SENTIMENT ANALYSIS FOR MULTILINGUAL LLMS

Shunqiang Feng
Department of Computer Science
University of Virginia
Charlottesville, VA 22904, USA
mpp7ez@virginia.edu

Zihan Zhao
Department of Computer Science
University of Virginia
Charlottesville, VA 22904, USA
rxy6cc@virginia.edu

ABSTRACT

Multilingual large language models (LLMs) are increasingly used in global applications, yet their internal consistency across languages remains underexplored—particularly in sentiment-sensitive tasks. Subtle linguistic and cultural variations can cause the same prompt to elicit different emotional tones across languages, complicating the evaluation of multilingual LLM outputs. This study introduces a cross-linguistic sentiment analysis pipeline that translates multilingual responses into English for unified sentiment scoring using an English sentiment analyzer. Through this method, we uncover three key findings: (1) Multilingual LLMs exhibit notable differences across languages, including tendencies toward neutral hedging and divergent attitudes toward the same event; (2) Fine-tuning can significantly reduce hedging and align multilingual outputs more closely with real-world perspectives; and (3) The internal behavior of multilingual LLMs is highly complex, with different subtopics under a shared theme producing varying sentiment trends. These insights highlight the importance of robust evaluation and targeted fine-tuning to ensure consistent and culturally aware performance in multilingual language models.

1 INTRODUCTION

In an increasingly interconnected world, media portrayals of events often differ substantially across linguistic and cultural contexts, reflecting regional biases and divergent political narratives. Understanding these differences is critical for identifying media bias, misinformation, and ideological framing. Prior studies have demonstrated that misleading content—particularly that which exhibits strong framing, sensationalism, or heightened sentiment—tends to disseminate more rapidly than accurate, factual reporting Vosoughi et al. (2018). For individuals in public-facing roles, such as policymakers, scholars, and industry leaders, anticipating how their statements may be represented in the media is vital to ensuring transparent and equitable communication with the public. While media bias and sentiment analysis have garnered significant attention across various disciplines, most research efforts remain confined to monolingual settings—predominantly English—overlooking the complexity of cross-linguistic variation. This gap is especially consequential in a global landscape shaped by multilingual nations and multinational institutions, such as India and the European Union.

In this study, we extend previous research and utilize existing datasets to design and evaluate large language model (LLM)-based systems capable of: (1) generating reactions to prompts concerning various topics across multiple languages, and (2) employing sentiment analysis to infer the likely tone and framing of these responses in relation to future or hypothetical events (e.g., predicting how media in a given language might respond to a policy announcement). Through this work, we engage with state-of-the-art techniques in sentiment analysis and media bias detection, leveraging multilingual pretrained language models (MPLMs) and LLMs to support our objectives. Our methodology involves: (1) reviewing prior literature and baseline approaches to identify appropriate models and techniques, (2) assembling and enhancing existing labeled multilingual media datasets—drawing from sources such as Azizov et al. (2024), Maity et al. (2023), Łukasz Augustyniak et al. (2023), Sales et al. (2021), and Vargas et al. (2023)—as well as collecting additional data, (3) training and fine-tuning our models to generate media-style reactions and perform sentiment analysis across languages, and (4) evaluating model performance relative to established baselines, using sentiment accuracy across topics as the primary evaluation metric.

The fine-tuned model demonstrated marked improvements in sentiment alignment compared to the baseline, achieving perfect polarity accuracy in English and notable performance gains in Japanese, French, and Chinese. Fine-tuning effectively reduced neutral hedging, resulting in generated sentiment that more closely mirrored that of real-world news articles—particularly in the context of culturally and politically sensitive topics. Furthermore, cross-linguistic analysis revealed significant variation in emotional tone across languages and topics, underscoring the impact of linguistic and cultural framing on media representation.

2 Related Work

Recent advancements in machine learning and natural language processing have led to substantial progress in the development of large language models (LLMs), including GPT (Floridi Chiriatti, 2020), Llama (Touvron et al., 2023), and Deepseek (Guo et al., 2025). Within this domain, sentiment analysis remains a pivotal yet complex research area, particularly in multilingual settings where cultural and political nuances significantly influence interpretation (Gupta et al., 2024). Traditional sentiment analysis approaches, which often rely on manually crafted features and rule-based systems, face challenges in capturing the nuanced emotional content of text and adapting to diverse linguistic and cultural contexts (Nadkarni et al., 2011). In contrast, modern LLMs such as GPT-3 (Floridi Chiriatti, 2020) have demonstrated the ability to learn emotional representations directly from large-scale text corpora, thereby enhancing sentiment analysis and facilitating the investigation of cross-linguistic variation and embedded biases in language models (Zhan et al., 2024).

Although there has been growing interest in examining media bias in cross-linguistic contexts, existing research remains relatively limited in breadth. For example, Azizov et al. (2024) introduced a multilingual corpus and labeled dataset to investigate political bias at both the media and article levels across ten languages, leveraging multilingual pretrained language models (MPLMs) and zero-shot LLMs for evaluation. In a related effort, Sales et al. (2021) explored bias detection in four languages using labeled datasets but employed smaller-scale techniques, such as translated subjectivity lexicons, rather than large language models. These studies underscore the promise of multilingual models in addressing media bias; however, they also expose key limitations—particularly in the scalability of traditional methods and the lack of comprehensive frameworks capable of effectively capturing bias across diverse linguistic and media landscapes.

To support the generative component of our framework, we employ Qwen-2.5-3B, a multilingual instruction-tuned language model developed by Alibaba. As an open-source successor in the Qwen series, Qwen-2.5-3B offers robust performance across more than 20 languages and has demonstrated strong results on a variety of multilingual benchmarks Qwen et al. (2025). Its chat-optimized fine-tuning, combined with native proficiency in languages such as Chinese and English, renders it particularly well-suited for prompt-based text generation tasks. Furthermore, its moderate scale—comprising 3 billion parameters—enables efficient fine-tuning on custom datasets while maintaining output coherence and factual reliability. These attributes position Qwen-2.5-3B as a highly effective backbone for generating culturally contextualized media content within our cross-linguistic analysis framework.

For sentiment analysis, we utilize the cardiffnlp/twitter-roberta-base-sentiment model Barbieri et al. (2020), a distilled RoBERTa-based classifier trained on a large corpus of social media text. Although originally optimized for short, informal content such as tweets, the model exhibits strong generalization to generated text due to its robustness in capturing nuanced tone and subjective expressions. It outputs probabilistic predictions across three sentiment classes—positive, neutral, and negative—making it well-suited for downstream evaluation of LLM-generated responses. Its high accuracy and reliable calibration under diverse and noisy input conditions align closely with the demands of multilingual sentiment analysis.

Nevertheless, cross-lingual sentiment prediction remains inherently challenging, often complicated by cultural asymmetries and translation-induced artifacts. Earlier methods based on sentiment lexicons Dehkharghani et al. (2012) or zero-shot transfer through English-centric LLMs such as GPT-3 Brown et al. (2020) frequently suffer from bias and inconsistent performance across languages. To address these issues, our framework combines Qwen-2.5-3B for context-sensitive and culturally informed text generation with a back-translation strategy that enables sentiment analysis to be performed on English renderings using a robust pretrained classifier. This design facilitates scal-

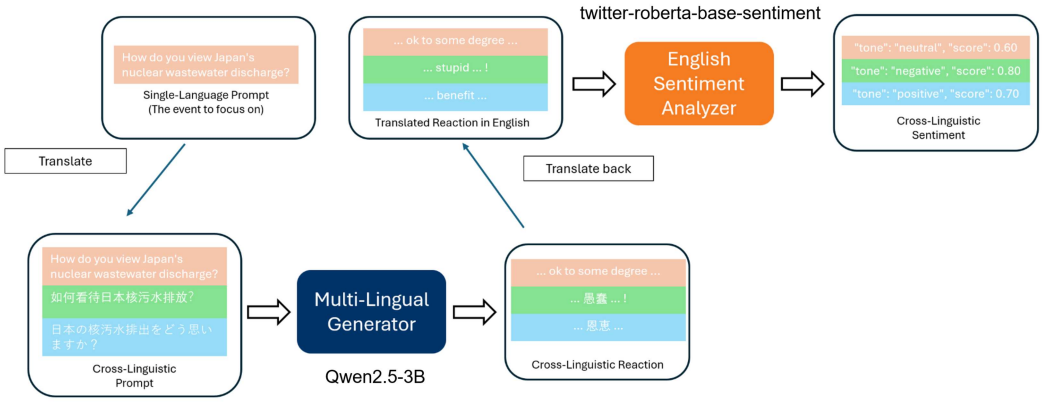

Figure 1: The overview framework

able, language-agnostic sentiment evaluation and supports culturally adaptive modeling of media responses across linguistic boundaries.

3 Problem Setup

This project focuses on predicting cross-linguistic media reactions and sentiment analysis in a globalized media environment. The motivation arises from the need to ensure objective reporting and maintain public trust, as biased framing accelerates misinformation spread Vosoughi et al. (2018). Traditional methods often neglect linguistic diversity, limiting their applicability in multilingual regions Sales et al. (2021); Azizov et al. (2024). The framework accepts a **Cross-Linguistic Prompt** as input in a single language at a time (e.g., "How do you view Japan's nuclear wastewater discharge?" in English). Outputs include:

- **Cross-Linguistic Reaction**: Generated media responses in the input language (e.g., "Chinese media might criticize Japan's nuclear wastewater discharge as an ecological threat").
- **Cross-Linguistic Sentiment**: Sentiment analysis with polarity (e.g., negative) and a sentiment score (e.g., 0.75).

The goals are to (1) predict media reactions across languages and (2) analyze sentiments in diverse linguistic contexts.

4 Method

The framework employs a modular pipeline: **Cross-Linguistic Prompt** $\rightarrow$ **Generator** $\rightarrow$ **Sentiment Analyzer** $\rightarrow$ **Cross-Linguistic Sentiment** as figure 1.

4.1 Model Design

- **Generator**: Utilizes **Qwen-2.5-3B**, a powerful multilingual chat model developed by Alibaba. It is instruction-tuned, optimized for dialogue-style inputs, and supports multi-turn generation with native support for Chinese, English, and over 20 other languages. Its relatively compact size (3B parameters) makes it efficient for fine-tuning and inference, while still maintaining competitive generation quality across domains. We found Qwen's adherence to structured prompts and stable generation behavior particularly suitable for multilingual headline-to-article tasks.
- **Sentiment Analyzer**: Employs `twitter-roberta-base-sentiment`, a RoBERTa-based model fine-tuned on millions of social media posts for nuanced sentiment detection. Despite its origin in the Twitter domain, the model generalizes well to short and medium-length news content. It outputs probability distributions over

`positive`, `neutral`, and `negative` classes. Its strengths lie in detecting subtle opinion tones, handling informal phrasing, and providing interpretable, discrete sentiment scores, which are ideal for cross-lingual comparison and aggregation.

4.2 FINE-TUNING/INFERENCE ALGORITHMS

- **Fine-Tuning**: We fine-tune the Qwen-2.5-3B-Instruct model on our multilingual news dataset using the HuggingFace `Trainer` and `Accelerate` framework. Each training instance follows a structured prompt format: a system message provides task instruction, and a user message provides the article title. The tokenizer applies Qwen's chat template to encode both source and target sequences with truncation and padding.

  Training is conducted on all five languages jointly, using shuffled datasets from JSONL files. We tokenize and filter out malformed examples, then split the data into 90% training and 10% validation. The model is trained for 8 epochs with BF16 precision, a batch size of 32, and gradient accumulation. We use a learning rate of 2×10^{-5}, with checkpointing and evaluation at each epoch. Accelerate ensures efficient utilization of multiple GPUs with minimal memory overhead.
- **Inference**: Our multilingual inference pipeline consists of four stages:
  1. **Prompt Translation**: Given an English topic, we use Google Translate API to translate it into four target languages (Chinese, Japanese, French, Spanish), preserving the original English query.
  2. **Cross-lingual Generation**: For each translated prompt, we apply the fine-tuned Qwen model to generate a 300-word news article, following the chat-based prompt structure. The same system prompt is used during inference as in training.
  3. **Back-translation**: The generated articles (except English) are translated back into English to normalize sentiment interpretation across languages.
  4. **Sentiment Evaluation**: The back-translated texts are analyzed using a zero-shot sentiment classifier, `cardiffnlp/twitter-roberta-base-sentiment`. It outputs a distribution over `positive`, `neutral`, and `negative` classes. Probabilistic scores are used to measure tone consistency across language variants.

  This full-cycle pipeline enables controlled, comparable sentiment analysis of model outputs across diverse languages, using only English as the analysis anchor.

4.3 PRE-/POST-PROCESSING STEPS

We adopt structured pre- and post-processing procedures to ensure data quality and enable consistent cross-lingual evaluation.

Pre-processing: During training, raw news data is filtered to remove empty or malformed entries. Each sample is assigned a language tag and formatted into a chat-style prompt using Qwen's instruction format, where the title is treated as input and the article body as the target. Tokenization is applied with truncation and padding to fixed lengths.

Post-processing: In inference, each English prompt is translated into four target languages. After generation, all non-English outputs are back-translated into English to standardize comparison. Sentiment analysis is then performed on the English texts using a pretrained RoBERTa model, yielding positive, neutral, and negative scores. These results are aggregated to assess cross-lingual sentiment divergence.

4.4 CONTRIBUTIONS

This method offers:

- **Novelty**: Integrates Qwen's generative capabilities with RoBERTa's sentiment analysis across different languages, extending prior work Sales et al. (2021); Azizov et al. (2024) and introducing a new pipeline for a novel purpose.
- **Effectiveness**: Captures linguistic nuances using Qwen for reaction generation and RoBERTa for sentiment analysis, leveraging their pre-training on diverse datasets.

- **Efficiency**: Utilizes pre-trained models, minimizing training overhead while achieving robust performance.
- **Simplicity**: Employs a modular pipeline, facilitating easy adaptation to new languages and tasks.

This framework enhances cross-linguistic media reaction prediction and sentiment analysis for global communication.

5 EXPERIMENTAL SETUP

5.1 DATASET DESCRIPTION

To train and evaluate multilingual news generation and sentiment analysis, we constructed a large-scale cross-lingual news dataset by crawling latest online media sources in five languages: Chinese, English, Spanish, Japanese, and French. For each article, we retained four key metadata fields:

- **URL —** original article link;
- **Publish Time —** timestamp of publication;
- **Title —** used as the prompt for generation;
- **Content —** used as the generation target or reference.

The sources include high-quality news agencies across regions and ideologies, such as *Xinhua, CNN, El Mundo, Yomiuri, Liberation*, among others. Table 1 shows the number of processed examples per language.

Table 1: Multilingual news corpus statistics.

Language	News Source	Size
Chinese	Xinhua, Sina, People...	21109
English	CNN, BBC, The Guardian...	19436
Spanish	El Mundo, abc, publico...	15399
Japanese	Yomiuri, Nikkei, Tokyo-np...	10757
French	Libération, FranceSoir, BFMTV...	20942

To prepare the dataset for training, we apply a language-aware preprocessing function. Each input is constructed as a system-user message pair where the `title` is used as the generation prompt, and the `content` is used as the expected target response. The model input follows the chat template structure used by Qwen:

```
System: ``You are Qwen. You are a helpful assistant and
expected to write a news report with the given title.''
User: [Title]
```

Tokenization is applied separately to the input and target texts using Qwen's chat template and tokenizer, with truncation and padding to predefined `max_source_length` and `max_target_length`. The preprocessing function also performs validation to filter out malformed inputs. Each sample results in three token sequences: `input_ids`, `attention_mask`, and `labels`, suitable for supervised fine-tuning.

During inference, we use the same template but modify the system message to explicitly instruct the model to write a 300-word article:

```
System: ``You are Qwen. You are a helpful assistant and
expected to write a 300 words news report with the given
topic.''
User: [Topic prompt]
```

This structure ensures alignment between training and inference, and enables consistent evaluation across multilingual prompts.

5.2 INFRASTRUCTURE

We used the provided Rivanna environment with A100 GPUs for our fine tuning, testing, and evaluation.

5.3 EVALUATION ON A GROUND TRUTH DATASET

To build the ground-truth set we first selected five newsworthy topics that had developed within the past week (from the time of submission) and thus were not part of the fine-tuning dataset: one each for culture, U.S. politics, technology, social justice and the environment—so that the reference sentiment would span a wide variety of polarity ranges. For every topic we harvested twenty-five short paragraphs (approximately 300 words) from distinct, reputable outlets published in the target languages: English, French, Spanish, Japanese and Chinese. The final dataset consists of 125 paragraphs organized by topic for each of the five languages. The sources included wire services, national papers, specialist tech sites and regional broadcasters to avoid editorial monoculture. Duplicate wire rewrites and opinion pieces were excluded. Only straight news reports published within the seven day window were kept. Each paragraph was lightly edited for length but never for tone, preserving its original sentiment signal. Finally, topic labels, language codes and a numeric article ID were attached so the evaluation pipeline could aggregate sentiment per topic and per language without ambiguity.

To score sentiment, we employed a multilingual RoBERTa model that returns a polarity score and a discrete class label (negative, neutral, positive). Each reference paragraph is analysed individually, then averaged to give the mean reference polarity for its topic. The generated paragraph is analysed once, and the two scores are compared. From these values we derive per-topic metrics: the generated polarity, the reference mean, the signed difference and its absolute value. Converting the polarities to class labels allows a straightforward accuracy test allowing us to answer the question: does the generator's label match the majority polarity of the news corpus?

Our procedure aggregates results in two ways. First, overall accuracy and a 3x3 confusion matrix show how often a language's outputs fall into the correct sentiment class or drift to "neutral" or the opposite sign. Second, by measuring the absolute polarity gap between a generated paragraph and each of its twenty five references, we obtain a distribution that highlights where the generator sits inside (or outside) the "real" news sentiment cloud. Box-plots visualize these distributions, and a heat map of mean absolute error pinpoints topic–language pairs with the largest divergences.

6 RESULTS

6.1 EFFECT OF FINE TUNING

To evaluate our fine-tuning procedure we follow the evaluation procedure given in section 5.3 above before (which we call baseline) and after fine-tuning for all languages (except Spanish, see section 7).

For the baseline model, the English generator matched the reference sentiment on only three of the five topics, yielding an accuracy of 0.60, while Japanese, French and Chinese trailed further at 0.40, 0.20 and 0.20 (see Figure 2 left). We found from experimentts that all classification errors across languages stem from the model defaulting to a neutral prediction in cases where the reference corpus exhibits a clearly positive or negative sentiment. Notably, no instances of direct polarity inversion were observed. The mean-difference heat-map (Figure 3 left) confirms the pattern: red cells for technology and US-politics in French and Chinese, a large gap for cultural news in all three non-English languages, and a uniformly high gap for social-issue stories because all languages, including English, lost confidence about the positive framing of the protests and charge dismissals.

On the other hand, the fine-tuned configuration reached perfect accuracy in English and pushed Japanese to 0.80 and both French and Chinese to 0.60 (see Figure 2 right). Those gains came almost entirely from eliminating the neutral hedging just described. On cultural and technology news, the fine-tuned model preserved the positive sign across languages, cutting the mean absolute distance by roughly 0.65 in French and Chinese and by more than 0.60 in Japanese. In politics in French, the model likewise moved from a near neutral polarity to a clear negative, pulling its

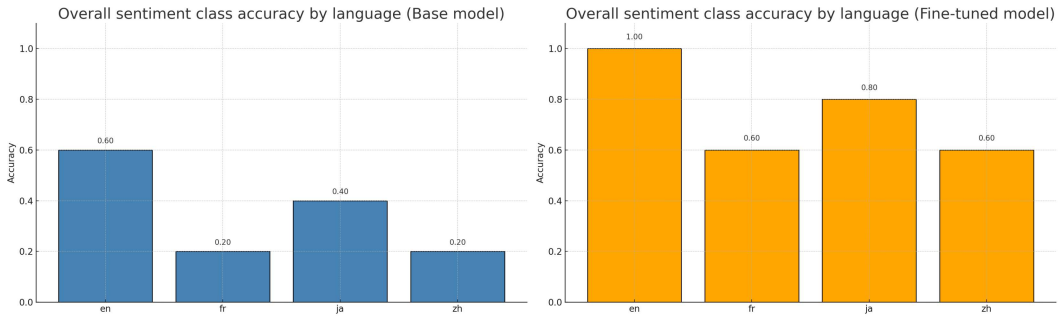

Figure 2: Overall sentiment accuracy across 5 topic areas for each language before fine-tuning (left) and after fine-tuning (right)

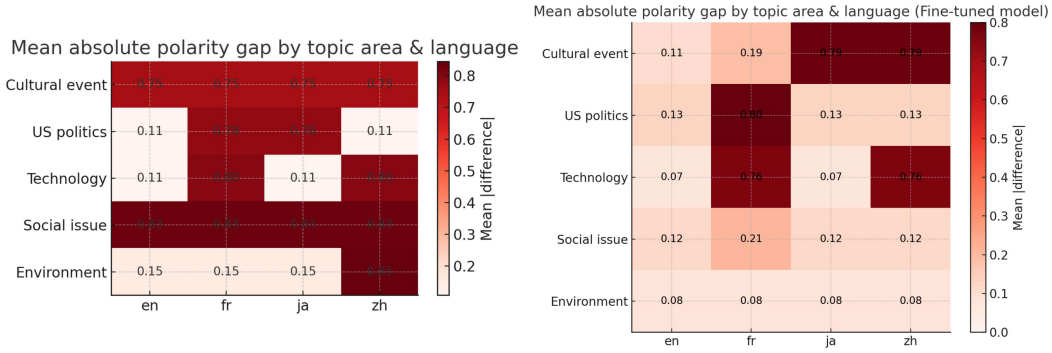

Figure 3: Baseline before fine-tuning (left) and after fine-tuning (right). The heat-map shows how far each language's generated sentiment diverges from real news across topic areas, spotlighting neutral-hedging errors—especially in cultural, technology and U.S.-political stories—and thus pinpoints which language-domain pairs need further tuning.

average gap down by the same order of magnitude (Figure 3 right). The social-issue topic saw the starkest improvement: every language switched from a neutral guess to the properly positive stance, reducing the mean gap from about 0.83 to 0.12.

The comparison suggests that the baseline model's main failure mode is not misreading sentiment but under committing whenever domain cues are subtle. The fine-tuning step evidently injected enough domain-specific signals, or simply adjusted the neutral band, to push scores past the threshold and into the correct discrete class. The English generator needed only minor recalibration. French and Chinese required larger magnitude shifts, but even a small push was enough because their errors were close to the decision boundary.

6.2 Multilingual Sentiment Divergence by Example-Level Analysis

To quantify the sentiment variations across languages in real-world contexts, we conduct a fine-grained analysis on 5 trending topics, each accompanied by 10 representative examples, totaling 50 unique prompts. For each example, we generate outputs in five languages (English, Spanish, French, Japanese, and Chinese), followed by sentiment annotation in the positive/neutral/negative scheme. We define a composite attitude score for each output as: $Attitude = Positive - Negative$. This score provides a clearer picture of the overall sentiment tendency in each language for a given scenario.

Figure 4 presents a heatmap where each row corresponds to a specific example, grouped by topic (demarcated with black horizontal lines), and each column represents a language. The color intensity reflects the composite attitude score. Positive values (in red shades) denote favorable sentiment, while negative values (in blue shades) indicate unfavorable attitudes.

From the heatmap, we observe several noteworthy patterns:

- Topic-level sentiment trends emerge clearly. For example, Cultural events consistently receive high positive attitudes across all languages (e.g., "Literature and book fairs" up to 0.88), reflecting their general acceptance and low controversy. In contrast, Political events show predominantly negative scores (e.g., "Human rights issues" and "Middle East politics"), indicating potential cultural and ideological friction in how these topics are framed.
- Language-specific sentiment variation is visible. Japanese (ja) and Chinese (zh) responses tend to show softer or more moderate sentiment polarity, possibly due to the translation or LLM alignment biases. English (en) and French (fr) often exhibit higher contrast, both positive and negative, hinting at more polarized reactions.
- Technological topics are mostly seen positively, especially those tied to innovation like "Quantum computing" and "Virtual/Augmented Reality." However, tech-related topics that intersect with politics or safety—like "AI regulation" and "Data privacy"—show more neutral or mixed responses.
- Environmental and social issues show the broadest variance. Topics such as "Plastic pollution" or "Youth unemployment" receive sharply negative sentiments, with consistent disapproval across languages, signaling global sensitivity.
- Sentiment divergence across languages can be subtle yet systemic. For instance, "Gender equality movements" gets a slightly negative score in English (0.17) but neutral-positive elsewhere, which may reflect differing sociocultural framing in training data.

The heatmap confirms that multilingual LLMs, even when conditioned on identical prompts, embed latent cultural and linguistic biases that lead to non-trivial divergences in sentiment framing. These findings underscore the importance of language-specific calibration in multilingual content generation and cross-lingual media analysis.

6.3 Topic-Level Cross-Language Sentiment Analysis

To explore how multilingual LLMs vary in their emotional response across different domains, we conduct topic-level sentiment aggregation and sensitivity analysis. For each `(topic, language)` pair, we compute the mean values of `positive`, `neutral`, and `negative` sentiment scores across all examples, and visualize them as three heatmaps in Figure 5–7.

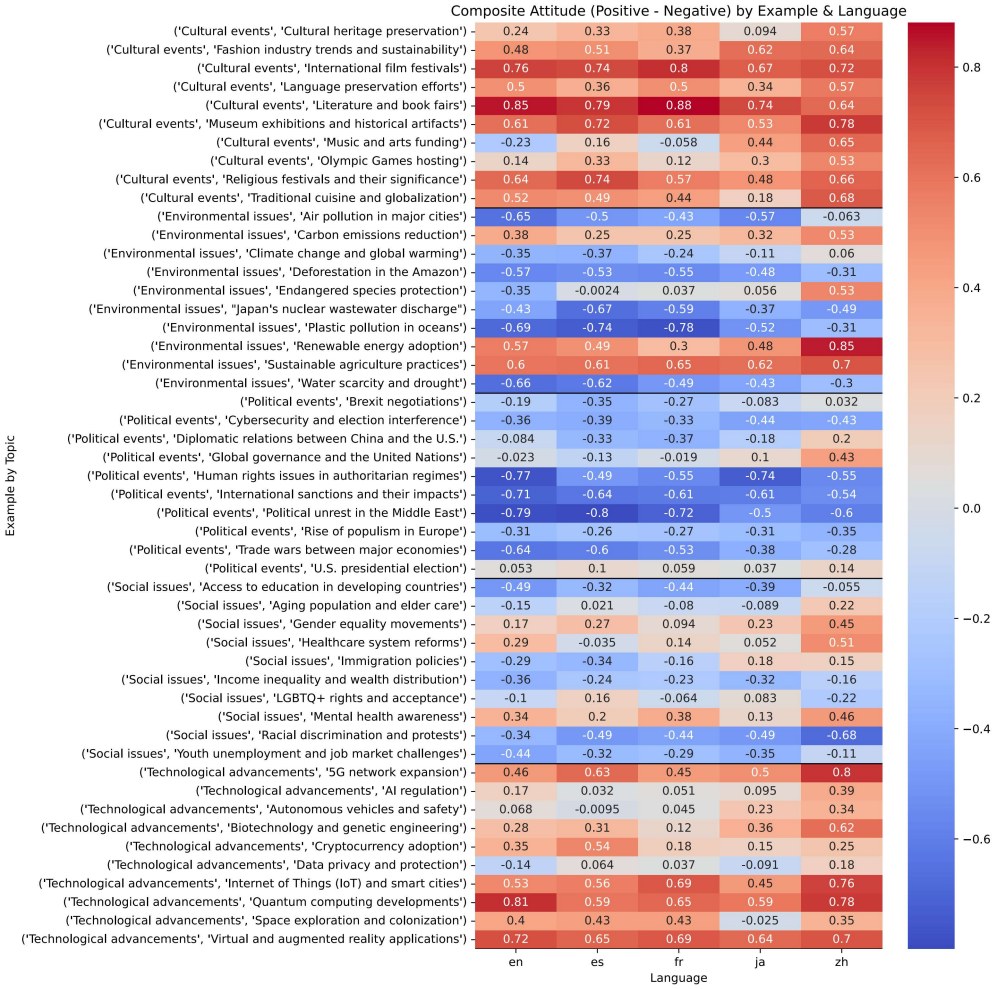

Figure 4: Composite Attitude (Positive - Negative) by Example & Language. Each cell represents the average sentiment score of a language's response to a specific example.

- **Positive sentiment** (Figure 5): Cultural and technological topics, such as *Literature and Book Fairs* and *Virtual Reality*, show consistently high positivity across languages, especially in zh and es outputs.
- **Negative sentiment** (Figure 6): Political and environmental topics, like *Middle East Politics* and *Plastic Pollution*, yield stronger negative responses, with English and French outputs showing the most negativity.
- **Neutral sentiment** (Figure 7): Topics involving social justice and geopolitics, such as *Immigration* or *Human Rights*, show higher neutrality, indicating that multilingual LLMs may tend to hedge or neutralize emotionally sensitive content.

Beyond average sentiment, we introduce a *language sensitivity score* to quantify cross-language divergence for each example. For each `(topic, example)` pair, we treat the three-dimensional sentiment scores from five languages as vectors and compute all pairwise Euclidean distances (10 pairs). Their sum represents the emotional divergence for that example.

By averaging these values per topic, we generate a topic-level sensitivity bar chart (Figure 8). We observe that **Environmental Issues** and **Cultural Events** show the highest cross-language variation, suggesting greater cultural or linguistic framing effects. In contrast, **Technological Advancements** show the lowest sensitivity, implying more uniform interpretations across languages.

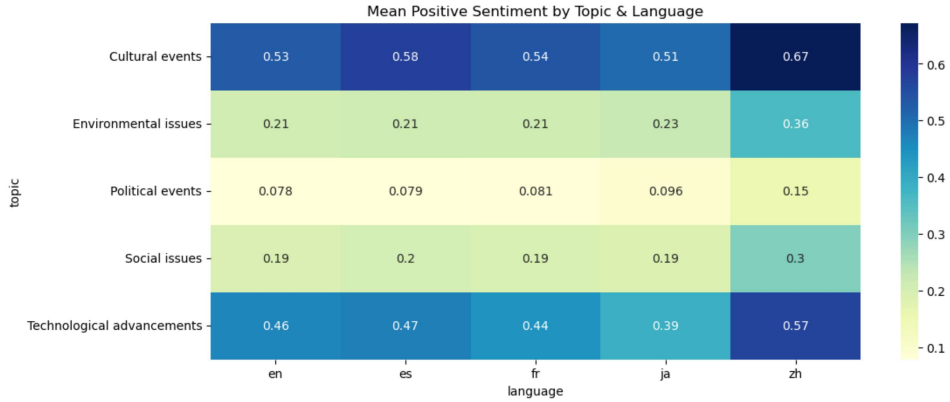

Figure 5: Mean Positive Sentiment by Topic & Language. Cultural and technological topics elicit high positive sentiment, especially in zh and es.

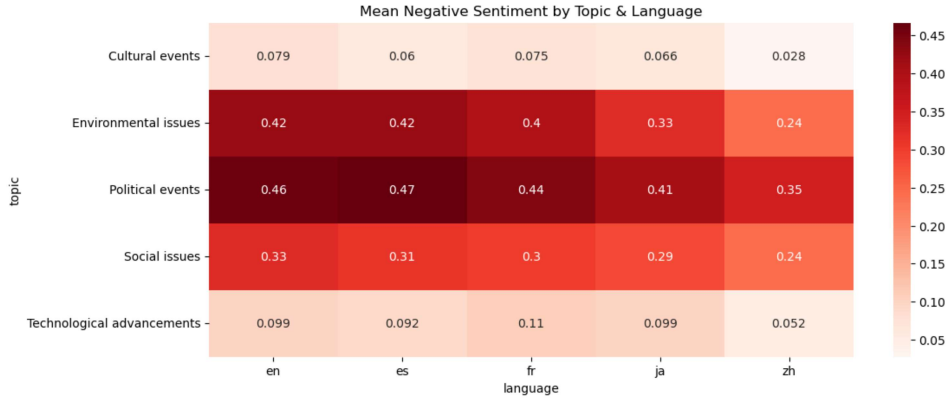

Figure 6: Mean Negative Sentiment by Topic & Language. Political and environmental issues yield the most negative sentiment, especially in en and fr.

To visualize how sensitive each example is across languages, we generate a topic-colored word cloud (Figure 9). In this visualization, the font size of each example corresponds to its cross-language divergence score, while the color reflects its topic domain. Notably, examples such as *Endangered species protection*, *Renewable energy adoption*, and *Music and arts funding* emerge as highly variable across languages. This suggests that even within the same topic, different examples elicit differing levels of multilingual disagreement—highlighting the fine-grained nature of LLM behavior.

7 Challenges and Solutions

- **Language Scope Optimization:** Initially, our datasets and models supported a wide range of languages—sometimes up to a hundred—but the overlap across datasets was incomplete. Only six languages were consistently represented in all sources. To maintain consistency and manageability, we refined our dataset selection to include only the target languages shared across the board.
- **Dataset Accessibility:** The SAFARI dataset, cited in Azizov et al. (2024), was labeled as publicly accessible; however, we were unable to retrieve it and received no reply from the authors. Alternative public datasets were also found to be incompatible with our needs. As a solution, we developed our own datasets for fine-tuning and evaluation by scraping diverse news websites. Due to a power outage in Spain during this period, the Spanish segment initially contained broken or missing links and was excluded from the figures in Section 6.1. (The issue has since been resolved for the final submission.)

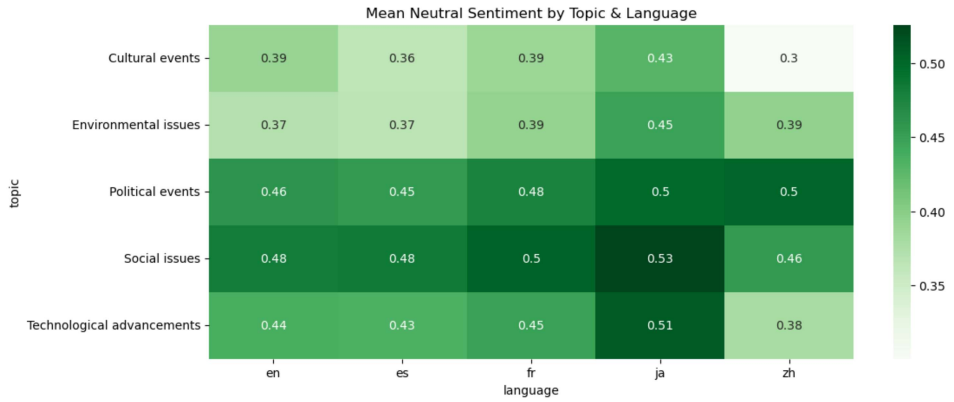

Figure 7: Mean Neutral Sentiment by Topic & Language. Social and political topics result in more neutral outputs, especially in ja and fr.

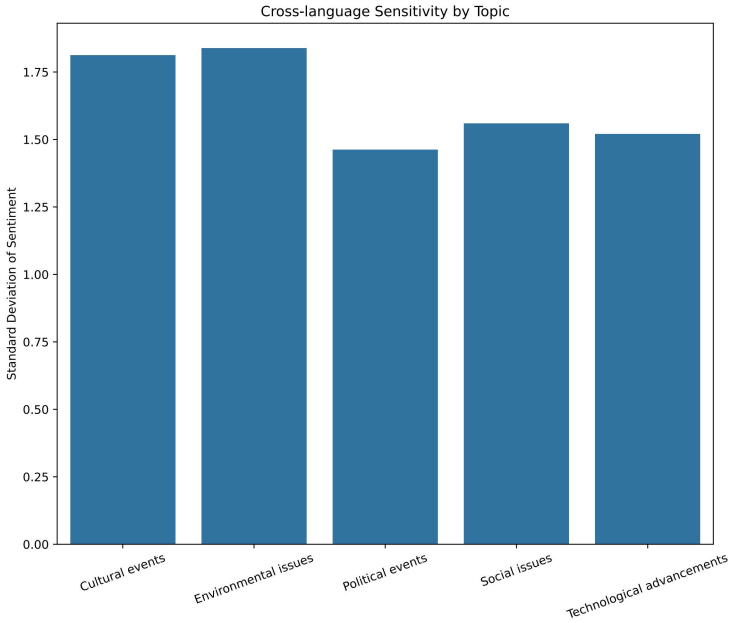

Figure 8: Cross-language Sensitivity by Topic. Measured as the average pairwise Euclidean distance of sentiment vectors across five languages. Environmental and cultural topics show highest variance.

- **Limitations of Smaller Baseline Models:** Although mT5 remains a commonly used model, it is relatively outdated, and many of its supporting libraries are no longer functional. Its baseline performance was also insufficient for our purposes. We experimented with other options such as BLOOMZ-1B and XGLM, but ultimately selected Qwen and LLaMA3-8B-multilingual as our main models. The results presented in this paper are based on Qwen.

- **Back-Translation for Sentiment Consistency:** To ensure fair and standardized sentiment evaluation across languages, we translated all multilingual outputs into English before performing sentiment analysis. This approach helps mitigate language-specific biases during scoring. To preserve the original emotional tone of each response, we utilized the Google Translation API, known for its reliable and consistent multilingual translation quality.

Figure 9: Topic-colored word cloud of example-level cross-language sensitivity. Font size reflects divergence across languages; color denotes topic category.

- **Learning Curve in Multilingual Model Fine-Tuning:** As this was our initial experience with fine-tuning multilingual LLMs and working with associated evaluation techniques, we faced a steep learning curve. Nonetheless, a thorough review of existing literature equipped us with the necessary knowledge to approach this task effectively.

8 Conclusion

This work highlights the capability of fine-tuned multilingual LLMs in effectively capturing cross-linguistic media sentiments and performing sentiment analysis across a wide range of languages and topics. By employing Qwen-2.5-3B for text generation and integrating a RoBERTa-based sentiment classifier within a modular framework, we achieve notable improvements in aligning generated sentiments with real-world news content. This approach is particularly effective in addressing the baseline models' tendency to produce neutral and hedged responses. Our experimental results demonstrate that fine-tuning significantly boosts both accuracy and sentiment coherence, with the greatest improvements observed in non-English language outputs. Furthermore, our topic-level analysis uncovers consistent and complex variations in emotional tone across languages, emphasizing the influence of cultural and linguistic factors in shaping media narratives.

For future work, this framework could be extended by incorporating additional languages and investigating how sentiment evolves over time in response to changing news events. Another valuable research direction involves examining the inherent biases of LLMs across different languages, assessing whether identical prompts lead to varied outputs due to imbalances in training data. Finally, expanding this system to process multimodal content or region-specific dialects may further enhance its effectiveness for real-time media tracking and policy impact evaluation.

References

Dilshod Azizov, Zain Muhammad Mujahid, Hilal AlQuabeh, Preslav Nakov, and Shangsong Liang. SAFARI: Cross-lingual bias and factuality detection in news media and news articles. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12217–12231, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.712. URL https://aclanthology.org/2024.findings-emnlp.712/.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020. URL https://arxiv.org/abs/2010.12421.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Rahim Dehkharghani, Berrin Yanikoglu, Dilek Tapucu, and Yucel Saygin. Adaptation and use of subjectivity lexicons for domain dependent sentiment classification. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 669–673, 2012. doi: 10.1109/ICDMW.2012.121.

Ankita Maity, Anubhav Sharma, Rudra Dhar, Tushar Abhishek, Manish Gupta, and Vasudeva Varma. Multilingual bias detection and mitigation for indian languages, 2023. URL https://arxiv.org/abs/2312.15181.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Allan Sales, Albin Zehe, Leandro Balby Marinho, Adriano Veloso, Andreas Hotho, and Janna Omeliyanenko. Assessing media bias in cross-linguistic and cross-national populations. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):561–572, May 2021. doi: 10.1609/icwsm.v15i1.18084. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18084.

Francielle Vargas, Kokil Jaidka, Thiago Pardo, and Fabrício Benevenuto. Predicting sentence-level factuality of news and bias of media outlets. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 1197–1206, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria. URL https://aclanthology.org/2023.ranlp-1.127/.

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. doi: 10.1126/science.aap9559. URL https://www.science.org/doi/abs/10.1126/science.aap9559.

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark, 2023. URL https://arxiv.org/abs/2306.07902.