, or climate scientists in Capacity and Resources).

12. Political: In general, any political considerations surrounding an issue. It includes political actions, maneuvering, efforts or stances towards an issue (e.g. partisan filibusters, lobbyist involvement, deal-making and vote trading), mentions of political entities or parties (e.g., Democrats, Republicans, Libertarians, Green Party). When a headline mentions “both sides” this refers to politics.

13. Policy prescription and evaluation: The analysis of whether hypothetical policies will work or existing policies are effective. What is/isn’t currently allowed and what should/shouldn’t be done? “Policy” encompasses formal government regulation (e.g., federal or state laws) as well as regulation by businesses (e.g., sports arenas not allowing the sale of alcohol). This frame dimension—perhaps more than any other—is likely to appear frequently across texts. Yet care should be given to only use this code category as the primary frame when the main thrust of an article is really about policy, for example when it describes the success and failure of existing policies or proposes policy solutions to a problem.

14. External regulation and reputation: In general, the country’s external relations with another nation; the external relations of a state with another. This frame includes: i) trade agreements and outcomes; ii) comparisons of policy outcomes between different groups or regions; iii) the perception of one nation, state, and/or group by another (for example, international criticisms of the United States maintaining capital punishment); iv) border relations, interstate or international efforts to achieve policy goals; v) alliances or disputes between groups.

15. Other: Any frame signal that does not fit in the first 14 dimensions.

A.7 Further results

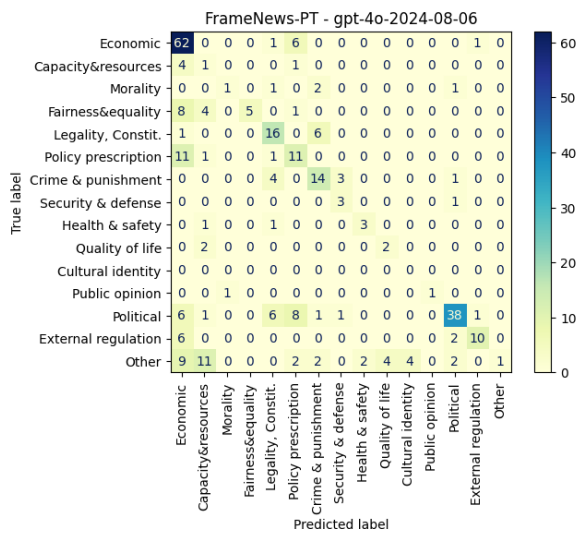


Figure 8: Confusion matrix of the best prompt and best model (chatGPT-4o) in comparison with annotator 1.

Model	Global F ₁	F ₁ (ann 1)	F ₁ (ann 2)	Accuracy
gpt-4o-2024-08-06	0.50 ±0.04	0.51 ±0.04	0.49 ±0.03	0.59 ±0.02
Qwen2.5-7B-Instruct	0.45 ±0.04	0.46 ±0.04	0.43 ±0.03	0.54 ±0.02
Llama-3.1-8B-Instruct	0.36 ±0.04	0.37 ±0.03	0.34 ±0.04	0.44 ±0.04
gemma-3-4b-it	0.35 ±0.03	0.36 ±0.02	0.34 ±0.04	0.43 ±0.02
Qwen2.5-1.5B-Instruct	0.29 ±0.09	0.30 ±0.09	0.28 ±0.08	0.41 ±0.08
Qwen2.5-3B-Instruct	0.33 ±0.08	0.34 ±0.08	0.32 ±0.07	0.40 ±0.07
Llama-3.2-3B-Instruct	0.11 ±0.07	0.12 ±0.06	0.10 ±0.07	0.26 ±0.06
Llama-3.2-1B-Instruct	0.16 ±0.02	0.17 ±0.02	0.15 ±0.02	0.24 ±0.03

(a) FrameNews-PT zero-shot results. We report F₁ scores averaged on both annotators (global F₁), against annotator 1 (ann1) and annotator 2 (ann 2).

Model	Global F ₁	Accuracy
gpt-4o-2024-08-06	0.47 ±0.01	0.46 ±0.01
Qwen2.5-7B-Instruct	0.39 ±0.02	0.40 ±0.01
Llama-3.1-8B-Instruct	0.35 ±0.02	0.37 ±0.02
gemma-3-4b-it	0.34 ±0.01	0.37 ±0.03
Qwen2.5-3B-Instruct	0.30 ±0.05	0.30 ±0.04
Llama-3.2-3B-Instruct	0.16 ±0.12	0.27 ±0.09
Qwen2.5-1.5B-Instruct	0.21 ±0.12	0.23 ±0.11
Llama-3.2-1B-Instruct	0.06 ±0.06	0.12 ±0.07

(b) MFC zero-shot results. We report scores against the gold label.

Table 4: Zero-shot results with generative models on both datasets. Accuracy is calculated by considering the output as correct if it matches at least one annotation. The standard deviation is calculated based on the three prompt templates used to evaluate each model. Best results are **in bold**.