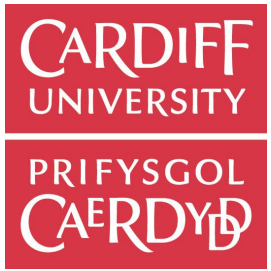
Practical and Specialised NLP Solutions:

The Case of Social Media

Jose Camacho-Collados



Cambridge, 30 May 2024

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Outline

- LLMs: **Issues and open problems**
- **Social media** NLP landscape
  - **Benchmarks**
  - **Challenges** (temporal, biases)

Cardiff NLP

About me

- Professor at **Cardiff University** (Wales, UK)
  - **UKRI Future Leaders Fellow**
  - Co-founder and head of the **Cardiff NLP group**.

- Areas of expertise: **Semantics, resources, multilinguality, social media**
  - Co-author of "**Embeddings in NLP**" book
  - General chair of *SEM-2024

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

- Young group (3 years old), growing fast (30+ lab members)
- **Website:** cardiffnlp.github.io
- Activities: hybrid seminars, workshops, hackathons, etc.
- **Twitter:** @Cardiff_NLP
- Open-source contributions

Cardiff NLP Workshop (1-2 July 2024)

Andreas Vlachos
University of Cambridge

Asahi Ushio
Amazon Tokyo

Anna Rogers
IT University of Copenhagen

Arkaitz Zubiaga
Queen Mary University of London

Javad Hosseini
Google Deepmind, UK

Nafise Sadat Moosavi
University of Sheffield

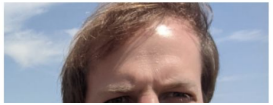
Emanuele Bugliarello
Google Research, France

Marie-Francine Moens
KU Leuven, Belgium

www.cardiffnlpworkshop.org

Registration open until June 5th!

Cardiff NLP Workshop (1-2 July 2024)



www.cardiffnlpworkshop.org

- A Tutorial on "Building RAG applications"
- Panel on "Career in NLP after graduation"
- Poster session
- Networking opportunities
- Welsh delights

Registration open until June 5th!

Andreas Vlach
University of Cambr

Javad Hosseini
Google Deepmind, UK

Nafise Sadat Moosavi
University of Sheffield

Emanuele Bugliarello
Google Research, France

Marie-Francine Moens
KU Leuven, Belgium

Cardiff NLP

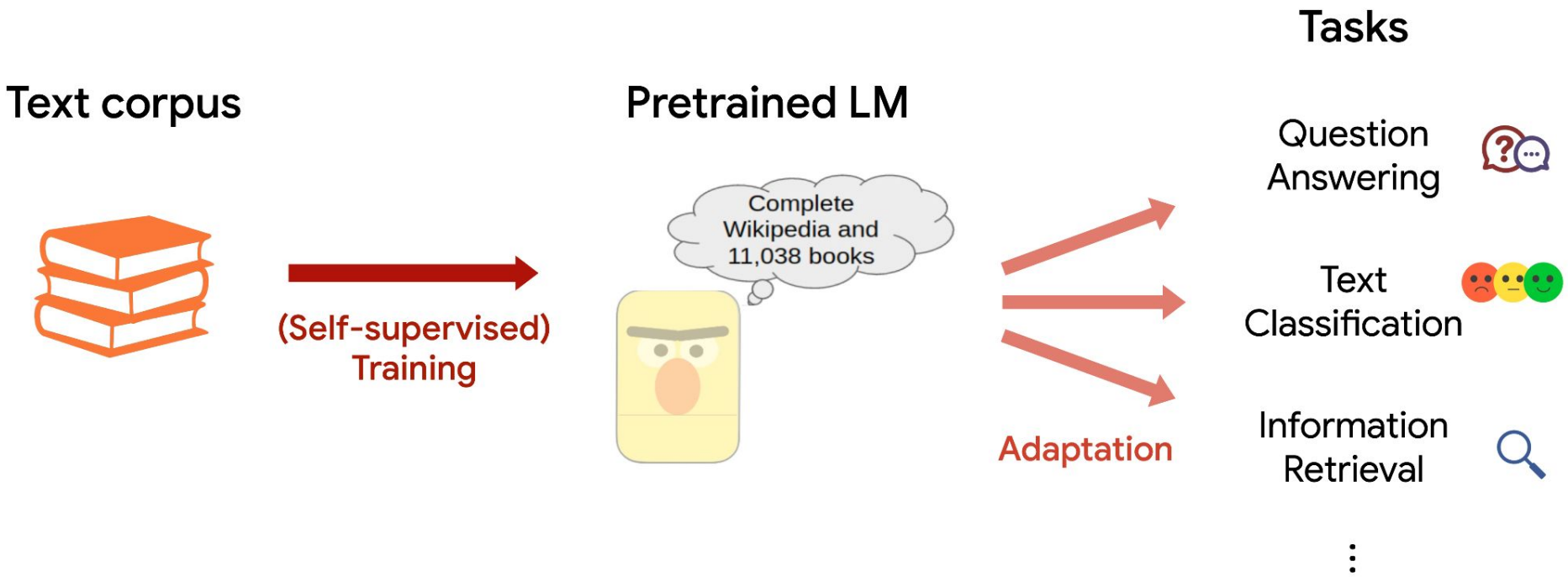
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

(Large) Language Models

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Language models (LMs)



Slide credit: Stanford AI

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Language models (LMs)

Text corpus

(Self-supervised) Training

Pretrained LM

Complete Wikipedia and 11,038 books

LLaMA GPT

Adaptation

Tasks

Question Answering

Text Classification

Information Retrieval

⋮

Slide credit: Stanford AI

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Areas for improvement

- **Evaluation**/reliability

- **Practicality** (are LLMs always the right choice for NLP problems?)

Analysing survey responses: LLMs to the rescue!

I hope all is well.

With the new Vice Chancellor at Cardiff University -- there is now an activity to seek consultation from a number of staff members at the University about their views on what Cardiff University should focus on going forward.

These consultations will comprise of text comments sent by various members of staff. My colleague, [redacted] has asked whether we have expertise in COMSC to process these text comments. All of you came to mind, given the significant expertise we now have in NLP in COMSC.

Could you please help out [redacted] process these responses please. [redacted] does not want to use ChatGPT or other chatbots -- as we are unclear on where this data goes.

Your help will be appreciated. Kindly respond [redacted].

Analysing survey responses: LLMs to the rescue!

We would hope to have a couple of thousand responses.

Rather than give them pre-framed options, we'd like to give them the freedom to write free text - if we can make sense of it.

I thought AI could help. And [redacted] suggested you.

Analysing survey responses: LLMs to the rescue!

We would hope to have a couple of thousand responses.

Rather than give them pre-framed options, we'd like to give them the freedom to write free text - if we can make sense of it.

I thought AI could help. And [redacted] suggested you.

If we are going to launch in October, we need to crack this quickly.

Email date: 26 September

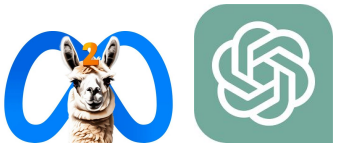
Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

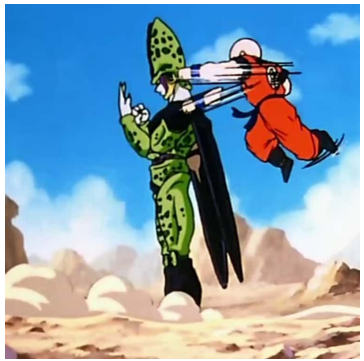
Text Classification: Is In-Context Learning enough?

(Edwards and Camacho-Collados, LREC/COLING 2024)

LLMs



In-context learning

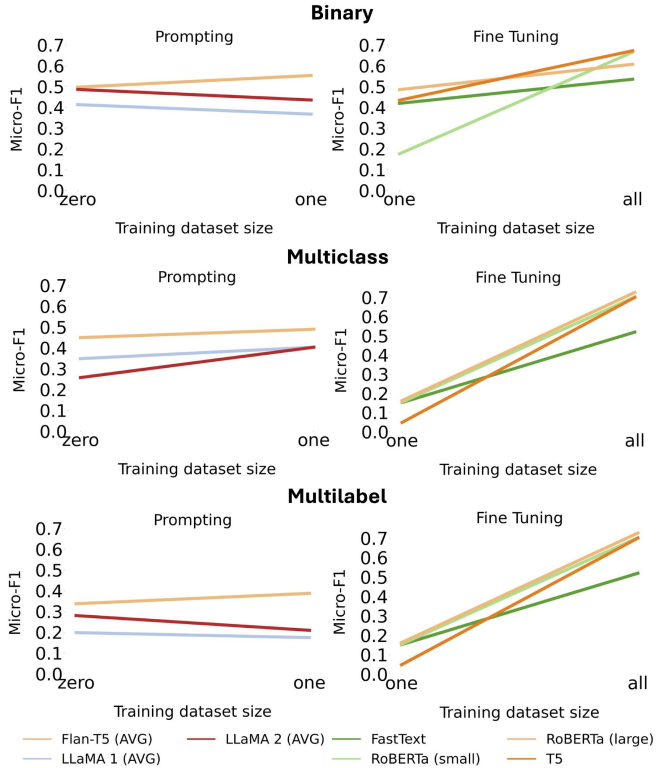


Fine-tuning

Small LMs (masked)

Text Classification: Is In-Context Learning enough?

(Edwards and Camacho-Collados, LREC/COLING 2024)



TLDR:

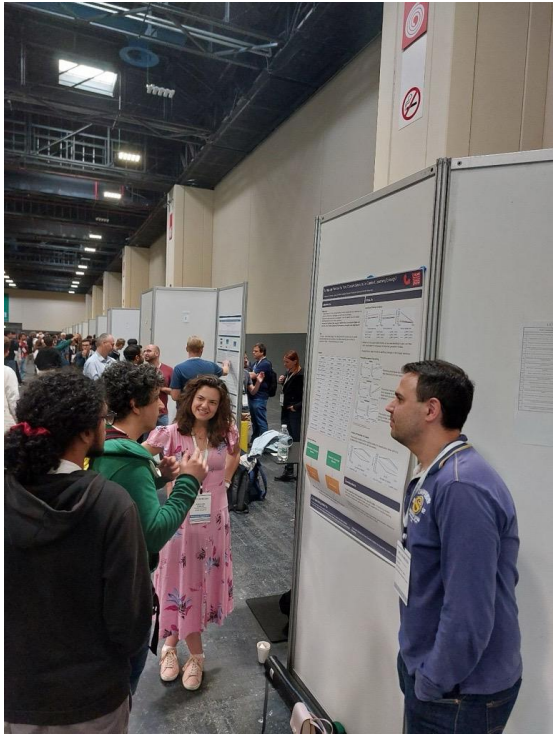
Fine-tuning smaller models (e.g. RoBERTa) led to better results than **LLMs with In-Context Learning**

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Text Classification:
Is In-Context Learning enough?

(Edwards and Camacho-Collados, LREC/COLING 2024)



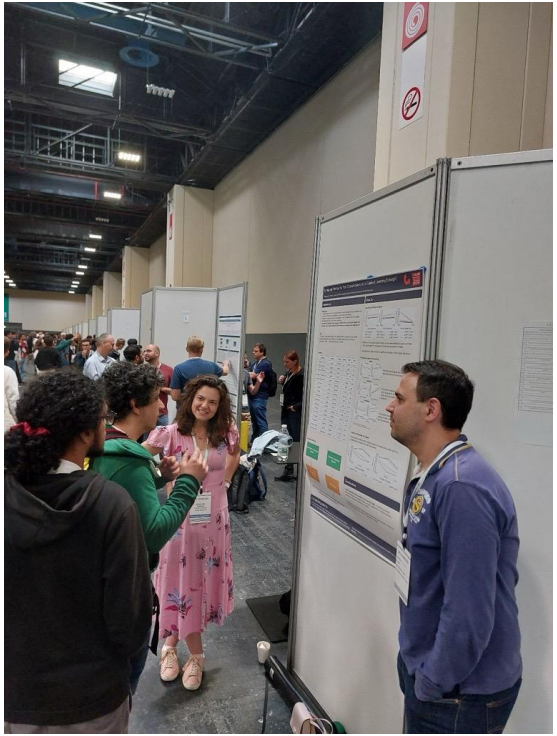
Text Classification: Is In-Context Learning enough?

(Edwards and Camacho-Collados, LREC/COLING 2024)

The LLM Skepticals

"I've been trying to use LLMs for task X, but a simple BERT classifier was always better!"

"I've been going crazy as I thought I was the only one for which this happened!"



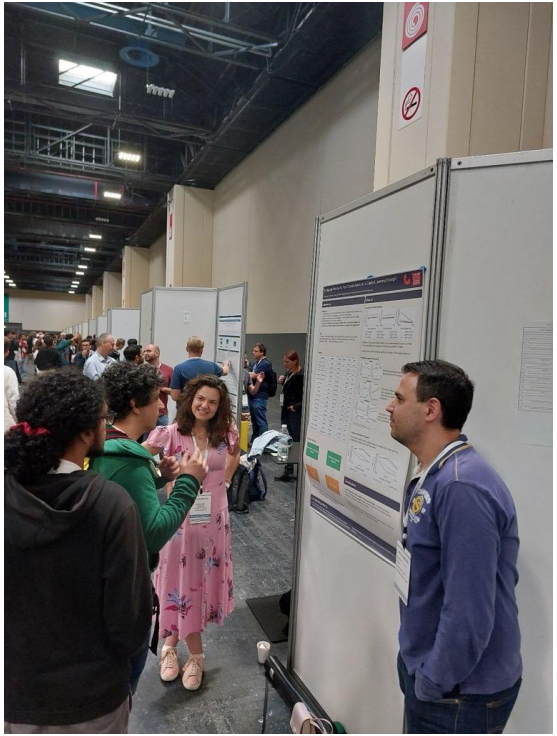
Text Classification: Is In-Context Learning enough?

(Edwards and Camacho-Collados, LREC/COLING 2024)

The LLM Skepticals

"I've been trying to use LLMs for task X, but a simple BERT classifier was always better!"

"I've been going crazy as I thought I was the only one for which this happened!"



The LLM Believers

"If you do better prompt engineering, LLMs will have better performance"

"The setting was not fair"

"You haven't used LLaMA-3 or GPT4-o"

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Social media
as a case domain

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Social media is a challenging domain

- Why?
  - **Informal** grammar
  - **Multilingual** (code-switching, etc.)
  - Irregular **vocabulary**
    - Emoji 😀, abbreviations, typos, hashtags, mentions…
  - Tweets are often **not standalone messages**
    - RTs, mentions, replies, threads, pictures…
  - **Dynamic**, constantly changing

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Social media: My story

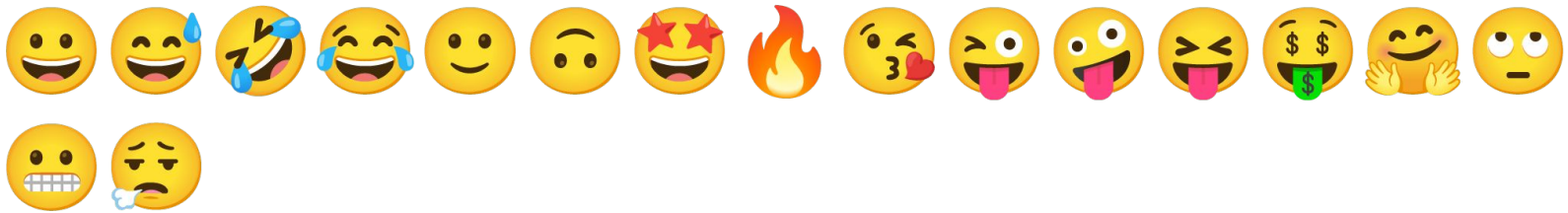
- Started as a side project
- Interested from the NLP research point of view: interesting and challenging domain, practical.

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Social media: My story

- Started as a side project
- Interested from the NLP research point of view: interesting and challenging domain, practical.
- I liked **emoji**:



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetEval:

Language Models and Evaluation Benchmark

TweetEval, the language model

(Barbieri et al. EMNLP Findings 2020)

- How?
  - RoBERTa architecture
  - Continue from RoBERTa checkpoint (BERTweet is from scratch)
  - Train on social media data (Twitter)

Specializing a LM on social media

- Why?

Haaland! That was <mask>

Specializing a LM on social media

- Why?

Cardiff NLP

TweetEval, the benchmark

(Barbieri et al. EMNLP Findings 2020)

- Why?

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetEval, the benchmark

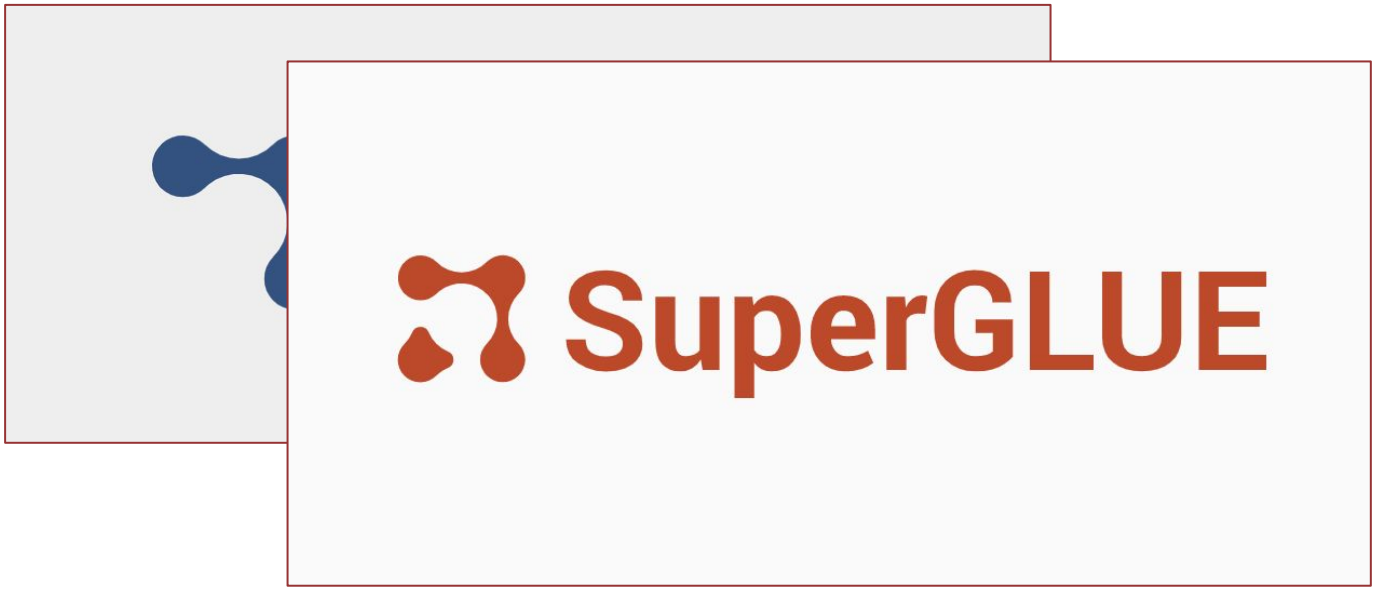
- Why?

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetEval, the benchmark

- Why?



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetEval, the benchmark

- Why?

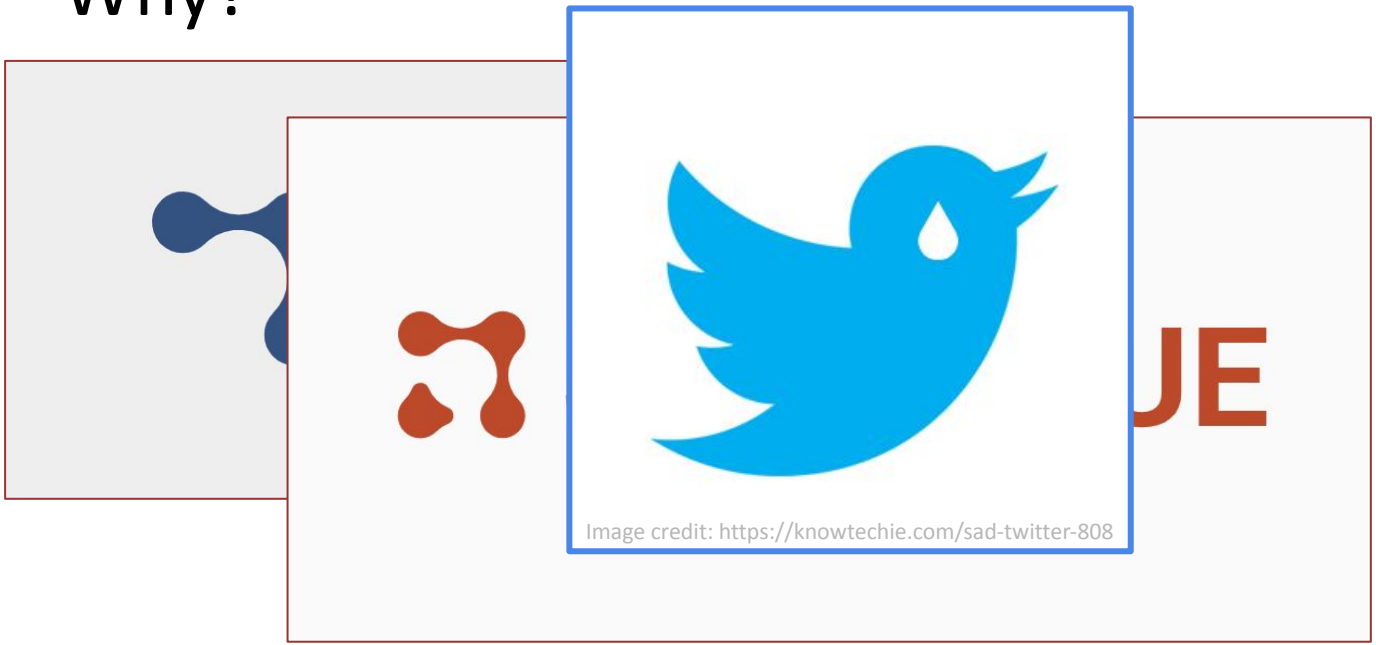


Image credit: https://knowtechie.com/sad-twitter-808

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetEval, the benchmark

➢ What?

Task	Lab	Train	Val	Test
Emoji prediction	20	45,000	5,000	50,000
Emotion rec.	4	3257	374	1421
Hate speech det.	2	9,000	1,000	2,970
Irony detection	2	2,862	955	784
Offensive lg. id.	2	11,916	1,324	860
Sent. analysis	3	45,389	2,000	11,906
Stance detection	3	2620	294	1249
Stance/Abortion	3	587	66	280
Stance/Atheism	3	461	52	220
Stance/Climate	3	355	40	169
Stance/Feminism	3	597	67	285
Stance/H. Clinton	3	620	69	295

SuperTweetEval, the benchmark

(Antypas et al. EMNLP Findings 2023)

➢ What?

Task (Dataset)	Train	Valid.	Test
TWEETNER7	4,616	576	2,807
TWEETEMOTION	6,838	886	3,259
TWEETQG	9,489	1,086	1,203
TWEETNERD	20,164	4,100	20,075
TWEETSENTIMENT	26,632	4,000	12,379
TEMPOWIC	1,427	395	1,472
TWEETEMOJI100	50,000	5,000	50,000
TWEETINTIMACY	1,191	396	396
TWEETQA	9,489	1,086	1,203
TWEETTOPIC	4,585	573	1,679
TWEETHATE	5,019	716	1,433
TWEETSIM	450	100	450

An extended and more challenging benchmark in the age of LLMs!

SuperTweetEval, the benchmark

(Antypas et al. EMNLP Findings 2023)

12 diverse NLP tasks

Task (Dataset)	Example Input	Example Output
NER (TWEETNER7)	**Tweet**: Winter solstice 2019 : A short day that 's long on ancient traditions url via @CNN_Travel	Winter solstice 2019: event @CNN_Travel: product
Emotion Classification (TWEETEMOTION)	**Tweet**: Whatever you decide to do make sure it makes you #happy.	joy, love, optimism
Question Generation (TWEETQG)	**Tweet**: 5 years in 5 seconds. Darren Booth (@darbooth) January 25, 2013 **Context**: vine	what site does the link take you to?
Name Entity Disambiguation (TWEETNERD)	**Tweet**: hella excited for ios 15 because siri reads notifications out loud to you [...] **Target**: siri **Definition**: intelligent personal assistant on various Apple devices	True
Sentiment Classification (TWEETSENTIMENT)	**Tweet**: #ArianaGrande Ari By Ariana Grande 80% Full url #Singer #Actress url **Target**: #ArianaGrande	negative or neutral
Meaning Shift Detection (TEMPOWIC)	**Tweet 1**: The minute I can walk well I'm going to delta pot **Tweet 2**: Then this new delta variant out im vaccinated but stilllll likeee' **Target**: delta	False
Emoji Classification (TWEETEMOJI100)	**Tweet**: SpiderMAtS back at it	🔥
Intimacy Analysis (TWEETINTIMACY)	**Tweet**: @user SKY scored 4 less runs just lol	1.20
Question Answering (TWEETQA)	**Tweet**: 5 years in 5 seconds. Darren Booth (@user) January 25, 2013 **Question**: which measurements of time are mentioned?	years and seconds
Topic Classification (TWEETTOPIC)	**Tweet**: Sweet, #IOWAvsISU is a nationally televised night game! Nebraska getting bumped to @FOX_Business is just a bonus.	film_tv_&_video, sports
Hate Speech Detection (TWEETHATE)	**Tweet**: Support Black Trans youth url	not_hate
Tweet Similarity (TWEETSIM)	**Tweet 1**: I wish kayvee all the best #bbnaija **Tweet 2**: Sammie about to cry to the housemates all night #bbnaija	2.33

Cardiff NLP

SuperTweetEval, the benchmark

(Antypas et al. EMNLP Findings 2023)

🤗 Already available at

huggingface.co/datasets/cardiffnlp/super_tweeteval

Includes **generative**, **regression** and **classification** tasks

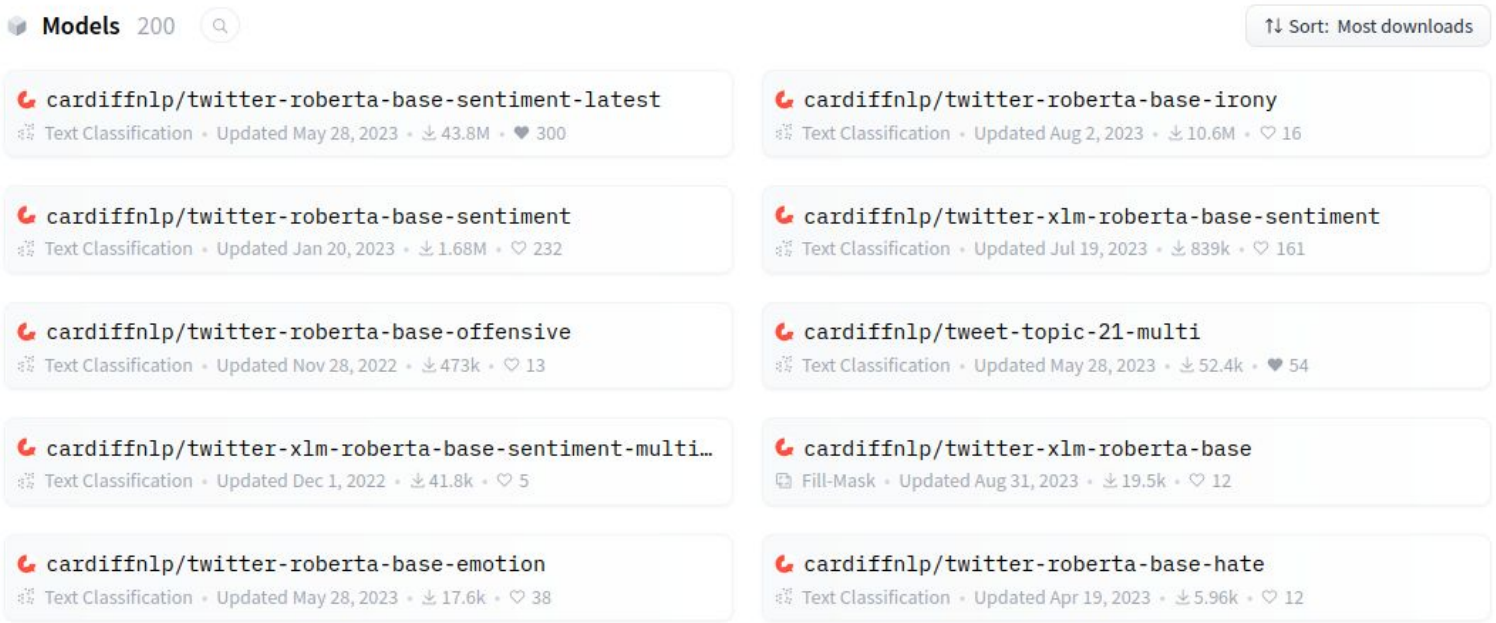
Also tasks with **temporal splits**!

Results? Smaller specialized models with supervision still better than LLMs (including ChatGPT)

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

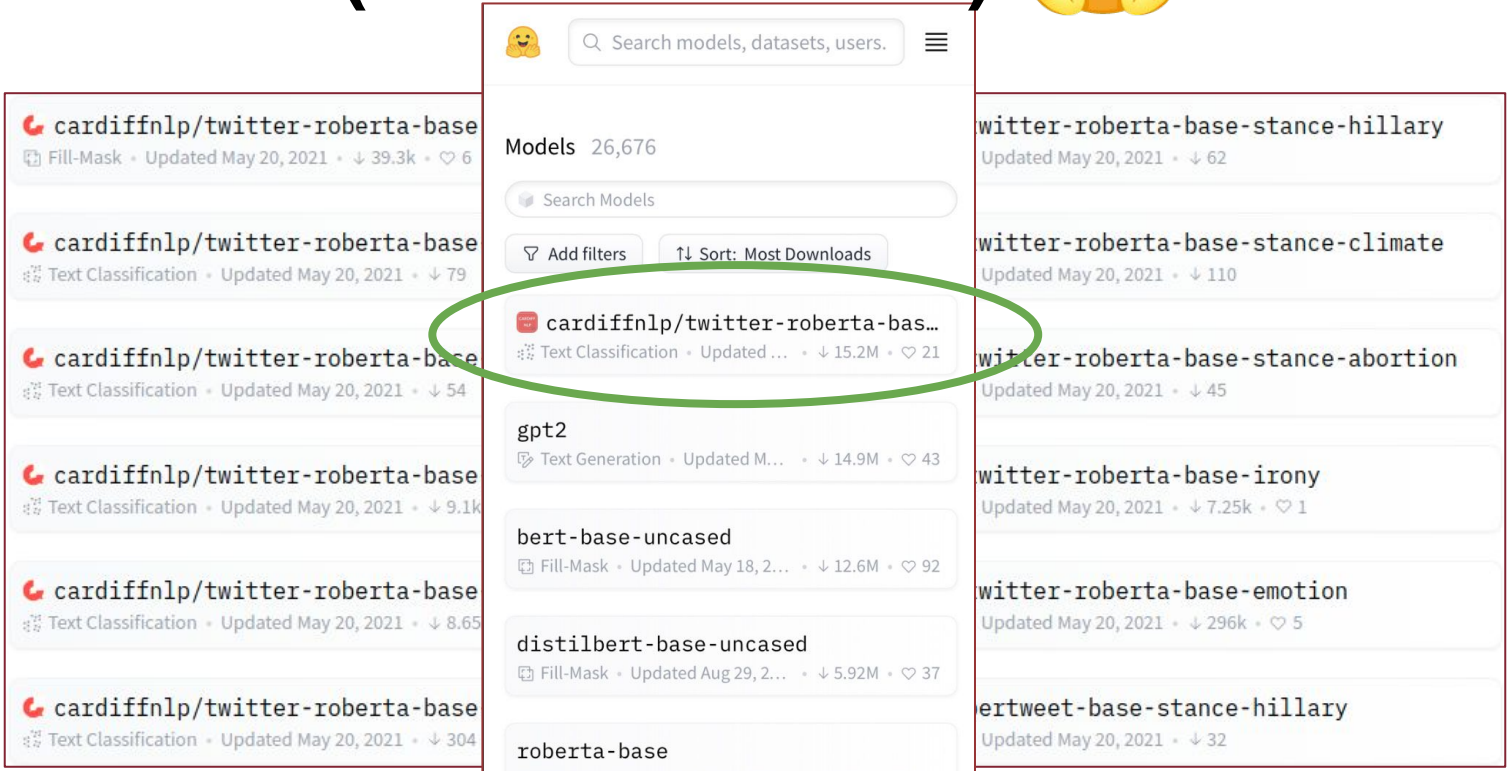
Specialized language models (+fine-tuned) 🤗



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

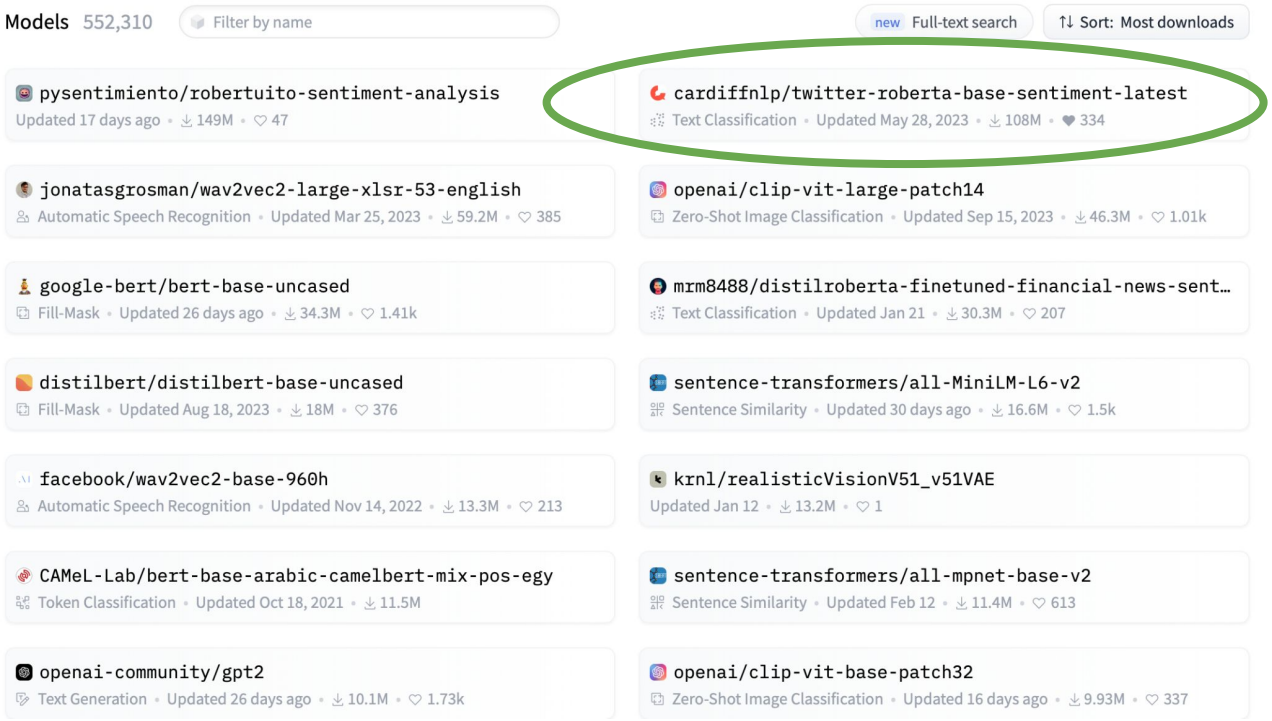
Specialized language models (+fine-tuned) 🤗



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

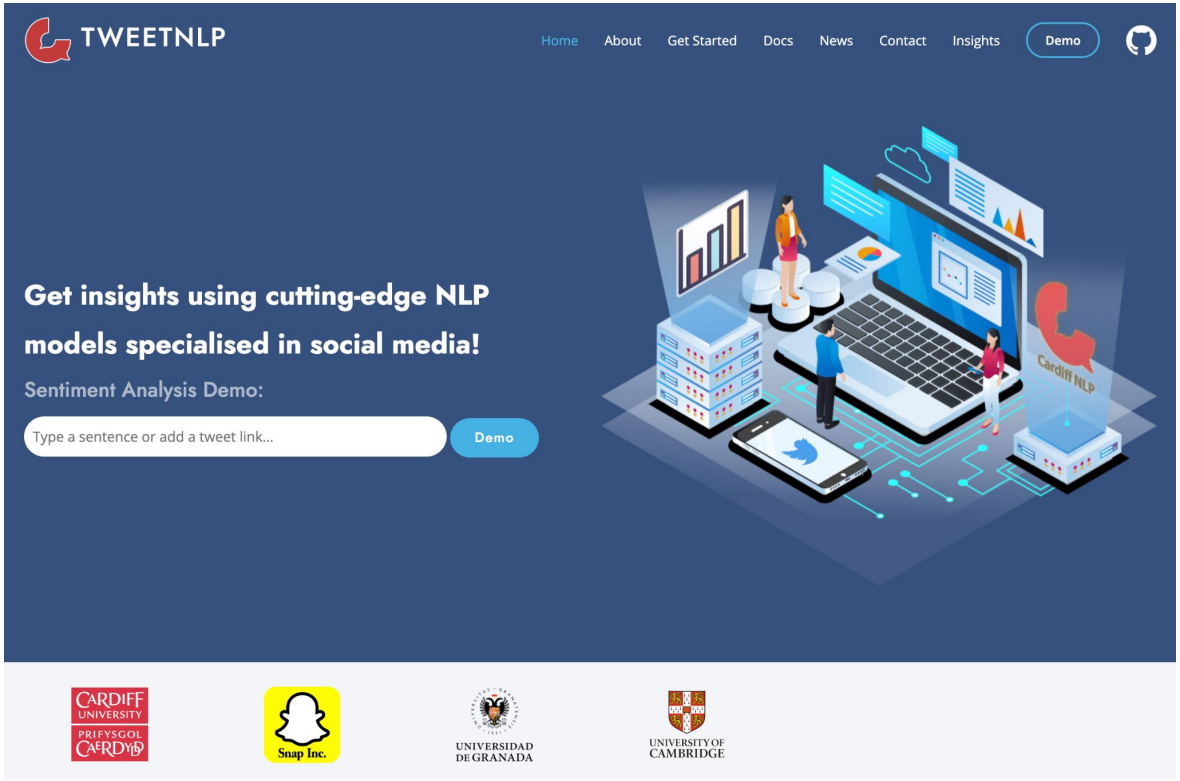
Specialized language models (+fine-tuned) 🤗



Two months ago, over 100 million downloads on a month 🤯

Cardiff NLP

TweetNLP (tweetnlp.org)



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetNLP - the team grows!

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 (Camacho-Collados et al. EMNLP Demo 2022)

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, **tutorials** and **Python API**.

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

TweetNLP Python library

Includes pre-trained models, inference, fine-tuning, evaluation…

```
import tweetnlp

# ENGLISH MODEL
model = tweetnlp.load_model('sentiment')  # Or `model = tweetnlp.Sentiment()`
model.sentiment("Yes, including Medicare and social security saving👍")  # Or
>>> {'label': 'positive'}
```

https://github.com/cardiffnlp/tweetnlp

TWEETNLP

Sentiment analysis

Type a sentence or a tweet to get insights (tweet URLs are also accepted)

Predictions are based on an English or a multilingual model. Languages supported are:

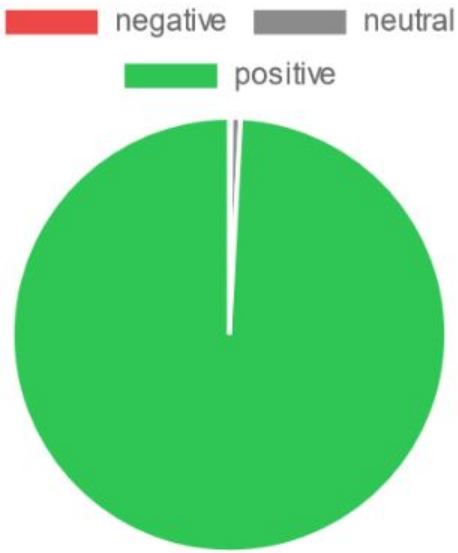
Today is a lovely day! 😊

For Example: Today is a lovely day!, I really don't like eating vegetables, https://twitter.com/Cardiff_NLP/status/1485518987807137792

Sentiment

English

GO!



Word prediction

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

- Sentence/Tweet Classification
- Hashtag Analysis
- **Word Prediction**
- Sentence/Tweet similarity
- Named Entity Recognition
- Question Answering/Generation

Type a sentence or a tweet with a masked word (<mask>) to predict the most likely word. every three months to select from.

Haaland! That was <mask>

For Example: I keep forgetting to bring a <mask>., COVID is a <mask>., So glad I'm <mask> vaccinated, Looking forward to watching <mask> Game tonight!

Latest June 2022 - English | GO!

fast (5%)

quick (4%)

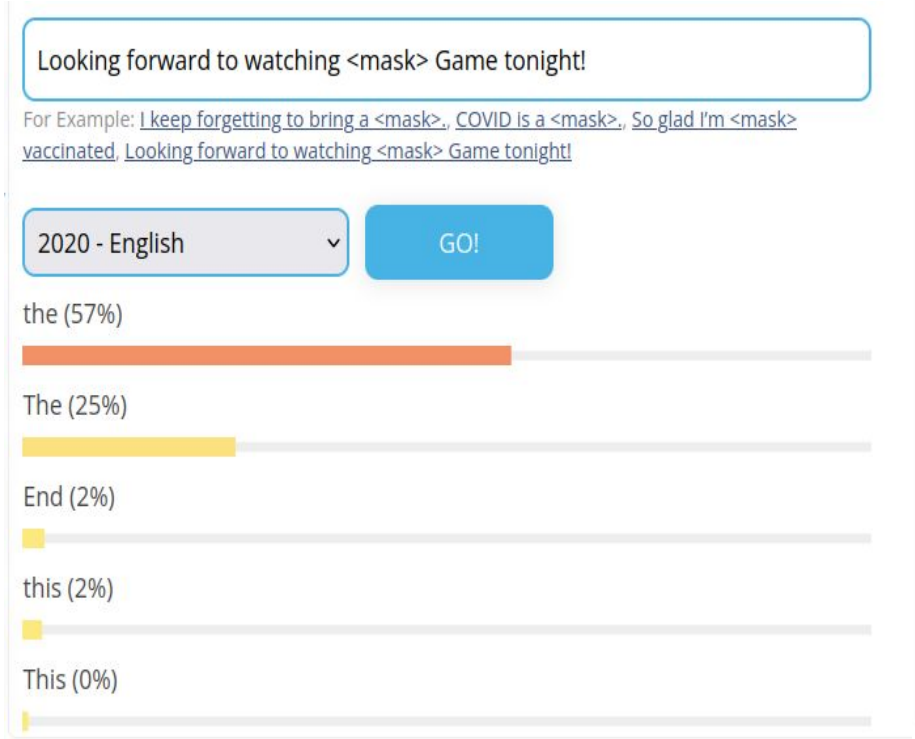
it (3%)

close (3%)

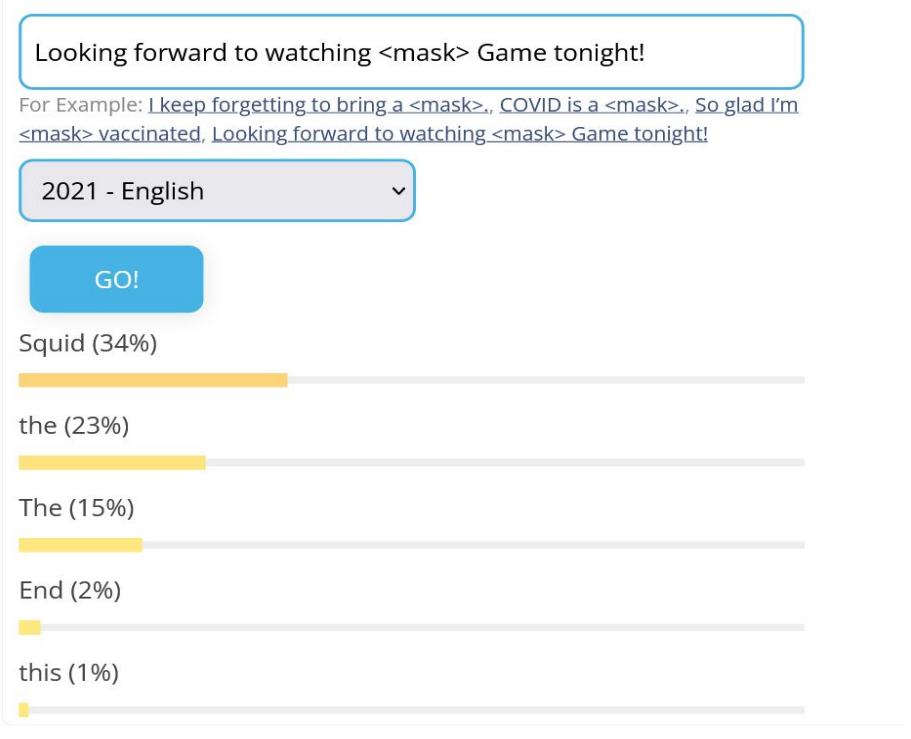
amazing (2%)

Word prediction (different years)

2020 model



2021 model



Cardiff NLP

Topic classification

(Antypas and Ushio et al. COLING 2022)

Type a sentence or a tweet to get insights (tweet URLs are also accepted)

Predictions are based on English (all tasks) or a multilingual model (sentiment). Languages supported are:

https://twitter.com/livescore/status/1632652988228710402

For Example: Today is a lovely day!, I really don't like eating vegetables, https://twitter.com/Cardiff_NLP/status/1485518987807137792

Topic classification | English

Note: Tweets get classified into one or more of 19 topics.

GO!

sports (99%)

gaming (2%)

news & social concern (1%)

celebrity & pop culture (1%)

diaries & daily life (1%)

Tweet

LiveScore @livescore

Liverpool have outscored Manchester United 18-1 in the last eight matches at Anfield 😴🔥 https://t.co/KQC7Bgvei7

Named Entity Recognition (NER)

(Ushio et al. AACL 2022)

Type a sentence or tweet link to get named entities.

https://twitter.com/BBCWorld/status/1532399905217597440

GO!

For Example: My name is Wolfgang and I live in Berlin ,Paradise is a song by Coldplay ,https://twitter.com/BBCWorld/status/1532399905217597440

Why **Johnny Depp** person lost in the **UK** location but won in the **US** location https://t.co/X5xheiDw2C

Tweet

BBC News (World) @BBCWorld

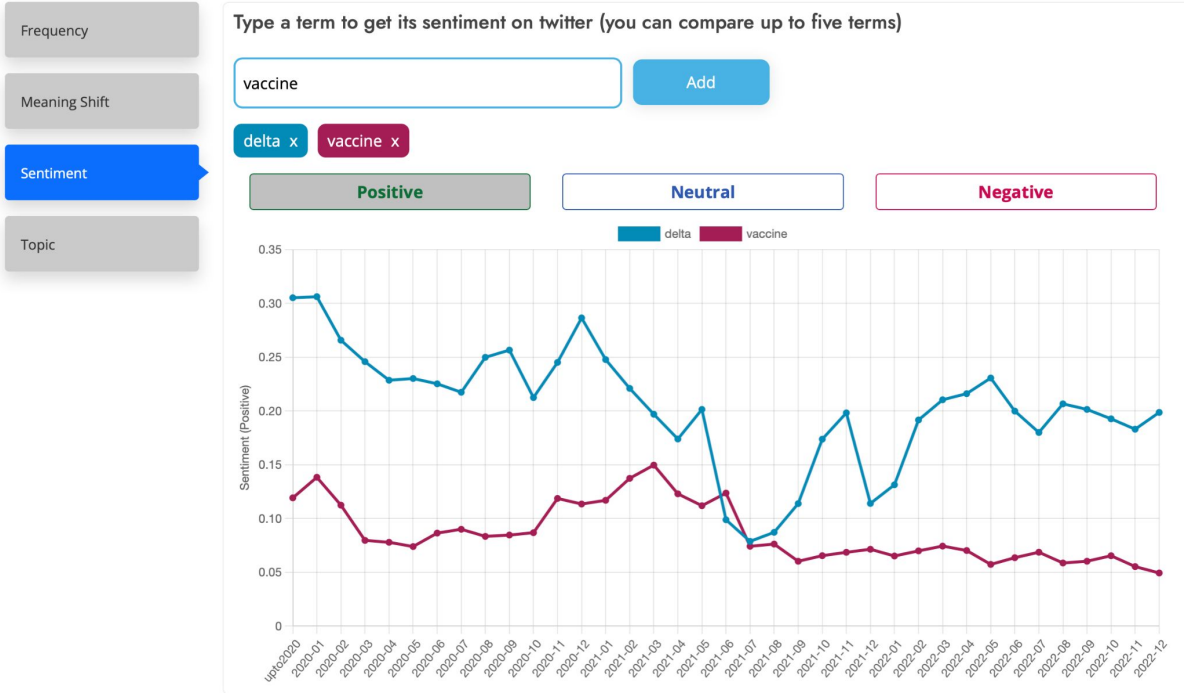
Why Johnny Depp lost in the UK but won in the US https://t.co/X5xheiDw2C

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

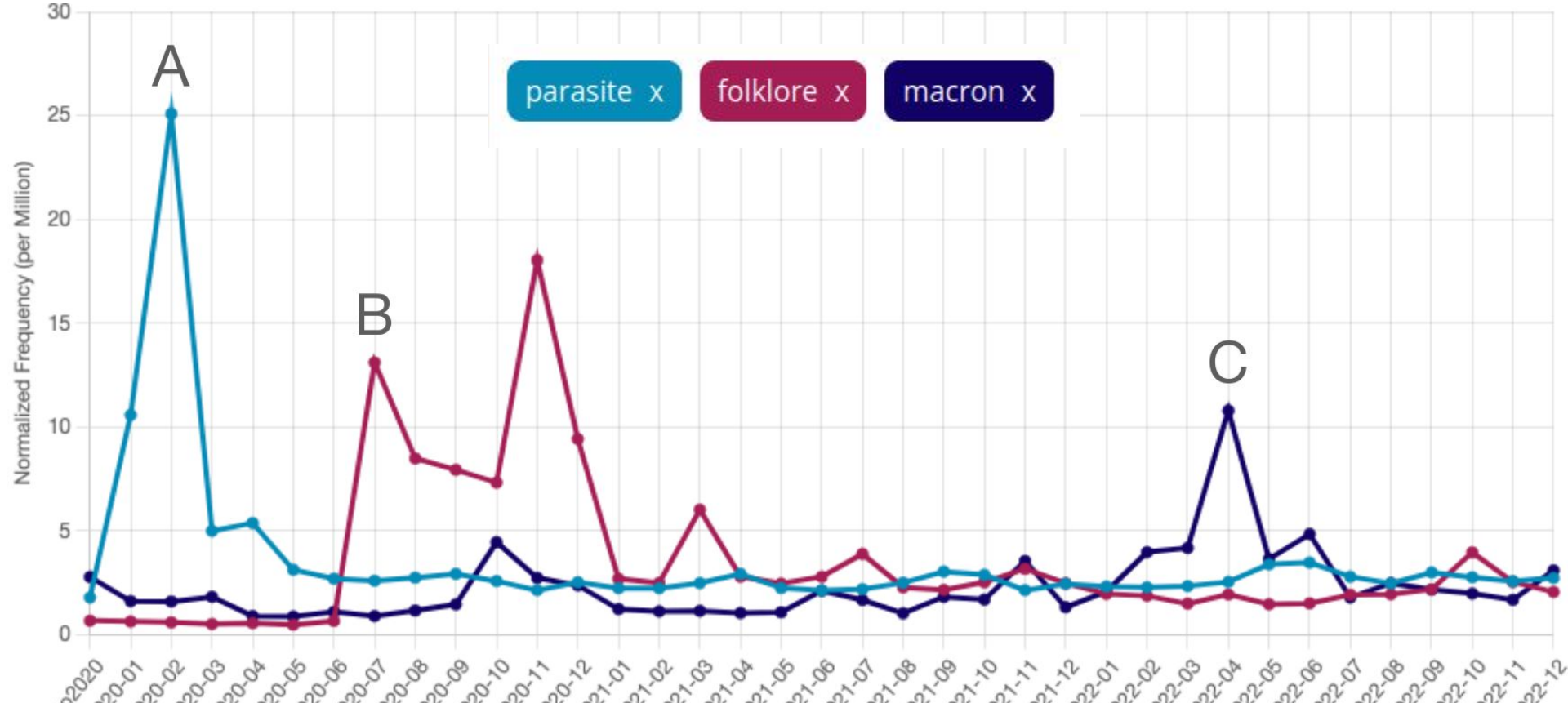
Tweet Insights

tweetnlp.org/insights



Tweet Insights

(Loureiro et al. 2023)



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Temporal challenges in NLP

Language is **changing** all the time.

New terms being introduced (e.g. *COVID-19*) or terms acquired new meanings (e.g. *Karen*).

Language models are not constantly updated.

This is especially true in **social media**, which is very dynamic.

NER and Topic Classification

(Antypas et al. COLING 2022; Ushio et al. AACL 2022)

Two datasets with temporal splits (i.e. training and test sets from different time periods):

- **TweetNER7** (Ushio et al. 2022) for NER
- **TweetTopic** (Antypas and Ushio et al. 2022) for topic classification

NER and Topic Classification

(Antypas et al. COLING 2022; Ushio et al. AACL 2022)

Two datasets with temporal splits (i.e. training and test sets from different time periods):

- **TweetNER7** (Ushio et al. 2022) for NER
- **TweetTopic** (Antypas and Ushio et al. 2022) for topic classification

Conclusion: Performance on temporal test splits lower than when dates are shuffled.

NER and Topic Classification

(Antypas et al. COLING 2022; Ushio et al. AACL 2022)

Two datasets with temporal splits (i.e. training and test sets from different time periods):

- **TweetNER7** (Ushio et al. 2022) for NER
- **TweetTopic** (Antypas and Ushio et al. 2022) for topic classification

Conclusion: Performance on temporal test splits lower than when dates are shuffled.

LongEval series at CLEF to evaluate performance drop over time

Temporal challenges

(Ushio and Camacho-Collados, arXiv 2024)

- **Analysis** of main sources of performance drop:
  - Pre-training data? Not really
  - Training data? YES
  - Nature of the domain/task? YES, entity- or event-driven

    particularly affected (e.g. NER, entity disambiguation, hate speech detection)

Temporal challenges

(Ushio and Camacho-Collados, arXiv 2024)

- **Analysis** of main sources of performance drop:
  - Pre-training data? Not really
  - Training data? YES
  - Nature of the domain/task? YES, entity- or event-driven

    particularly affected (e.g. NER, entity disambiguation, **hate speech detection**)

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Hate speech detection

⚠️ **Warning: Offensive language**

Hate speech detection

#YesAllWomen should stay in the kitchen

Hitler didn't finish it. Can u. If a n****r ur Jew confronts u in the street what then.

ditrty stinky sp*c -URL-

i am locked in for a month and will probably lose my job. i can't pay rent. all they are worried about is what name i call it. they eat dogs and bats are we really shocked. #chinavirus

The driving age for females should be like 25, y'all can't drive for sh*t 😂

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Challenges in hate speech detection

In addition to those specific to social media, some challenges:

- Limited resources (not diverse)
- Culturally specific, not a global definition (inherently subjective)

Cardiff NLP

Cross-dataset analysis

(Antypas and Camacho-Collados, 2023)

Macro-F1 results on 13 hate speech datasets.

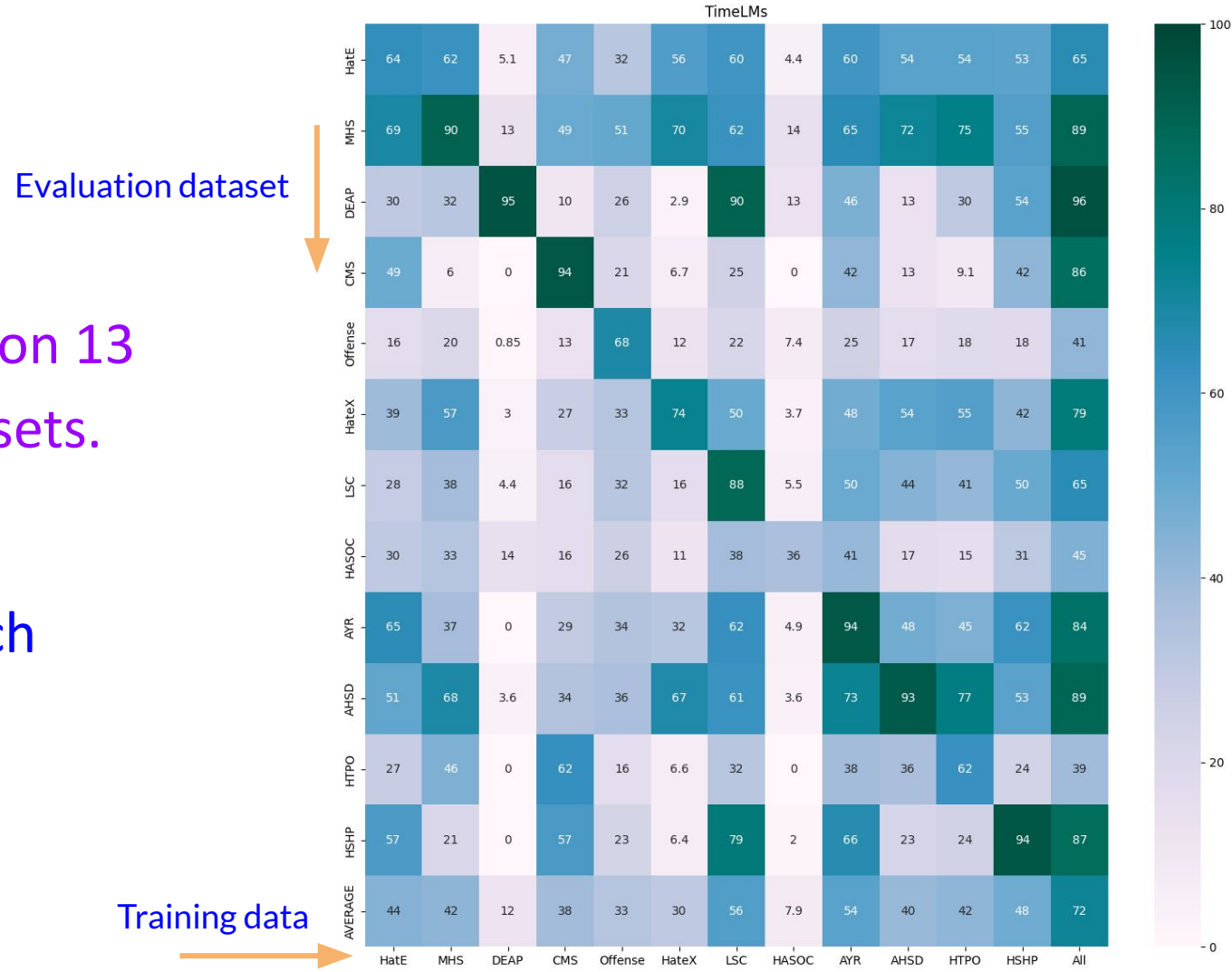
Specialised LM fine-tuned on each dataset.

Cardiff NLP

Results

Macro-F1 results on 13 hate speech datasets.

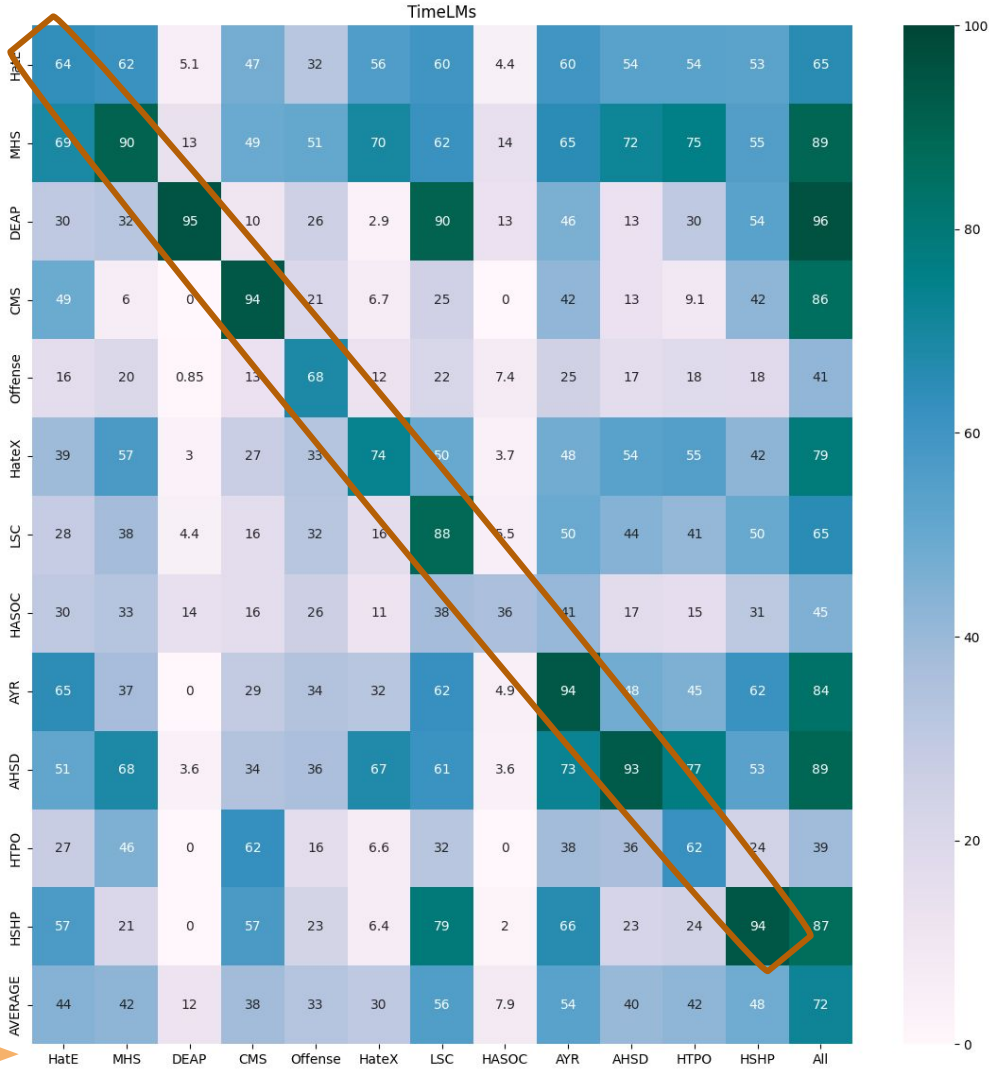
Specialised LM fine-tuned on each dataset.



Cardiff NLP

Results

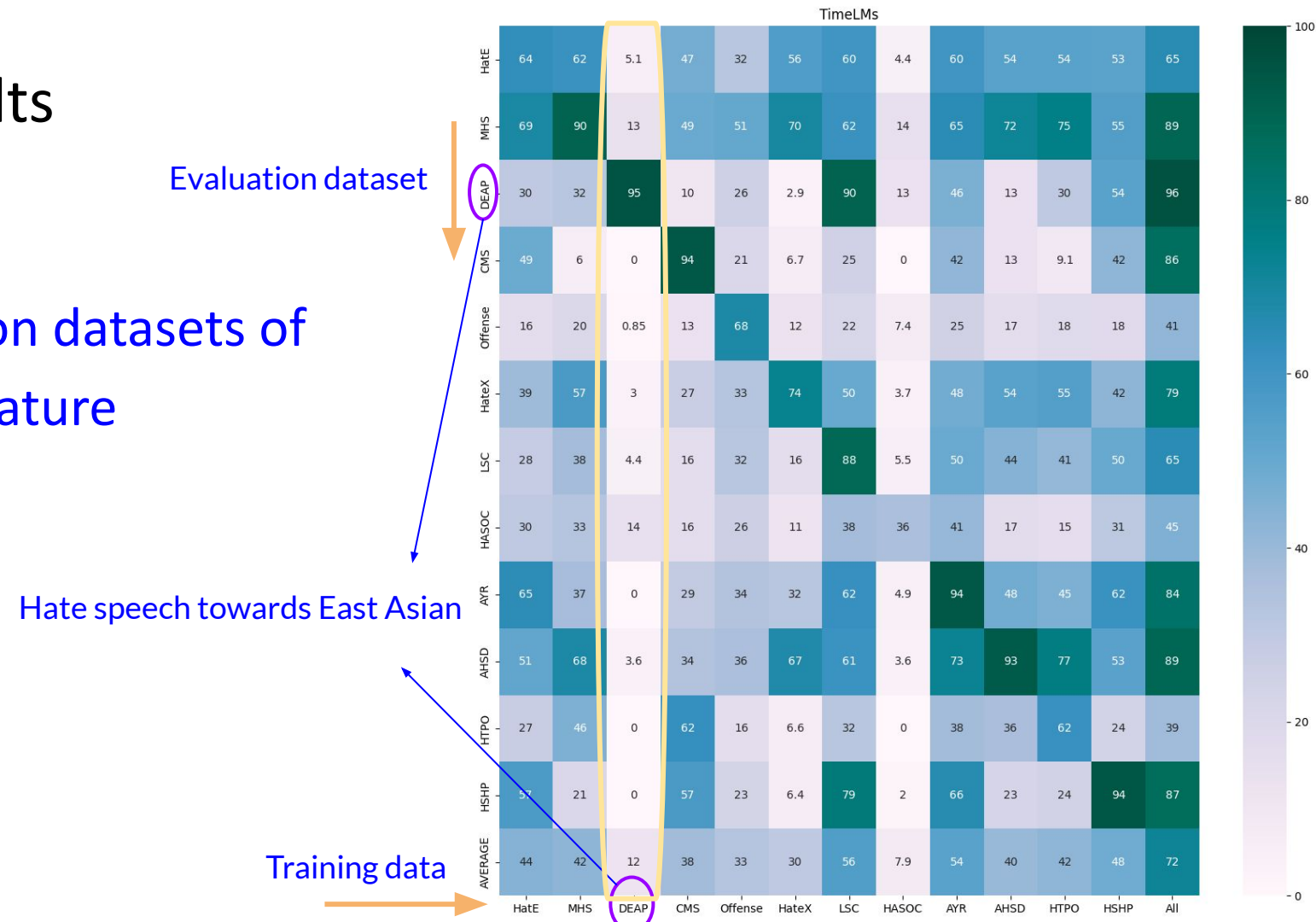
Best results when trained and evaluated on the same dataset



Cardiff NLP

Results

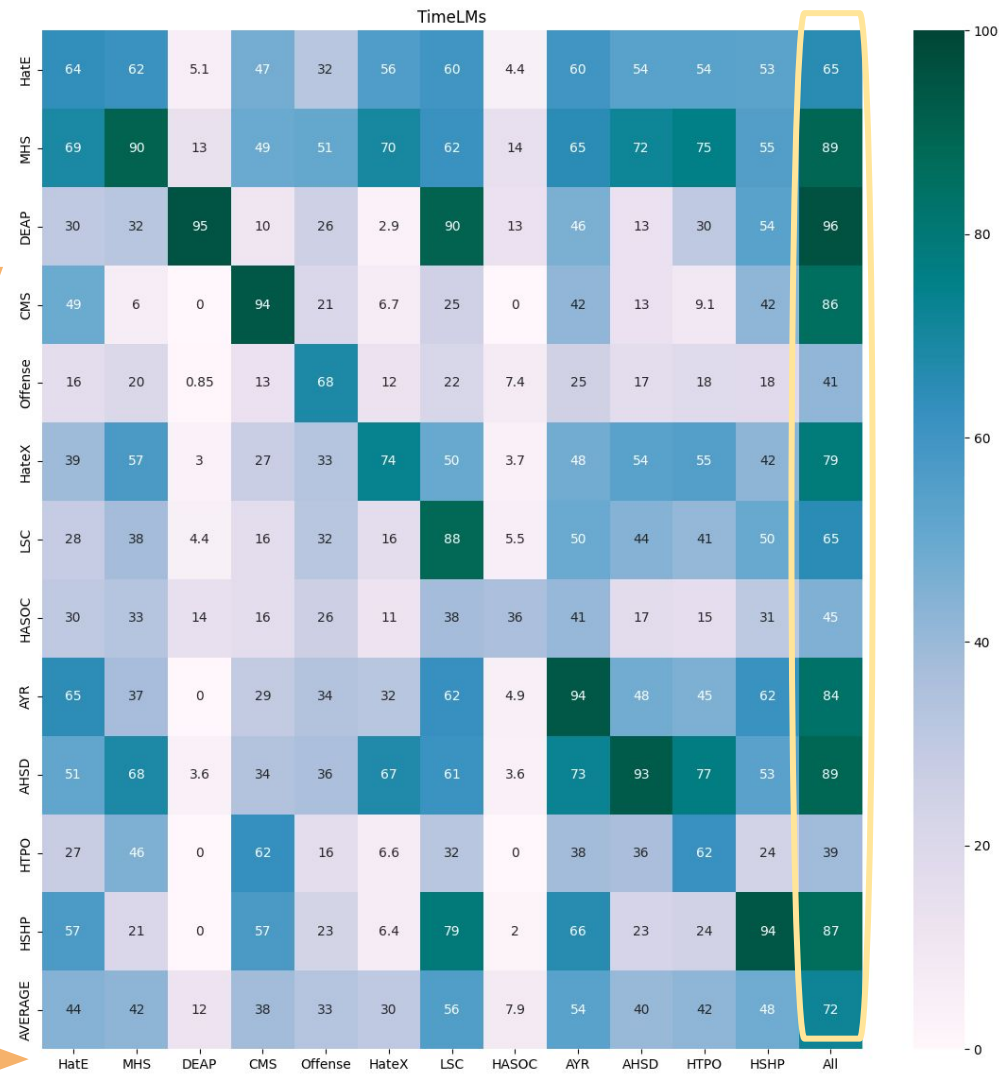
Not good on datasets of different nature



Cardiff NLP

Results

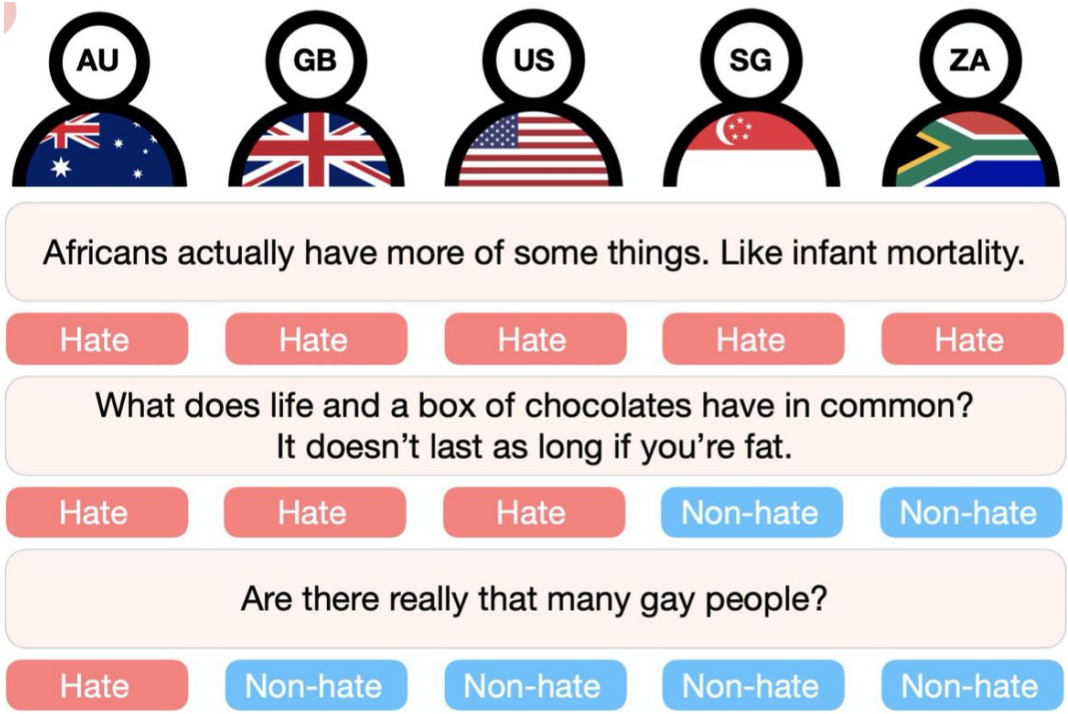
Model trained on all datasets is more robust.



Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

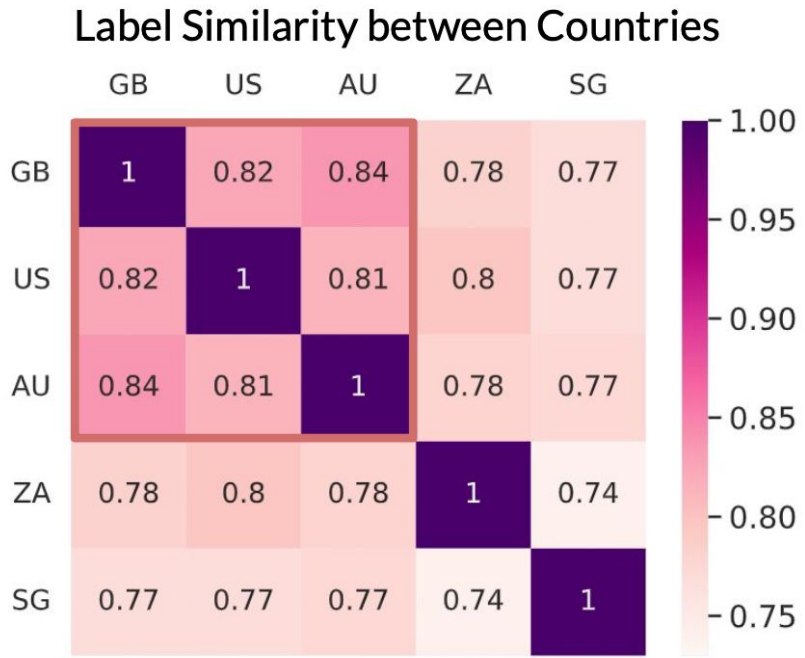
Hate speech dataset annotated by people from 5 different countries



Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

UK, US and Australia annotations are similar.

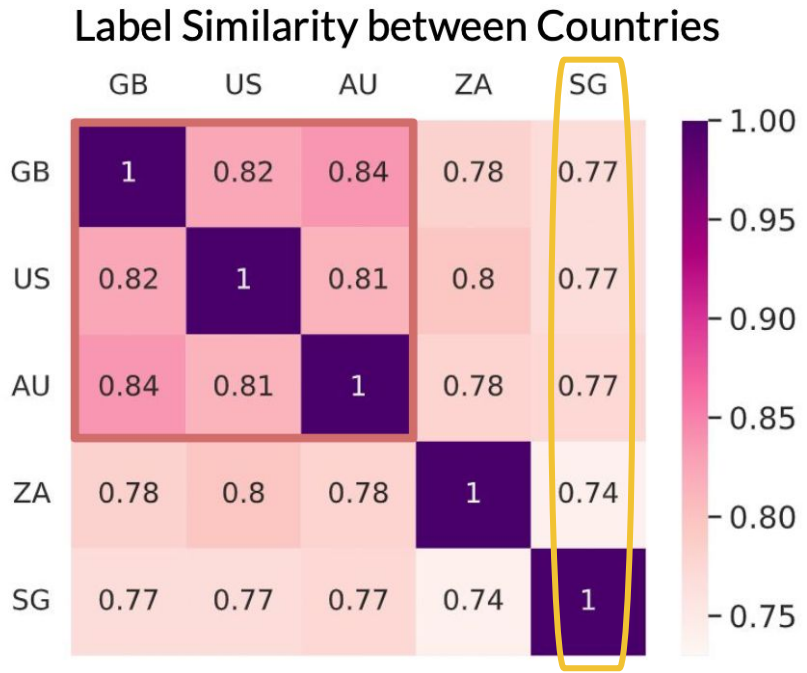


Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

UK, US and Australia annotations are similar.

Singapore and South Africa differ.



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

What about LLMs?

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

Work in progress

(other multidisciplinary collaborations)

- **Polarisation** (e.g. in politics) - *trigger words*.
- Analysing **earthquakes responses** in social media.
- Early health care interventions: **depression detection** on social media (Twitter, Reddit)
- Finding outbreaks and adherence/sentiment to **health interventions** (e.g. COVID) using social media

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Conclusion

LLMs may not be the best solution for all problems.

Specialized language models are a good solution to domain-specific tasks (plus: no need for huge models!)

Challenges remain (temporal awareness, biases, etc.)

Applications are endless, **huge opportunities** for NLPers.

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Conclusion

LLMs may not be the best solution for all problems.

Specialized language models are a good solution to domain-specific tasks (plus: no need for huge models!)

Challenges remain (temporal awareness, biases, etc.)

Applications are endless, **huge opportunities** for NLPers.

Caveat: Perhaps in a few months a new LLM solves everything!

Summary of resources

TweetNLP

github.com/cardiffnlp/tweetnlp

tweetnlp.org

All models available in the Hugging Face hub:

https://huggingface.co/cardiffnlp 🤗

Thank you!

@CamachoCollados