NLP and social media:

Language Modelling, Benchmarking and Temporal Challenges

Jose Camacho-Collados

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Manchester, 6 February 2024

Cardiff NLP

Outline

- **Specialized language models** in social media
- **Evaluation**
- **Temporal** challenges and **applications**

Outline

- **Specialized language models** in social media
- **Evaluation**
- **Temporal** challenges and **applications**

★ **TweetNLP**: A social media NLP platform

About me

- Professor at **Cardiff University** (Wales, UK)
  - **UKRI Future Leaders Fellow** (£1.4M funding, 4+ years)
  - 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

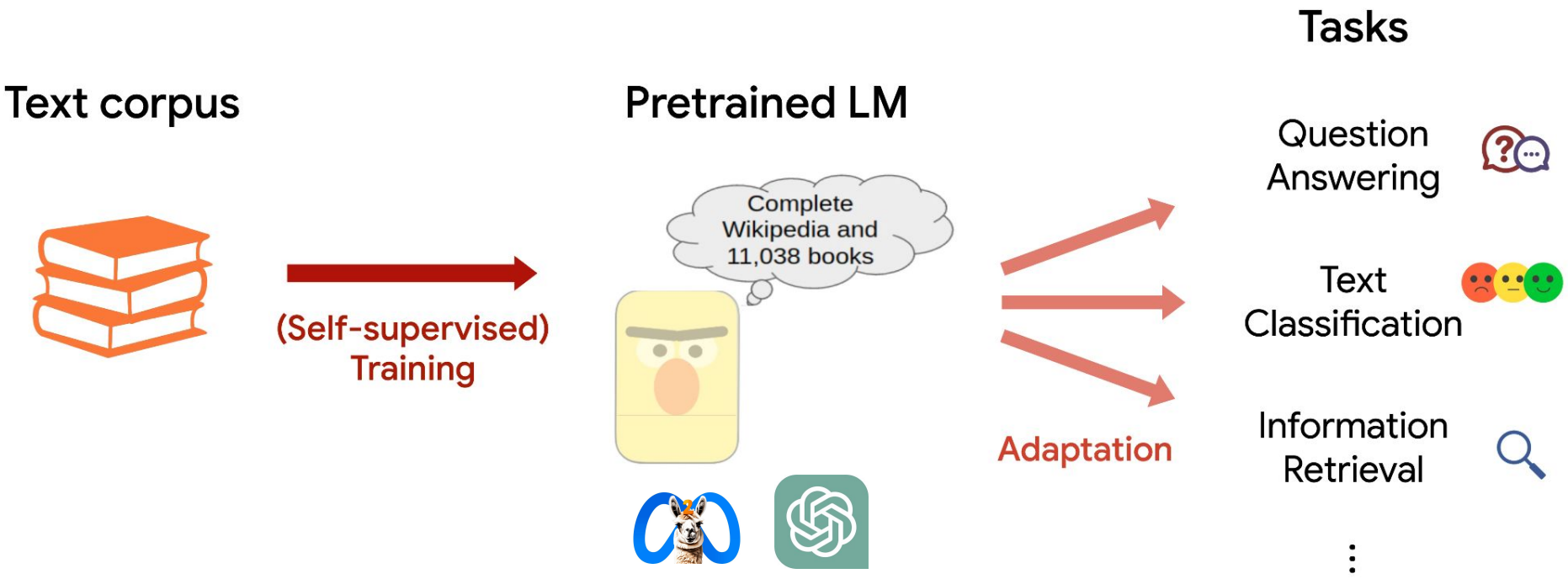
Cardiff NLP

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

Cardiff NLP Workshop (1-2 July 2024)

Specialized Language Models

Language models (LMs)



Slide credit: Stanford AI

Specialized language models

- LMs can be specialized by:
  - augmenting with external information
  - pretraining on domain-specific corpora

Specialized language models

- LMs can be specialized by:
  - augmenting with external information
  - **pretraining on domain-specific corpora**

Specialized language models

- Pretraining on domain-specific corpora

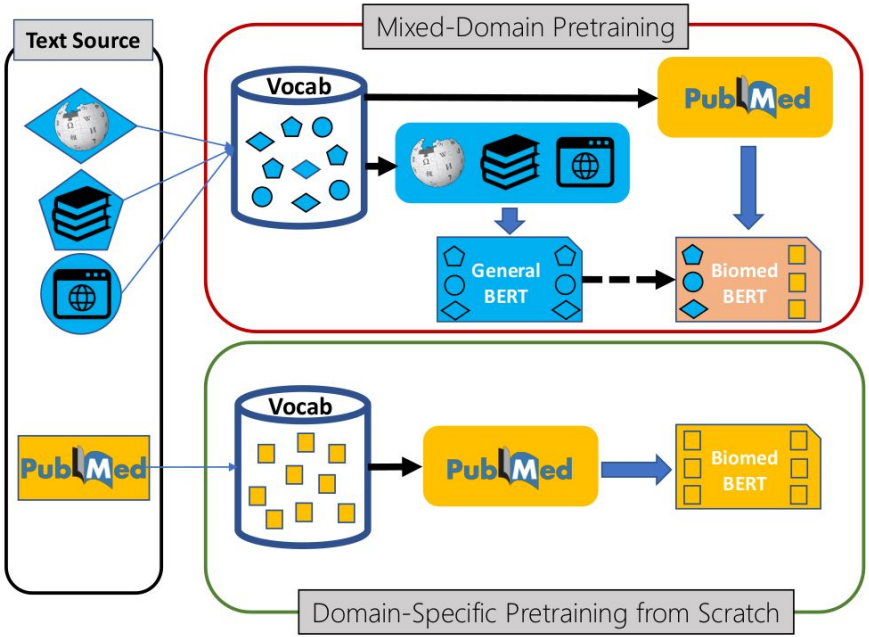


image credit: Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2021). **Domain-specific language model pretraining for biomedical natural language processing**. ACM Transactions on Computing for Healthcare (HEALTH), 3(1), 1-23.

Specializing a LM on social media

Specializing a LM on social media

- Why?

Specializing a LM on social media

- 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…
  - And because social media is **important**

Specializing a LM on social media

- Why?

Haaland! That was <mask>

Specializing a LM on social media

➢ Why?

TweetEval:

Language Models and Evaluation Benchmark

TweetEval, the language model

(Barbieri et al. EMNLP Findings 2020)

- How?
  - RoBERTa architecture
  - Continue from RoBERTa checkpoint vs from scratch (BERTweet is from scratch)

TweetEval, the benchmark

(Barbieri et al. EMNLP Findings 2020)

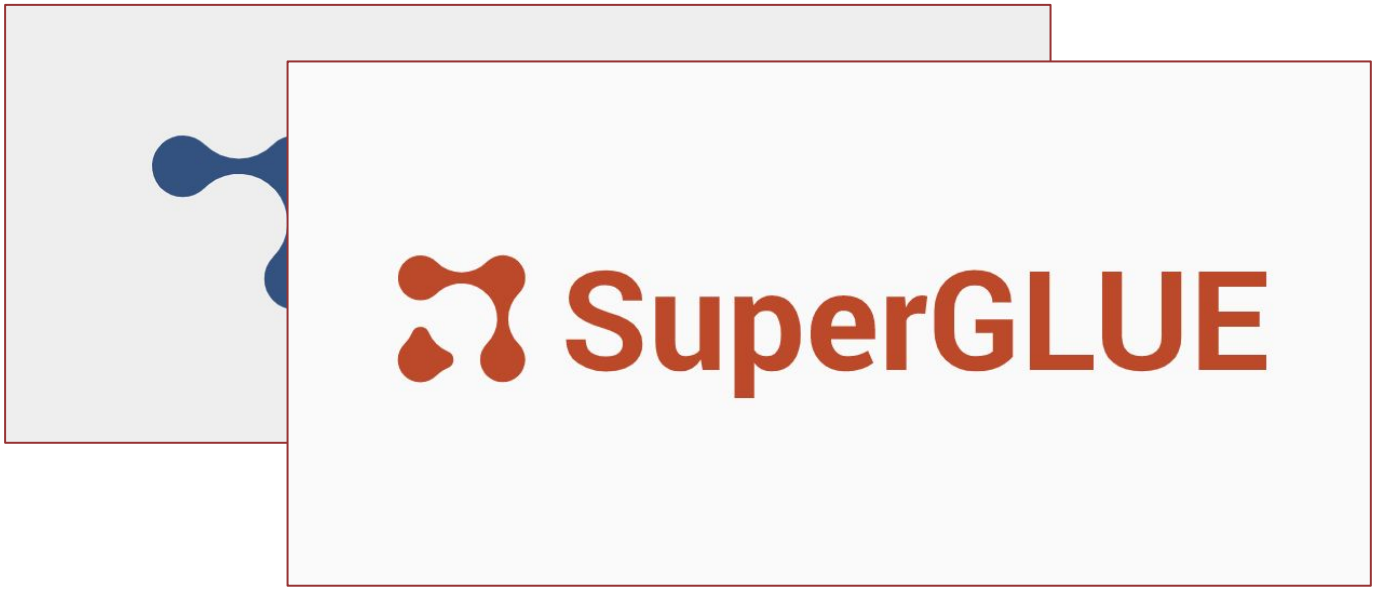
- Why?

TweetEval, the benchmark

- Why?

TweetEval, the benchmark

- Why?



TweetEval, the benchmark

- Why?

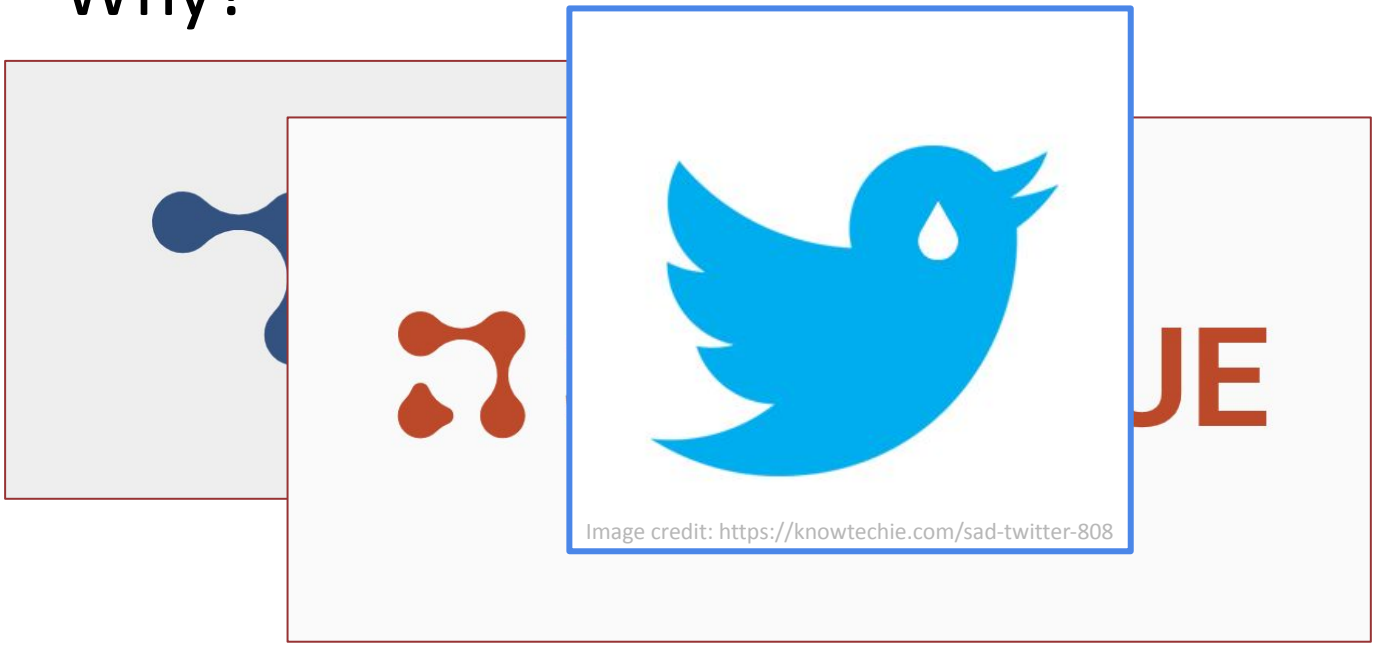


Image credit: https://knowtechie.com/sad-twitter-808

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

SuperTweetEval, the benchmark

(Antypas et al. EMNLP Findings 2023)

🤗 Already available at

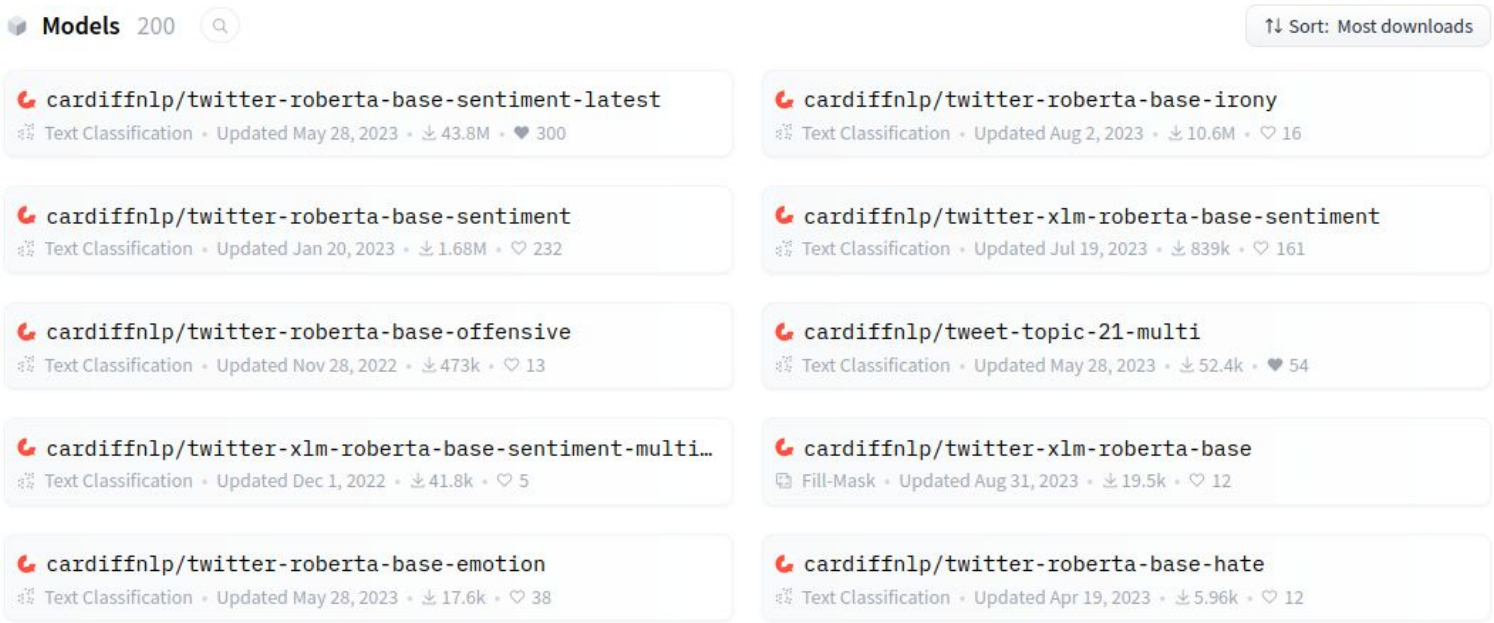
huggingface.co/datasets/cardiffnlp/super_tweeteval

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

Also tasks with **temporal splits**!

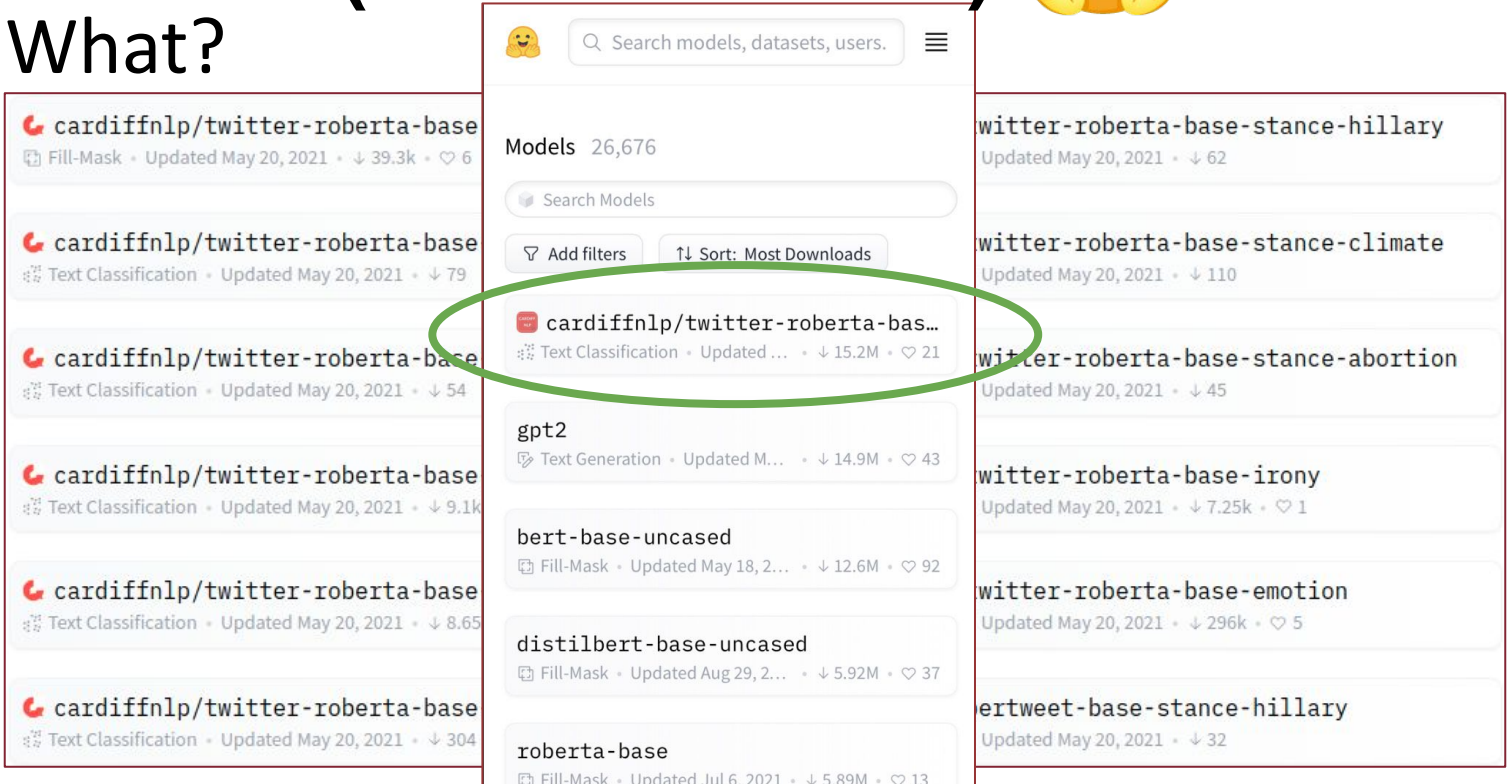
Results? Smaller specialized models with supervision still better than LLMs (including ChatGPT)

Specialized language models (+fine-tuned) 🤗



Specialized language models (+fine-tuned) 🤗

- What?



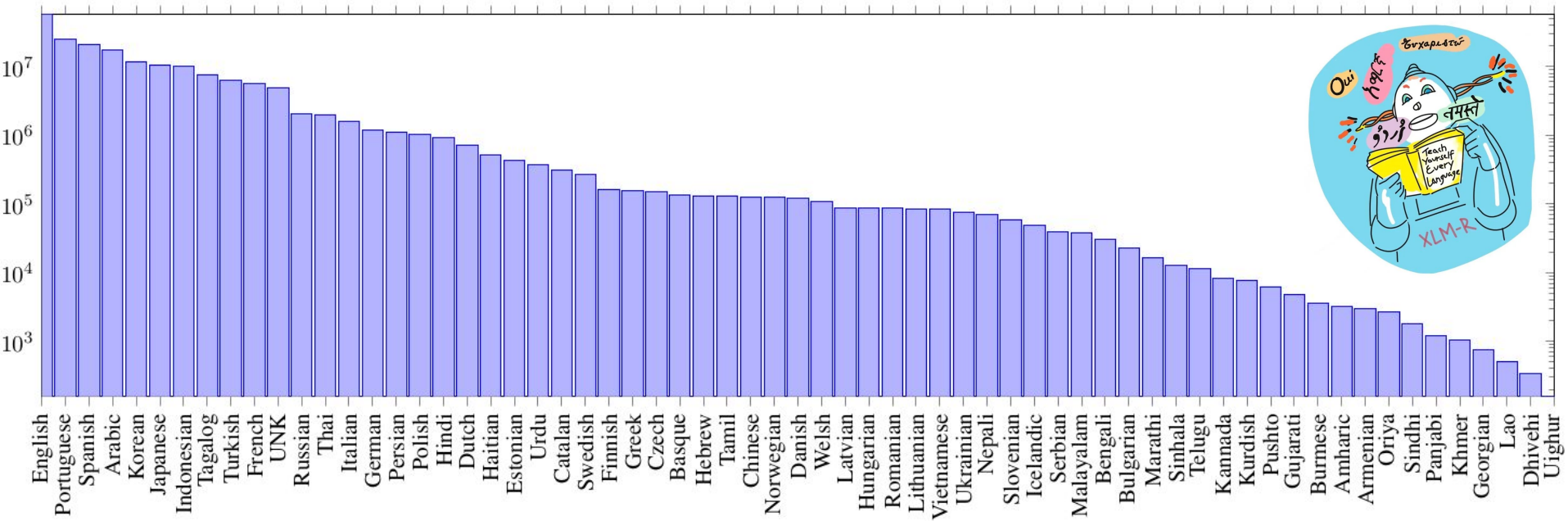
XLM-T: Multilingual Language Model for Twitter

XLM-T: Multilingual Twitter LM

- Why?
  - Same as in TweetEval: no multilingual LMs, and no unified multilingual benchmarks

XLM-T: Multilingual Twitter LM

- What?



XLM-T: Multilingual Twitter LM

- How?
  - XLM-R architecture
  - Continue from XLM-R checkpoint

Updated large-size 2023 model now available!

TimeLMs: Diachronic Language Models

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*).

Popular models are **old** (e.g. BERT, 2018).

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

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*).

Popular models are **old** (e.g. BERT, 2018).

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

Solution?

TimeLMs initiative

(Loureiro et al. ACL Demo 2022)

Commitment to train **new language models** every quarter (three months)

Twitter as our training corpus 

RoBERTa models **from 2019** until now already available

Language models improve over time

Models	2020-Q1	2020-Q2	2020-Q3	2020-Q4	2021-Q1	2021-Q2	2021-Q3	2021-Q4	Change
Barbieri et al., 2020	9.420	9.602	9.631	9.651	9.832	9.924	10.073	10.247	N/A
2019-90M	4.823	4.936	4.936	4.928	5.093	5.179	5.273	5.362	N/A
2020-Q1	4.521	4.625	4.699	4.692	4.862	4.952	5.043	5.140	-
2020-Q2	4.441	4.439	4.548	4.554	4.716	4.801	4.902	5.005	-4.01%
2020-Q3	4.534	4.525	4.450	4.487	4.652	4.738	4.831	4.945	-2.15%
2020-Q4	4.533	4.524	4.429	4.361	4.571	4.672	4.763	4.859	-2.81%
2021-Q1	4.509	4.499	4.399	4.334	4.439	4.574	4.668	4.767	-2.89%
2021-Q2	4.499	4.481	4.376	4.319	4.411	4.445	4.570	4.675	-2.83%
2021-Q3	4.471	4.455	4.335	4.280	4.366	4.394	4.422	4.565	-3.26%
2021-Q4	4.467	4.455	4.330	4.263	4.351	4.381	**4.402**	**4.463**	-2.24%
2021-124M	**4.319**	**4.297**	**4.279**	**4.219**	**4.322**	**4.361**	4.404	4.489	N/A

Perplexity-based evaluation

Language models improve over time

Models	2020-Q1	2020-Q2	2020-Q3	2020-Q4	2021-Q1	2021-Q2	2021-Q3	2021-Q4	Change
Barbieri et al., 2020	9.420	9.602	9.631	9.651	9.832	9.924	10.073	10.247	N/A
2019-90M	4.823	4.936	4.936	4.928	5.093	5.179	5.273	5.362	N/A
2020-Q1	4.521	4.625	4.699	4.692	4.862	4.952	5.043	5.140	
2020-Q2	4.441	4.439	4.548	4.554	4.716	4.801	4.902	5.005	-4.01%
2020-Q3	4.534	4.525	4.450	4.487	4.652	4.738	4.831	4.945	-2.15%
2020-Q4	4.533	4.524	4.429	4.361	4.571	4.672	4.763	4.859	-2.81%
2021-Q1	4.509	4.499	4.399	4.334	4.439	4.574	4.668	4.767	-2.89%
2021-Q2	4.499	4.481	4.376	4.319	4.411	4.445	4.570	4.675	-2.83%
2021-Q3	4.471	4.455	4.335	4.280	4.366	4.394	4.422	4.565	-3.26%
2021-Q4	4.467	4.455	4.330	4.263	4.351	4.381	**4.402**	**4.463**	-2.24%
2021-124M	**4.319**	**4.297**	**4.279**	**4.219**	**4.322**	**4.361**	4.404	4.489	N/A

TimeLMs

Sample predictions with models trained at different time periods.

Model	So glad I'm <mask> vaccinated.	I keep forgetting to bring a <mask>.	Looking forward to watching <mask> Game tonight!
2020-Q1	not getting self	bag purse charger	the The this
2020-Q2	not getting fully	mask bag purse	The the End
2020-Q3	not getting fully	mask bag purse	the The End
2020-Q4	not getting fully	bag purse charger	the The End
2021-Q1	getting not fully	purse charger bag	the The End
2021-Q2	fully getting not	bag charger lighter	the The this
2021-Q3	fully getting not	charger bag purse	the The This
2021-Q4	fully getting not	bag lighter charger	Squid the The

TimeLMs

Sample predictions with models trained at different time periods.

Model	So glad I'm <mask> vaccinated.	I keep forgetting to bring a <mask>.	Looking forward to watching <mask> Game tonight!
2020-Q1	not getting self	bag purse charger	the The this
2020-Q2	not getting fully	mask bag purse	The the End
2020-Q3	not getting fully	mask bag purse	the The End
2020-Q4	not getting fully	bag purse charger	the The End
2021-Q1	getting not fully	purse charger bag	the The End
2021-Q2	fully getting not	bag charger lighter	the The this
2021-Q3	fully getting not	charger bag purse	the The This
2021-Q4	fully getting not	bag lighter charger	Squid the The

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):

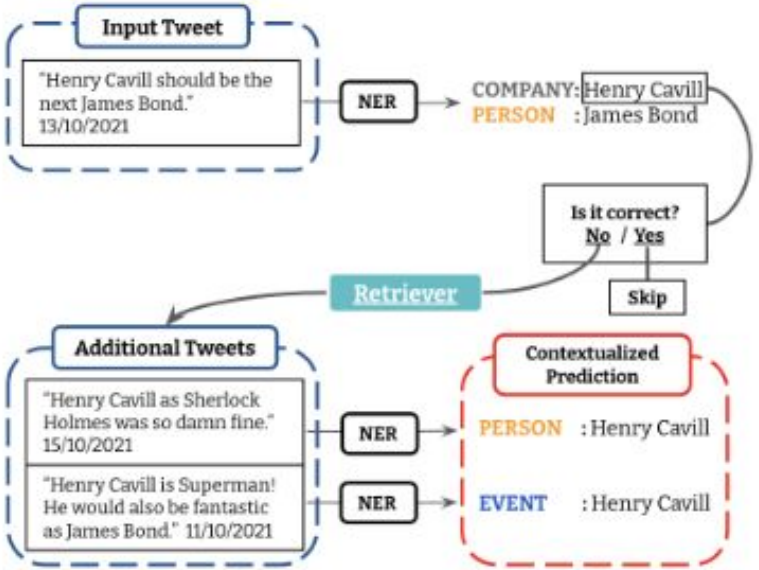
- **TweetNER7** (Ushio et al. 2022) for NER
- **TweetTopic** (Antypas and Ushio et al. 2022) for topic classification

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

LongEval series at CLEF to evaluate performance drop over time

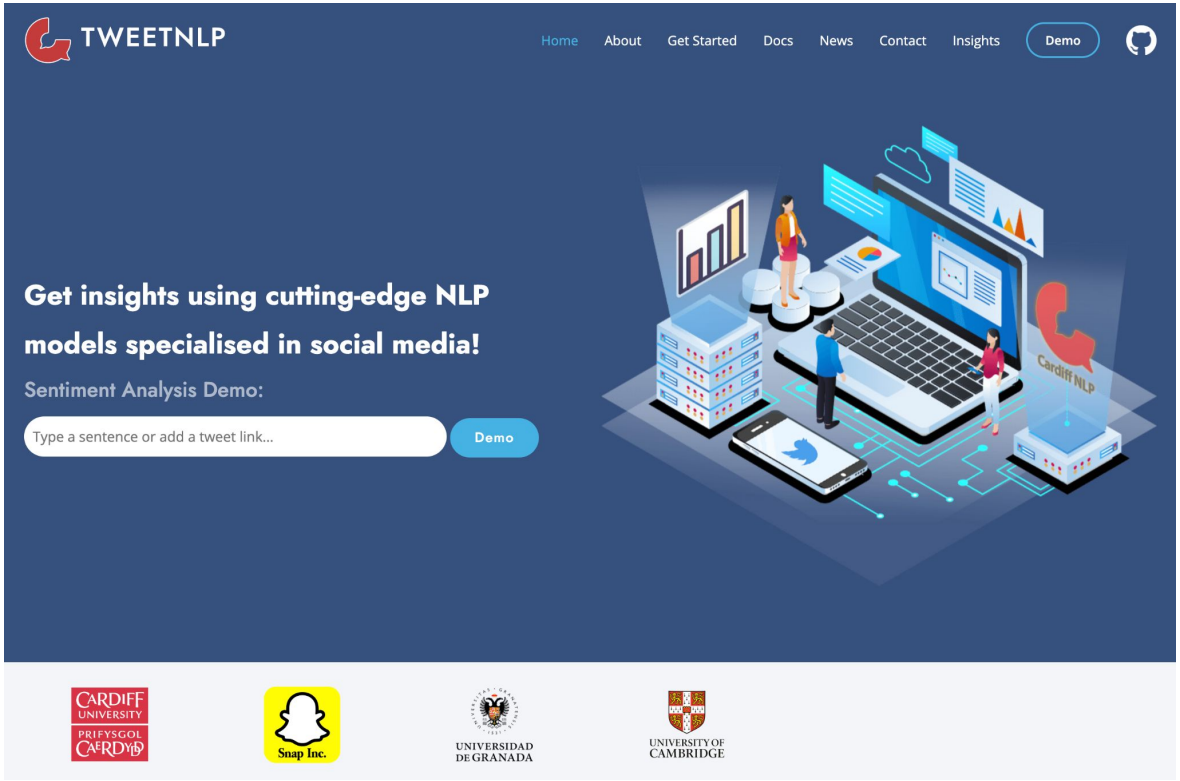
Temporal challenges: Ongoing work

- **Analysis** of main sources of performance drop (ongoing):
  - pre-training data?
  - training data?
  - nature of the domain/task?
- **Development of new methods** to integrate contextual information or non-labeled data in predictive models



Ushio et al. (AACL 2022)

TweetNLP (tweetnlp.org)



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:

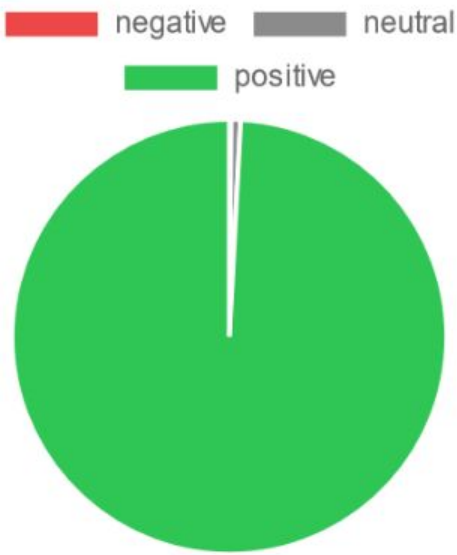
Today is a lovely day! 😊

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

Sentiment

English

GO!



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%)

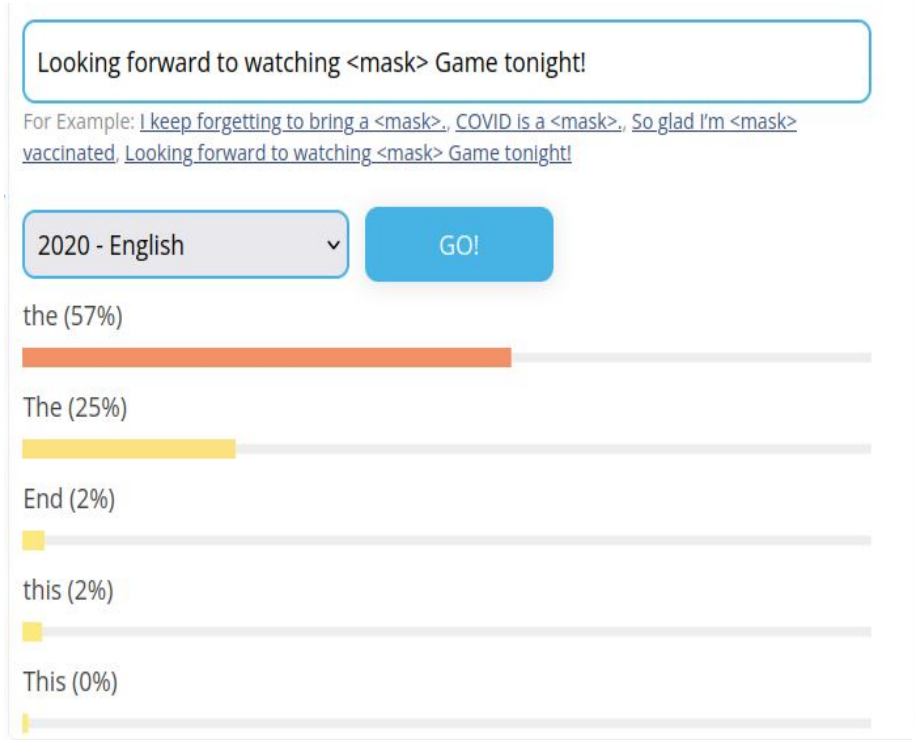
it (3%)

close (3%)

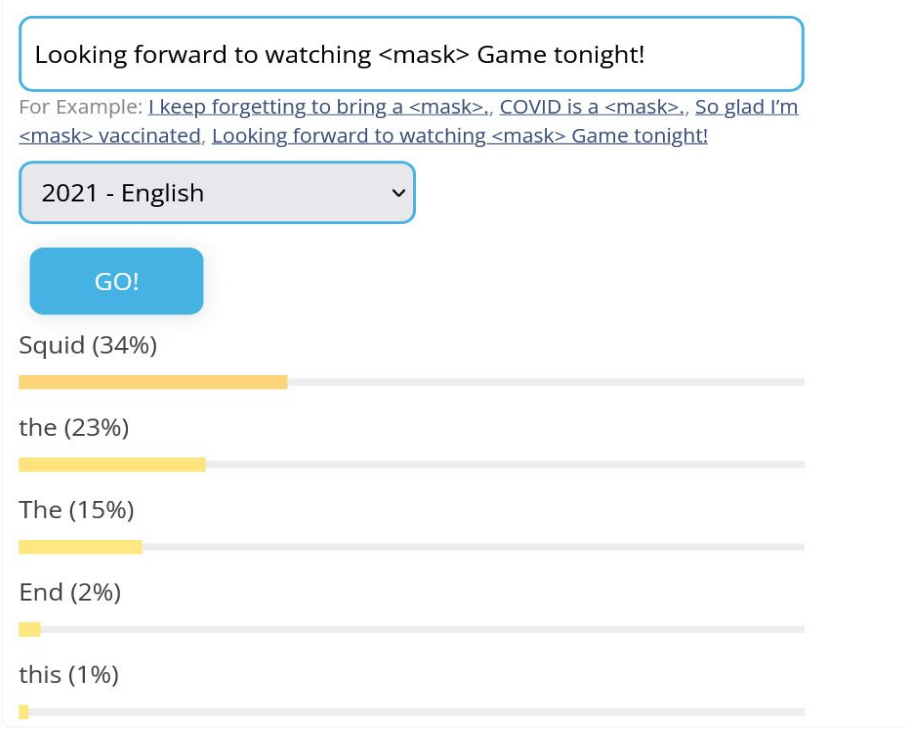
amazing (2%)

Word prediction (different years)

2020 model



2021 model



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

Question answering & generation

(Ushio et al. EMNLP 2022, ACL Findings 2023)

Jarry, who won yesterday in Cincinnati, withdrew from the tournament to follow the birth o

For Example: Beyonce further expanded her acting career, starring as blues singer Etta James in the 2008 musical biopic, Cadillac Records.

who withdrew from the tournament?

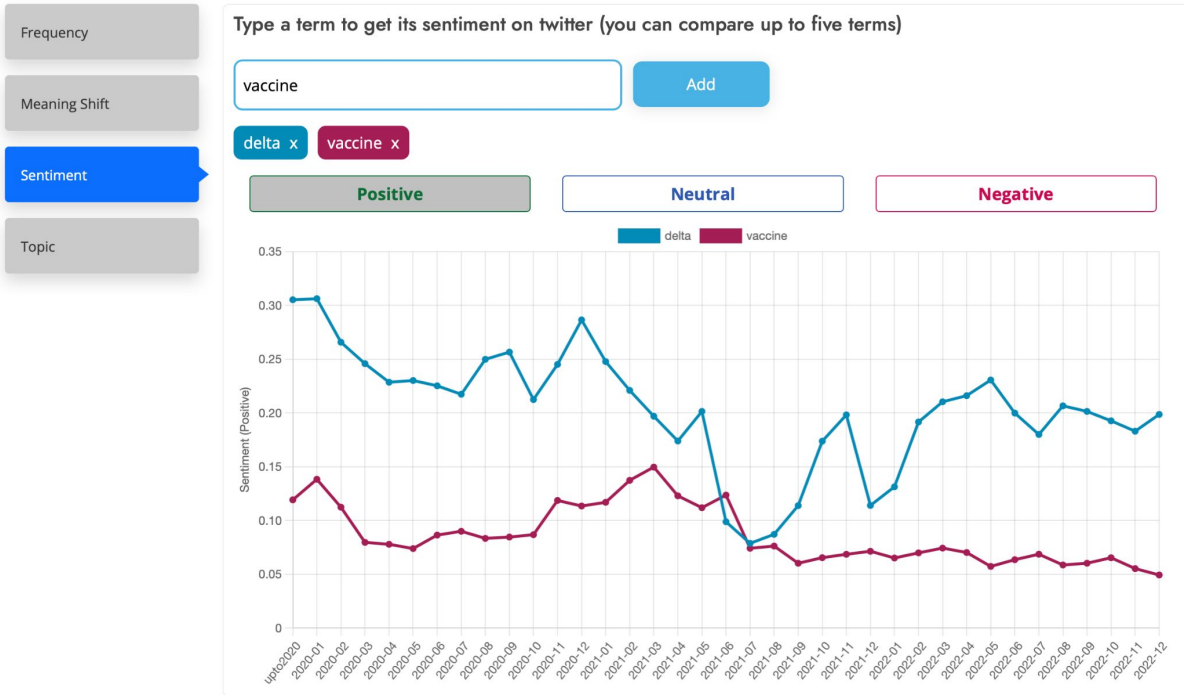
For Example: what did beyonce do in the musical biopic?

Output

jarry

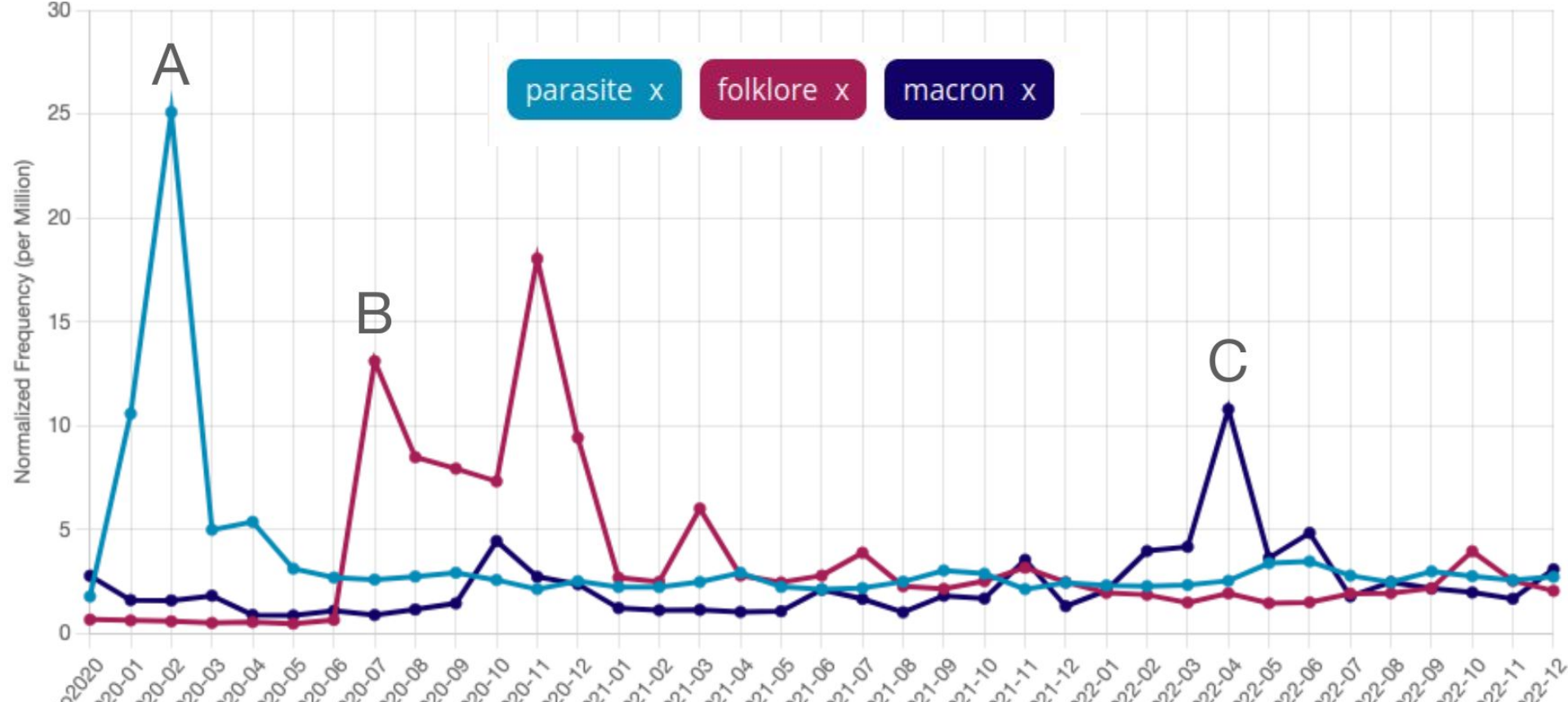
Tweet Insights

tweetnlp.org/insights



Tweet Insights

(Loureiro et al. 2023)



Other social media applications

Politics, sentiment and virality

Dimosthenis Antypas

Alun Preece

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD
Crime and Security Research Institute
Sefydliad Ymchwil Trosedd a Diogelwch

Negativity spreads faster: A large-scale multilingual twitter analysis on the role of sentiment in political communication (*Online Social Networks and Media* journal, 2023)

Politics, sentiment and virality

Collected a corpus of Twitter messages from MPs in **Greece, Spain and UK** (focus on 2021, 400K tweets)

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

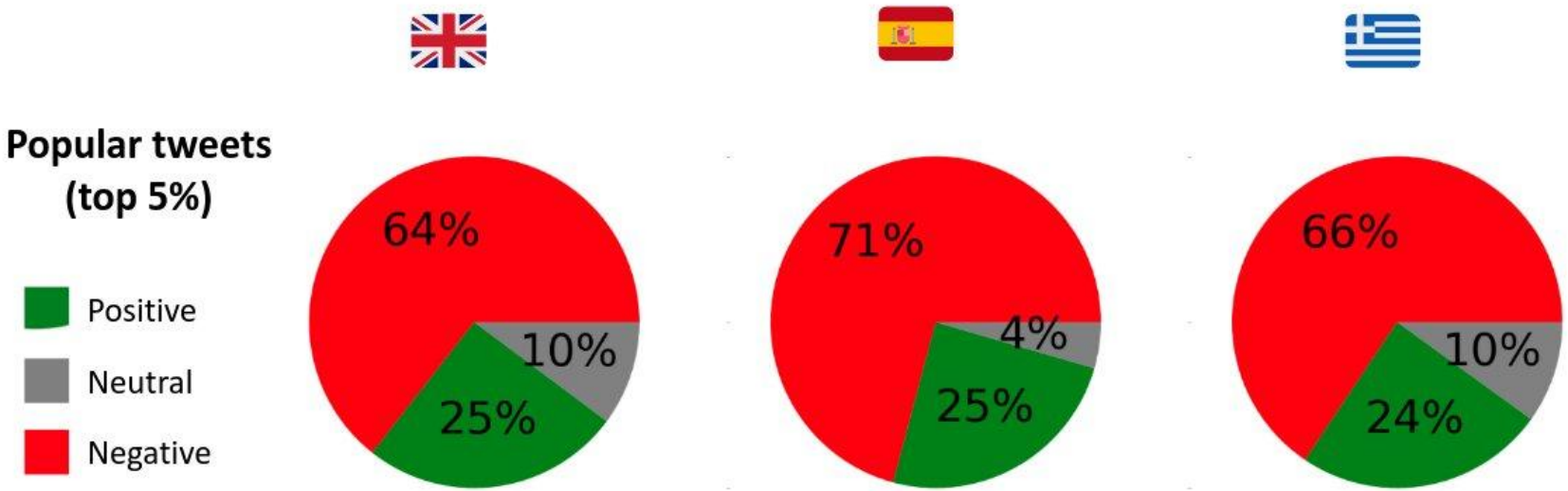
Politics, sentiment and virality

Collected a corpus of Twitter messages from MPs in **Greece, Spain and UK** (focus on 2021, 400K tweets)

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

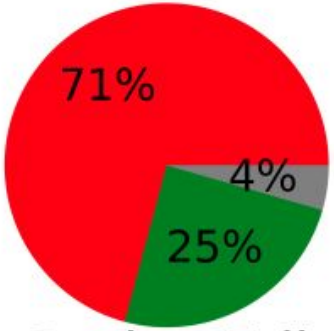
➡️ **Conclusion:** Tweets negatively charged -> More popular ⬆️

Sentiment of MPs' tweets

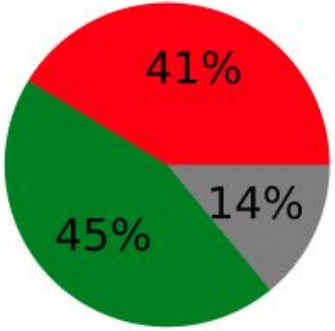


Viral tweets are overwhelmingly negative

Sentiment of Spanish MP tweets



***Popular* tweets**
(top 5%)



***Unpopular* tweets**
(bottom 35%)

Viral tweets (top 5%)

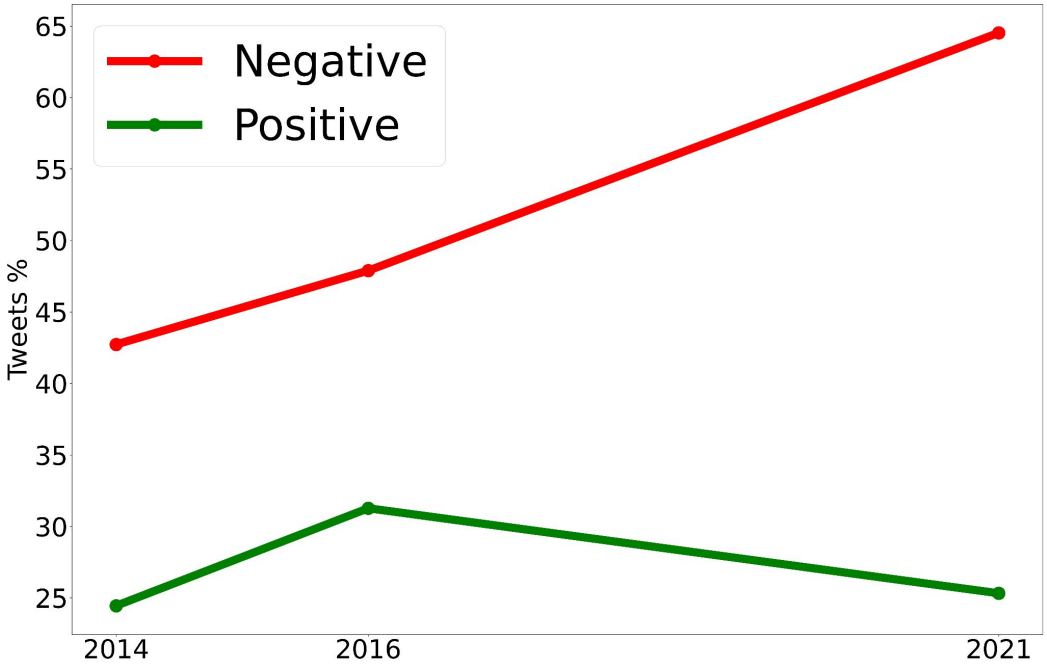
71% are negative!

Positive

Neutral

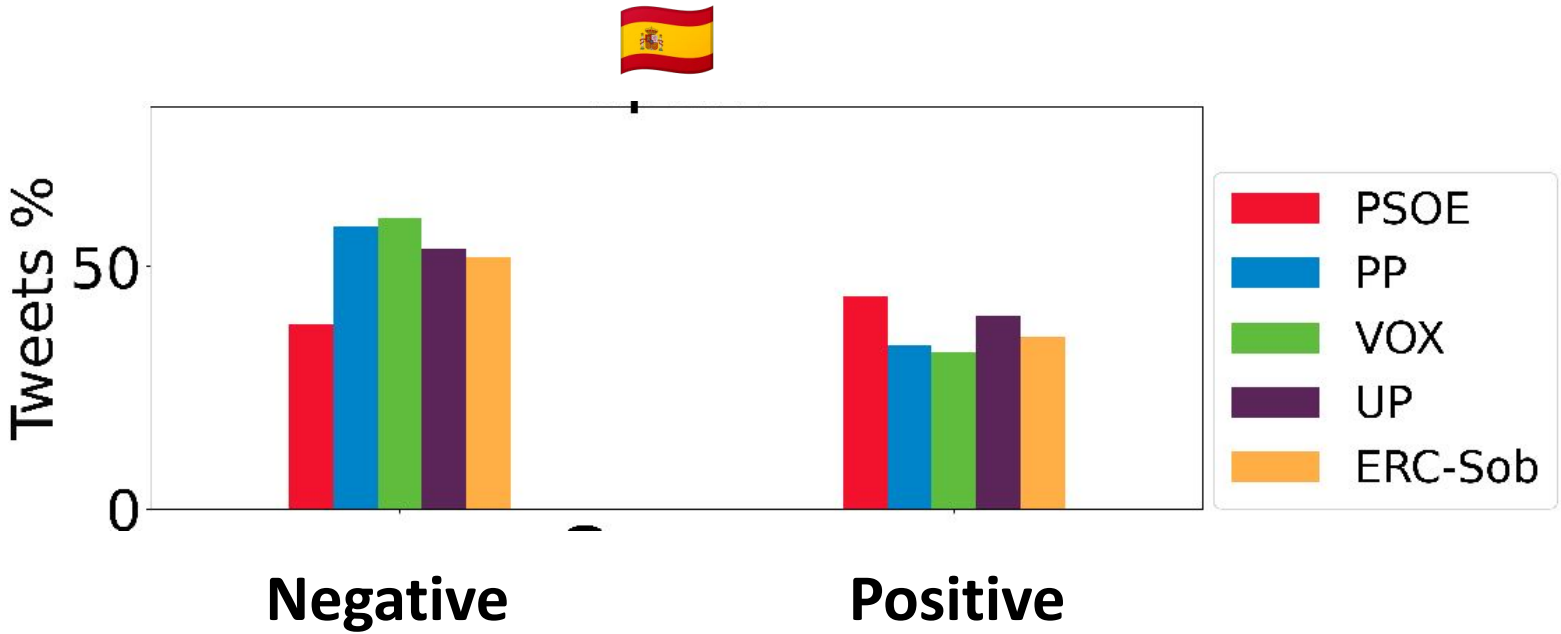
Negative

Sentiment over time



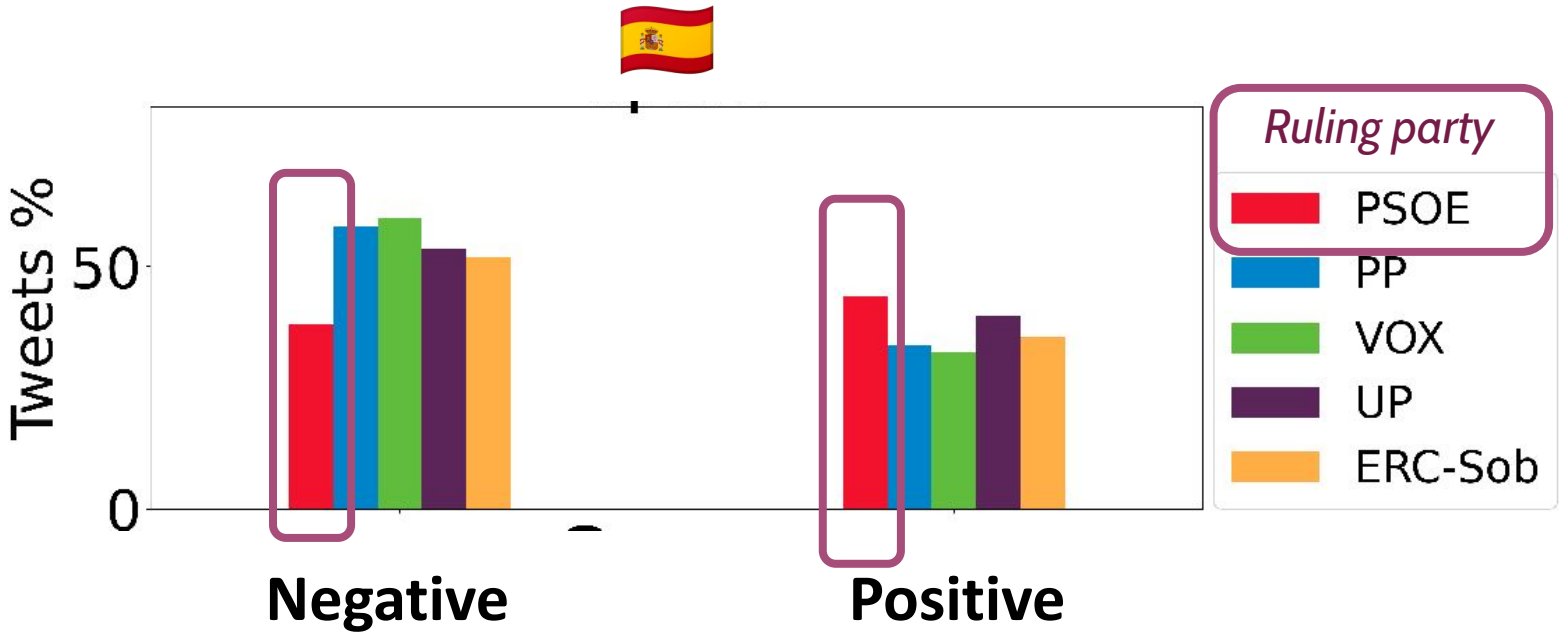
Tweets by MPs are becoming more negative over time (UK)

Government vs. opposition (Spain)



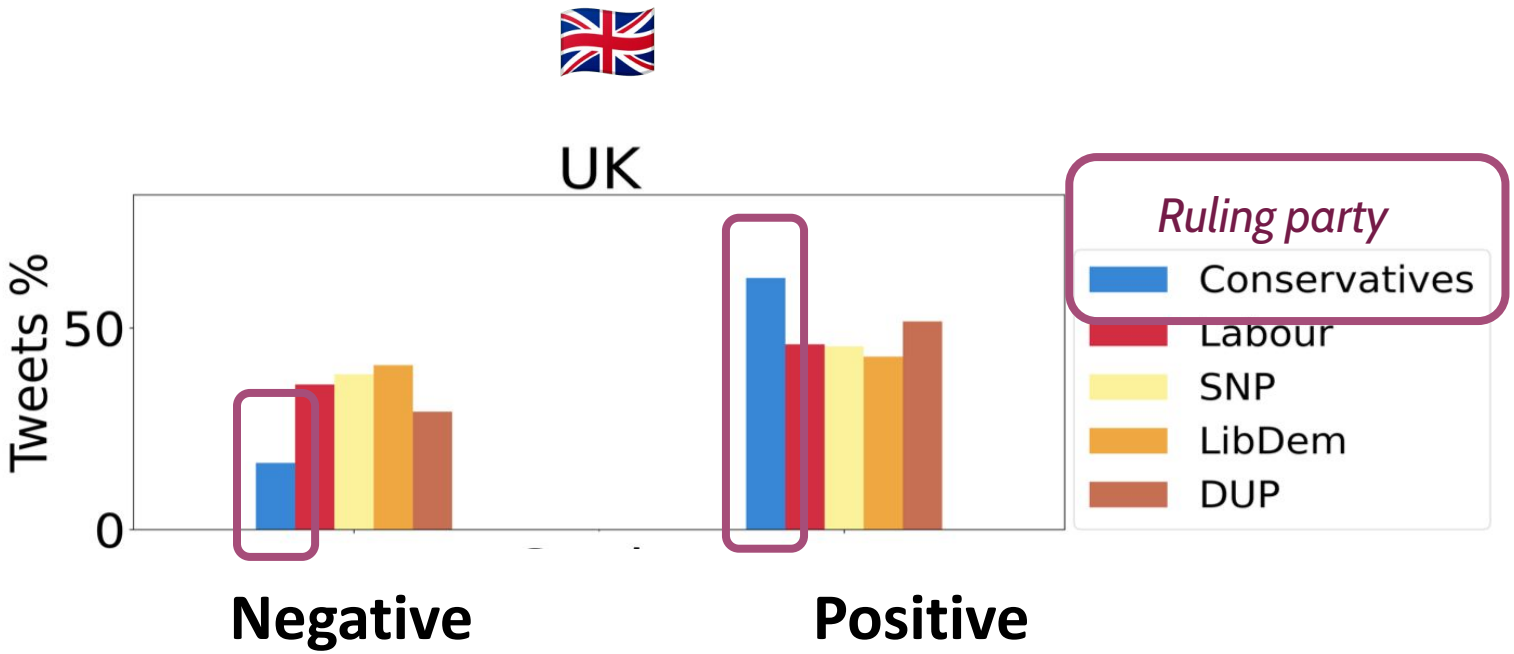
MPs from the government party are **more positive and less negative** -> *this also holds in other countries with different ideologies*

Government vs. opposition (Spain)



MPs from the government party are **more positive and less negative** ->
this also holds in other countries with different ideologies

Government vs. opposition (UK)



MPs from the government party are **more positive and less negative**

Government vs. opposition (Spain, 2021)

Pedro Sánchez
Prime Minister

13% vs. 63%

Negative tweets

Pablo Casado
Leader of the opposition

Government vs. opposition (UK, 2021)

Boris Johnson
Prime Minister

5% vs. 35%

Negative tweets

Keir Starmer
Leader of the opposition

Work in progress
(other multidisciplinary collaborations)

- Analysing **disinformation** at scale (with a special focus on textual content)
- Early health care interventions: **depression detection** on social media (Twitter, Reddit)
- Analysing outbreaks and adherence/sentiment to **health interventions** (e.g. COVID) using social media

Conclusion

Social media entails many **challenges.**

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

Temporal adaptation is needed, and can only be partially solved with updated models.

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

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