Predicting cross-linguistic media reaction and sentiment with LLMs

Zihan Zhao, Shunqiang Feng

04/28/2025

Motivation

- Media Diversity
  - Media coverage differs across languages and cultures, reflecting regional biases and political framing.
- Information Spread
  - Misleading and emotionally charged content spreads faster than factual reporting (Vosoughi et al., 2018)
- Public Communication
  - Public figures must anticipate media framing to ensure accurate and fair communication

Background

- Cross-Linguistic Sentiment Analysis Gaps
  - A substantial body of work explored sentiment analysis
    - Politics
    - Social Media
    - Product Reviews
  - Most focus on just English content
    - This limitation is particularly significant in an increasingly globalized world, especially in contexts with numerous multilingual entities and countries, such as the European Union and India.

Related Work

Significant advancements in large language models (LLMs) like GPT [1], Llama [2], and Deepseek [3] have marked progress in machine learning and NLP

Breakthrough in LLMs

Sentiment Analysis Challenges

Sentiment analysis, a key LLM research area, is particularly challenging in multilingual contexts due to cultural and political influences.

Traditional approaches rely on hand-crafted features and rules, struggling to capture complex emotions and adapt across diverse cultural contexts [4].

Advancements with Multi-Language LLM

Recent multi-language LLM advancements, such as XGLM[7], XLM-R[8], and mT5[9], have enhanced cross-lingual performance through pre-training and fine-tuning on diverse multilingual datasets.

Cross-Lingual Media Sentiment

Studies like Azizov et al. [5] provide datasets to analyze political bias in ten languages using MPLMs and zero-shot LLMs, while Sales et al. (2021) [6] focus on bias in four languages with smaller-scale methods like translated lexicons.

Claim / Target Task

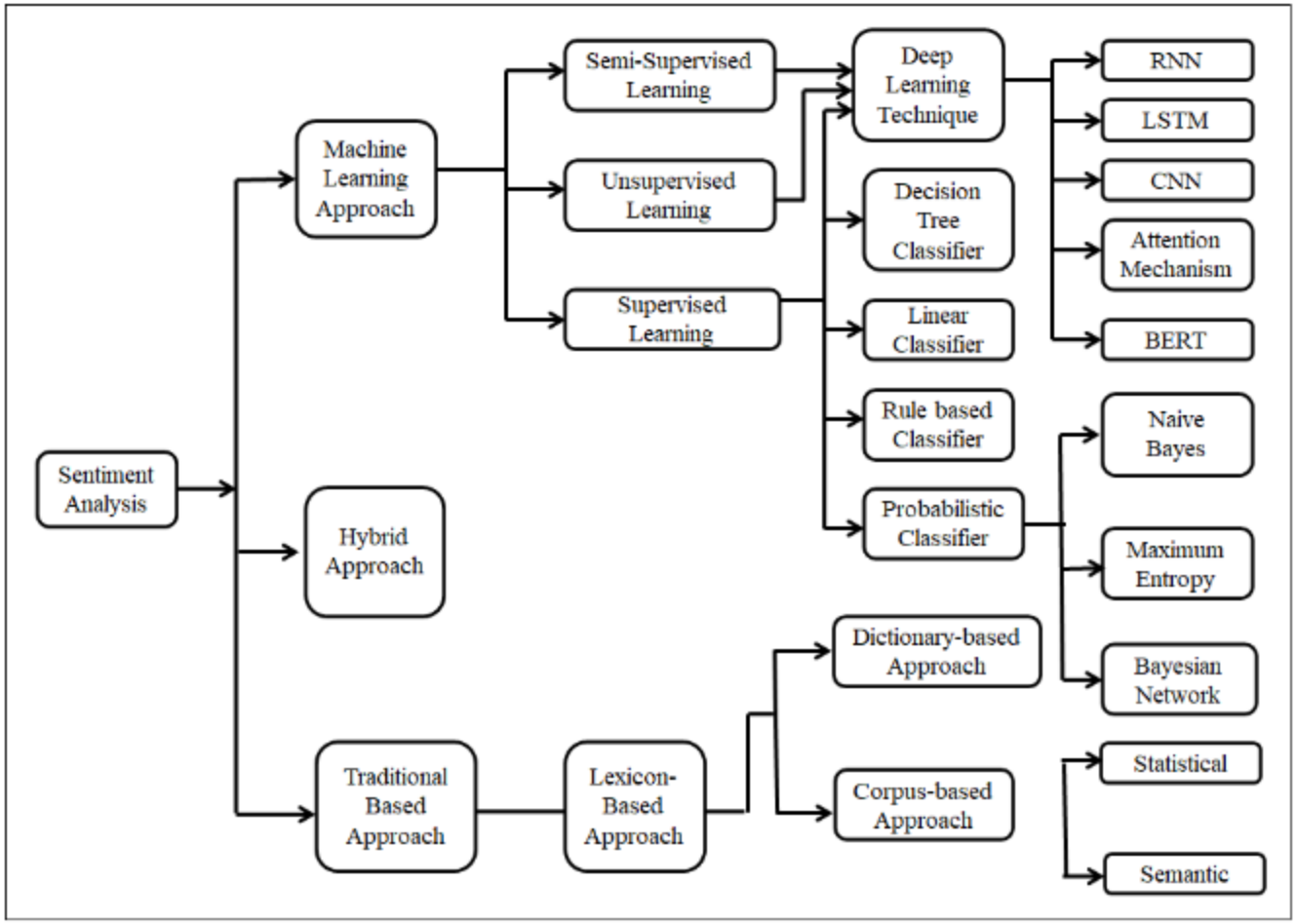
- **Target**

Our goal is to explore the **intrinsic differences between languages** in multilingual LLM, which is characterized through sentiment analysis of the multilingual text generated by these models.

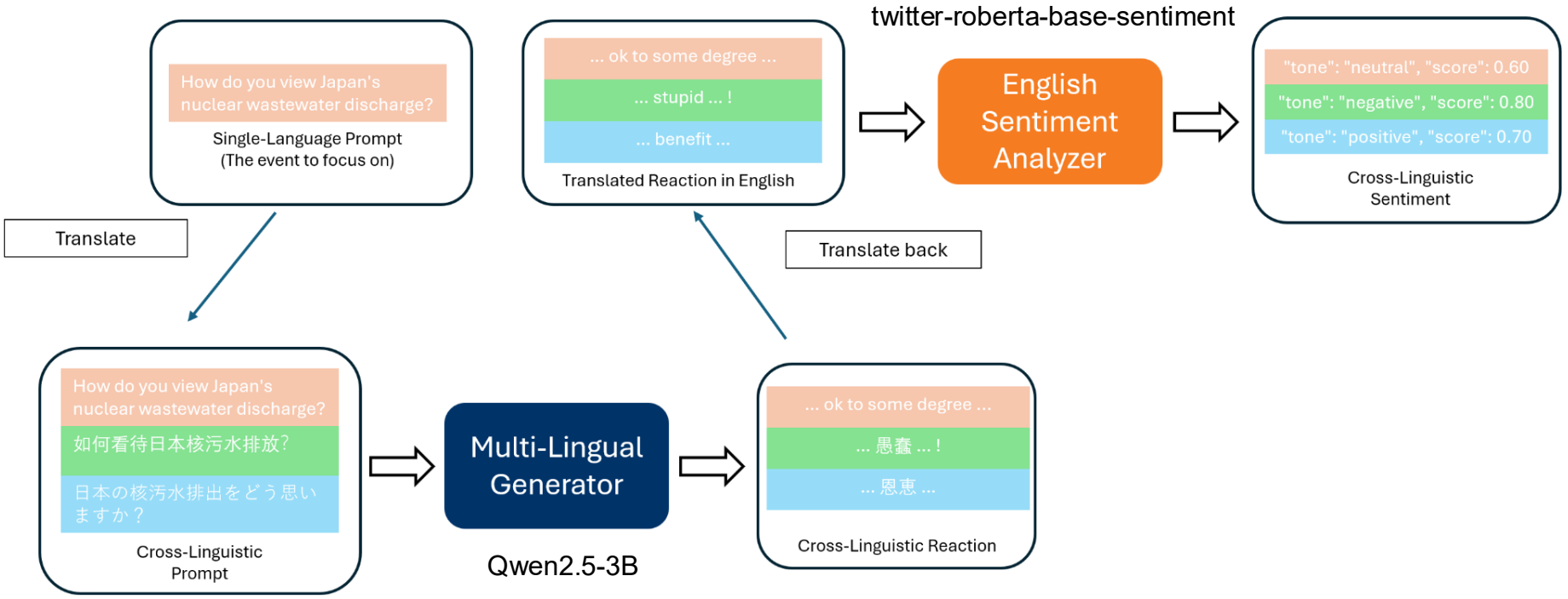
- **Questions to Answer**

1. Do intrinsic differences exist between languages in multilingual LLMs?
2. If differences exist, in which areas are they more sensitive, and where do larger disparities occur?
3. If differences exist, what are the fundamental **reasons** behind them?
4. …

An Intuitive Figure Showing WHY Claim



Proposed Solution



Implementation

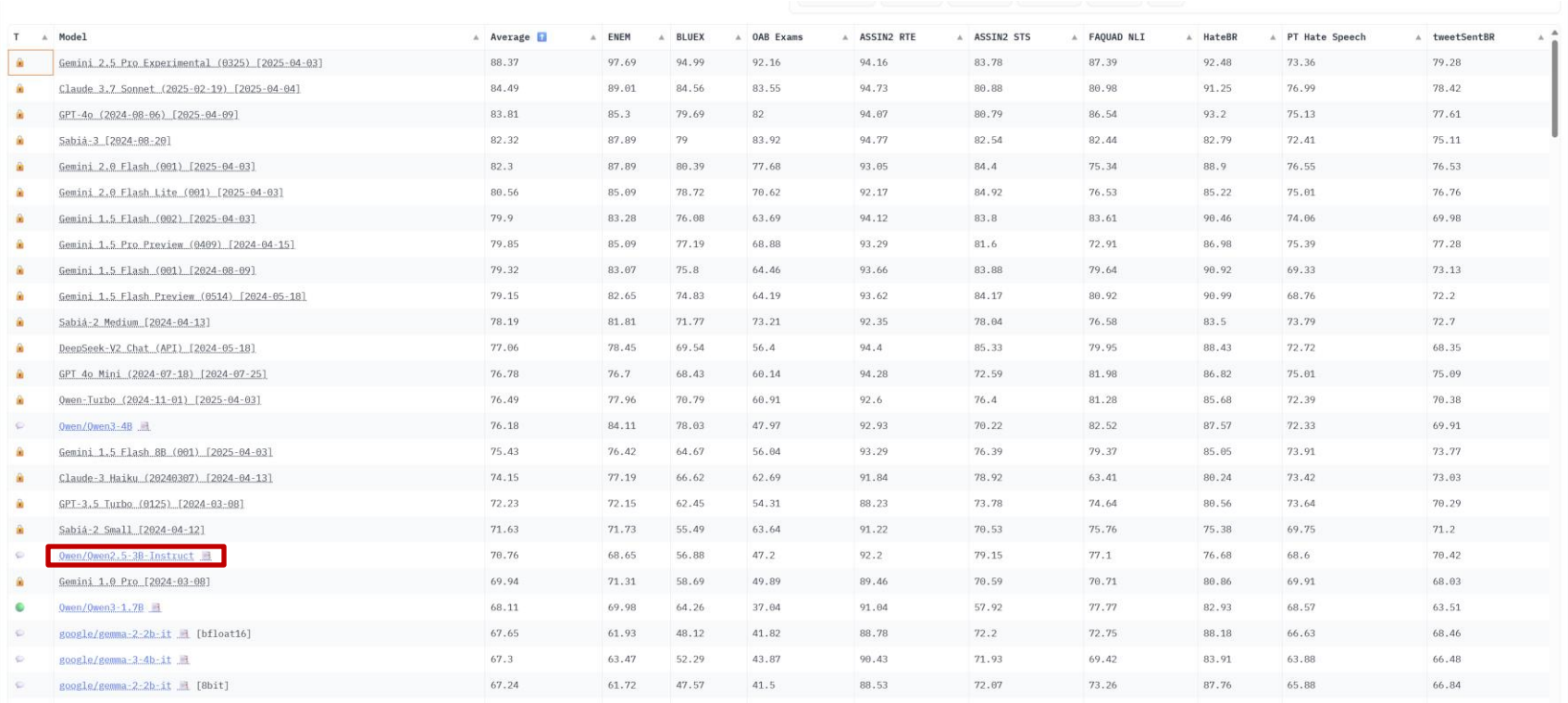
- Translation
  - Ensure consistent meaning across languages for input queries and unify sentiment analysis for different language outputs.

Translation 1	Translate the English prompt (baseline) into multiple languages.
Translation 2	Translate output reactions from different languages back into English for sentiment analysis.

  - We chose Google Translate API for reducing training costs caused by excessive LLM use.
- Generator [Fine-Tune]
  - **Models Explored:**
    - mT5-base
    - Llama-3-8B
    - …
  - **Final Model Chosen: XGLM-7.5B**
  - **Selection Criteria:**
    - Model performance
    - Language coverage
    - Training cost
- Sentiment Analyzer [Pre-Train]
  - **DistilBERT**

Implementation [Update]

- Generator [Fine-Tune]
  - **Final Model Chosen: XGLM-7.5B → Qwen2.5-3B**
  - **Why:**
    - XGLM was proposed in 2021, and the training support is not up to time [A100 is not enough…].
    - Qwen2.5 was introduced in 2024, with a smaller model size, better training support, and lower computational cost.
    - Qwen2.5 is powerful!



T	Model	Average	ENEM	BLUEX	OAB Exams	ASSIN2 RTE	ASSIN2 STS	FAQUAD NLI	HateBR	PT Hate Speech	tweetSentBR
	Gemini 2.5 Pro Experimental (0325) [2025-04-03]	88.37	97.69	94.99	92.16	94.16	83.78	87.39	92.48	73.36	79.28
	Claude 3.7 Sonnet (2025-02-19) [2025-04-04]	84.49	89.01	84.56	83.55	94.73	80.88	80.98	91.25	76.99	78.42
	GPT-4o (2024-08-06) [2025-04-09]	83.81	85.3	79.69	82	94.07	80.79	86.54	93.2	75.13	77.61
	Sabiá-3 [2024-08-20]	82.32	87.89	79	83.92	94.77	82.54	82.44	82.79	72.41	75.11
	Gemini 2.0 Flash (001) [2025-04-03]	82.3	87.89	80.39	77.68	93.05	84.4	75.34	88.9	76.55	76.53
	Gemini 2.0 Flash Lite (001) [2025-04-03]	80.56	85.09	78.72	70.62	92.17	84.92	76.53	85.22	75.01	76.76
	Gemini 1.5 Flash (002) [2025-04-03]	79.9	83.28	76.08	63.69	94.12	83.8	83.61	90.46	74.06	69.98
	Gemini 1.5 Pro Preview (0409) [2024-04-15]	79.85	85.09	77.19	68.88	93.29	81.6	72.91	86.98	75.39	77.28
	Gemini 1.5 Flash (001) [2024-08-09]	79.32	83.07	75.8	64.46	93.66	83.88	79.64	90.92	69.33	73.13
	Gemini 1.5 Flash Preview (0514) [2024-05-18]	79.15	82.65	74.83	64.19	93.62	84.17	80.92	90.99	68.76	72.2
	Sabiá-2 Medium [2024-04-13]	78.19	81.81	71.77	73.21	92.35	78.04	76.58	83.5	73.79	72.7
	DeepSeek-V2 Chat (API) [2024-05-18]	77.06	78.45	69.54	56.4	94.4	85.33	79.95	88.43	72.72	68.35
	GPT 4o Mini (2024-07-18) [2024-07-25]	76.78	76.7	68.43	60.14	94.28	72.59	81.98	86.82	75.01	75.09
	Qwen-Turbo (2024-11-01) [2025-04-03]	76.49	77.96	70.79	60.91	92.6	76.4	81.28	85.68	72.39	70.38
	Qwen/Qwen3-4B	76.18	84.11	78.03	47.97	92.93	70.22	82.52	87.57	72.33	69.91
	Gemini 1.5 Flash 8B (001) [2025-04-03]	75.43	76.42	64.67	56.04	93.29	76.39	79.37	85.05	73.91	73.77
	Claude-3 Haiku (20240307) [2024-04-13]	74.15	77.19	66.62	62.69	91.84	78.92	63.41	80.24	73.42	73.03
	GPT-3.5 Turbo (0125) [2024-03-08]	72.23	72.15	62.45	54.31	88.23	73.78	74.64	80.56	73.64	70.29
	Sabiá-2 Small [2024-04-12]	71.63	71.73	55.49	63.64	91.22	70.53	75.76	75.38	69.75	71.2
	Qwen/Qwen2.5-3B-Instruct	70.76	68.65	56.88	47.2	92.2	79.15	77.1	76.68	68.6	70.42
	Gemini 1.0 Pro [2024-03-08]	69.94	71.31	58.69	49.89	89.46	70.59	70.71	80.86	69.91	68.03
	Qwen/Qwen3-1.7B	68.11	69.98	64.26	37.04	91.04	57.92	77.77	82.93	68.57	63.51
	google/gemma-2-2b-it [bfloat16]	67.65	61.93	48.12	41.82	88.78	72.2	72.75	88.18	66.63	68.46
	google/gemma-3-4b-it	67.3	63.47	52.29	43.87	90.43	71.93	69.42	83.91	63.88	66.48
	google/gemma-2-2b-it [8bit]	67.24	61.72	47.57	41.5	88.53	72.07	73.26	87.76	65.88	66.84

LLM leaderboard

Implementation [Update]

- Template
  - Fine-Tuning

```
        messages = [
            {
                "role": "system",
                "content": "You are Qwen. You are a helpful assistant and expected to write a news report with the given title."
            },
            {"role": "user", "content": title}
        ]
```

  - Inference

```
        messages = [
        {
            "role": "system",
            "content": "You are Qwen. You are a helpful assistant and expected to write a 300 words news report with the given topic."
        },
        {"role": "user", "content": prompt}
    ]
```

- Fine-Tuning Setting
  - Datasets of 5 languages in total
  - 8 Epochs
  - A100 * 0.5 day

Implementation [Update]

- Sentiment Analyzer [publicly available model]

1. **twitter-roberta-base-sentiment**
    1. a RoBERTa-base model trained on ~58M tweets and fine-tuned for sentiment analysis using the TweetEval benchmark.
2. More Parameters → More Powerful
3. Multi-Class Output for further analysis.

Output:

```
1) positive 0.8466
2) neutral 0.1458
3) negative 0.0076
```

Data Summary

- Web Crawler
  - URL
  - Publish Time
  - Title
  - Content

Language	News Source	Size
Chinese	Xinhua, Sina, People …	21109
English	CNN, BBC, The Guardian …	19436
Spanish	Elmundo, abc, publico …	15399
Japan	Yomiuri, nikkei, tokyo-np …	10757
French	Liberation, francesoir, bfmtv …	20942

Experimental Results

- Multilingual text input

query = "About Japan's nuclear wastewater discharge, I think"

- Model output

```
Loading checkpoint shards: 100%|██████████| 7/7 [00:20<00:00,  2.96s/it]
Device set to use cuda:0
==> 1. The original input is <==
en: About Japan's nuclear wasterwater discharge, I think
zh: 关于日本核废水排放，我认为
ja: 日本の核廃棄物排出については、
fr: À propos du rejet des eaux usées nucléaires du Japon, je pense
es: Sobre el vertido de aguas residuales nucleares de Japón, creo que...
==> 2. The generated text is <==
en: About Japan's nuclear wasterwater discharge, I think it's wrong to think that we're not releasing any radioactive waste into the sea, since there are so many spent fuel pools in Fukushima-Daiichi (I just can't call the oth
zh: 关于日本核废水排放,我认为,首先要搞清楚日本是如何把核废水处理掉的。日本为了让核废水尽可能地得到有效地利用,在核电站的处理设施上安装了很多自动化的设备。核废水的处理是通过分离,沉淀,除去水中杂质来实现的。日本目前还有五座核电站,全部运
ja: 日本の核廃棄物排出については、核兵器廃棄物保管施設として東芝の山口県熊毛郡上関町の山口廃棄物センター(仮称)が建設される予定だが、ここに日本の核廃棄物が最終的に運ばれることには異論の余地はない。
fr: À propos du rejet des eaux usées nucléaires du Japon, je pense que la France, comme la plupart des pays européens, a des centrales nucléaires à haut risque. Ce qui, bien sûr, ne veut pas dire qu'on n'a pas besoin d'avoir d
es: Sobre el vertido de aguas residuales nucleares de Japón, creo que... no es tan grave (si se pueden salvar los reactores a escala mundial) y puede pasar como pasó con el accidente de Chernobyl (que sí, es un desastre nuclea
==> 3. The generated text in English is <==
en: About Japan's nuclear wasterwater discharge, I think it's wrong to think that we're not releasing any radioactive waste into the sea, since there are so many spent fuel pools in Fukushima-Daiichi (I just can't call the oth
zh: Regarding the discharge of nuclear wastewater from Japan, I think we should first understand how Japan treats the wastewater. In order to make the wastewater as effective as possible, Japan has installed a lot of automated
ja: Regarding Japan&#39;s nuclear waste disposal, Toshiba plans to build the Yamaguchi Waste Center (tentative name) in Kaminoseki Town, Kumage County, Yamaguchi Prefecture as a nuclear weapons waste storage facility, and ther
fr: Regarding the discharge of nuclear wastewater from Japan, I think that France, like most European countries, has high-risk nuclear power plants. Which, of course, does not mean that we don&#39;t need to have nuclear power
es: Regarding the Japanese nuclear wastewater spill, I think... it&#39;s not that serious (if reactors can be saved worldwide) and it could happen again like the Chernobyl accident (which, yes, is a nuclear disaster of giganti
==> 4. The sentiment analyse result is <==
en: {'label': 'NEGATIVE', 'score': 0.996428906917572}
zh: {'label': 'NEGATIVE', 'score': 0.8066835403442383}
ja: {'label': 'NEGATIVE', 'score': 0.9895998239517212}
fr: {'label': 'NEGATIVE', 'score': 0.9874590635299683}
es: {'label': 'POSITIVE', 'score': 0.9934046864509583}
```

Experimental Analysis

- There can be significant differences in the multilingual understanding of the same news event
  - Culture
  - History
  - Geopolitics
- Limitations
  - Google Translator Bias
  - Sentiment Analyzer Bias
  - Limited Sentiment Categories
    - Positive/Negative/Neutral

Experimental Analysis [Update]

- Collect 5 topics, each topic with 10 samples

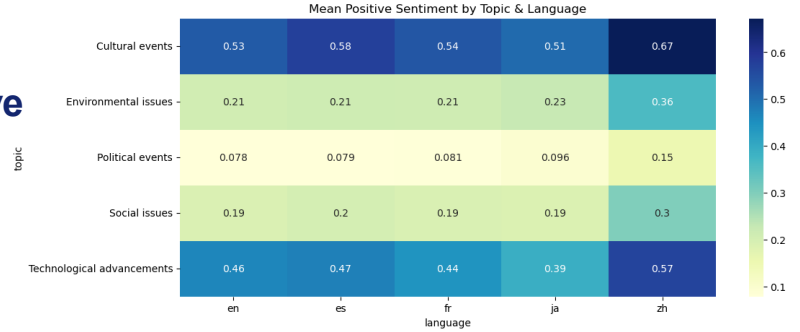
Environmental issues	Political events	Social issues	Technological advancements	Cultural events
Japan's nuclear wastewater discharge	U.S. presidential election	Gender equality movements	AI regulation	Olympic Games hosting
Deforestation in the Amazon	Brexit negotiations	Immigration policies	Cryptocurrency adoption	Cultural heritage preservation
Plastic pollution in oceans	Trade wars between major economies	Racial discrimination and protests	Autonomous vehicles and safety	International film festivals
Climate change and global warming	Political unrest in the Middle East	LGBTQ+ rights and acceptance	Data privacy and protection	Music and arts funding
Renewable energy adoption	Rise of populism in Europe	Income inequality and wealth distribution	5G network expansion	Language preservation efforts
Air pollution in major cities	Cybersecurity and election interference	Access to education in developing countries	Biotechnology and genetic engineering	Traditional cuisine and globalization
Water scarcity and drought	International sanctions and their impacts	Healthcare system reforms	Space exploration and colonization	Religious festivals and their significance
Endangered species protection	Diplomatic relations between China and the U.S.	Mental health awareness	Quantum computing developments	Museum exhibitions and historical artifacts
Sustainable agriculture practices	Human rights issues in authoritarian regimes	Aging population and elder care	Internet of Things (IoT) and smart cities	Literature and book fairs
Carbon emissions reduction	Global governance and the United Nations	Youth unemployment and job market challenges	Virtual and augmented reality applications	Fashion industry trends and sustainability

- For each example, run 25 times

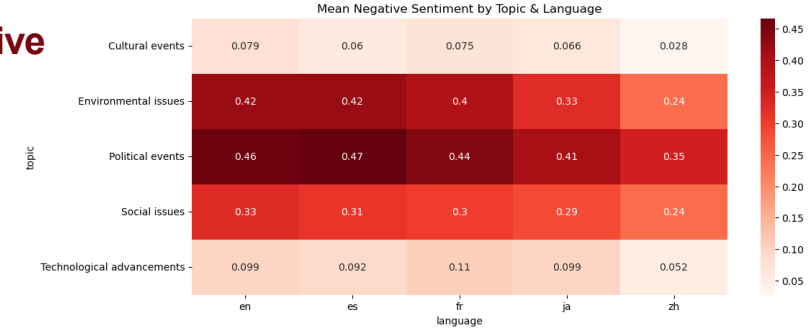
Experimental Analysis [Update]

- Topic Level Analysis

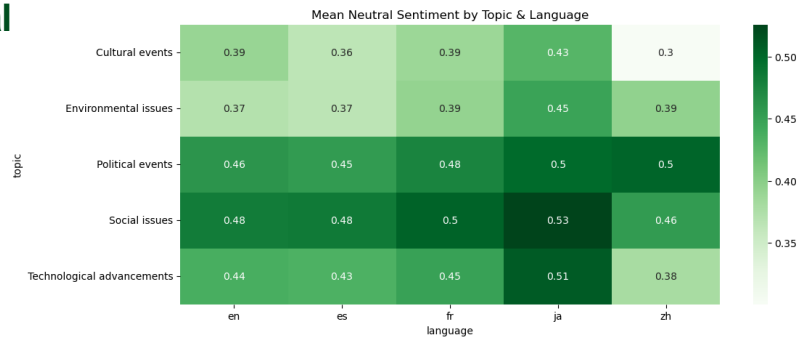
Positive



Negative



Neutral



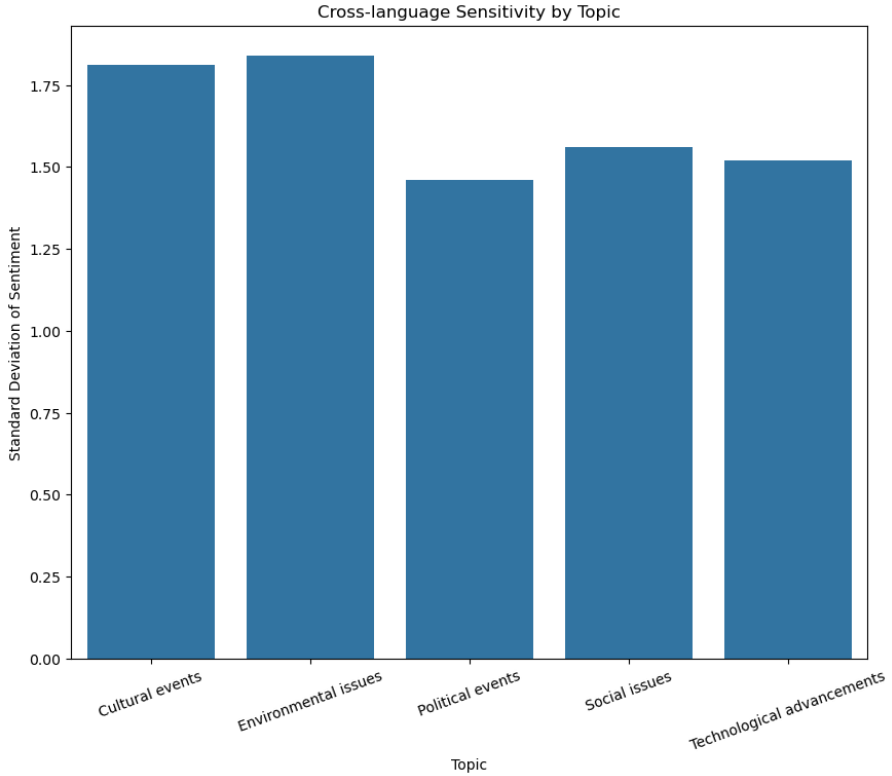
- **Cultural events** received the **most positive sentiment** across all languages, especially in **Chinese (0.67)**.
- **Political events** showed the **highest negative sentiment** and **lowest positive sentiment** in all languages.
- **Technological advancements** were viewed **positively and consistently**, with low negativity.
- **Environmental and social issues** had **mixed or neutral sentiment**, indicating complexity or division.

Experimental Analysis [Update]

- Topic Sensitivity analysis

Measuring Topic Sensitivity via Cross-Language Tone Divergence

1. For each example, compute the average tone scores across 25 inferences per language.
2. Then calculate the sum of **Euclidean distances** between the five language tone vectors.
3. The total distance across all examples within a topic defines its **sensitivity score**.

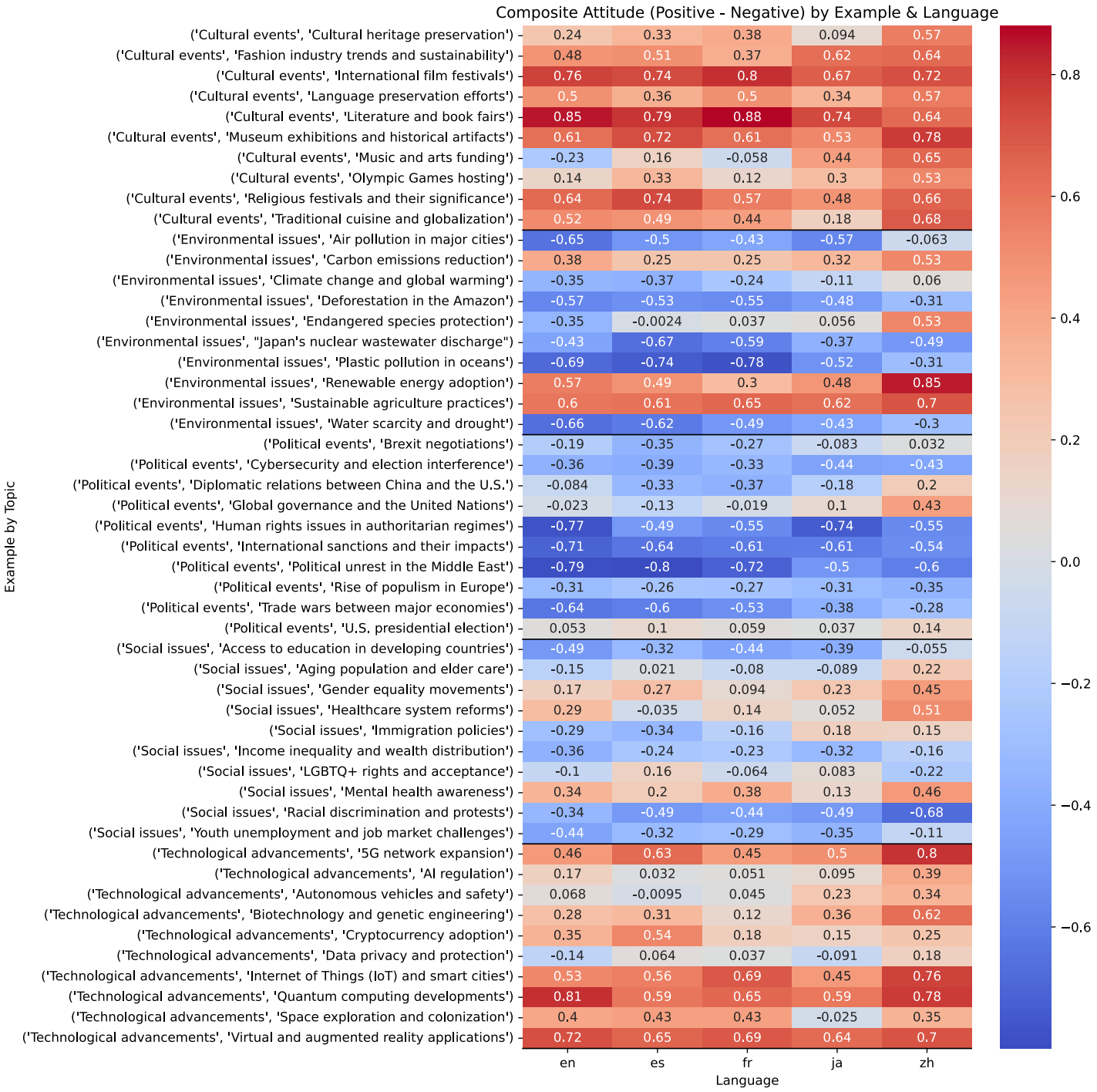


- **Cultural and Environmental Events** show the **highest cross-language sensitivity**, indicating strong variation in sentiment across languages—possibly due to cultural context or framing differences.
- **Political Events** have **lower sensitivity** than expected, suggesting that sentiment toward politics may be more **consistently perceived** across languages. [in fact, negative consistently]
- **Technological Advancements** show the **least variation**, implying a relatively **universal tone** in multilingual discourse on tech topics.

Experimental Analysis [Update]

- Example-level analysis

  - Composite Attitude
    1. For each example, compute the average tone scores of 25 inference
    2. composite attitude = mean positive – mean negative



[Update]

This heatmap shows the **net sentiment** for each example across five languages.

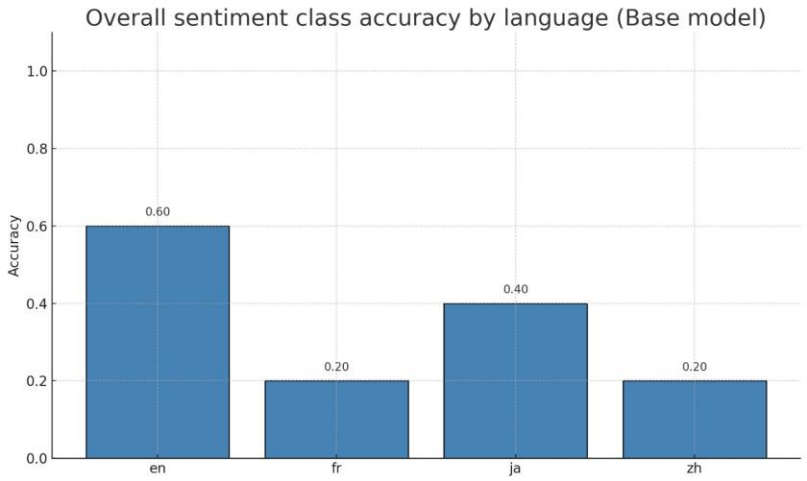
Redder values indicate **stronger positive attitudes**, while **bluer values** indicate **more negative perceptions**.

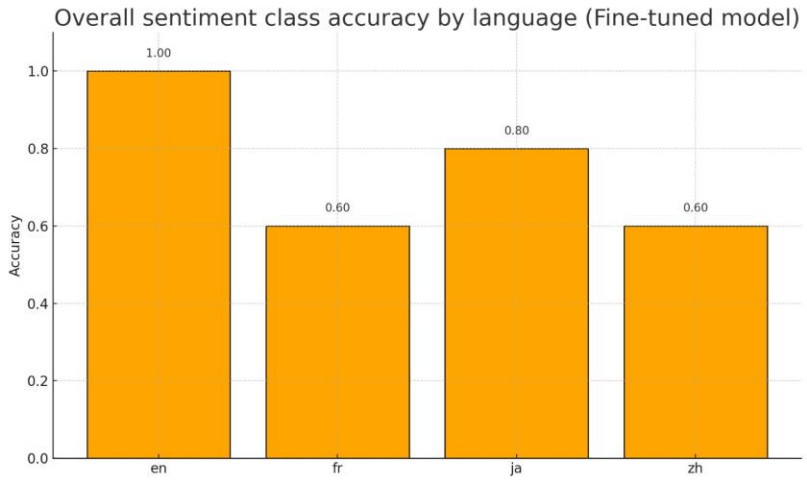
Notable variation across languages reveals **how different topics are emotionally framed** in different linguistic contexts.

Experimental Analysis [Update]

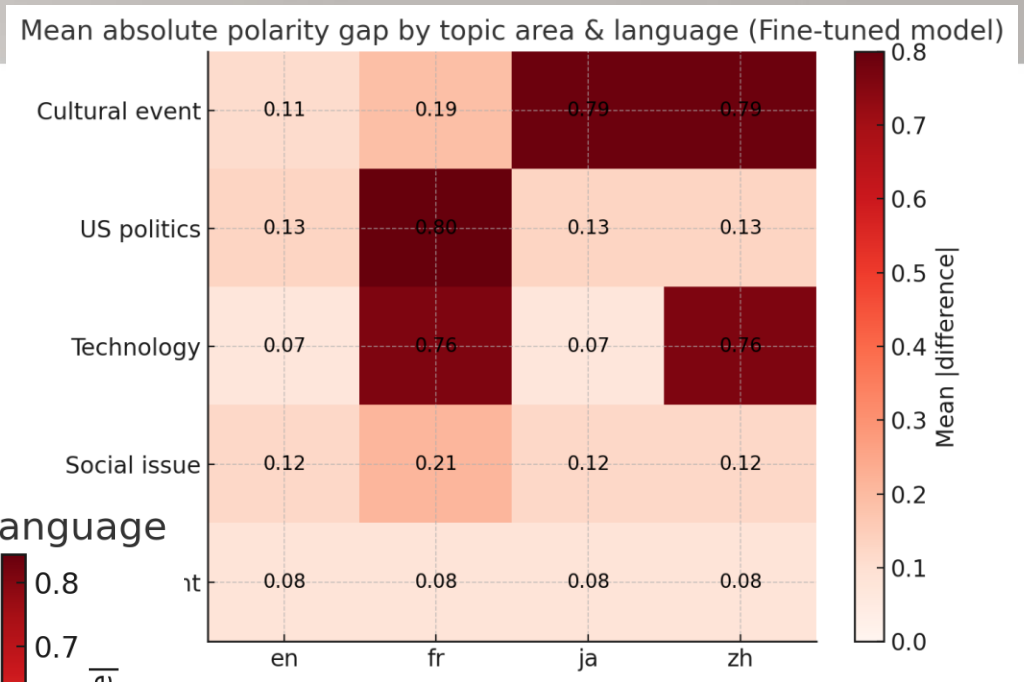
Word cloud showing the sensitivity of different examples

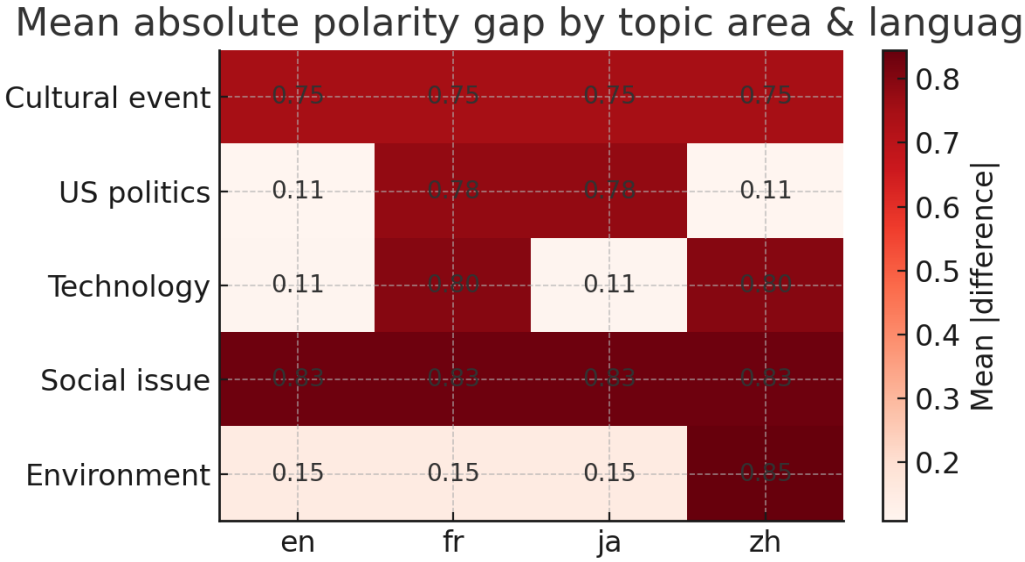
Effect of Fine Tuning [Update]





Effect of Fine Tuning [Update]





Conclusion and Future Work [Update]

- We Find:
  - Multilingual LLMs exhibit notable differences across languages, including issues like neutral hedging and divergent attitudes toward specific topics and examples;
  - Fine-tuning can improve model performance by reducing neutral hedging and producing responses that more closely reflect real-world perspectives;
  - The internal behavior of multilingual LLMs is highly complex; even under the same topic, different subtopics may trigger significant variations in sentiment across languages

Conclusion and Future Work

- We designed a multilingual generator-analyzer joint model aimed at analyzing the intrinsic sentiment differences within multilingual models
- Future Work
  - Subtle or ambiguous sentiment understanding
    - Nuanced expressions, like sarcasm, irony, etc.
  - Domain-specific sentiment
    - Meaning of "sick" in different contexts (gaming vs. healthcare)
  - Granularity of sentiment
    - Beyond "positive/negative/neutral" labeling

References

- [1] GPT-3: Its Nature, Scope, Limits, and Consequences
- [2] Llama 2: Open Foundation and Fine-Tuned Chat Models
- [3] DeepSeek-Coder: When the Large Language Model Meets Programming -- The Rise of Code Intelligence
- [4] Natural language processing: an introduction
- [5] SAFARI: Cross-lingual Bias and Factuality Detection in News Media and News Articles
- [6] Assessing media bias in cross-linguistic and cross-national populations.
- [7] Few-shot Learning with Multilingual Language Models
- [8] Unsupervised Cross-lingual Representation Learning at Scale
- [9] mT5: A massively multilingual pre-trained text-to-text transformer