Temporal Awareness in NLP:

The Case of Social Media and Specialized Language Models

Jose Camacho-Collados

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

XTempLLMs, COLM 2025, 10 October 2025

Cardiff NLP

About me

- Professor at **Cardiff University** (Wales, UK)
  - **UKRI Future Leaders Fellow**
  - Founder of the **Cardiff NLP group**

- Interested in NLP and computational social science
- Areas of “expertise”: **semantics, resources, multilinguality, social media**

Cardiff NLP

Outline

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

- **Social media:**
  - **NLP tasks, specialised models**
- **Temporal** challenges:
  - **Mismatch** between LM training data and test

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Motivation

Language models may not know about recent topics of conversation or events.

This can cause a temporal **mismatch between training and test data**.

There are existing solutions, but not for when analyzing real-time data at scale is required (e.g., **social media**).

Social Media

Cardiff NLP

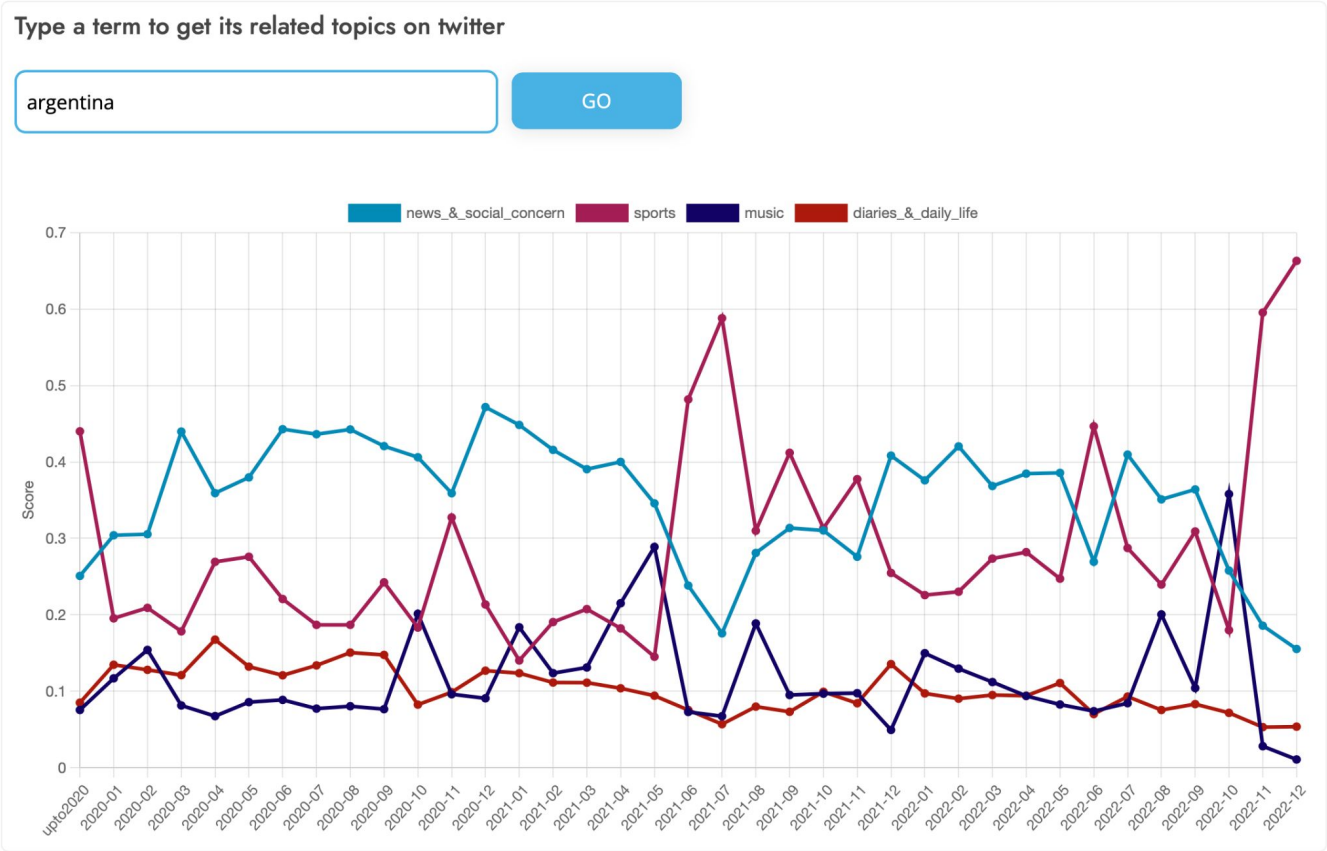
CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Why NLP research in social media matters

- Massive **real-world** data
- Understanding **public opinion**
- Detecting **harmful content**
- **Crisis** response and **public health**

Tweet Insights

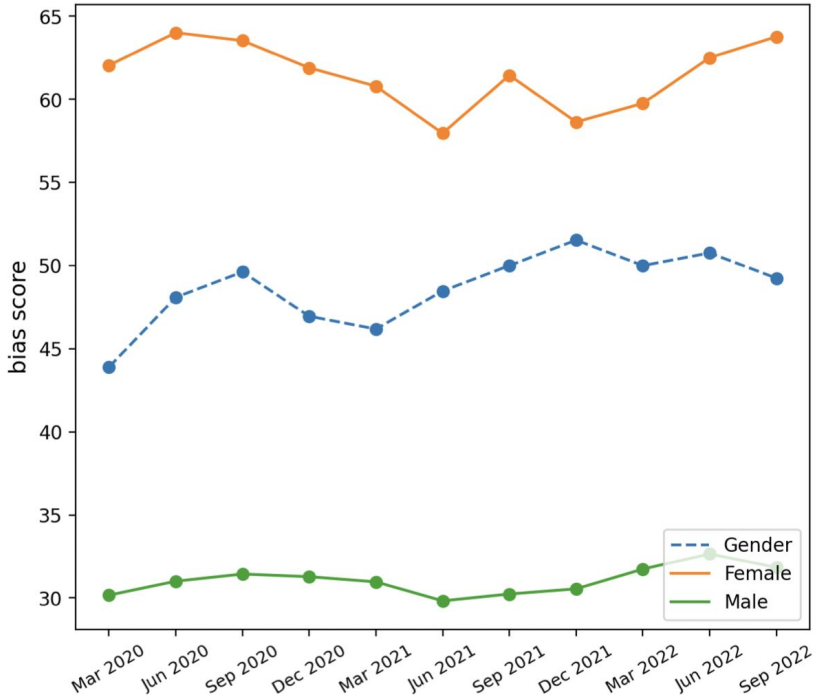
(Loureiro et al. 2023)



tweetnlp.org/insights

Measuring social biases over time

(Zhou et al. EMNLP 2024)



(a) Gender

Cardiff NLP

Social media: Why LLMs may not (always) be the best solution

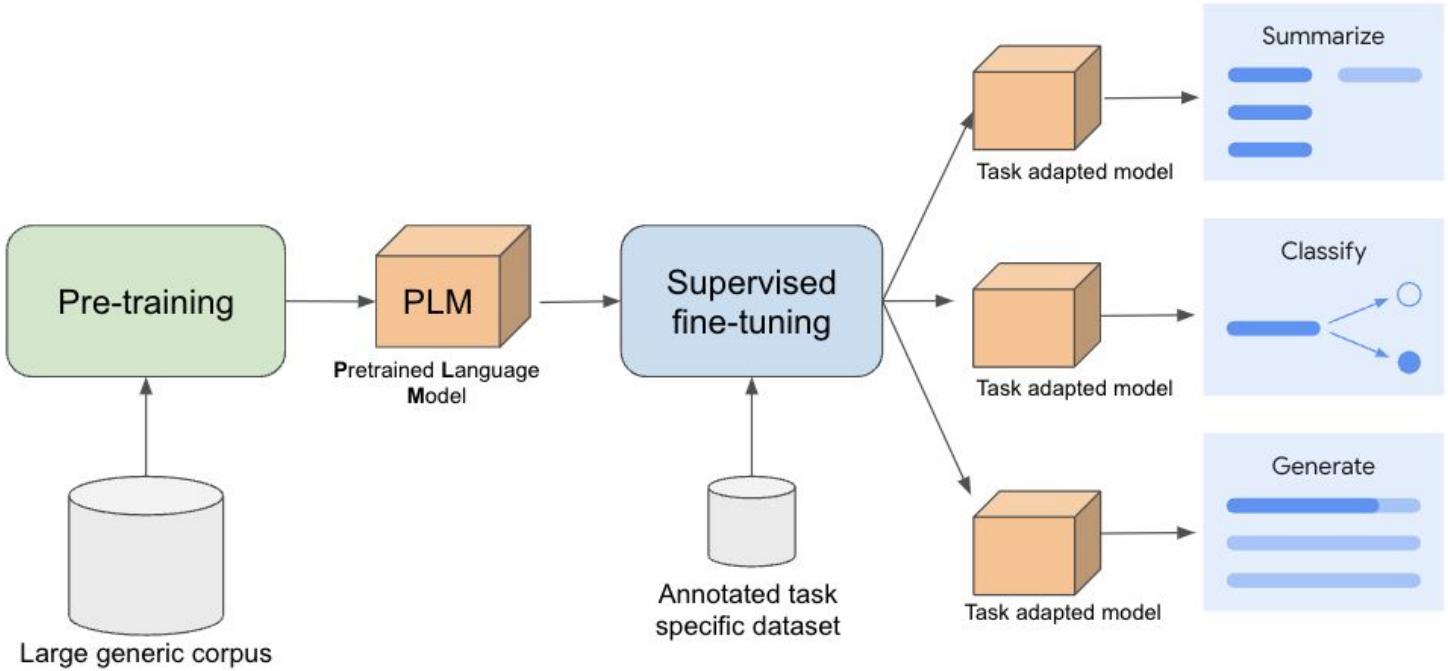
- Need for **efficient** solutions (real-time monitoring, large volume)
- Lack of **context**
- **Sensitive/private** data
- **Specialized and fine-tuned** models often work better for simple or classification problems (Edwards and Camacho-Collados 2024, Bucher and Martini, 2024)

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

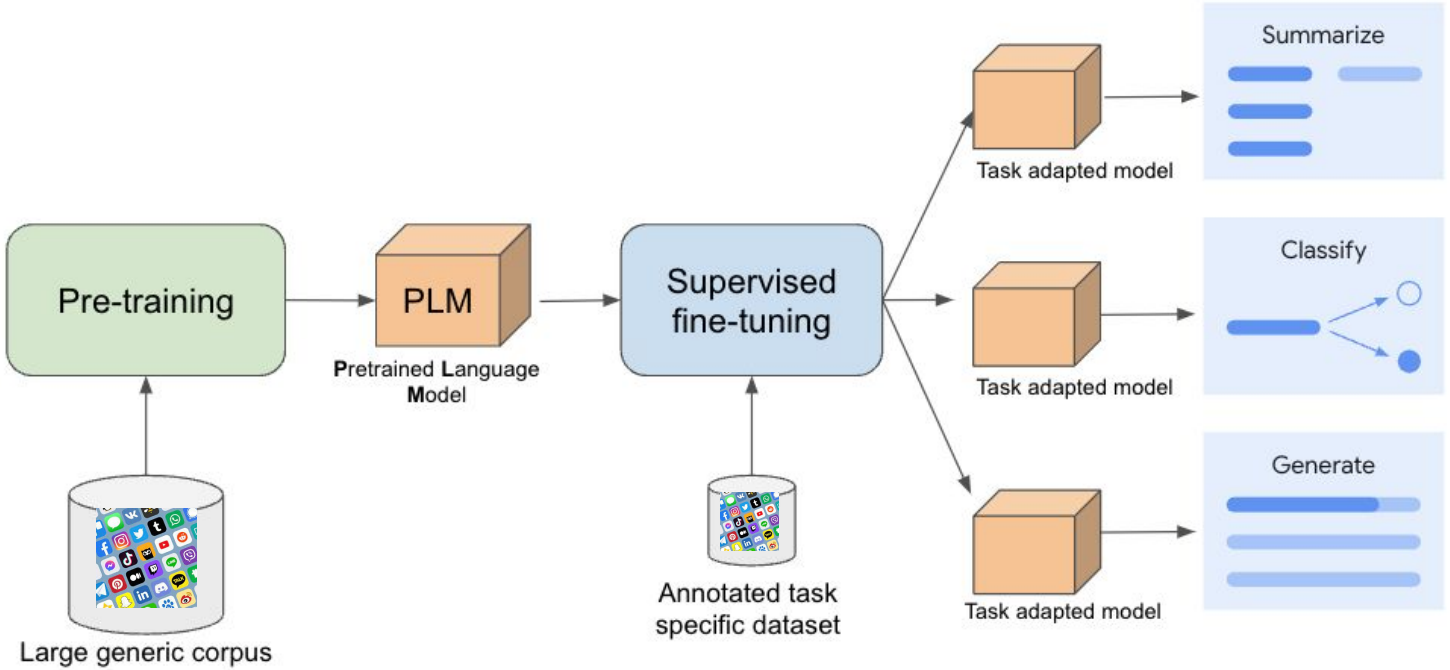
Specializing a LM on social media

LLM fine-tuning



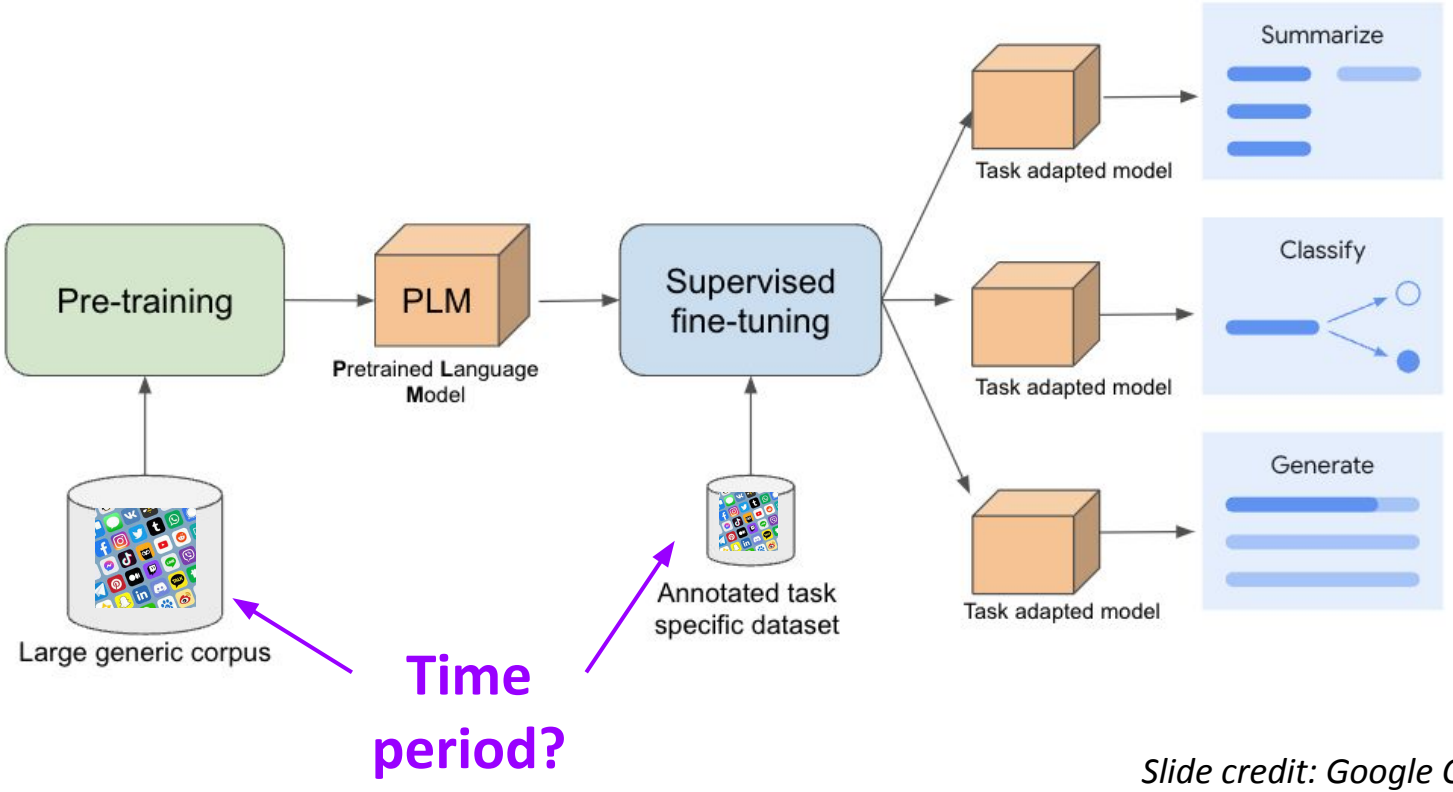
Slide credit: Google Cloud

LLM fine-tuning (social media)



Slide credit: Google Cloud

LLM fine-tuning (social media)



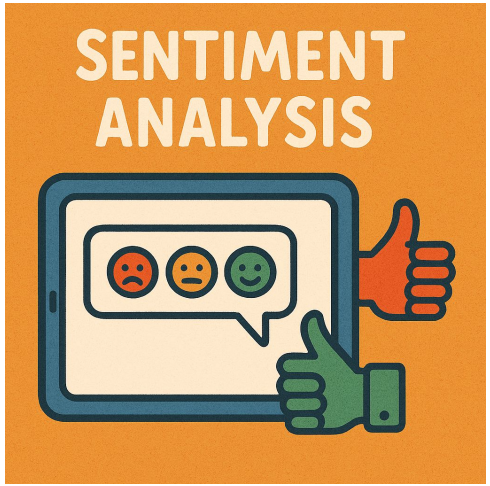
Slide credit: Google Cloud

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

NLP tasks on social media (some examples)

Sentiment analysis



Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Sentiment analysis

One of the most **popular** tasks on social media.

Deciding whether a post is **positive, negative** or **neutral** (+variations)

Indicator of **public opinion**.

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Hate speech detection

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Hate speech detection

An important task consisting on identifying **hateful content** on social media.

Usually framed as **classification** (hateful/not-hateful), including potential target groups.

Mostly needed in (almost) **real-time**.

Challenges in hate speech detection

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

- **Limited resources** (not diverse enough, generalisation issues)
- **Culturally specific**, not a global definition (inherently subjective)
- **Language and style variations** (**temporal?**)

Topic classification

Topic classification

Classify each tweet by topic domain.

Based on a taxonomy of 19 topics specially tailored to social media (Antypas et al. 2022):

- *Apple Removed More Than 30,000 Apps From The Chinese App Store*: **business** - **news & society** - **science & technology**
- *#copreps Football: End of the line for FLHS season:* **sports & games**

Topic Classification: Multilingual

(Antypas et al. EMNLP 2024)

`*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

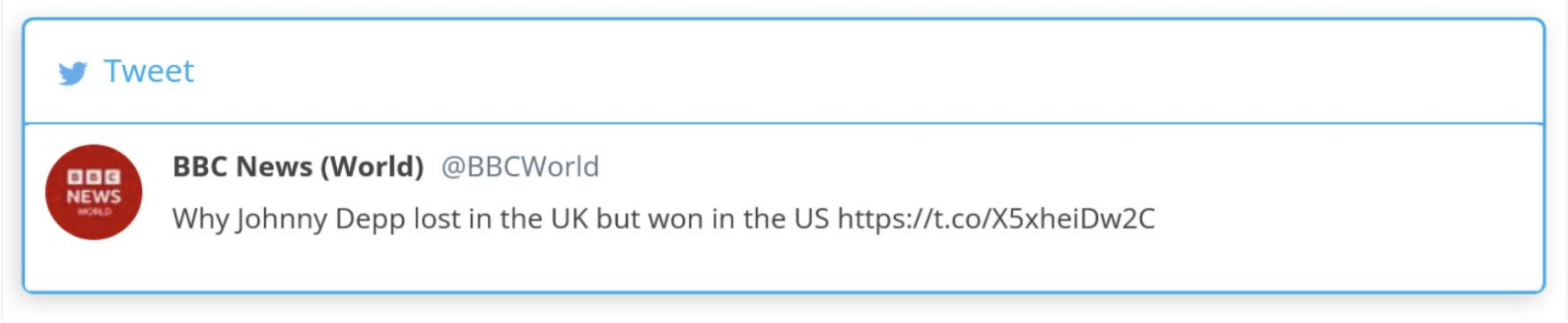
`*久々にな一一んもしないでいい日が二日もあるのでゆっくり富平井絆果と向き合うよ*`: Diaries & Daily Life, Gaming

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

Named Entity Recognition

Named Entity Recognition (NER)

Classical NLP task to identify entities in text.



Output NER:

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

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 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 et al. 2022) for topic classification

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

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Temporal challenges

Sources of performance drop can be due to:

- **Pre-training data**?
- **Training data**?
- Nature of the **domain/task**?
- *Other?*

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Temporal Generalization of Language Models

Why relevant/important?

In social media, many tasks need to be solved in real-time.

However, language models may have been trained in previous time periods.

Note: Not only the temporal aspect is relevant, but in many cases domain switches in general can affect results (e.g. hate speech detection: Yin and Zubiaga, 2021; Antypas and Camacho-Collados, 2023)

Are temporal shifts a problem?

Empirically it has been shown that models trained on closer time periods to the test dataset achieve better performance (e.g., evidence from NER and topic classification shown earlier).

However, we lack a comprehensive **understanding of why temporal shifts degrades performance**, and what can be done to solve it.

Research questions

(Ushio and Camacho-Collados, 2024)

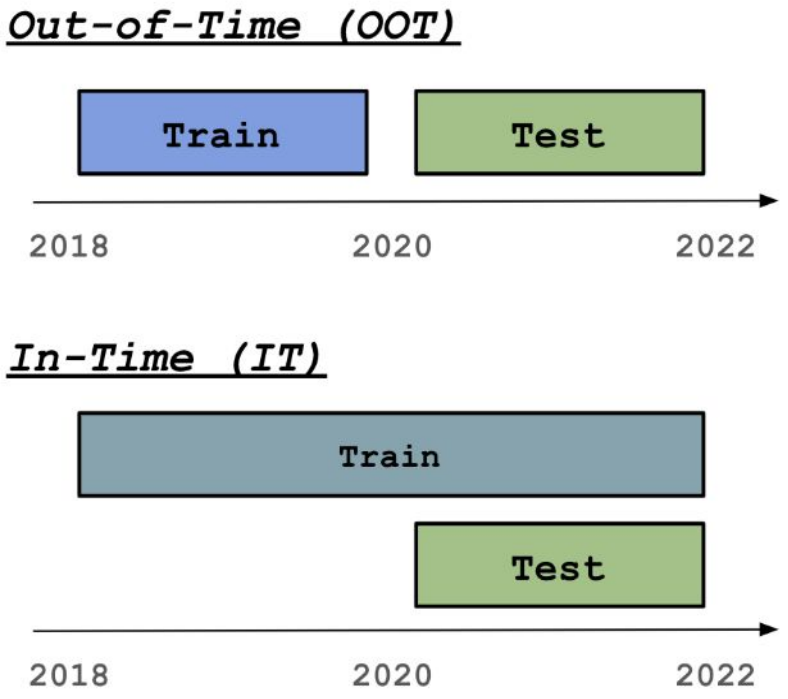
Are temporal shifts in social media detrimental for LM performance in NLP tasks?

If so, what are the causes of this temporal shift and can it be mitigated (by e.g. using LMs pre-trained on recent data)?

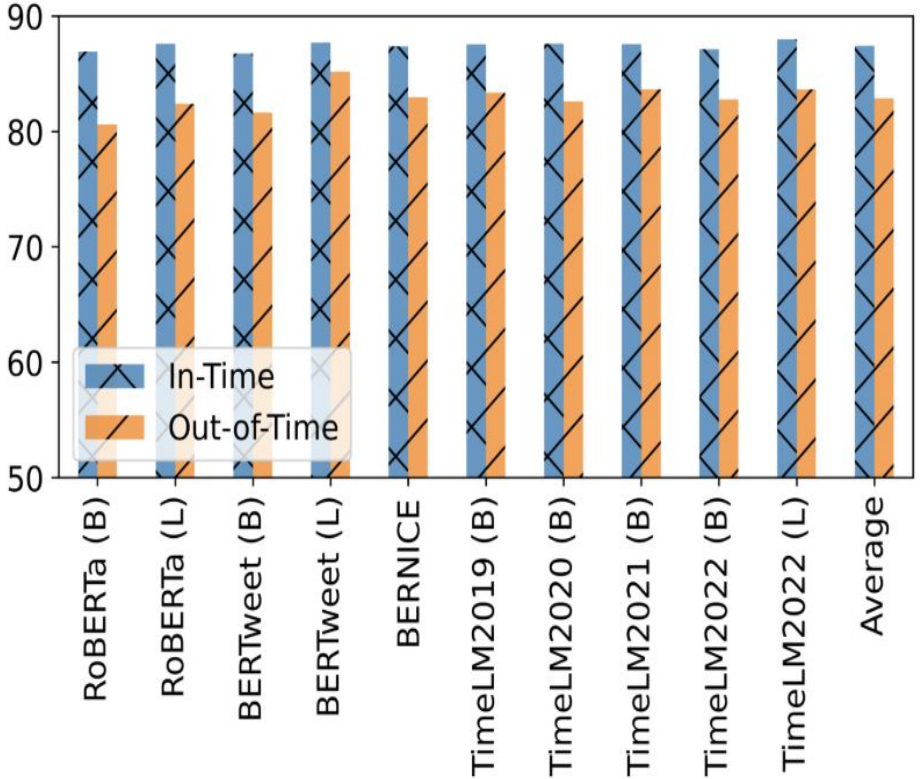
Evaluating temporal shifts

- ***Out-of-Time (OOT)***
  - The period of training split is prior to test split.
  - Model have no access to the instances from test period.

- ***In-Time (IT)***
  - The period of training split contains test split.



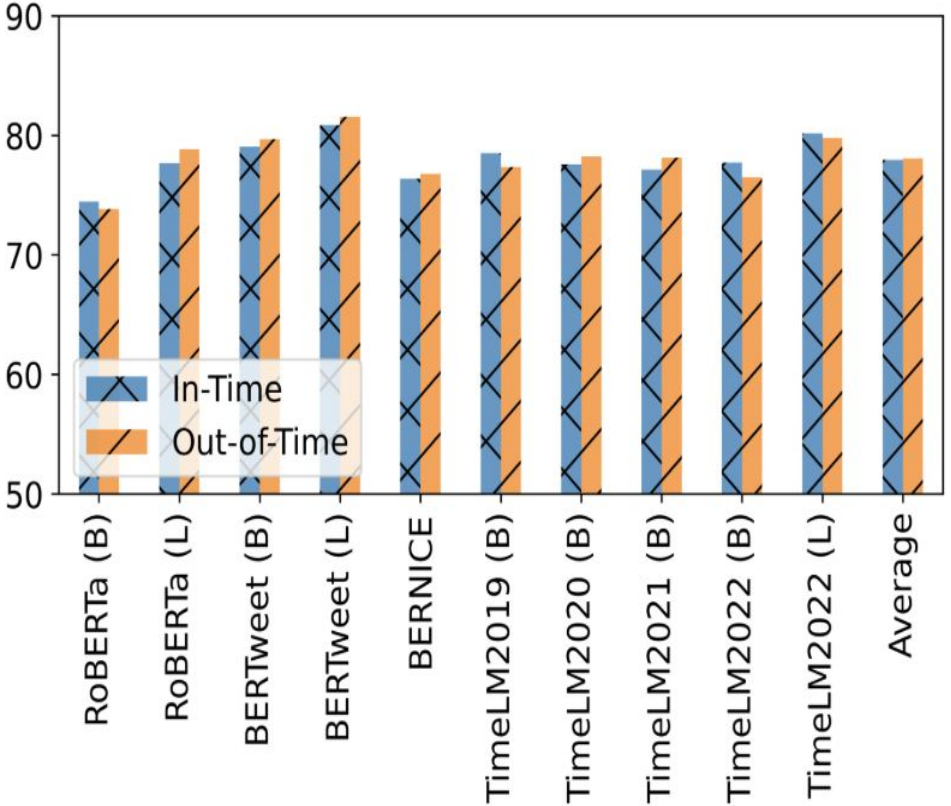
Hate speech detection



Better performance for in-time in all cases

Clear temporal effect

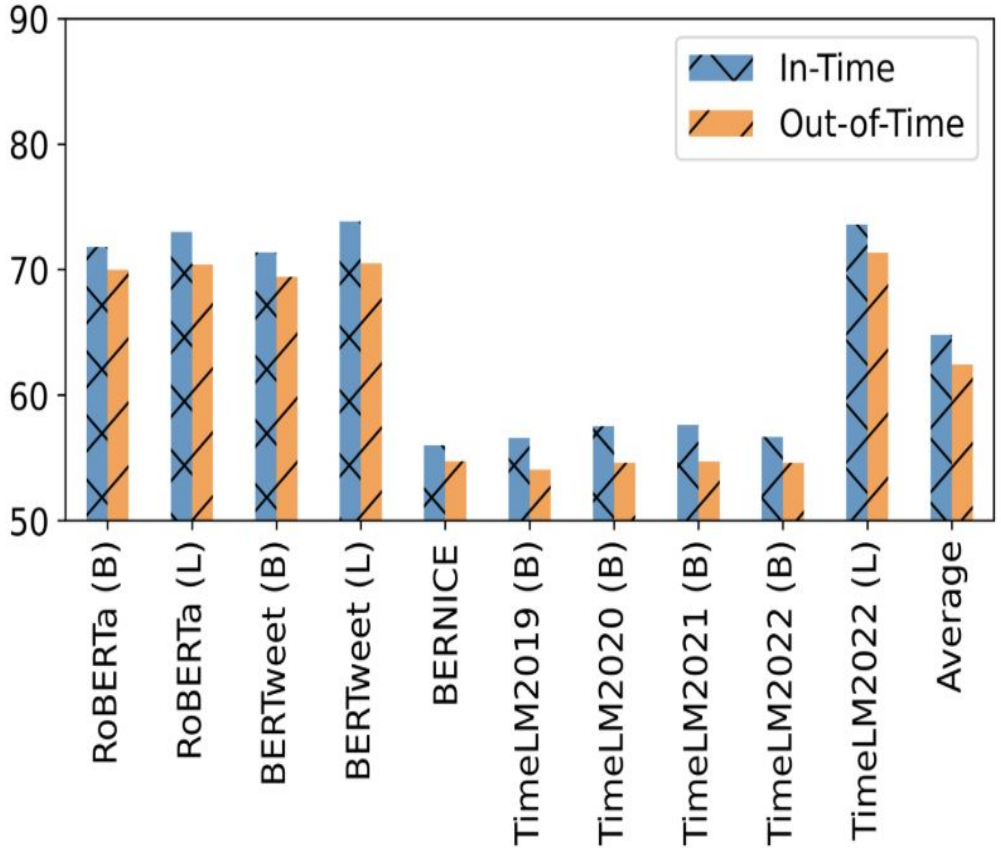
Sentiment



Pretty close results for in-time and out-of-time.

No temporal effect

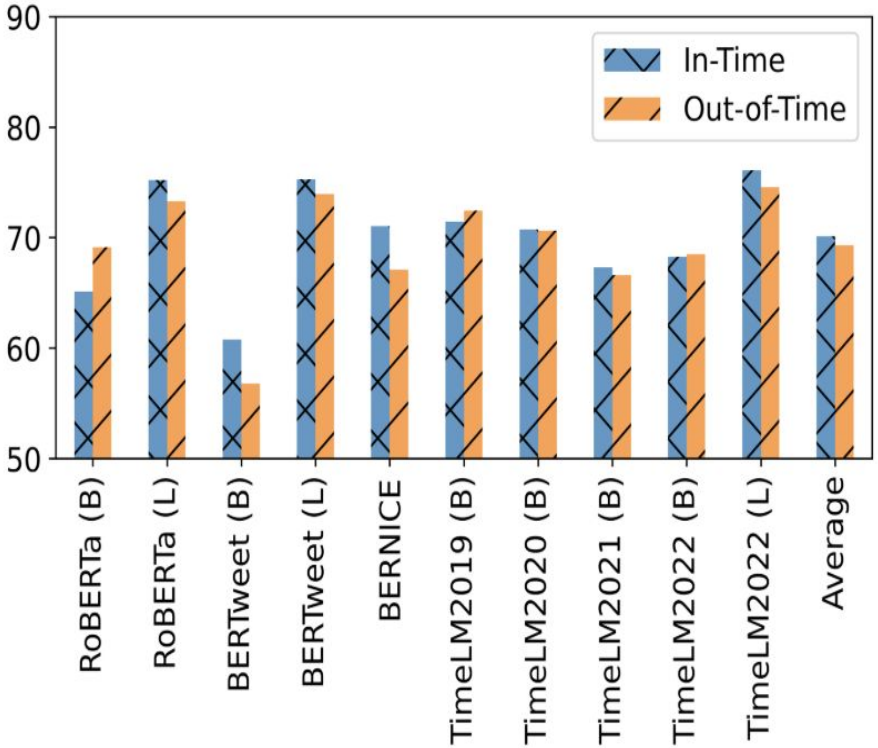
Named Entity Recognition



Better performance for in-time in all cases

Clear temporal effect

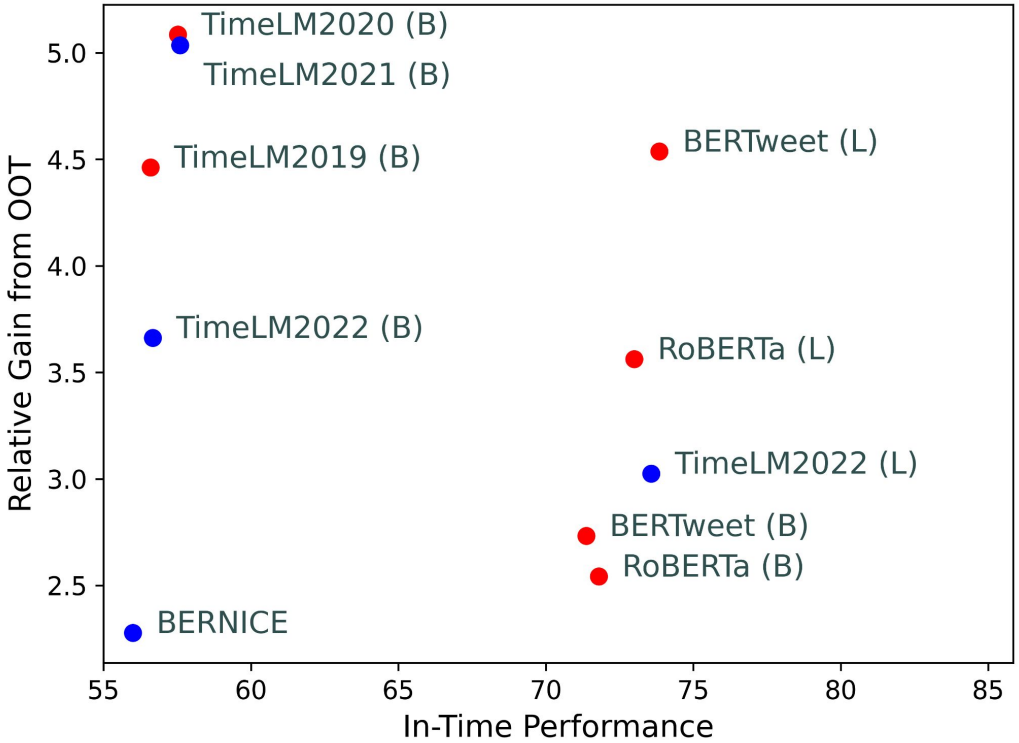
Topic classification



Mixed results for in-time and out-of-time.

Some temporal effects but more connected with data distribution

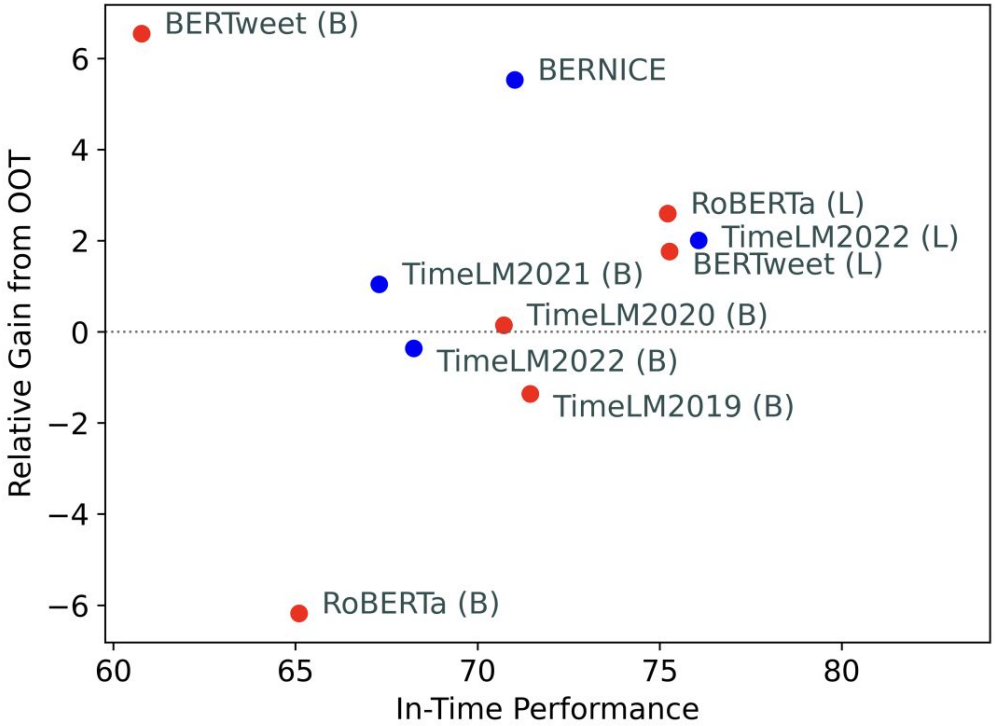
Effect of LM pre-training corpus



NER

- **Blue:** LMs with pre-training corpus including the test period
- **Red:** LMs without temporal overlap in.

Effect of LM pre-training corpus



Topic classification

- **Blue:** LMs with pre-training corpus including the test period
- **Red:** LMs without temporal overlap in.

Summary of results

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

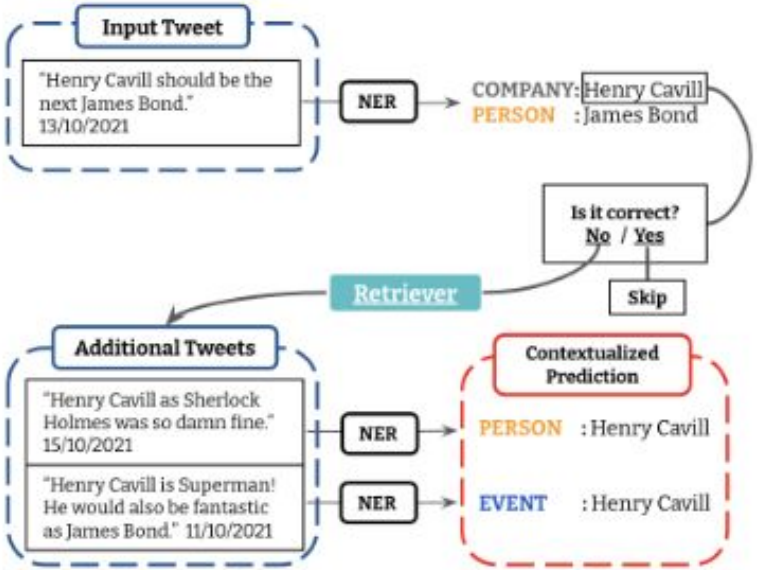
    particularly affected (e.g. NER, hate speech detection)

** Similar observations in Luu et al. (2022) and Agarwal and Nenkova (2022)*

Temporal challenges: Future work

- Development of new methods to **integrate contextual information** or non-labeled data in predictive models.

- **Semi-automatic labeling** of recent data, or to better leverage non-labeled data.



SuperTweetEval benchmark

(Antypas et al. EMNLP Findings 2023)

Unified benchmark with a range of **social media NLP tasks**, including regression, generation and classification.

⏳ Includes tasks with **temporal splits**!

SuperTweetEval benchmark

(Antypas et al. EMNLP Findings 2023)

Unified benchmark with a range of **social media NLP tasks**, including regression, generation and classification.

⏳ Includes tasks with **temporal splits**!

LongEval series at CLEF to evaluate performance consistency over time

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

SuperTweetEval, the benchmark

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 stillll 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

12 diverse NLP tasks

With temporal splits

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 stillllll 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

Conclusion

Social media entails many **challenges**, including immediacy.

Temporal adaptation is needed for some NLP tasks, and can only be partially solved with updated models.

Annotating **recent data** is a solution, but comes with costs -> Automate (part of) the process with LLMs?

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Thank you!