Multilinguality and Cultural Awareness in Language Models

Jose Camacho-Collados

Cardiff NLP

CARDIFF UNIVERSITY

PRIFYSGOL CAERDYDD

LoResLM

COLING 2025 • Abu Dhabi

Abu Dhabi, COLING LoResLM, 20 January 2025

About me

Professor at Cardiff University (Wales, UK)

Research interests: Mainly NLP, and in particular **semantics, resources, multilinguality, computational social science**

Co-author of “Embeddings in NLP”

Co-founder of the Cardiff NLP group

Cardiff NLP

Cardiff NLP

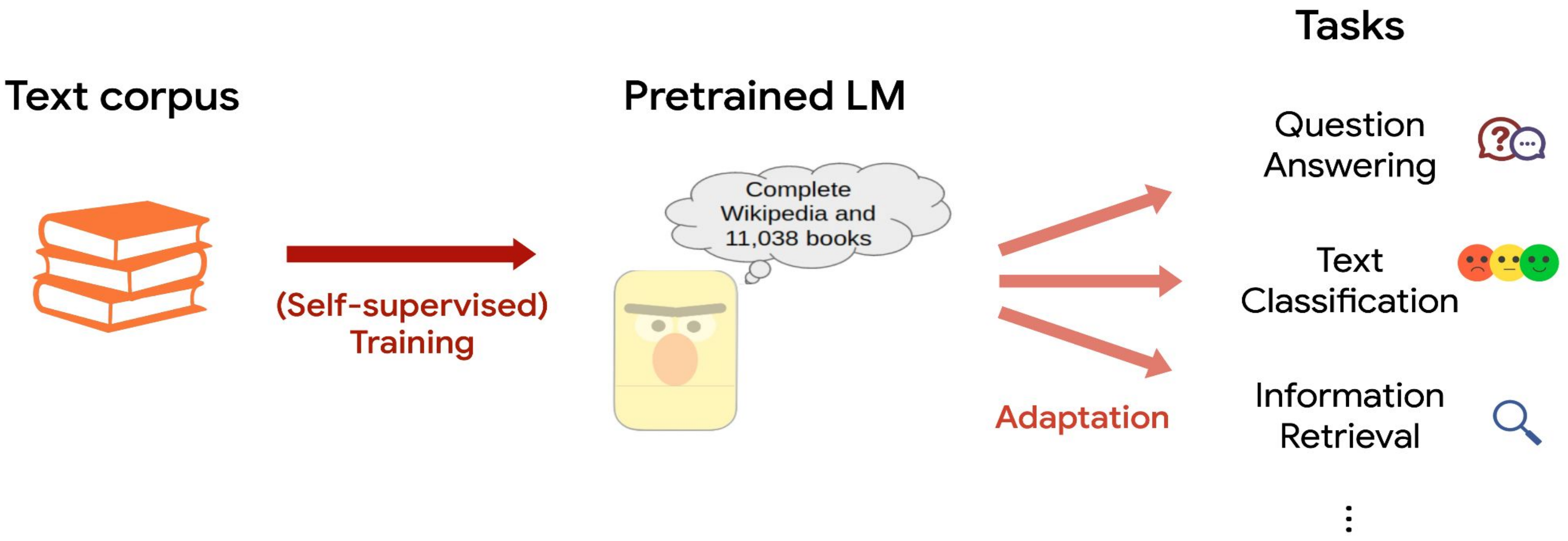
Research group at Cardiff University working on all aspects of NLP

- ➢ Young group (4 years old), growing fast (25+ lab members)
- ➢ **Website/Twitter:** cardiffnlp.github.io 🌐 @Cardiff_NLP
- ➢ **Activities:** seminars, workshops, hackathons, MSc NLP
- ➢ Interested in **multi-disciplinary** collaborations
- ➢ **Open-source** contributions

Today’s talk

- **Multilingual Language Models and Applications**
  - Examples
  - Opportunities (especially in computational social science)
- **Challenges/Issues**
  - Size
  - Language coverage
  - Cultural awareness

Language models: BERT, LLaMA, etc.



Text corpus

(Self-supervised) Training

Pretrained LM

Complete Wikipedia and 11,038 books

Adaptation

Tasks

Question Answering

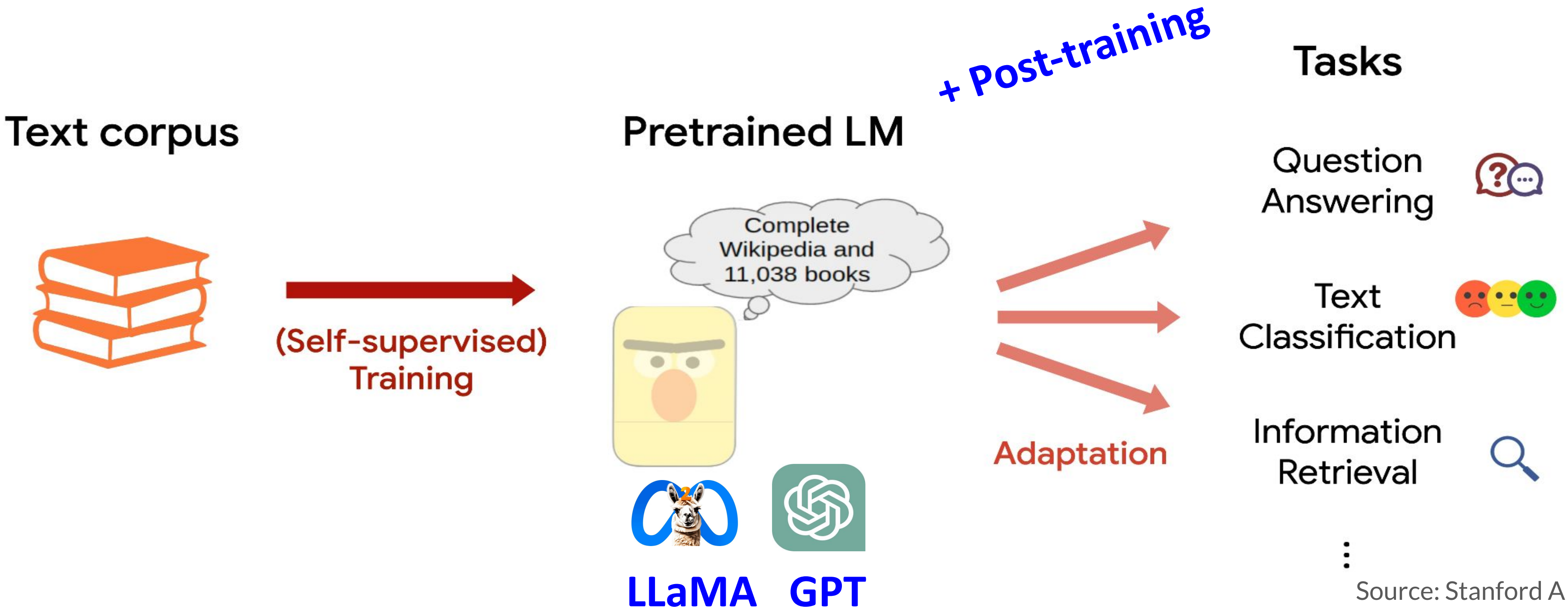
Text Classification

Information Retrieval

⋮

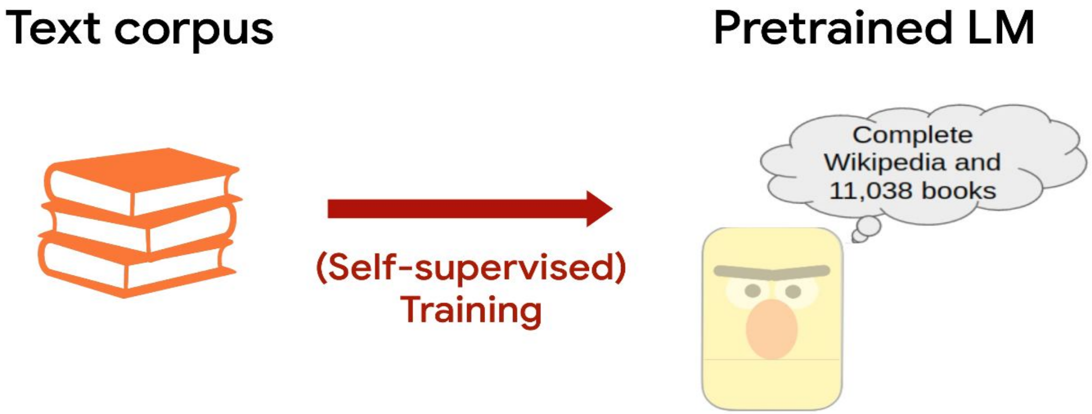
Source: Stanford AI

Language models: BERT, LLaMA, etc.



Source: Stanford AI

Multilingual Language Models



Text corpus

(Self-supervised) Training

Pretrained LM

Complete Wikipedia and 11,038 books

+ Post-training

mBERT, XLM-R and more recently LLMs such as **GPT-4, Mistral, Qwen**, etc.

If text corpus in many languages -> Multilingual language model

So, what are Multilingual Language Models?

Language models that can interpret (and generate) text in different languages.

Source image: Vermont Public

Most LLMs nowadays are (sort of) multilingual

GPT-4 understands most major languages (50+).

Mistral is “natively fluent in English, French, Spanish, German, and Italian”.

LLaMA 3.1 “supports Spanish, Portuguese, Italian, German, Thai, French and Hindi”.

Claude supports a “wide array of languages, including but not limited to English, French, German, Portuguese, Spanish, Japanese, Italian, Mandarin, Russian, Arabic, Hindi, and Korean”.

Qwen has a “multilingual support for over 29 languages”.

How/why do multilingual language models work?

How/why do multilingual language models work?

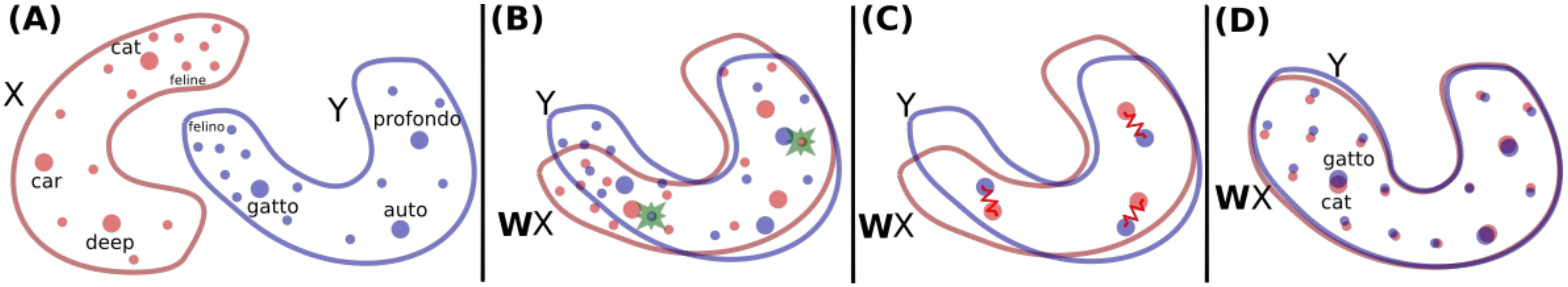
Magic!

Let’s get back a few years to the word embedding era…

Cross-lingual word embeddings

Cross-lingual word embeddings are vector spaces that include **words in different languages** - see Ruder et al. (JAIR 2019) if interested in knowing more.

🤯 They can be **learned from monolingual data** only! (Artetxe et al. ACL 2018, Conneau et al. ICLR 2018; inter alia)



Source image: Anastasopoulos

My motivation

Since these works, multilinguality and this research area has fascinated me.

How is it even possible to learn a multilingual embedding space without any parallel data? 🤯

At the same time, I was interested in **social media… which can be viewed as a multilingual corpus on its own!**

Source image: Anastasopoulos

Social media as a multilingual corpus

A very peculiar multilingual corpus: irregular vocabulary, informal language, code-switching, limited context…. And **emoji**! 😎🍕

Turns out that we can even learn cross-lingual word embeddings using just **emoji as anchors across languages**! 😂 (Camacho-Collados and Doval et al. ICWSM 2020)

Source image: Anastasopoulos

Social media as a multilingual corpus

A very peculiar multilingual corpus: irregular vocabulary, informal language, code-switching, limited context…. And **emoji**! 😎🍕

Turns out that we can even learn cross-lingual word embeddings using just **emoji as anchors across languages**! 😂 (Camacho-Collados and Doval et al. ICWSM 2020)

And… we can also learn **multilingual LMs**!

Source image: Anastasopoulos

Snap Inc.

XLM-T: A Multilingual Language Model Specialised on Social Media

(Barbieri, Espinosa-Anke and Camacho-Collados, LREC 2022)

We developed and trained a multilingual language model on Twitter.

Motivation:

There were no multilingual LMs specialised on social media - now there are others (e.g. Bernice; DeLucia et al. EMNLP 2022).

XLM-T: How it was trained

RoBERTA/XLM-R architecture (Conneau et al. 2020)

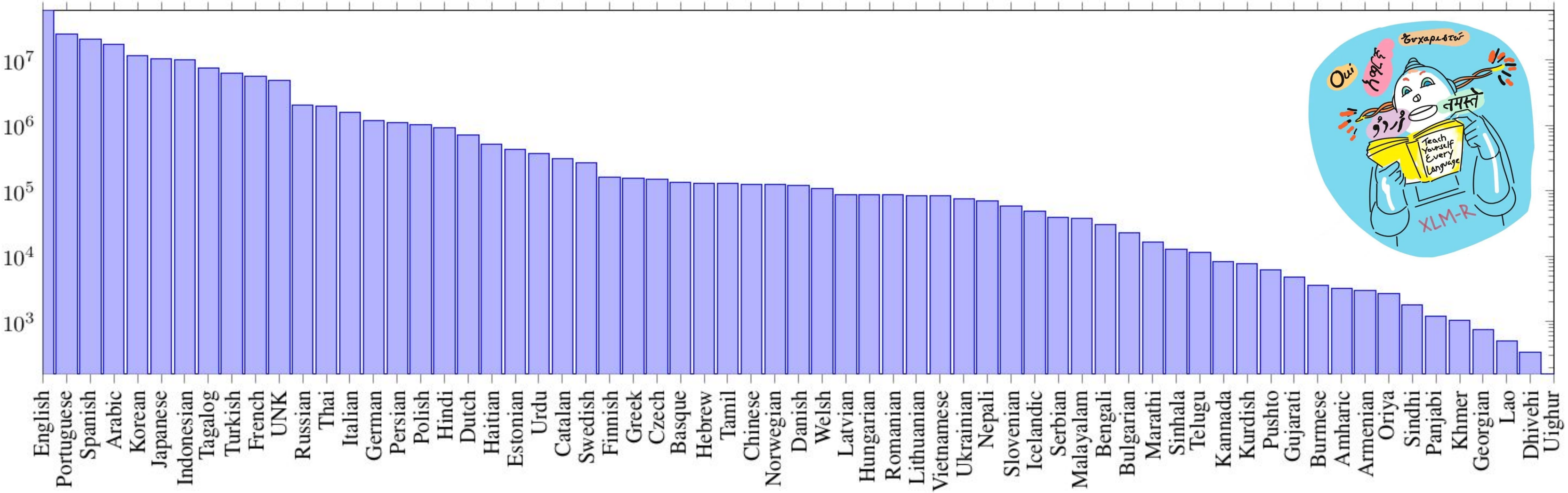
Use the XLM-R checkpoint (general-domain multilingual LM) as the initial reference

Then, continue training on millions of tweets from multiple languages

Source image: Grok

XLM-T: Training data

Distribution across languages (log-scale)



Some applications of XLM-T on social media

- ➢ Sentiment analysis
- ➢ Topic classification
- ➢ Hate speech detection
- ➢ Emoji prediction

….

Some applications of XLM-T on social media

- **Sentiment analysis**
- **Topic classification**
- Hate speech detection
- Emoji prediction

....

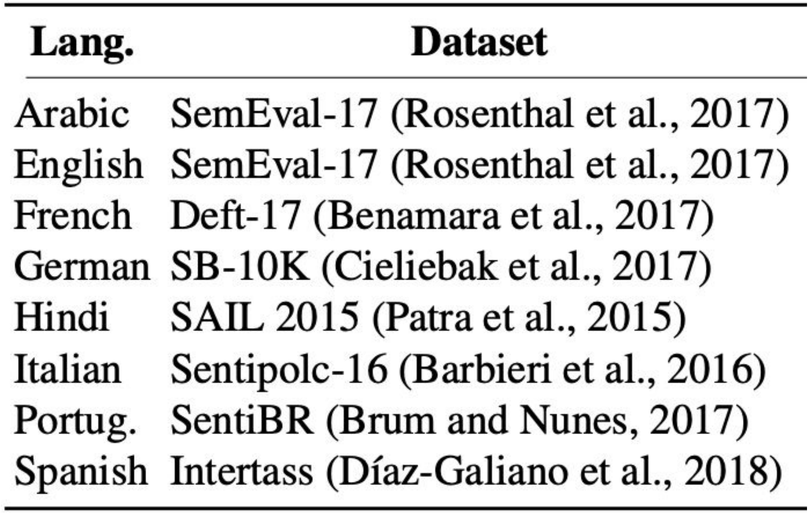
Sentiment analysis

A very popular task.

No **unified benchmark** on social media.

We collected and put together a unified benchmark for sentiment analysis.

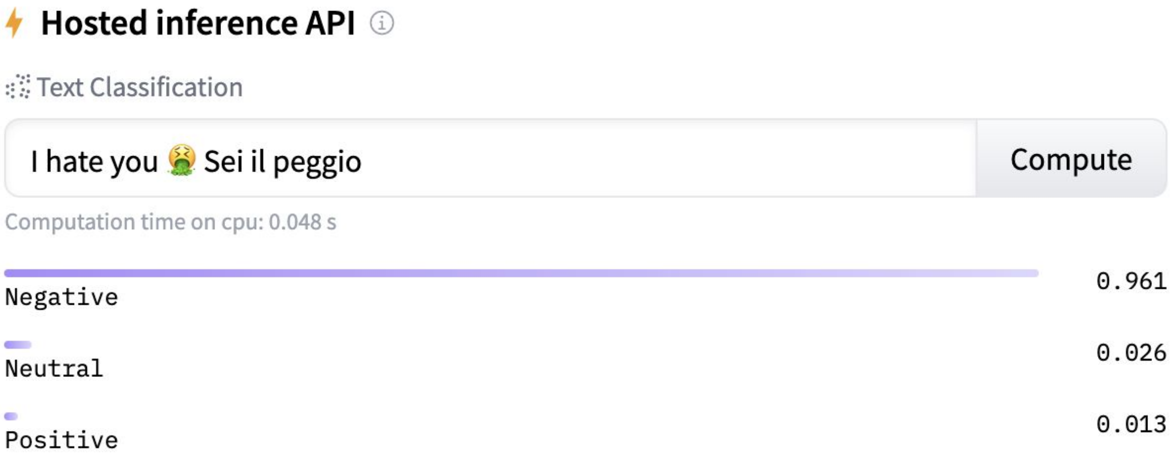
Then, we fine-tuned XLM-T on it!



Lang.	**Dataset**
Arabic	SemEval-17 (Rosenthal et al., 2017)
English	SemEval-17 (Rosenthal et al., 2017)
French	Deft-17 (Benamara et al., 2017)
German	SB-10K (Cieliebak et al., 2017)
Hindi	SAIL 2015 (Patra et al., 2015)
Italian	Sentipolc-16 (Barbieri et al., 2016)
Portug.	SentiBR (Brum and Nunes, 2017)
Spanish	Intertass (Díaz-Galiano et al., 2018)

🤗 XLM-T-Sentiment

XLM-T fine-tuned on sentiment analysis datasets from different languages.

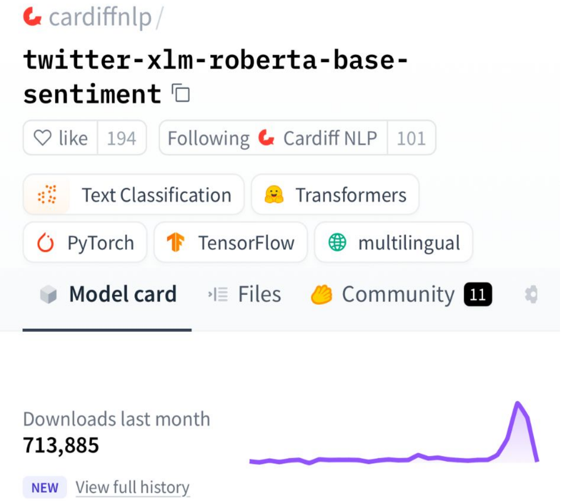


I hate you 🤮 Sei il peggio

Computation time on cpu: 0.048 s

```
Negative    0.961
Neutral     0.026
Positive    0.013
```

Multilingual sentiment analysis models are still very popular

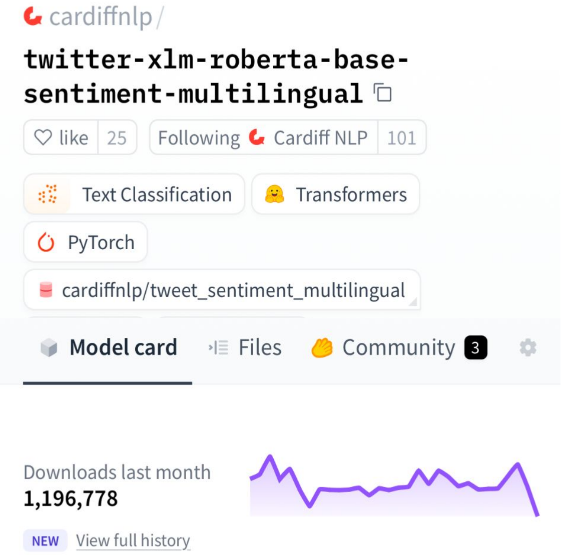


cardiffnlp /

twitter-xlm-roberta-base-sentiment

Downloads last month

713,885



cardiffnlp /

twitter-xlm-roberta-base-sentiment-multilingual

Downloads last month

1,196,778

Case study: Sentiment and virality

(Antypas, Preece and Camacho-Collados, OSNEM 2023)

Collected a corpus of Twitter messages from MPs in **Greece, Spain and UK** (~1M tweets). Then, we used **XLM-T-Sentiment** on all tweets!

Analysed the relation between **sentiment** (as provided by our Twitter-based models) and **virality** (measured by number of retweets and other metrics).

Case study: Sentiment and virality

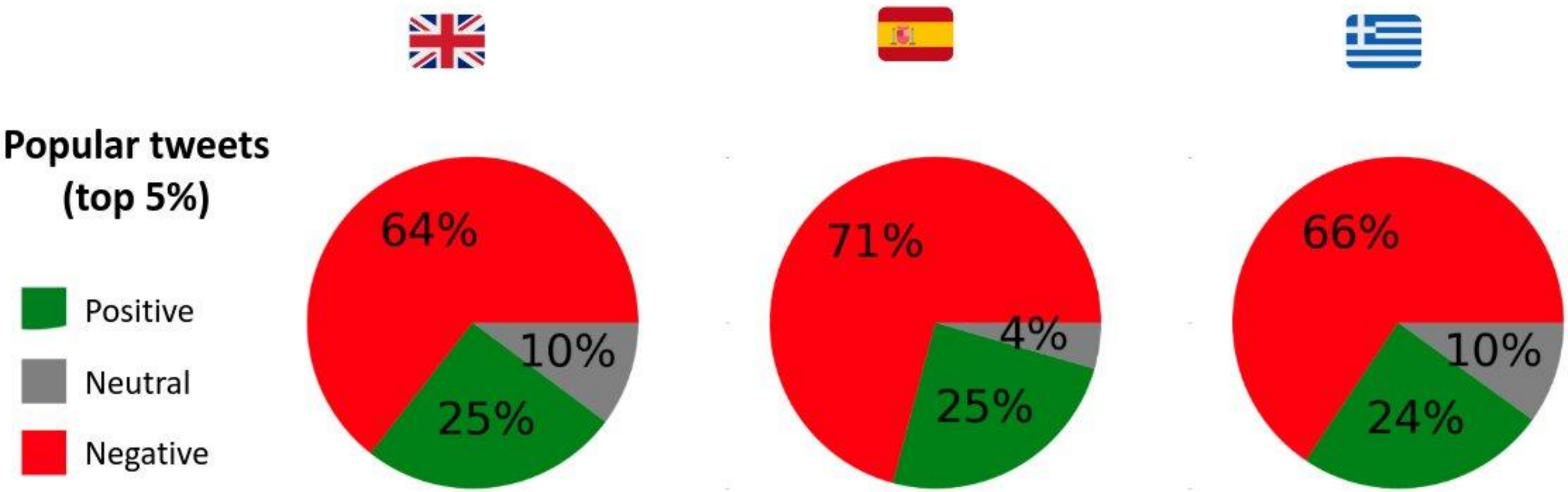
(Antypas, Preece and Camacho-Collados, OSNEM 2023)

Collected a corpus of Twitter messages from MPs in **Greece, Spain and UK** (~1M tweets). Then, we used **XLM-T-Sentiment** on all tweets!

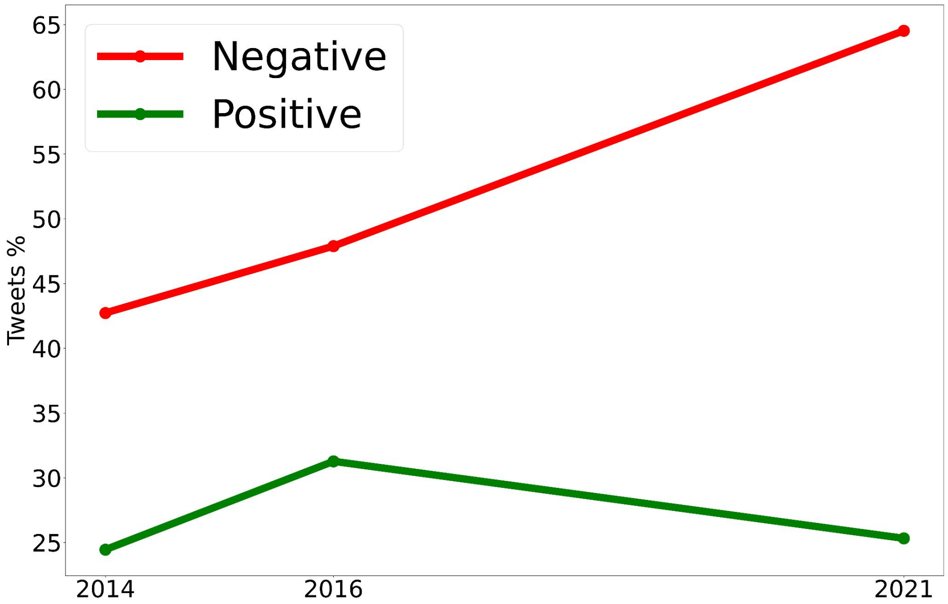
Analysed the relation between **sentiment** (as provided by our Twitter-based models) and **virality** (measured by number of retweets and other metrics).

➡️ **Conclusion:** Tweets negatively charged 👉 More popular ⬆️

Sentiment of MPs' tweets



Sentiment over time



Tweets by MPs are becoming more negative over time (UK)

Multilingual Tweet Topic Classification

(Antypas et al. EMNLP 2024)

Task: Associate each post with a topic

19 topics (Sports, Gaming, Music, Relationships, etc.)

Multilingual topic classification dataset annotated in English, Spanish, Japanese, Greek

Multilingual Tweet Topic Classification

(Antypas et al. EMNLP 2024)

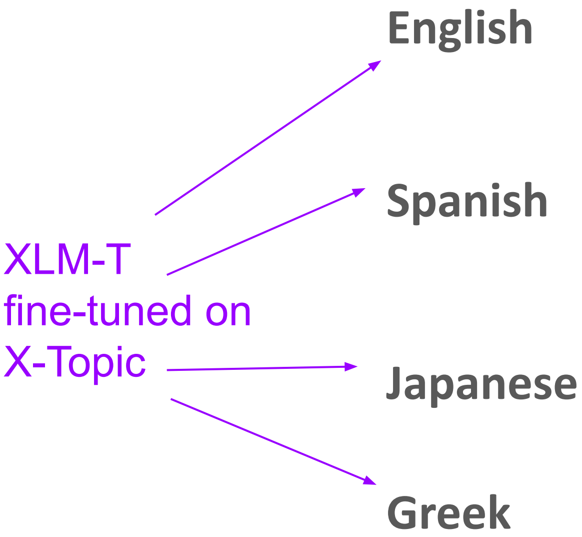
Task: Associate each post with a topic

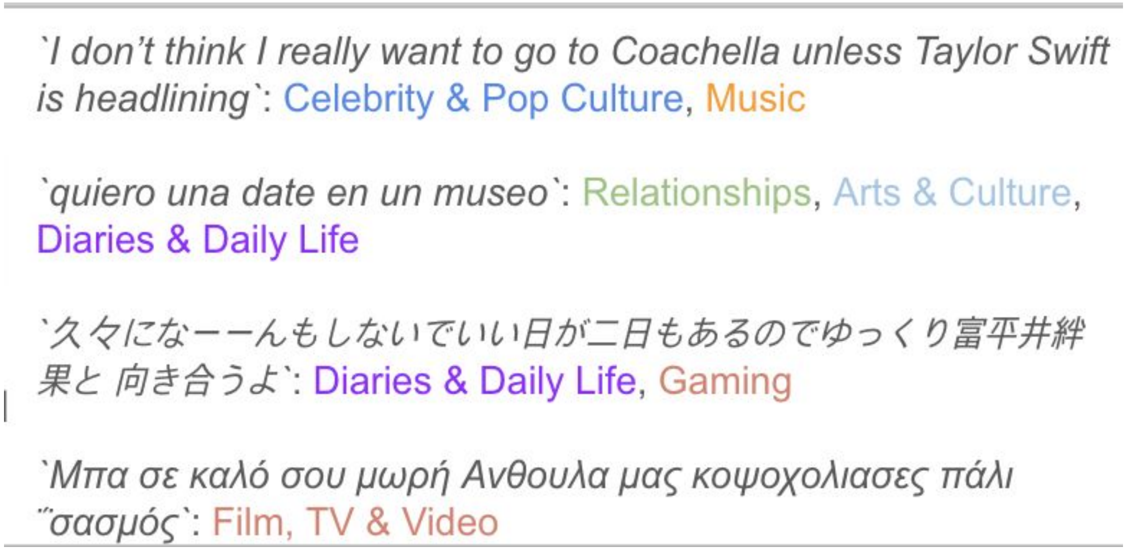
19 topics (Sports, Gaming, Music, Relationships, etc.)

Multilingual topic classification dataset annotated in English, Spanish, Japanese, Greek

XLM-T fine-tuned on X-Topic 👉🏼 Multilingual topic classification models

Multilingual Tweet Topic Classification





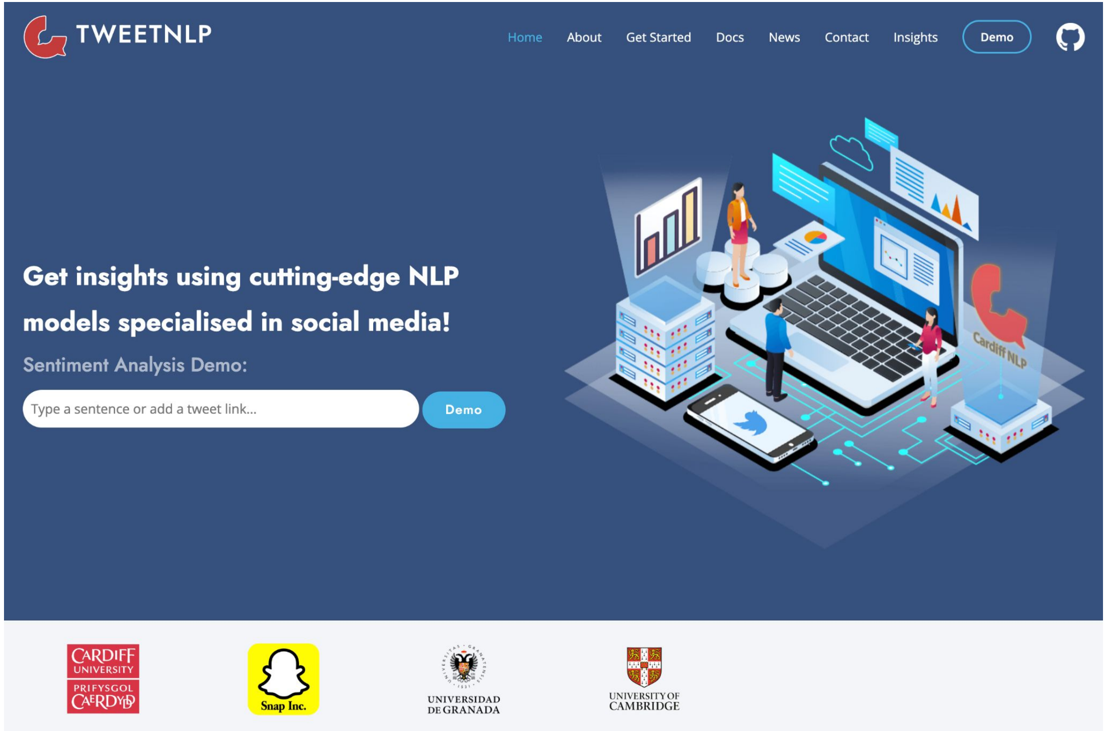
`I don't think I really want to go to Coachella unless Taylor Swift is headlining`: Celebrity & Pop Culture, Music

`quiero una date en un museo`: Relationships, Arts & Culture, Diaries & Daily Life

`久々にな——んもしないでいい日が二日もあるのでゆっくり富平井絆果と 向き合うよ`: Diaries & Daily Life, Gaming

`Μπα σε καλό σου μωρή Ανθουλα μας κοψοχολιασες πάλι ῎σασμός`: Film, TV & Video

TweetNLP (Camacho-Collados et al., EMNLP Demo 2022)



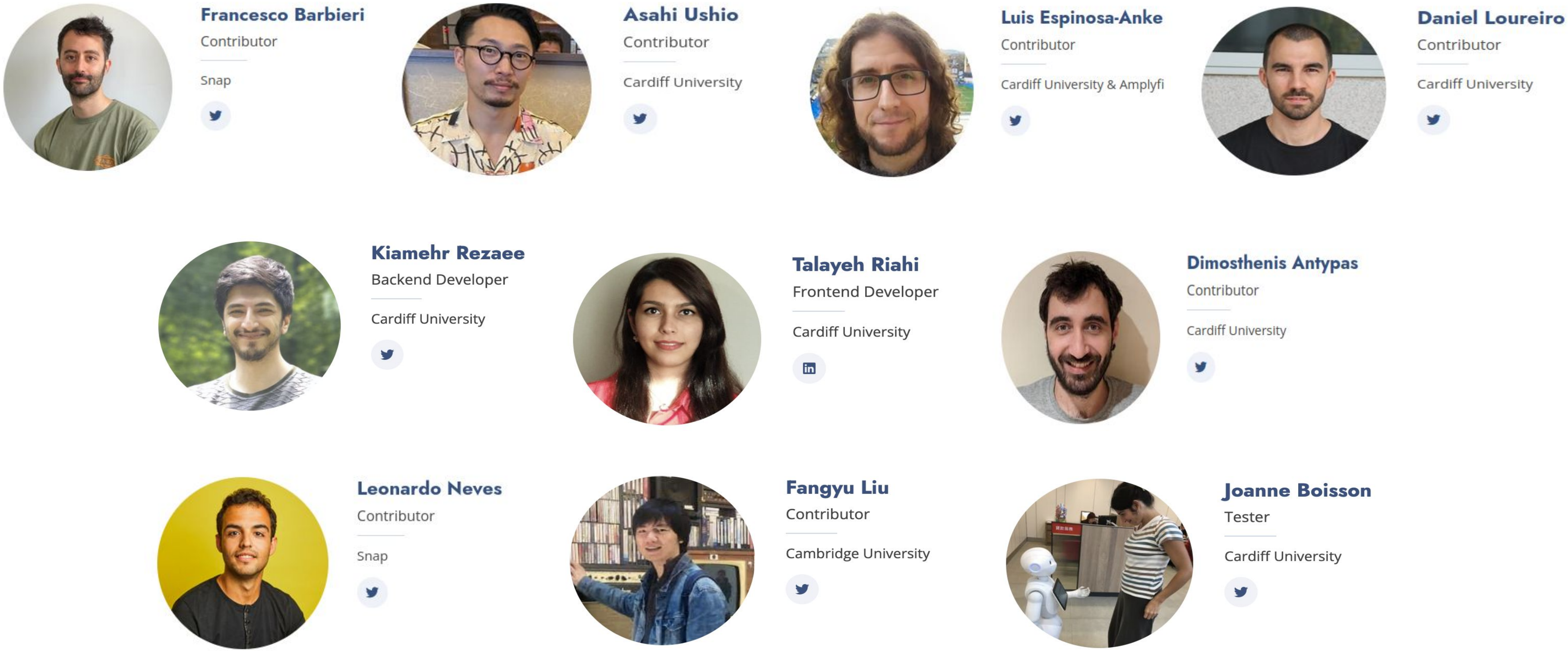
Get insights using cutting-edge NLP models specialised in social media!

Sentiment Analysis Demo:

Type a sentence or add a tweet link...

Demo

TweetNLP - the team



Francesco Barbieri
Contributor
Snap

Asahi Ushio
Contributor
Cardiff University

Luis Espinosa-Anke
Contributor
Cardiff University & Amplyfi

Daniel Loureiro
Contributor
Cardiff University

Kiamehr Rezaee
Backend Developer
Cardiff University

Talayeh Riahi
Frontend Developer
Cardiff University

Dimosthenis Antypas
Contributor
Cardiff University

Leonardo Neves
Contributor
Snap

Fangyu Liu
Contributor
Cambridge University

Joanne Boisson
Tester
Cardiff University

TweetNLP

Integration of all these resources, including **multilingual LMs** (needs to be extended to more languages)

A platform for **NLP specialised on social media**

Integration of all resources with relatively **small models**

NLP applications from sentiment analysis to hate speech detection and NER

Demo, models and Python API

Ok, so multilingual LMs can enable multiple applications, support many languages…

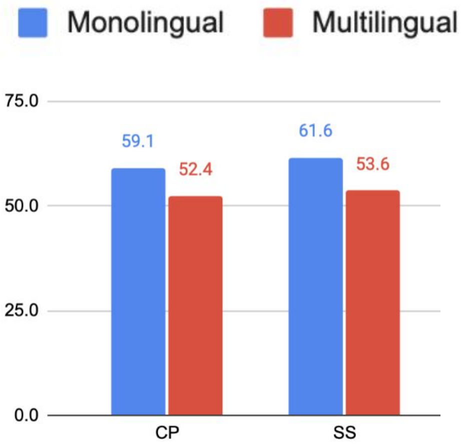
Anything else?

Multilingual LMs are less socially biased than monolingual LMs!

(Zhou et al. EMNLP 2023)

Since they have been trained in texts from multiple languages and sources, multilingual LMs **tend to be less biased than their monolingual counterparts**.

Results averaged across models in CP and SS social bias datasets



Conclusions in line with previous work (Liang et al. 2020; Ahn and Oh 2021)

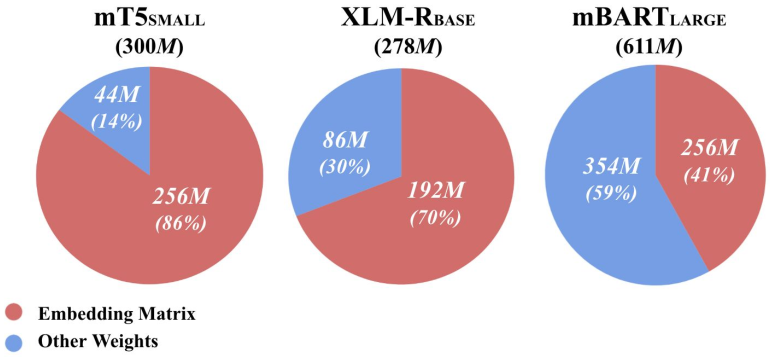
Ok, so multilingual LMs are great….

Any issues?

Issue 1: Multilingual LMs are generally big

Multilingual LMs need to be large enough to absorb information in different languages: more content to process, different scripts…

The embedding matrix is usually the largest component of multilingual LMs.



Vocabulary trimming: A "trick" to reduce size of multilingual LMs in language-specific tasks

(Ushio et al. EMNLP Findings 2023)

Multilingual LMs can be used or fine-tuned on specific languages (e.g. Korean, Japanese, etc.).

However, multilingual LMs are larger than language-specific LMs.

When we finetune a multilingual LM on a single language, we only need the vocabulary of that language….right?

💡 **Idea:** Keep only tokens used in your target language!

Issue 2: Language coverage

Data is mostly available in English and high-resource languages.

Current LMs are incredibly data-hungry, so this leads to obvious **performance variation across languages**.

Also, some languages are then more “multilingual” than others!

Solution? No obvious solution other than creating data for low-resource languages and develop models less dependant on data (hard)

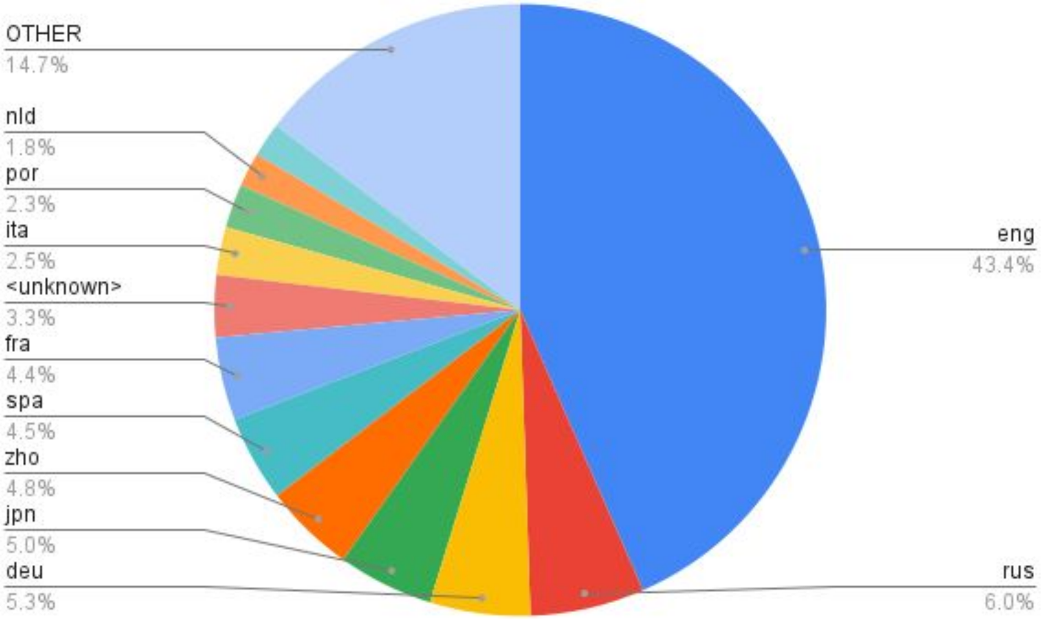
Common Crawl language distribution

Main source of pre-training data for (multilingual) LMs.

English: 43.4%

Top 10 languages: >82%

Rest of ~7,000 languages in the world: <15%



Issue 3: Cultural sensitivity and awareness

Are multilingual LMs sensitive to different cultures and contexts?

For instance, common traditions are different across countries.

While there are many “objective” usages of LMs, in many cases LLMs need to adapt to the context of the user (e.g. their region/country, and others).

Issue 3: Cultural sensitivity and awareness

Are multilingual LMs sensitive to different cultures and contexts?

For instance, common traditions are different across countries.

While there are many “objective” usages of LMs, in many cases LLMs need to adapt to the context of the user (e.g. their region/country, and others).

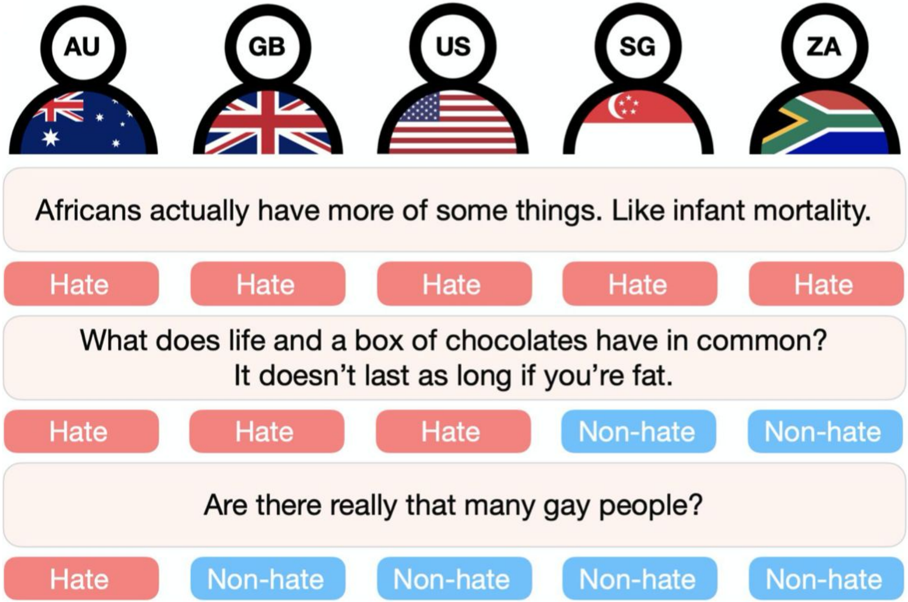
📝 **Note:** this is not an issue of multilingual LMs exclusively

Cross-cultural differences in English hate speech

(Lee, Jung and Myung et al. NAACL 2024)

KAIST

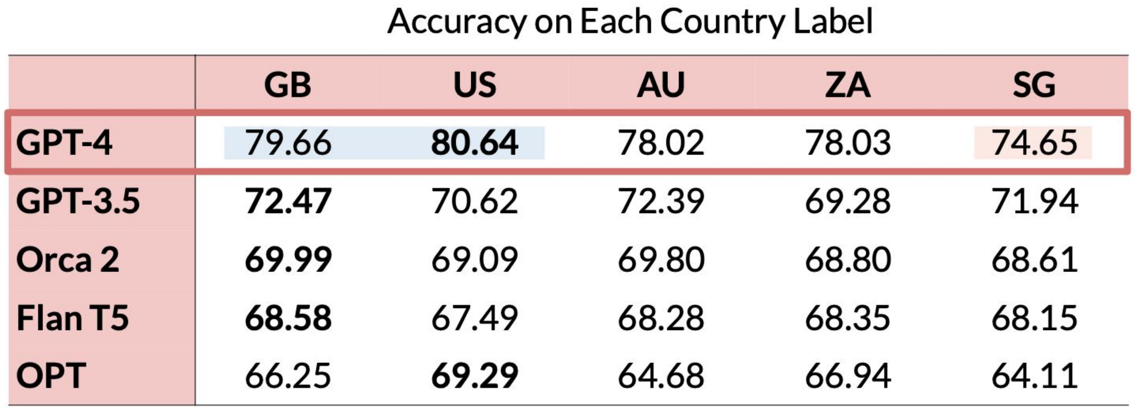
Hate speech dataset annotated by people from 5 different countries



AU	GB	US	SG	ZA
Africans actually have more of some things. Like infant mortality.				
Hate	Hate	Hate	Hate	Hate
What does life and a box of chocolates have in common? It doesn’t last as long if you’re fat.				
Hate	Hate	Hate	Non-hate	Non-hate
Are there really that many gay people?				
Hate	Non-hate	Non-hate	Non-hate	Non-hate

Results of LLMs prompted to detect "hate speech"

Significant differences between Western countries and Singapore



Accuracy on Each Country Label

	GB	**US**	**AU**	**ZA**	**SG**
GPT-4	79.66	**80.64**	78.02	78.03	74.65
GPT-3.5	**72.47**	70.62	72.39	69.28	71.94
Orca 2	**69.99**	69.09	69.80	68.80	68.61
Flan T5	**68.58**	67.49	68.28	68.35	68.15
OPT	66.25	**69.29**	64.68	66.94	64.11

OK, this was for English, what about for other languages?

The problem is even more marked when it comes to **different languages** (and especially low-resource languages!)

However, **hard to evaluate** - how to get relevant data for many languages and countries?

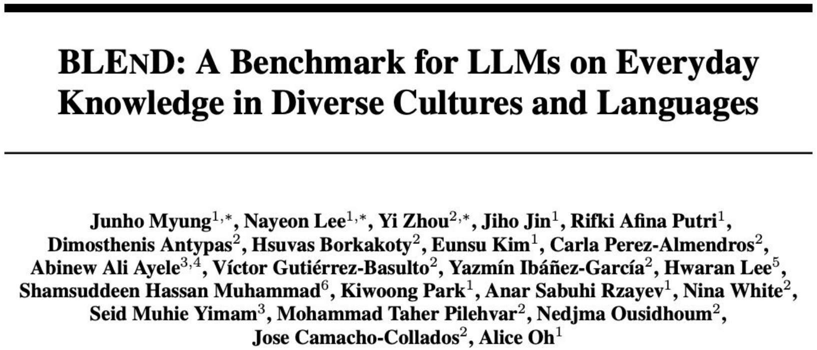
Let’s create a multilingual and multicultural benchmark!

BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages

(Myung, Lee and Zhou et al. NeurIPs D&B 2024)

Cardiff NLP

KAIST



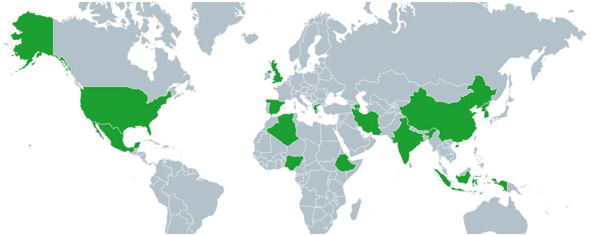
BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages

Junho Myung[1,*], Nayeon Lee[1,*], Yi Zhou[2,*], Jiho Jin[1], Rifki Afina Putri[1], Dimosthenis Antypas[2], Hsuvas Borkakoty[2], Eunsu Kim[1], Carla Perez-Almendros[2], Abinew Ali Ayele[3,4], Víctor Gutiérrez-Basulto[2], Yazmín Ibáñez-García[2], Hwaran Lee[5], Shamsuddeen Hassan Muhammad[6], Kiwoong Park[1], Anar Sabuhi Rzayev[1], Nina White[2], Seid Muhie Yimam[3], Mohammad Taher Pilehvar[2], Nedjma Ousidhoum[2], Jose Camacho-Collados[2], Alice Oh[1]

BLEND: Key Characteristics

Most cultural datasets rely heavily on social media or Wikipedia, which often overlook the **mundane everyday lifestyles of underrepresented cultures.**

In BLEND, we **manually** collect questions about everyday life from people from **16 countries and regions, in 13 different languages**



Languages Included:

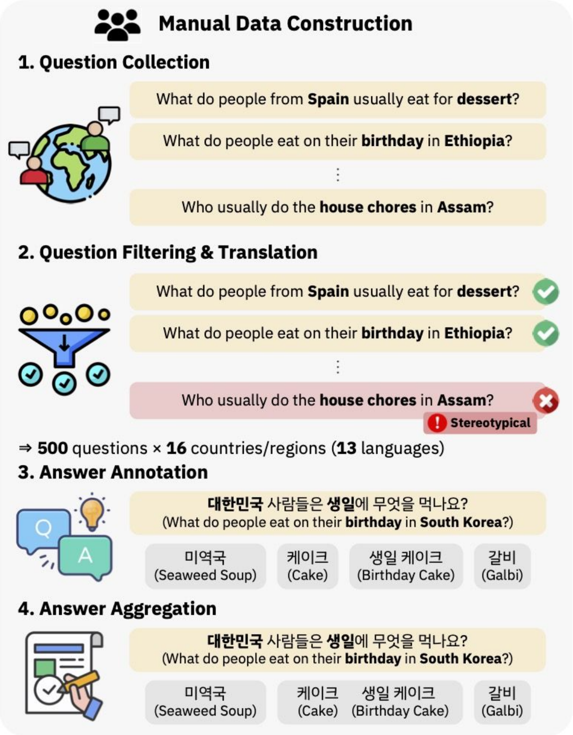
- ○ English
- ○ Chinese
- ○ Spanish
- ○ Indonesian
- ○ Korean
- ○ Greek
- ○ Persian
- ○ Arabic
- ○ Azerbaijani
- ○ Sundanese
- ○ Assamese
- ○ Hausa
- ○ Amharic

Construction of BLEND

Manual collection of question and answers from **native annotators in each country/region**

Filtering and aggregation steps are done to remove any duplicates and to ensure high quality

500 QA pairs are used to expand the benchmark into two tasks: **Short Answer Questions (SAQ), and Multiple Choice Questions (MCQ)**



Manual Data Construction

1. Question Collection

What do people from **Spain** usually eat for **dessert**?

What do people eat on their **birthday** in **Ethiopia**?

⋮

Who usually do the **house chores** in **Assam**?

2. Question Filtering & Translation

What do people from **Spain** usually eat for **dessert**? ✓

What do people eat on their **birthday** in **Ethiopia**? ✓

⋮

Who usually do the **house chores** in **Assam**? ✗ ! **Stereotypical**

⇒ **500** questions × **16** countries/regions (**13** languages)

3. Answer Annotation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

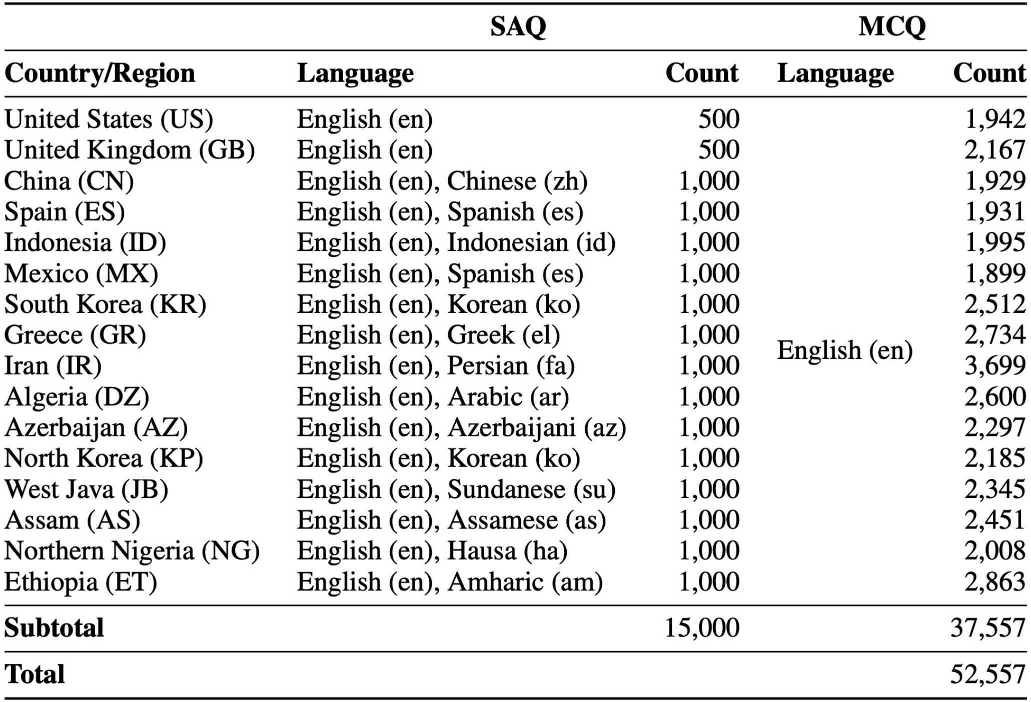
미역국 (Seaweed Soup) | 케이크 (Cake) | 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

4. Answer Aggregation

대한민국 사람들은 **생일**에 무엇을 먹나요?
(What do people eat on their **birthday** in **South Korea**?)

미역국 (Seaweed Soup) | 케이크 (Cake) 생일 케이크 (Birthday Cake) | 갈비 (Galbi)

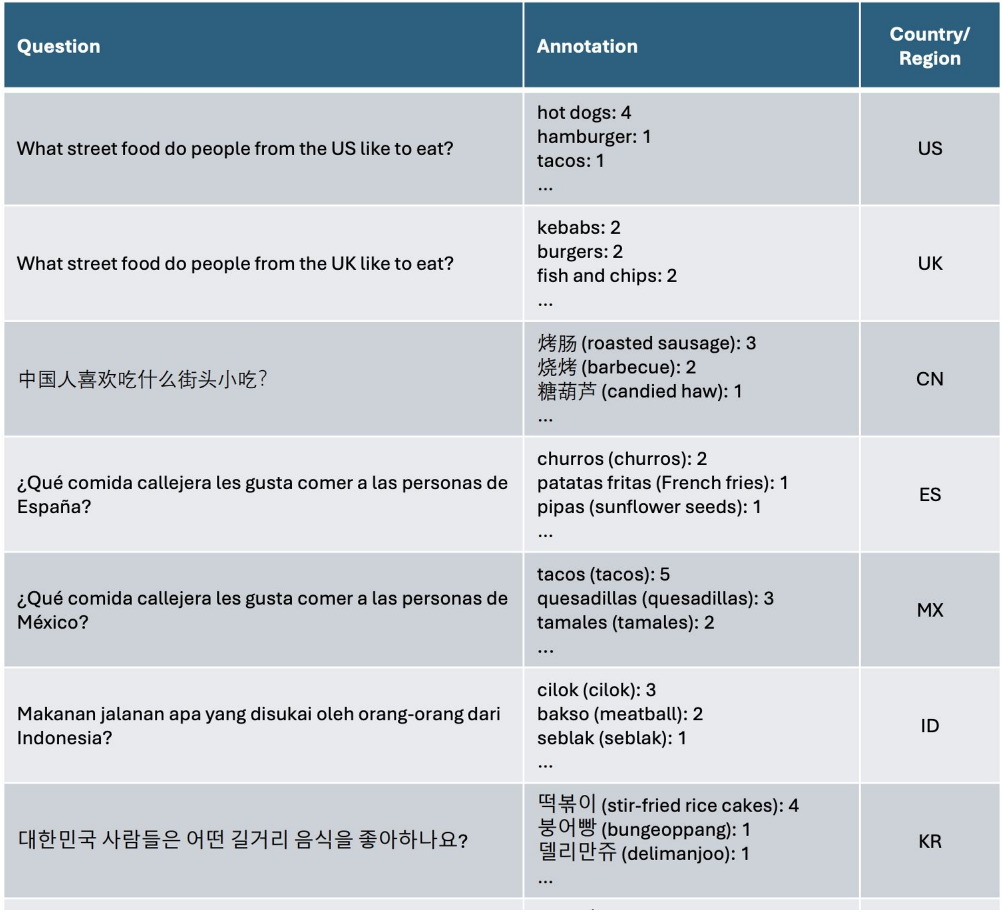
BLEnD: Statistics



	SAQ		MCQ	
Country/Region	**Language**	**Count**	**Language**	**Count**
United States (US)	English (en)	500	English (en)	1,942
United Kingdom (GB)	English (en)	500		2,167
China (CN)	English (en), Chinese (zh)	1,000		1,929
Spain (ES)	English (en), Spanish (es)	1,000		1,931
Indonesia (ID)	English (en), Indonesian (id)	1,000		1,995
Mexico (MX)	English (en), Spanish (es)	1,000		1,899
South Korea (KR)	English (en), Korean (ko)	1,000		2,512
Greece (GR)	English (en), Greek (el)	1,000		2,734
Iran (IR)	English (en), Persian (fa)	1,000		3,699
Algeria (DZ)	English (en), Arabic (ar)	1,000		2,600
Azerbaijan (AZ)	English (en), Azerbaijani (az)	1,000		2,297
North Korea (KP)	English (en), Korean (ko)	1,000		2,185
West Java (JB)	English (en), Sundanese (su)	1,000		2,345
Assam (AS)	English (en), Assamese (as)	1,000		2,451
Northern Nigeria (NG)	English (en), Hausa (ha)	1,000		2,008
Ethiopia (ET)	English (en), Amharic (am)	1,000		2,863
Subtotal		15,000		37,557
Total				52,557

Example in BLEnD: *"What street food do people like to eat?"*

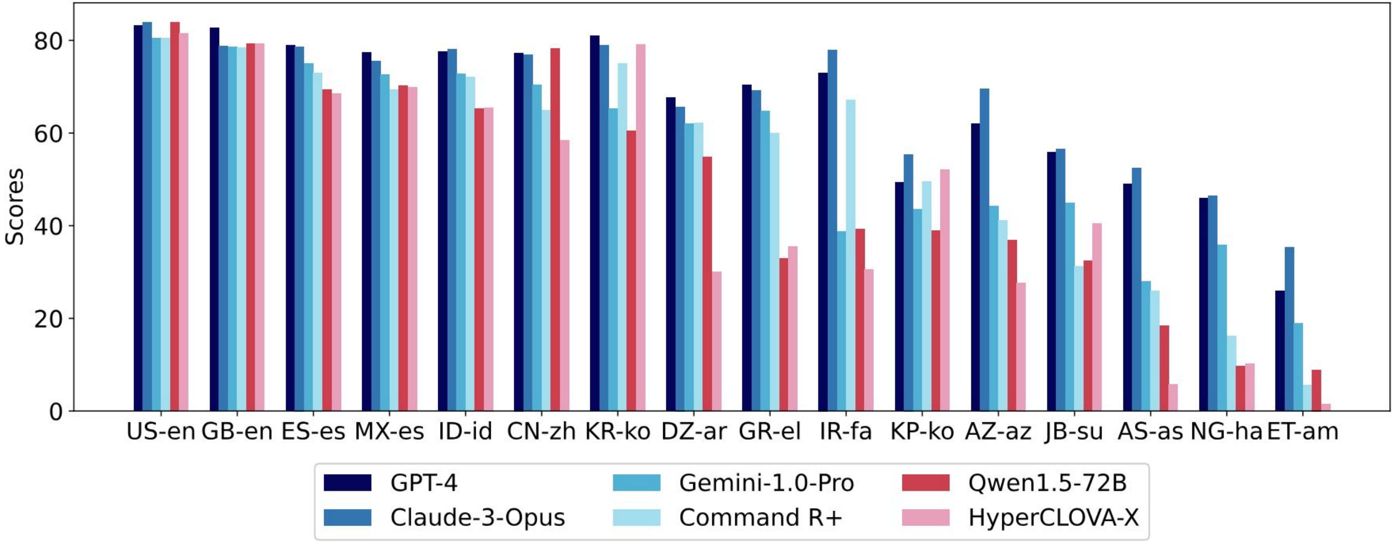
Answers for this simple question vary a lot across countries!



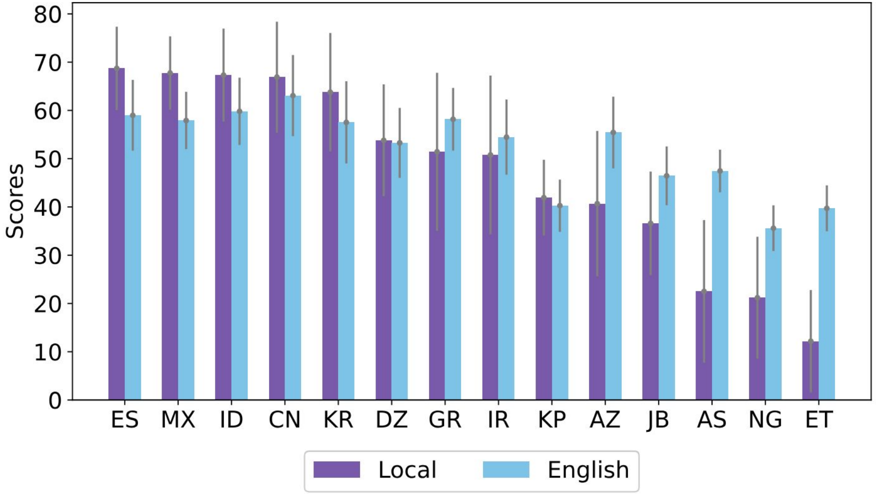
Question	Annotation	Country/ Region
What street food do people from the US like to eat?	hot dogs: 4 hamburger: 1 tacos: 1 ...	US
What street food do people from the UK like to eat?	kebabs: 2 burgers: 2 fish and chips: 2 ...	UK
中国人喜欢吃什么街头小吃？	烤肠 (roasted sausage): 3 烧烤 (barbecue): 2 糖葫芦 (candied haw): 1 ...	CN
¿Qué comida callejera les gusta comer a las personas de España?	churros (churros): 2 patatas fritas (French fries): 1 pipas (sunflower seeds): 1 ...	ES
¿Qué comida callejera les gusta comer a las personas de México?	tacos (tacos): 5 quesadillas (quesadillas): 3 tamales (tamales): 2 ...	MX
Makanan jalanan apa yang disukai oleh orang-orang dari Indonesia?	cilok (cilok): 3 bakso (meatball): 2 seblak (seblak): 1 ...	ID
대한민국 사람들은 어떤 길거리 음식을 좋아하나요?	떡볶이 (stir-fried rice cakes): 4 붕어빵 (bungeoppang): 1 델리만쥬 (delimanjoo): 1 ...	KR

LLMs' Performance in Local Languages

Models show a significant **drop in performance for underrepresented cultures**, with a maximum performance difference of 57.3 percentage points between the US and Ethiopia



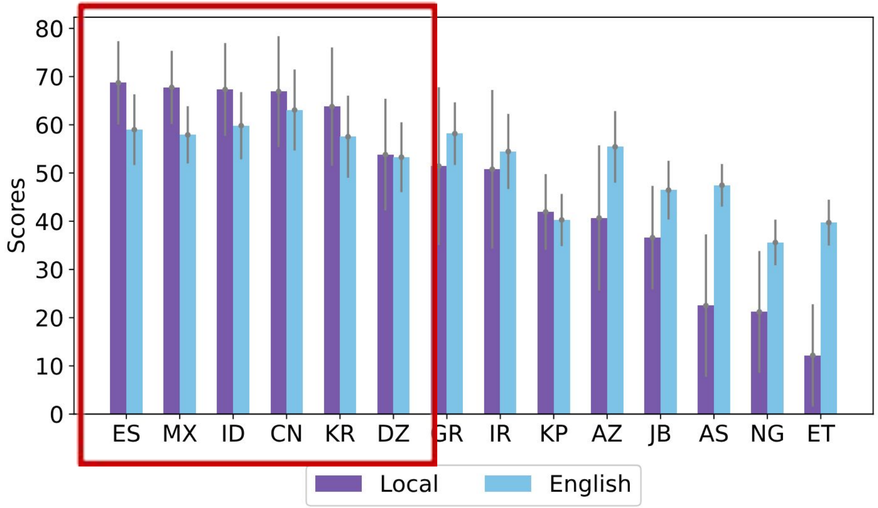
LLMs’ Performance in Local Languages vs English



Average Score for All Models; Models **prompted on English vs Local language** - same questions

LLMs' Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

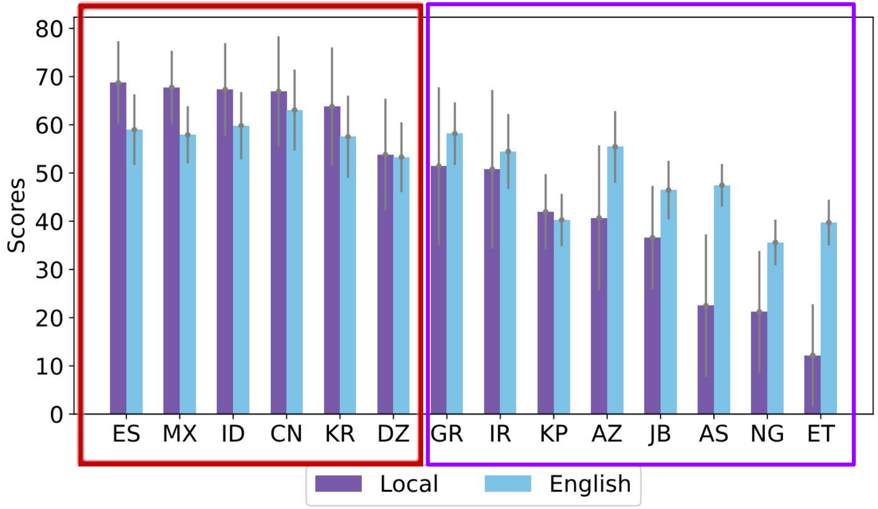


Average Score for All Models; Models **prompted on English vs Local language** - same questions

LLMs’ Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

For **low-resource languages** like Azerbaijani, Sundanese, and Amharic, models generally showed **better performance when prompted in English**

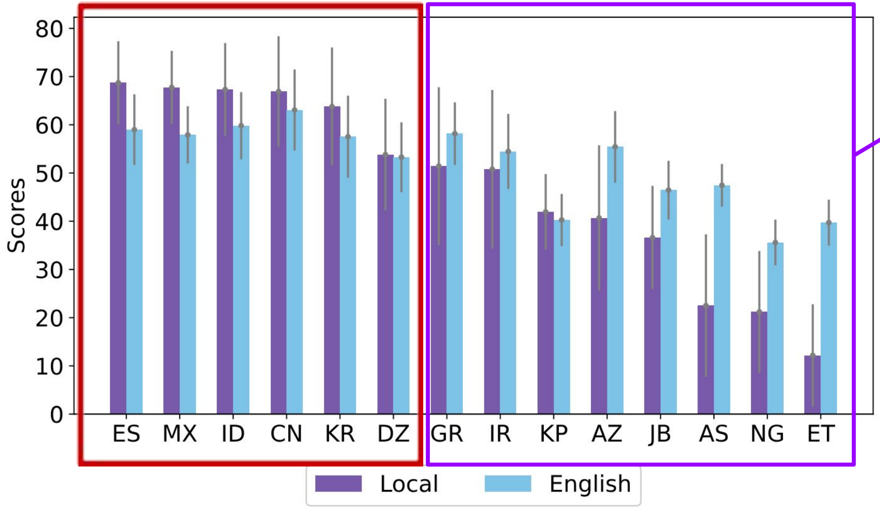


Average Score for All Models; Models **prompted on English vs Local language** - same questions

LLMs’ Performance in Local Languages vs English

For **high-resource languages** like Spanish and Chinese, models showed **better performance when prompted with their local languages**

For **low-resource languages** like Azerbaijani, Sundanese, and Amharic, models generally showed **better performance when prompted in English**



💡 **Research idea!**

Analyse whether to prompt in one language or another for different tasks

Key Findings from Human Evaluation

Most **stereotypical responses** came from questions related to **food or festivals.**

LLMs often mentioned the **most famous** food item (e.g. Kimchi in Korea) or festival in response to completely unrelated questions.

Hallucinations were common for questions asking for a name or a title of an entity:

➢ For instance, the model answered 'Ruslan Cfrov' as the most famous basketball player in Azerbaijan, even though **no such player exists**

➢ Models occasionally **answered questions in a different language**, particularly for low-resource languages like Azerbaijani

Conclusion

Multilingual LMs are incredible and magical creatures.

But many **questions remain**, from the theoretical and practitioner perspectives?

- How to balance language abilities and **cultural awareness**?
- Do we need **multilingual or monolingual** LMs for low-resource languages?
- Should we **prompt** in our native language or a high-resource one?
- How ***truly* multilingual** are multilingual LMs?

Interesting times for research in this area!

Thank you!

Most resources available in the Cardiff NLP Hugging Face page:

https://huggingface.co/cardiffnlp 🤗

@CamachoCollados