

Practical and Specialised NLP Solutions:

The Case of Social Media

Jose Camacho-Collados



Cardiff NLP



Cambridge, 30 May 2024

Outline




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

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

- Young group (3 years old), growing fast (30+ lab members)
- **Website:** cardiffnlp.github.io 
- Activities: hybrid seminars, workshops, hackathons, etc.
- **Twitter:** [@Cardiff_NLP](https://twitter.com/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

www.cardiffnlpworkshop.org

Registration open
until June 5th!



Javad Hosseini
Google Deepmind, UK



Nafise Sadat Moosavi
University of Sheffield



Emanuele Bugliarello
Google Research, France



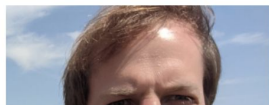
Marie-Francine Moens
KU Leuven, Belgium

Cardiff NLP Workshop

(1-2 July 2024)



Andreas Vlachos
University of Cambridge



www.cardiffnlpworkshop.org



A Tutorial on "Building RAG applications"



Panel on "Career in NLP after graduation"



Poster session

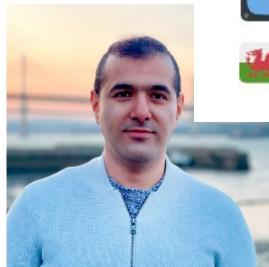


Networking opportunities



Welsh delights

Registration open
until June 5th!



Javad Hosseini
Google Deepmind, UK



Nafise Sadat Moosavi
University of Sheffield



Emanuele Bugliarello
Google Research, France



Marie-Francine Moens
KU Leuven, Belgium

(Large) Language Models

Language models (LMs)

Tasks

Question
Answering



Text
Classification



Information
Retrieval



⋮

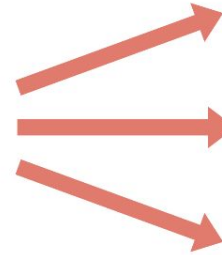
Text corpus



(Self-supervised)
Training



Pretrained LM



Adaptation

Language models (LMs)

Text corpus



(Self-supervised)
Training

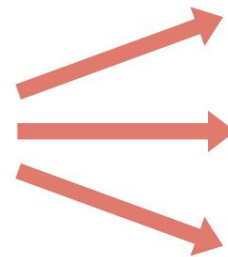
Pretrained LM



LLaMA



GPT



Adaptation

Tasks

Question
Answering



Text
Classification



Information
Retrieval



⋮

Slide credit: Stanford AI

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] [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  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.

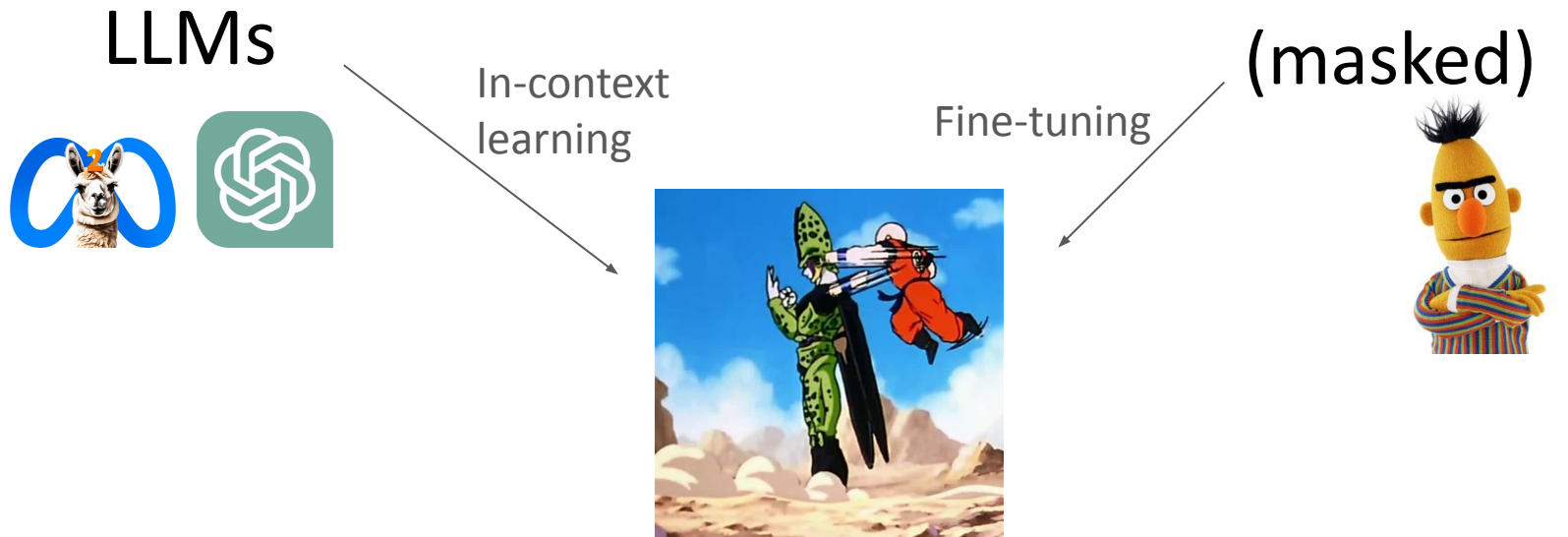
I thought AI could help. And  suggested you.

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

Email date: 26 September

Text Classification: Is In-Context Learning enough?

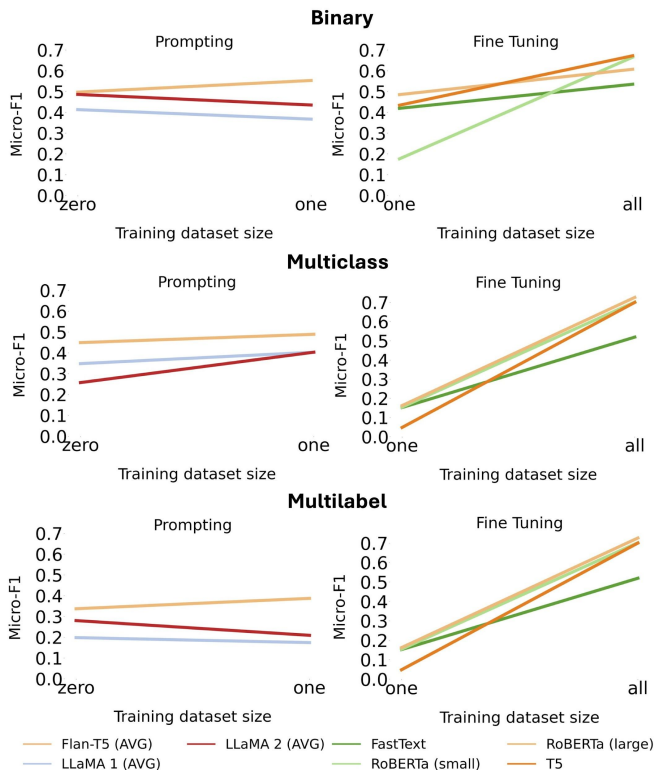
(Edwards and Camacho-Collados, LREC/COLING 2024)



Text Classification:

Is In-Context Learning enough?

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



TLDR:

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

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”

Social media as a case domain



Social media is a challenging domain

➤ Why?

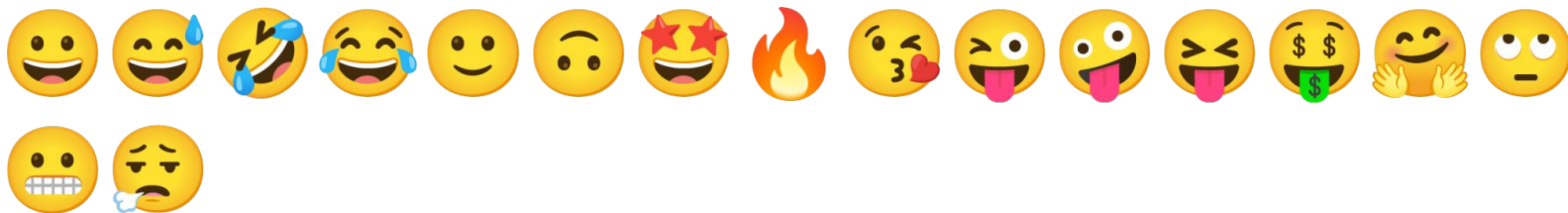
- **Informal** grammar
- **Multilingual** (code-switching, etc.)
- Irregular **vocabulary**
 - Emoji 😊, abbreviations, typos, hashtags, mentions...
- Tweets are often **not standalone messages**
 - RTs, mentions, replies, threads, pictures...
- **Dynamic**, constantly changing

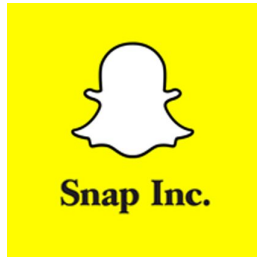
Social media: *My story*

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

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**:





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?



Haaland! That was <mask>

it, right, me,
him, good...



fast, quick, close,
amazing...



TweetEval, the benchmark

(Barbieri et al. EMNLP Findings 2020)

➤ Why?

TweetEval, the benchmark

➤ Why?



TweetEval, the benchmark

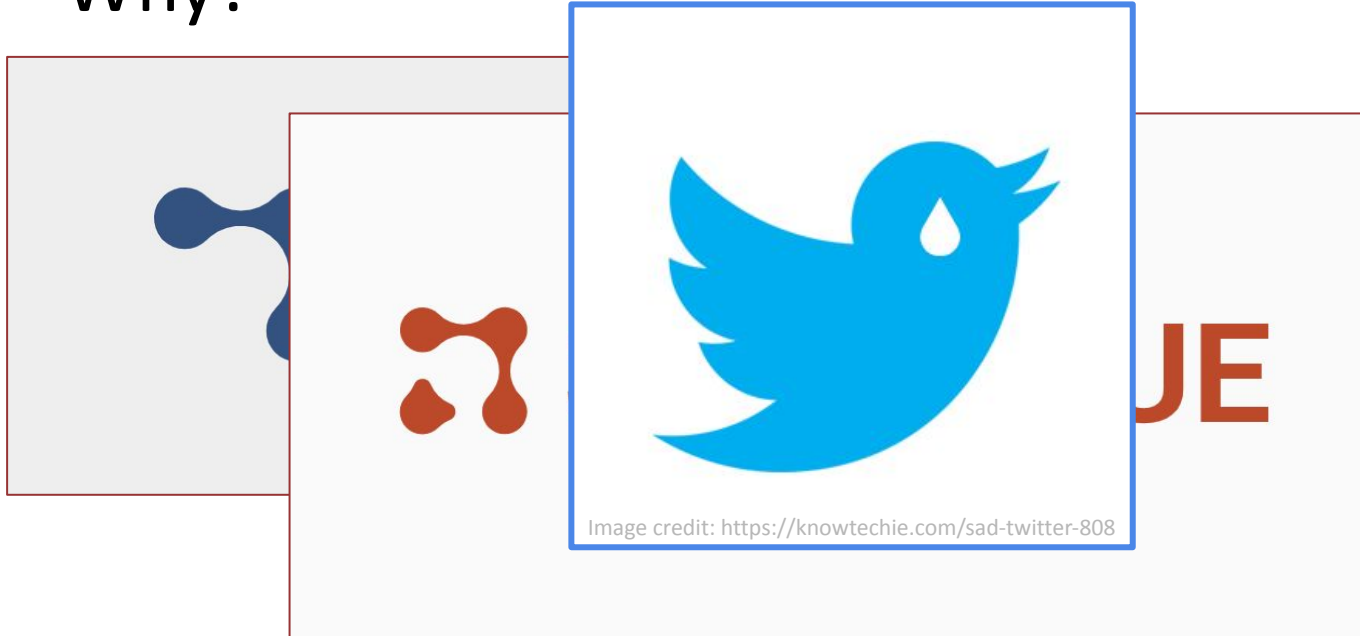
➤ Why?



 **SuperGLUE**

TweetEval, the benchmark

➤ Why?



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: SpiderMAIS 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

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)


Specialized language models (+fine-tuned) 🤗

Models 200


Sort: Most downloads

 **cardiffnlp/twitter-roberta-base-sentiment-latest**


 Text Classification • Updated May 28, 2023 •  43.8M •  300

 **cardiffnlp/twitter-roberta-base-irony**


 Text Classification • Updated Aug 2, 2023 •  10.6M •  16

 **cardiffnlp/twitter-roberta-base-sentiment**


 Text Classification • Updated Jan 20, 2023 •  1.68M •  232

 **cardiffnlp/twitter-xlm-roberta-base-sentiment**


 Text Classification • Updated Jul 19, 2023 •  839k •  161

 **cardiffnlp/twitter-roberta-base-offensive**


 Text Classification • Updated Nov 28, 2022 •  473k •  13

 **cardiffnlp/tweet-topic-21-multi**


 Text Classification • Updated May 28, 2023 •  52.4k •  54

 **cardiffnlp/twitter-xlm-roberta-base-sentiment-multi...**


 Text Classification • Updated Dec 1, 2022 •  41.8k •  5


 **cardiffnlp/twitter-xlm-roberta-base**

 Fill-Mask • Updated Aug 31, 2023 •  19.5k •  12

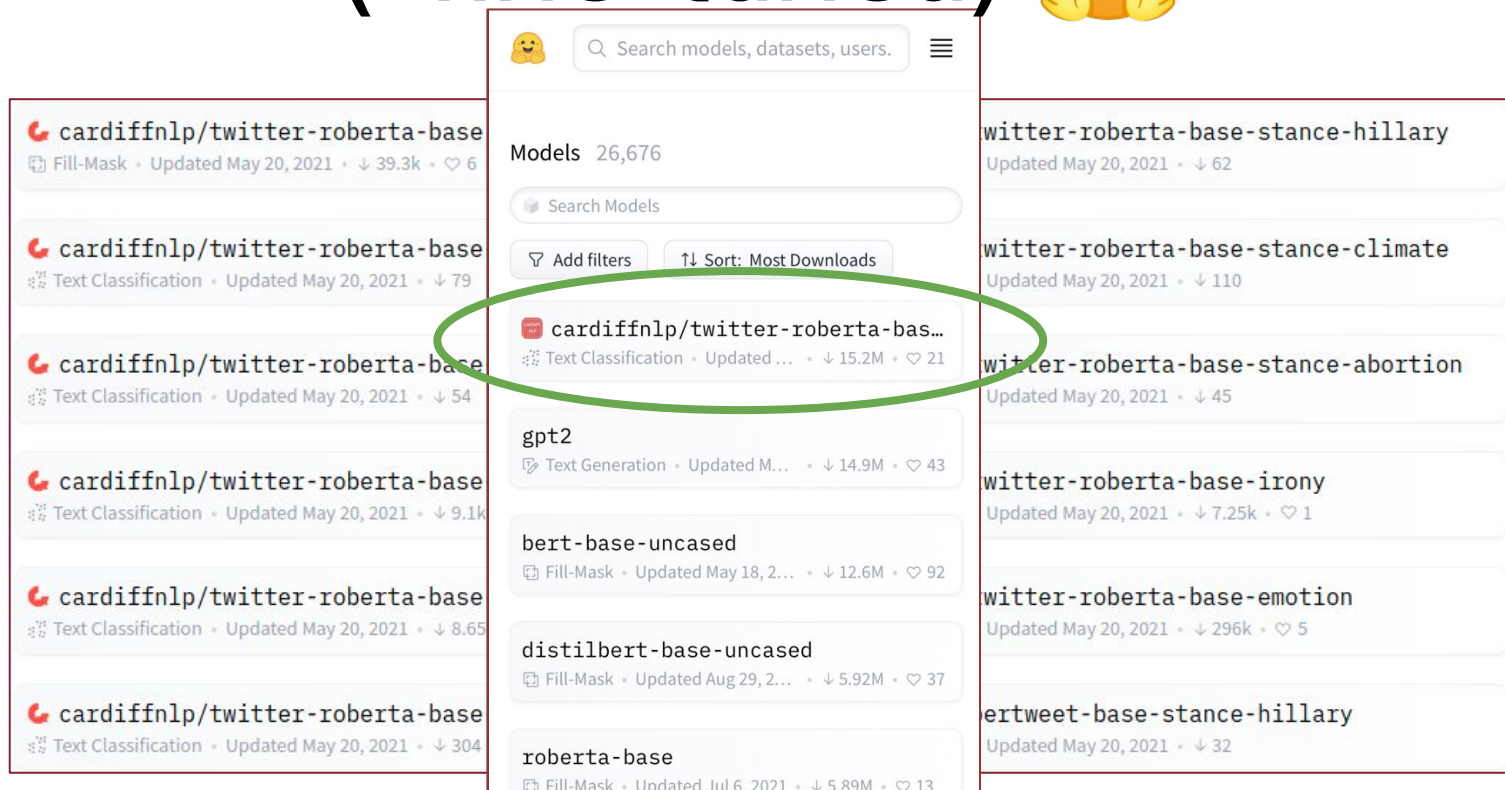
 **cardiffnlp/twitter-roberta-base-emotion**

 Text Classification • Updated May 28, 2023 •  17.6k •  38

 **cardiffnlp/twitter-roberta-base-hate**

 Text Classification • Updated Apr 19, 2023 •  5.96k •  12

Specialized language models (+fine-tuned) 🤗

















The screenshot displays the Hugging Face models interface. At the top, there is a search bar with the text "Search models, datasets, users." and a menu icon. Below the search bar, the text "Models 26,676" is visible. A search bar with the text "Search Models" is also present. Below this, there are two buttons: "Add filters" and "Sort: Most Downloads". The main list of models is shown, with the model "cardiffnlp/twitter-roberta-base-stance-hillary" highlighted by a green circle. Other models listed include "cardiffnlp/twitter-roberta-base-stance-climate", "cardiffnlp/twitter-roberta-base-stance-abortion", "cardiffnlp/twitter-roberta-base-irony", "cardiffnlp/twitter-roberta-base-emotion", "bert-base-uncased", "distilbert-base-uncased", and "roberta-base".

Model Name	Task	Updated	Downloads	Likes
cardiffnlp/twitter-roberta-base-stance-hillary	Text Classification	Updated May 20, 2021	↓ 39.3k	♥ 6
cardiffnlp/twitter-roberta-base-stance-climate	Text Classification	Updated May 20, 2021	↓ 79	
cardiffnlp/twitter-roberta-base-stance-abortion	Text Classification	Updated May 20, 2021	↓ 54	
cardiffnlp/twitter-roberta-base-irony	Text Classification	Updated May 20, 2021	↓ 9.1k	
cardiffnlp/twitter-roberta-base-emotion	Text Classification	Updated May 20, 2021	↓ 8.65	
cardiffnlp/twitter-roberta-base	Text Classification	Updated May 20, 2021	↓ 304	
gpt2	Text Generation	Updated M...	↓ 14.9M	♥ 43
bert-base-uncased	Fill-Mask	Updated May 18, 2...	↓ 12.6M	♥ 92
distilbert-base-uncased	Fill-Mask	Updated Aug 29, 2...	↓ 5.92M	♥ 37
roberta-base	Fill-Mask	Updated Jul 6, 2021	↓ 5.89M	♥ 13

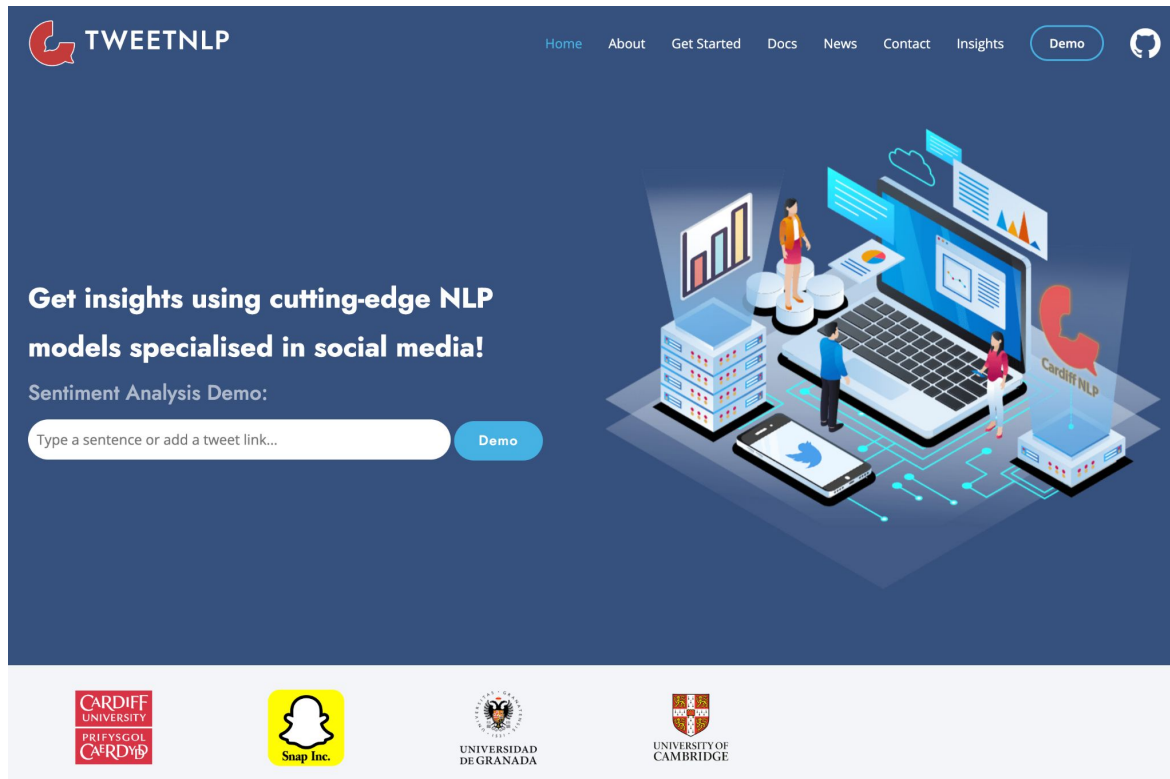
Specialized language models (+fine-tuned) 🤗

Models 552,310 new Full-text search Sort: Most downloads

 pysentimiento/robertuito-sentiment-analysis Updated 17 days ago • 149M • 47	 cardiffnlp/twitter-roberta-base-sentiment-latest Text Classification • Updated May 28, 2023 • 108M • 334
 jonatasgrosman/wav2vec2-large-xlsr-53-english Automatic Speech Recognition • Updated Mar 25, 2023 • 59.2M • 385	 openai/clip-vit-large-patch14 Zero-Shot Image Classification • Updated Sep 15, 2023 • 46.3M • 1.01k
 google-bert/bert-base-uncased Fill-Mask • Updated 26 days ago • 34.3M • 1.41k	 mr8488/distilroberta-finetuned-financial-news-sentiment-classification Text Classification • Updated Jan 21 • 30.3M • 207
 distilbert/distilbert-base-uncased Fill-Mask • Updated Aug 18, 2023 • 18M • 376	 sentence-transformers/all-MiniLM-L6-v2 Sentence Similarity • Updated 30 days ago • 16.6M • 1.5k
 facebook/wav2vec2-base-960h Automatic Speech Recognition • Updated Nov 14, 2022 • 13.3M • 213	 knl/realisticVisionV51_v51VAE Updated Jan 12 • 13.2M • 1
 CAMEL-Lab/bert-base-arabic-camelbert-mix-pos-egy Token Classification • Updated Oct 18, 2021 • 11.5M	 sentence-transformers/all-mpnet-base-v2 Sentence Similarity • Updated Feb 12 • 11.4M • 613
 openai-community/gpt2 Text Generation • Updated 26 days ago • 10.1M • 1.73k	 openai/clip-vit-base-patch32 Zero-Shot Image Classification • Updated 16 days ago • 9.93M • 337

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

TweetNLP (tweetnlp.org)



The screenshot shows the TweetNLP website interface. At the top left is the 'TWEETNLP' logo. To its right is a navigation menu with links: Home, About, Get Started, Docs, News, Contact, Insights, and a 'Demo' button. A user profile icon is also present. The main content area features the text 'Get insights using cutting-edge NLP models specialised in social media!' followed by 'Sentiment Analysis Demo:'. Below this is a text input field with the placeholder 'Type a sentence or add a tweet link...' and a 'Demo' button. On the right side of the main area is a large, colorful 3D isometric illustration depicting a person at a laptop, a smartphone with the Twitter logo, server racks, and various data visualizations like bar charts and pie charts. A red 'Cardiff NLP' logo is also part of the illustration. The footer contains logos for Cardiff University, Snap Inc., Universidad de Granada, and the University of Cambridge.

TWEETNLP

Home About Get Started Docs News Contact Insights Demo

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

Sentiment Analysis Demo:

Type a sentence or add a tweet link... Demo

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

SNAP INC.

UNIVERSIDAD DE GRANADA

UNIVERSITY OF CAMBRIDGE

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.



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! 😊

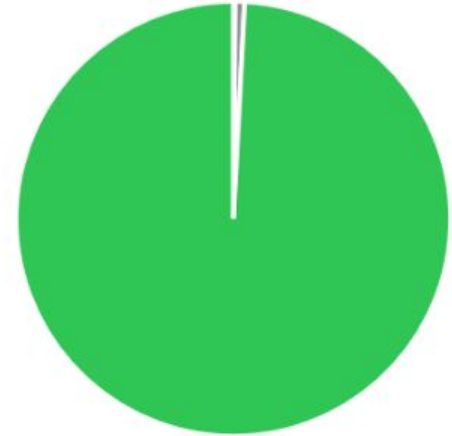
For Example: Today is a lovely day!, I really don't like eating vegetables.
https://twitter.com/Cardiff_NLP/status/1485518987807137792

Sentiment ▼

English ▼

GO!

negative neutral
positive



Word prediction

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

Looking forward to watching <mask> Game tonight!

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!

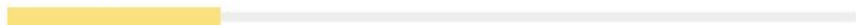
2020 - English

GO!

the (57%)



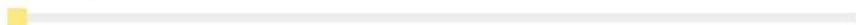
The (25%)



End (2%)



this (2%)



This (0%)



2021 model

Looking forward to watching <mask> Game tonight!

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!

2021 - English

GO!

Squid (34%)



the (23%)



The (15%)



End (2%)



this (1%)





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!

 Tweet



LiveScore @livescore

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

sports (99%)

gaming (2%)

news & social concern (1%)

celebrity & pop culture (1%)

diaries & daily life (1%)

Named Entity Recognition (NER)

(Ushio et al. ACL 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

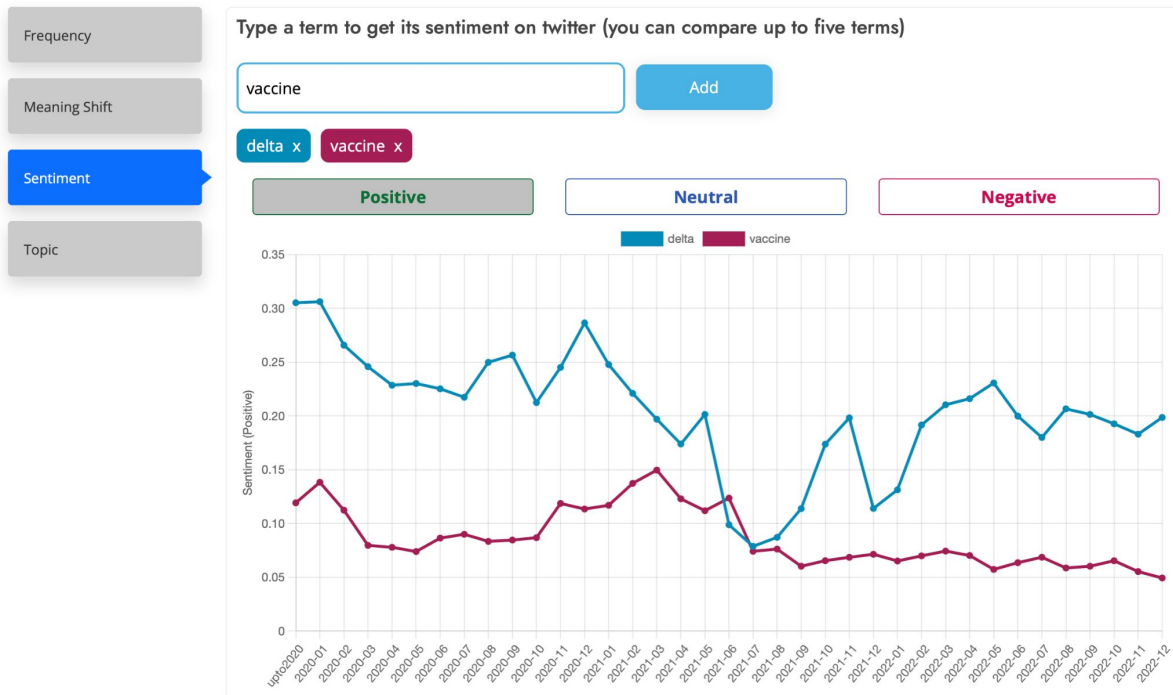


BBC News (World) @BBCWorld

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

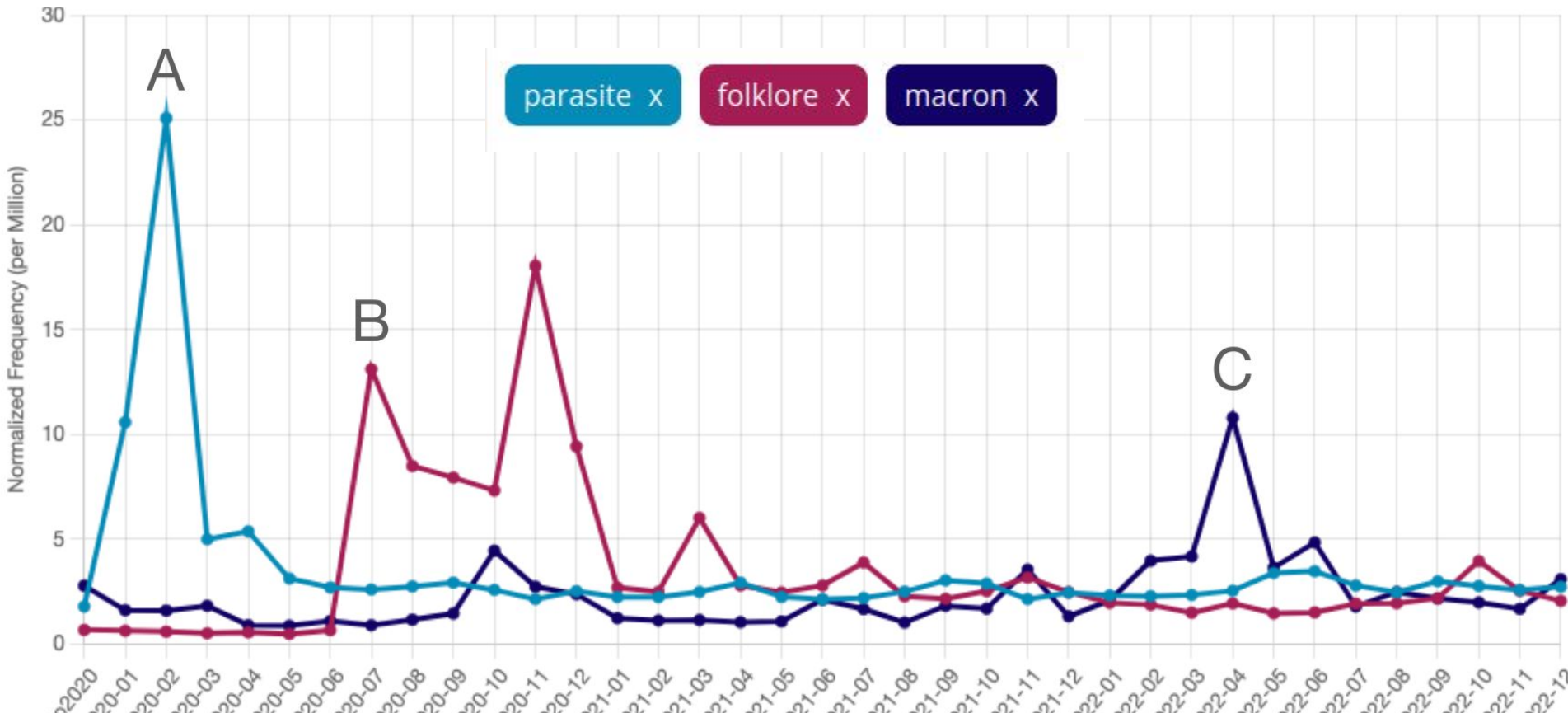
Tweet Insights

tweetnlp.org/insights



Tweet Insights

(Loureiro et al. 2023)



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. ACL 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**)

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 😂

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)

Cross-dataset analysis

(Antypas and Camacho-Collados, 2023)

Macro-F1 results on 13
hate speech datasets.

Specialised LM
fine-tuned on each
dataset.

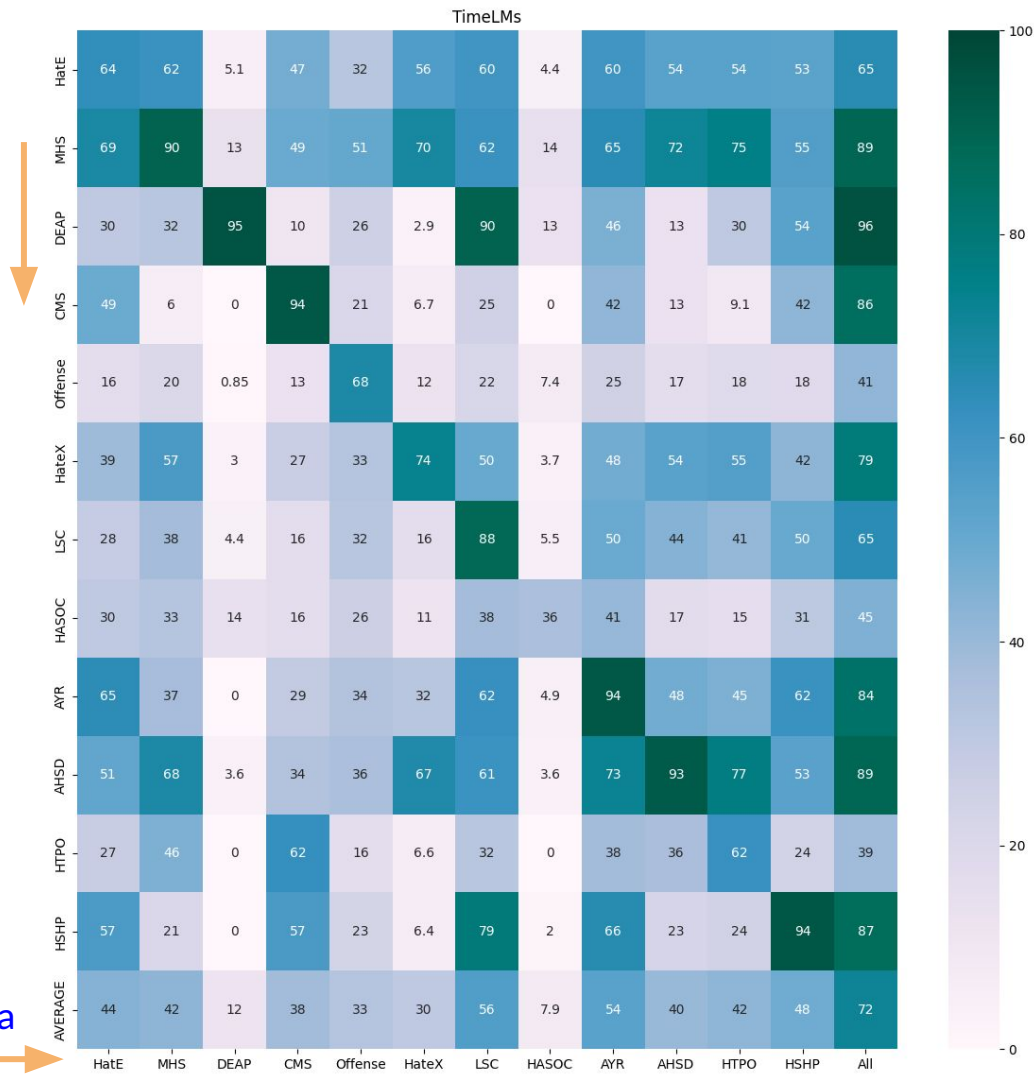
Results

Macro-F1 results on 13 hate speech datasets.

Specialised LM fine-tuned on each dataset.

Evaluation dataset

Training data

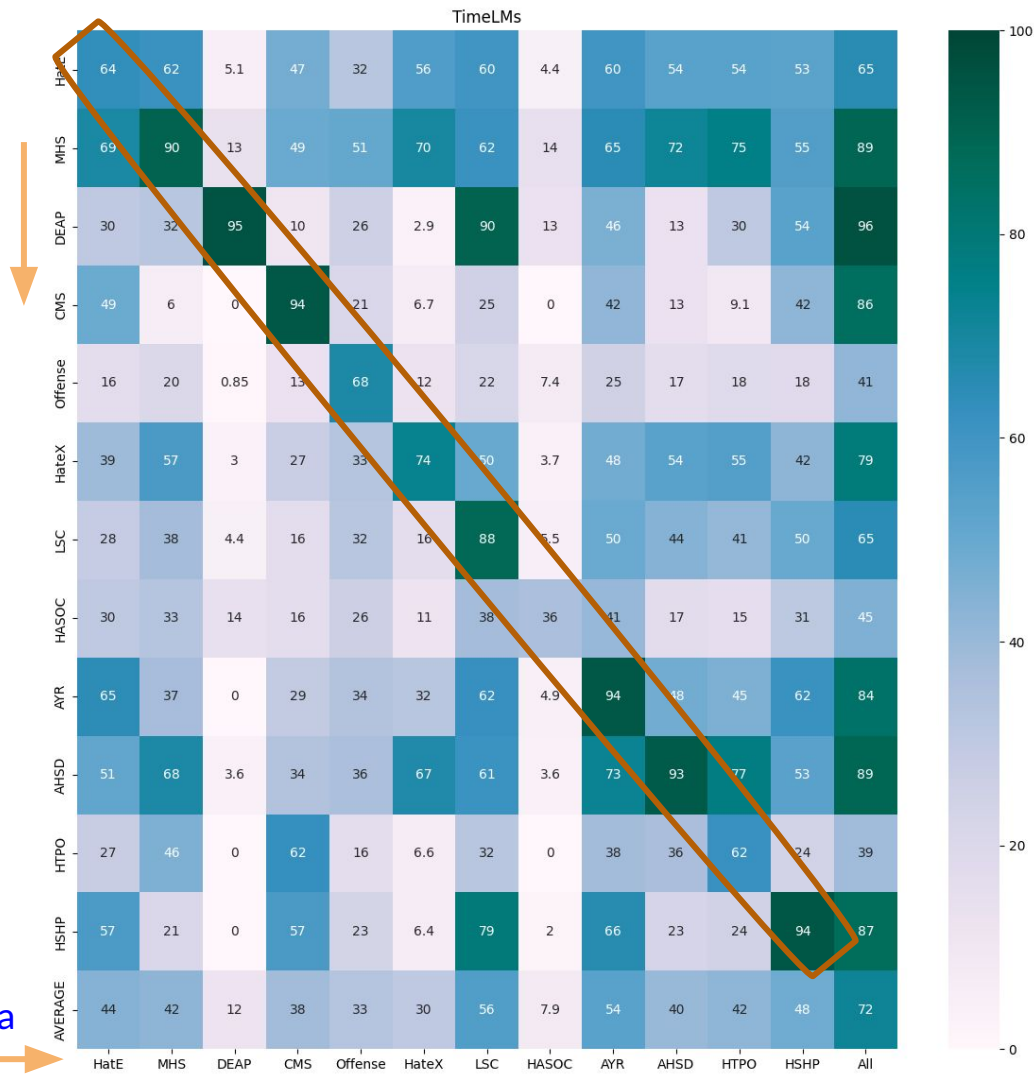


Results

Best results when
trained and evaluated on
the same dataset

Evaluation dataset

Training data



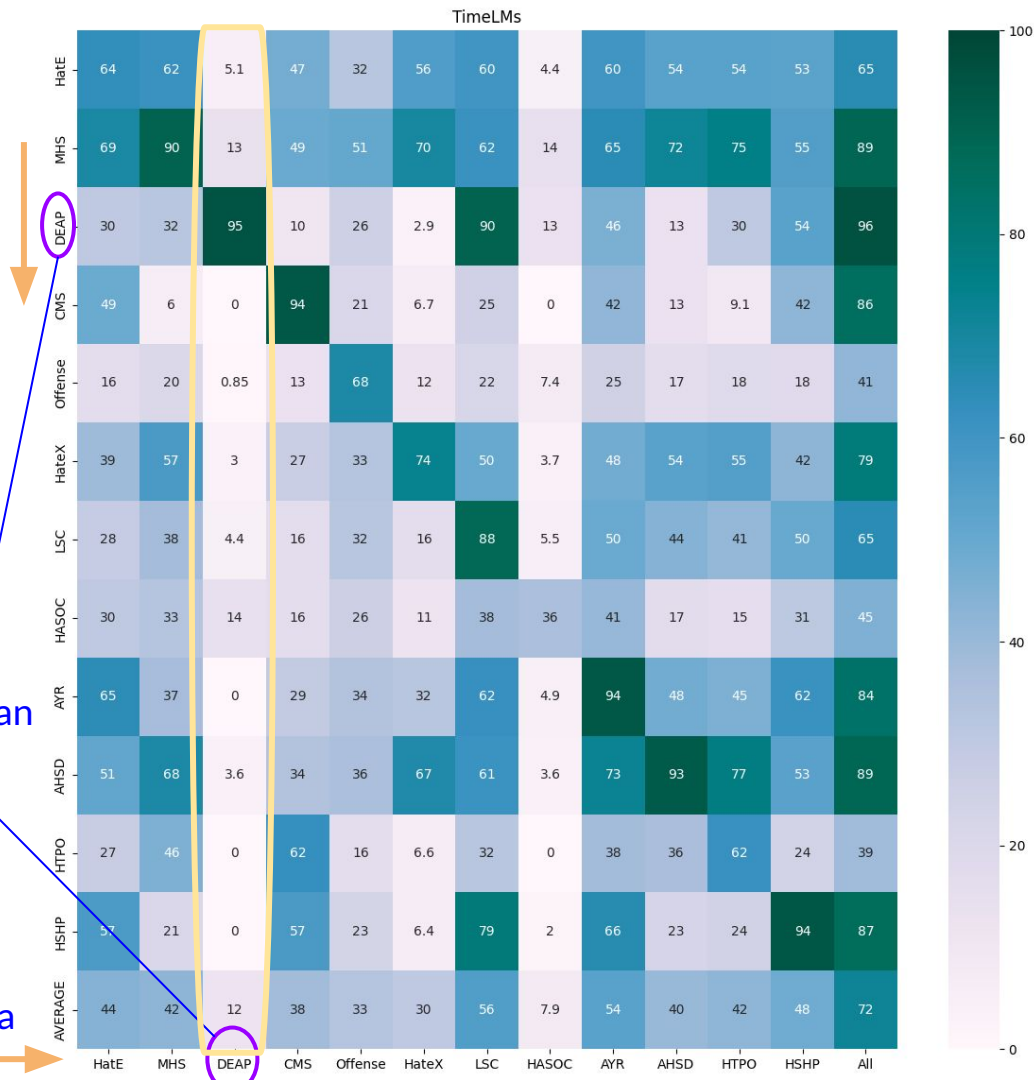
Results

Not good on datasets of different nature

Hate speech towards East Asian

Training data

Evaluation dataset

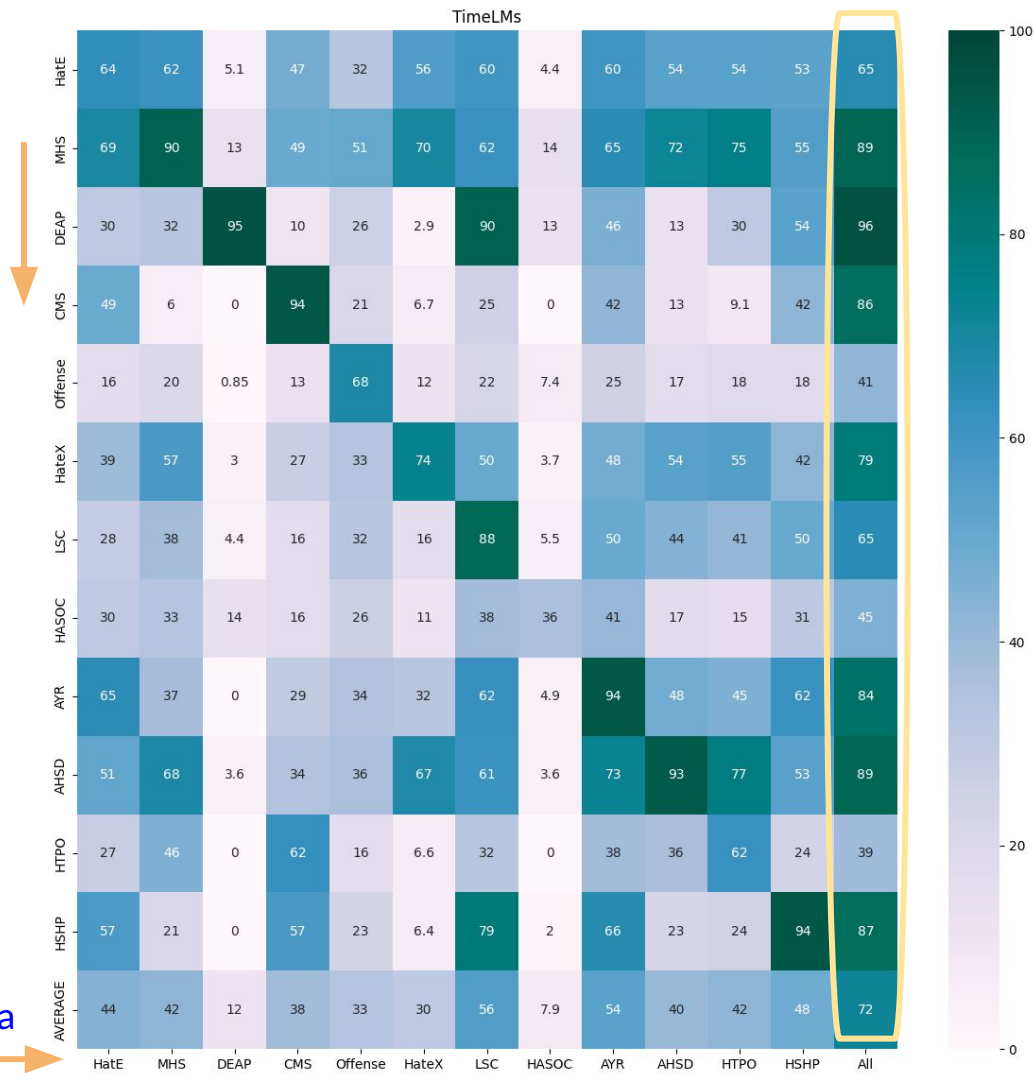


Results

Evaluation dataset

Model trained on all datasets is more robust.






Training data



Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

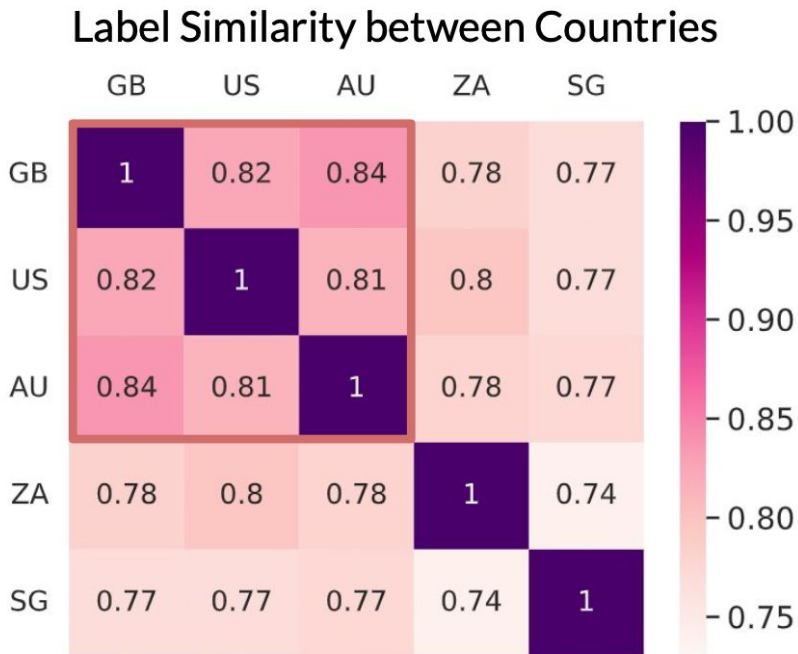
Hate speech dataset
annotated by people
from 5 different
countries

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

Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

UK, US and Australia
annotations are
similar.

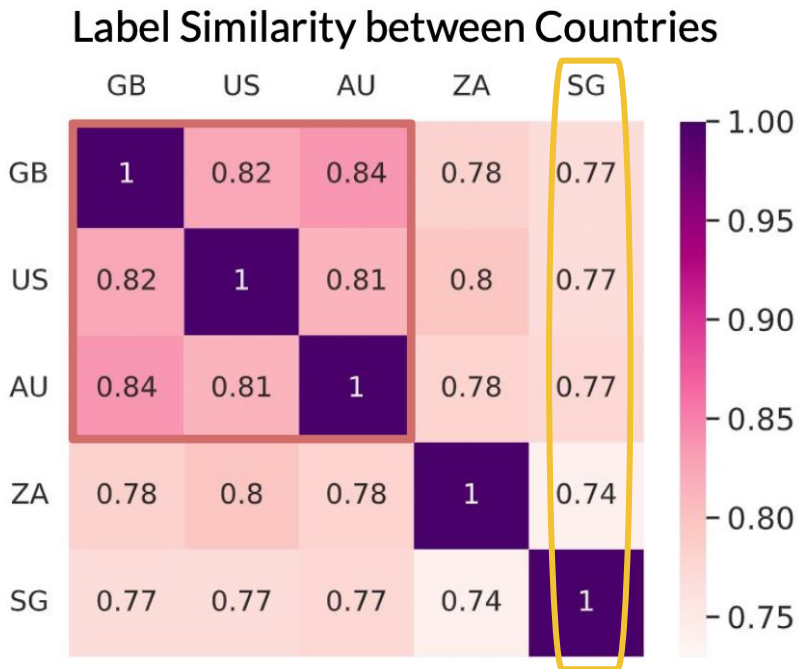


Cross-cultural differences in English hate speech?

(Lee et al., NAACL 2024)

UK, US and Australia
annotations are
similar.

Singapore and South
Africa differ.



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

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.

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!





Cardiff NLP

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