Language Models for Social Media
Challenges and Applications

Jose Camacho Collados

Cardiff NLP

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYD

Sheffield, 25 March 2022

Outline

- ➤ **Training** specialized language models

  - ○ **Challenges** of NLP in social media

- ➤ TweetEval: an evaluation **benchmark**

- ➤ **Multilingual** language models (XLM-T)

- ➤ **Temporal** adaptation (TimeLMs) and applications

About me

- Lecturer at **Cardiff University** (Wales, UK)
  - **UKRI Future Leaders Fellow** (£1.4M funding, 4+ years)
  - Co-founder and leader of the **Cardiff NLP group**.
- Areas of expertise: **Semantics, resources, multilinguality, social media**
- Co-author of "**Embeddings in NLP**" book
- **Previously:** Google Doctoral Fellow, PhD at Sapienza University (Italy, 2018)
  - Going even further back… studied **Mathematics** and an Erasmus Mundus Master in **NLP**.

Cardiff NLP

➢ Very young group (2 years old)

➢ Growing fast (100+ members, 20+ in the lab)

➢ **Website:** https://cardiffnlp.github.io/

➢ Activities welcome to public (seminars, etc.)

➢ **Twitter:** @Cardiff_NLP

Cardiff NLP Workshop 2022

➢ **Dates:** June 30 and July 1 (2 days) in Cardiff

➢ Especially targeted to **NLP PhD students** in Europe (but everyone is welcome)

➢ **Free registration**

➢ Mix of **invited speakers, tutorials** and **networking**

➢ More info coming soon! https://cardiffnlp.github.io/workshop/

The team

Cardiff NLP

Snap Inc.

Daniel Loureiro

Luis Espinosa Anke

Francesco Barbieri

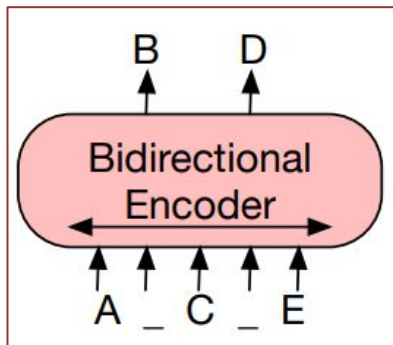Leonardo Neves

Jose Camacho Collados
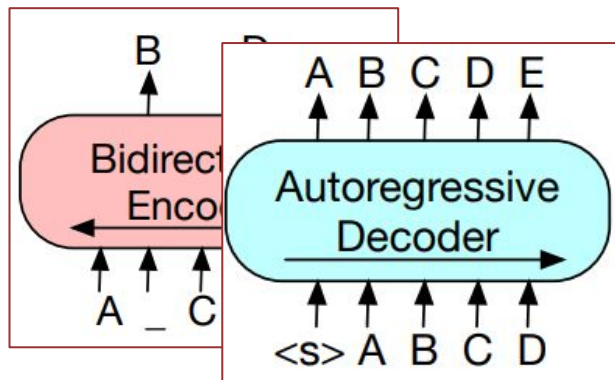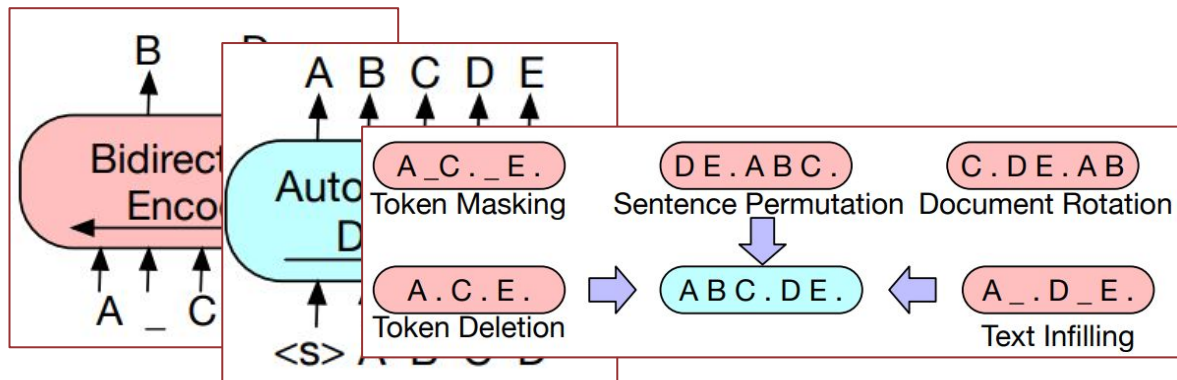
The team

Specialized LMs

Specialized language models

➢ Typical language model architectures

Specialized language models

➢ Typical language model architectures

Specialized language models

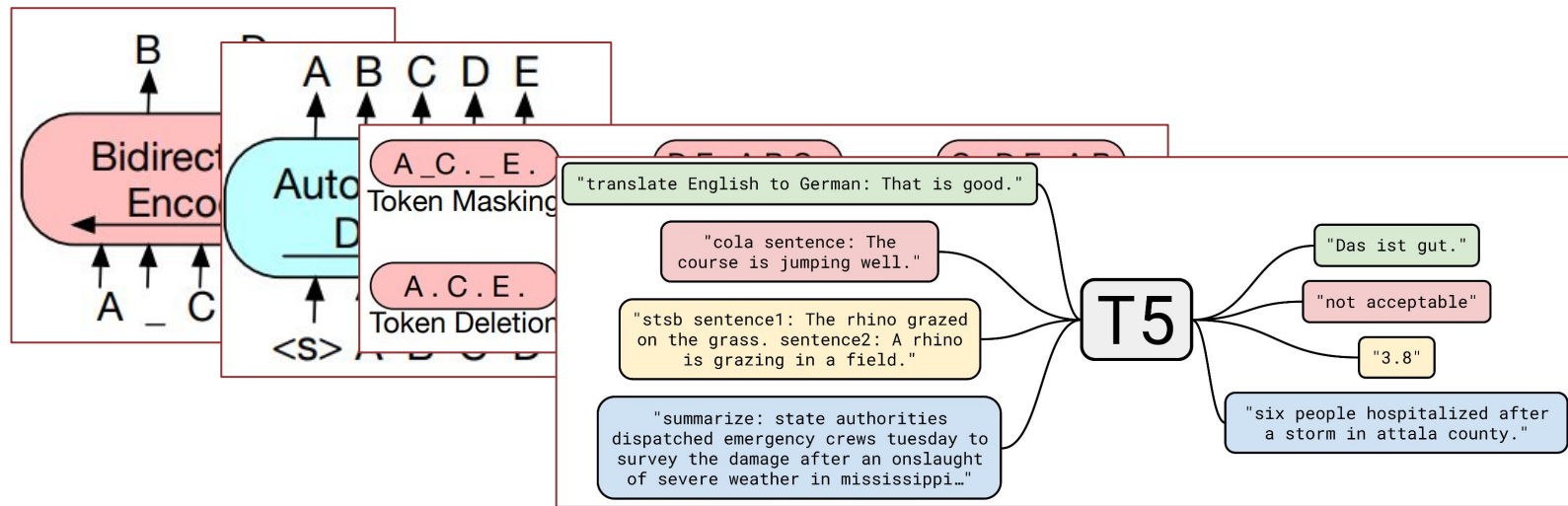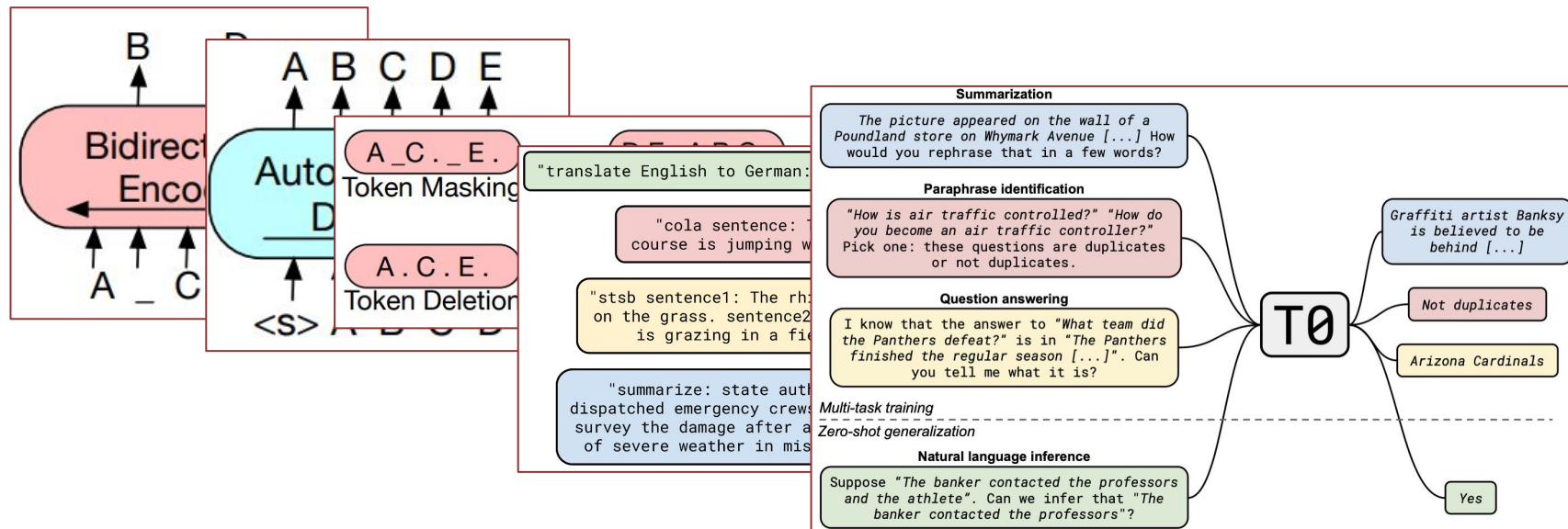➢ Typical language model architectures

Specialized language models

➢ Typical language model architectures

Specialized language models

➢ Typical language model architectures

Specialized language models

➢ Typical language model architectures
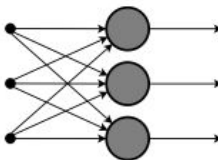
Specialized language models

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

Specialized language models

➢ Augmenting with external information

Task

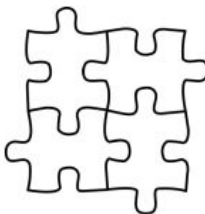Story ending,
Machine Comprehension
Social common sense
NLI

Neural Component

Pre/post pre-trained
language models

Knowledge Source

Knowledge bases,
extracted from text,
hand-crafted rules

Combination Method

Attention, pruning, word
embeddings, multi-task
learning

image credit: Maarten Sap, Vered
Shwartz, Antoine Bosselut, Yejin
Choi, and Dan Roth. 2020.
**Commonsense Reasoning for
Natural Language Processing**. In
Proceedings of the 58th Annual
Meeting of the Association for
Computational Linguistics: Tutorial
Abstracts, pages 27–33, Online.
Association for Computational
Linguistics.

Specialized language models
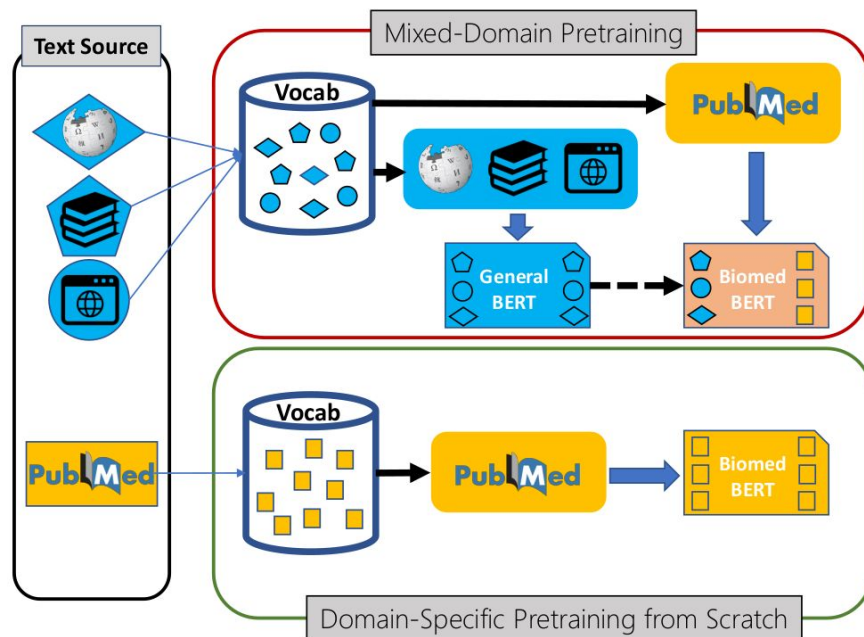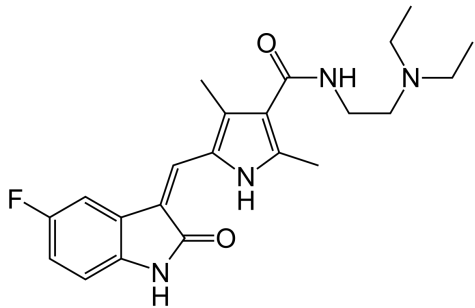
➢ Pretraining on domain-specific corpora

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.

Specialized language models

[MASK] is a tyrosine kinase inhibitor.

Specialized language models

[MASK] is a tyrosine kinase inhibitor.

it, he, this, she…

sunitinib, imatinib, gefitinib, sorafenib…

Specializing a LM on social media

Specializing a LM on social media

➢ Why?

- **Informal** grammar
- **Multilingual** (code-switching, etc.)
- Irregular **vocabulary**
  - Emoji, abbreviations, typographical errors, hashtags, mentions…
- Tweets are often **not standalone messages**
  - RTs, mentions, replies, threads, pictures…
- And because social media is **important**
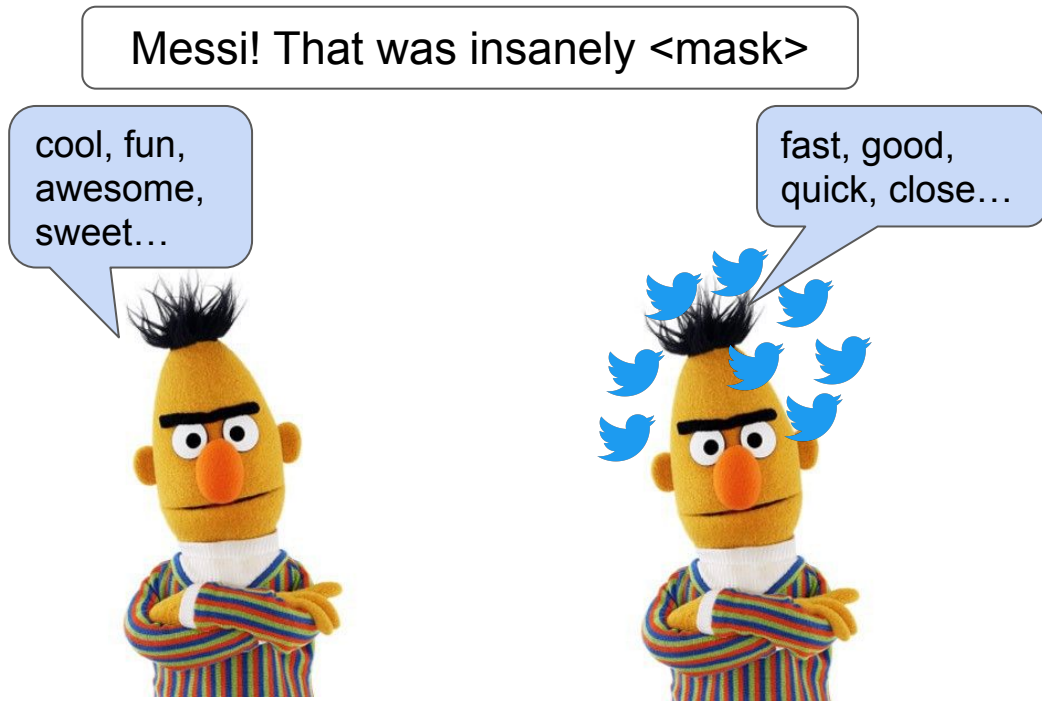
Specializing a LM on social media

➢ Why?

Messi! That was insanely <mask>

Specializing a LM on social media

➢ Why?

Messi! That was insanely <mask>

cool, fun, awesome, sweet…

fast, good, quick, close…

Specializing a LM on social media

➢ Why?

Specializing a LM on social media

➢ Why?

		Emoji	Emotion	Hate	Irony	Offensive	Sentiment	Stance	ALL
Val	SVM	25.0	63.8	73.1	63.4	72.7	68.4	67.9	62.0
	FastText	23.2	62.9	71.7	62.7	70.0	62.2	67.3	60.0
	BLSTM	19.4	62.6	72.1	60.6	72.1	61.9	63.4	58.9
	RoB-Bs	24.7±0.3 (24.3)	73.1±1.7 (74.9)	76.5±0.3 (76.6)	73.7±0.6 (73.7)	77.1±0.6 (77.6)	71.4±1.9 (72.7)	71.4±1.9 (73.9)	67.7
	RoB-RT	24.4±1.5 (**26.2**)	75.4±1.5 (**77.0**)	77.8±1.1 (**79.6**)	74.7±1.5 (**75.6**)	77.2±0.6 (**77.7**)	73.0±1.2 (**74.2**)	72.9±1.0 (**75.2**)	**69.4**
	RoB-Tw	23.4±1.1 (24.6)	67.6±0.9 (68.6)	74.3±2.0 (76.6)	70.0±0.3 (70.7)	76.1±0.6 (76.2)	70.5±1.0 (69.4)	68.3±2.4 (71.4)	65.4
Test	SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
	FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
	BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
	RoB-Bs	30.9±0.2 (30.8)	76.1±0.5 (76.6)	46.6±2.5 (44.9)	59.7±5.0 (55.2)	79.5±0.7 (78.7)	71.3±1.1 (72.0)	68±0.8 (70.9)	61.3
	RoB-RT	31.4±0.4 (**31.6**)	78.5±1.2 (**79.8**)	52.3±0.2 (**55.5**)	61.7±0.6 (62.5)	80.5±1.4 (**81.6**)	72.6±0.4 (**72.9**)	69.3±1.1 (**72.6**)	**65.2**
	RoB-Tw	29.3±0.4 (29.5)	72.0±0.9 (71.7)	46.9±2.9 (45.1)	65.4±3.1 (**65.1**)	77.1±1.3 (78.6)	69.1±1.2 (69.3)	66.7±1.0 (67.9)	61.0
	SotA	36.0*	-	65.1	70.5	82.9	68.5	71.0	-
Metric		M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG ($F^{(a)}, F^{(f)}$)	TE

cardiffnlp/twitter-roberta-base

Specializing a LM on social media

➢ Why?

		Emoji	Emotion	Hate	Irony
Val	SVM	25.0	63.8	73.1	63.4
	FastText	23.2	62.9	71.7	62.7
	BLSTM	19.4	62.6	72.1	60.6
	RoB-Bs	24.7±0.3 (24.3)	73.1±1.7 (74.9)	76.5±0.3 (76.6)	73.7±0.6 (73.)
	RoB-RT	24.4±1.5 (**26.2**)	75.4±1.5 (**77.0**)	77.8±1.1 (**79.6**)	74.7±1.5 (**75.**)
	RoB-Tw	23.4±1.1 (24.6)	67.6±0.9 (68.6)	74.3±2.0 (76.6)	70.0±0.3 (70.)
Test	SVM	29.3	64.7	36.7	61.7	52.3	62.9	67.3	53.5
	FastText	25.8	65.2	50.6	63.1	73.4	62.9	65.4	58.1
	BLSTM	24.7	66.0	52.6	62.8	71.7	58.3	59.4	56.5
	RoB-Bs	30.9±0.2 (30.8)	76.1±0.5 (76.6)	46.6±2.5 (44.9)	59.7±5.0 (55.2)	79.5±0.7 (78.7)	71.3±1.1 (72.0)	68±0.8 (70.9)	61.3
	RoB-RT	31.4±0.4 (**31.6**)	78.5±1.2 (**79.8**)	52.3±0.2 (**55.5**)	61.7±0.6 (62.5)	80.5±1.4 (**81.6**)	72.6±0.4 (**72.9**)	69.3±1.1 (**72.6**)	**65.2**
	RoB-Tw	29.3±0.4 (29.5)	72.0±0.9 (71.7)	46.9±2.9 (45.1)	65.4±3.1 (**65.1**)	77.1±1.3 (78.6)	69.1±1.2 (69.3)	66.7±1.0 (67.9)	61.0
	SotA	36.0*	-	65.1	70.5	82.9	68.5	71.0	-
Metric		M-F1	M-F1	M-F1	$F^{(i)}$	M-F1	M-Rec	AVG ($F^{(a)}$, $F^{(f)}$)	TE

	Model	F_1^{pos}		Accuracy	
		soft	hard	soft	hard
Our results	RoBERTa_large	73.2	71.9	76.5	75.1
	XLM-R_large	70.8	69.7	74.2	73.2
	RoBERTa_base	71.0	71.2	74.0	74.0
	XLM-R_base	66.6	66.2	70.8	70.8
	BERTweet	**74.6**	**74.3**	**78.2**	**78.2**
	Wu et al. (2018)	70.5		73.5	
	Baziotis et al. (2018)	67.2		73.2	

BERTweet

cardiffnlp/twitter-roberta-base

TweetEval:
Language Models and Evaluation Benchmark

TweetEval, the benchmark
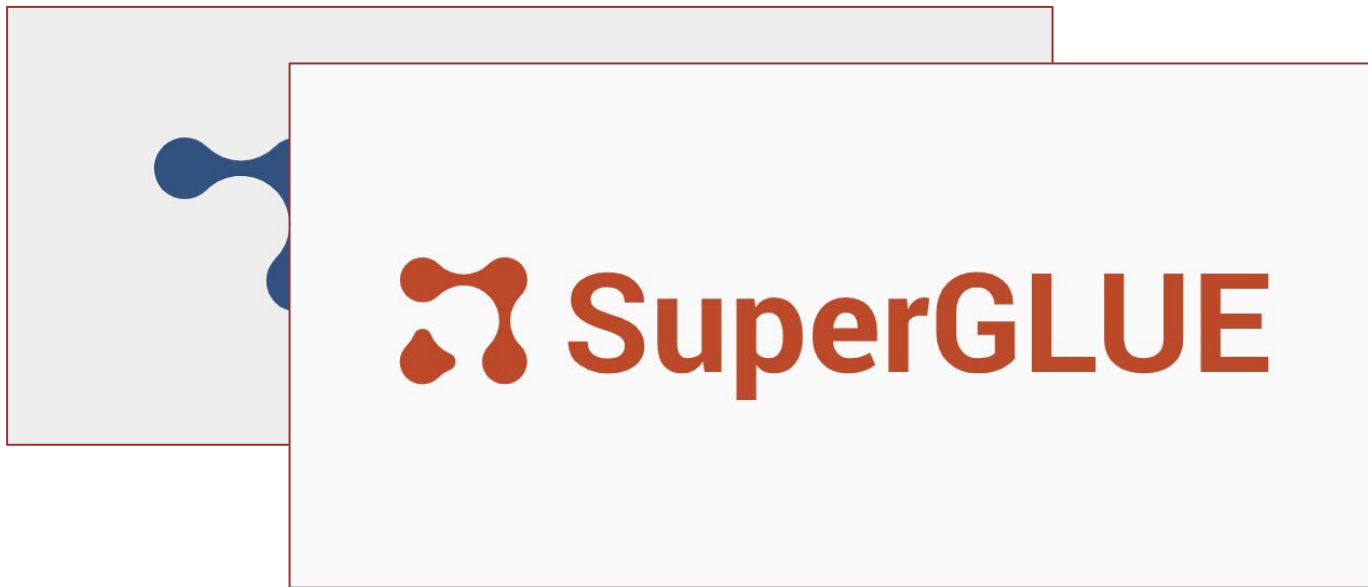
(Barbieri et al. EMNLP Findings 2020)
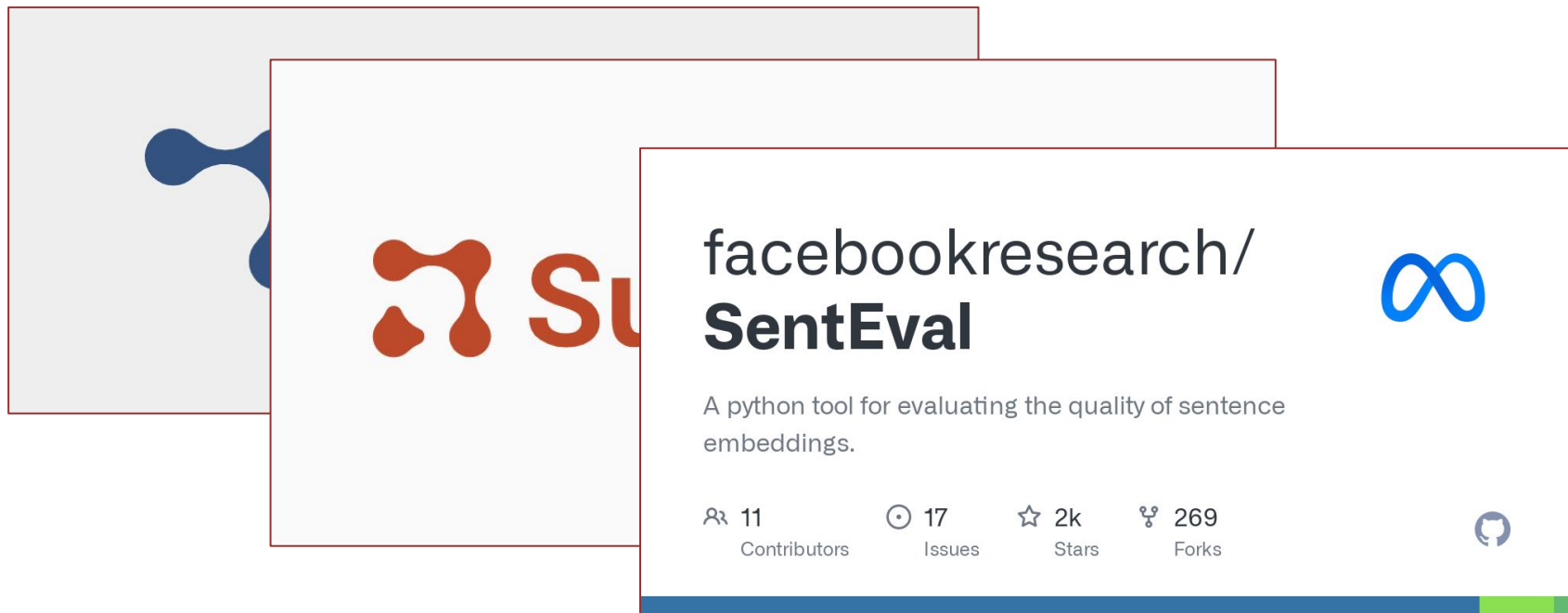
➢ Why?

TweetEval, the benchmark

➢ Why?

TweetEval, the benchmark

➢ Why?

TweetEval, the benchmark

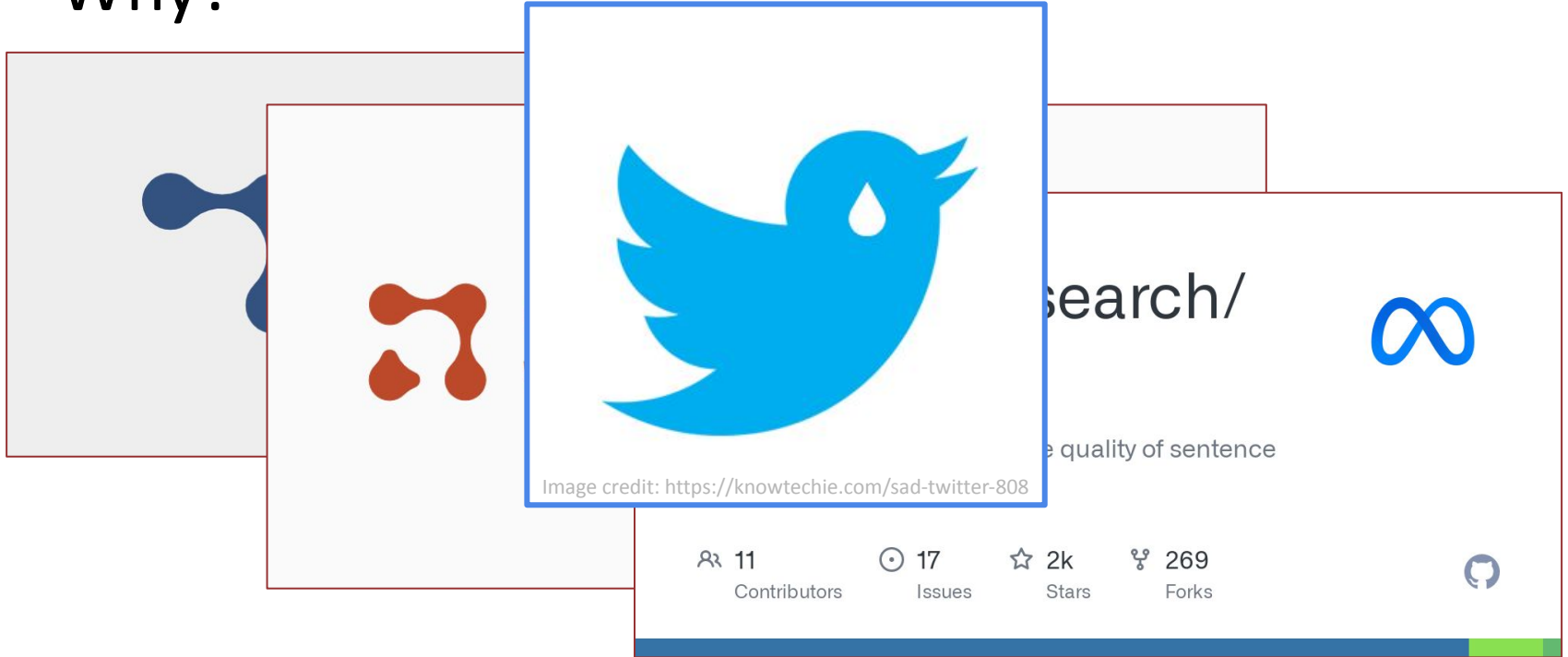➢ Why?

facebookresearch/
SentEval

A python tool for evaluating the quality of sentence embeddings.

👥 11   ⊙ 17   ☆ 2k   ⑂ 269
Contributors   Issues   Stars   Forks

TweetEval, the benchmark

➢ Why?

search/

e quality of sentence

👥 11
Contributors

⊙ 17
Issues

⭐ 2k
Stars

ᵞ 269
Forks

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

TweetEval, the language model

➢ What?

TweetEval, the language model

➤ What?

TweetEval, the language model

➢ How?

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

XLM-T:
Multilingual Language Model for Twitter

XLM-T

➢ Why?

XLM-T

➢ Why?

  ○ Same as in TweetEval, no multilingual LMs, and no unified multilingual benchmarks

XLM-T

➢ What?

Sentiment analysis benchmark

	FT Mono	XLM-R Mono	XLM-Tw Mono	XLM-R Multi	XLM-Tw Multi
Arabic	46.0	63.6	67.7	64.3	66.9
English	50.9	68.2	66.9	68.5	70.6
French	54.8	72.0	68.2	70.5	71.2
German	59.6	73.6	76.1	72.8	77.3
Hindi	37.1	36.6	40.3	53.4	56.4
Italian	54.7	71.5	70.9	68.6	69.1
Portuguese	55.1	67.1	76.0	69.8	75.4
Spanish	50.1	65.9	68.5	66.0	67.9
All lang.	*51.0*	*64.8*	*66.8*	*66.8*	*69.4*

XLM-T

➢ What?
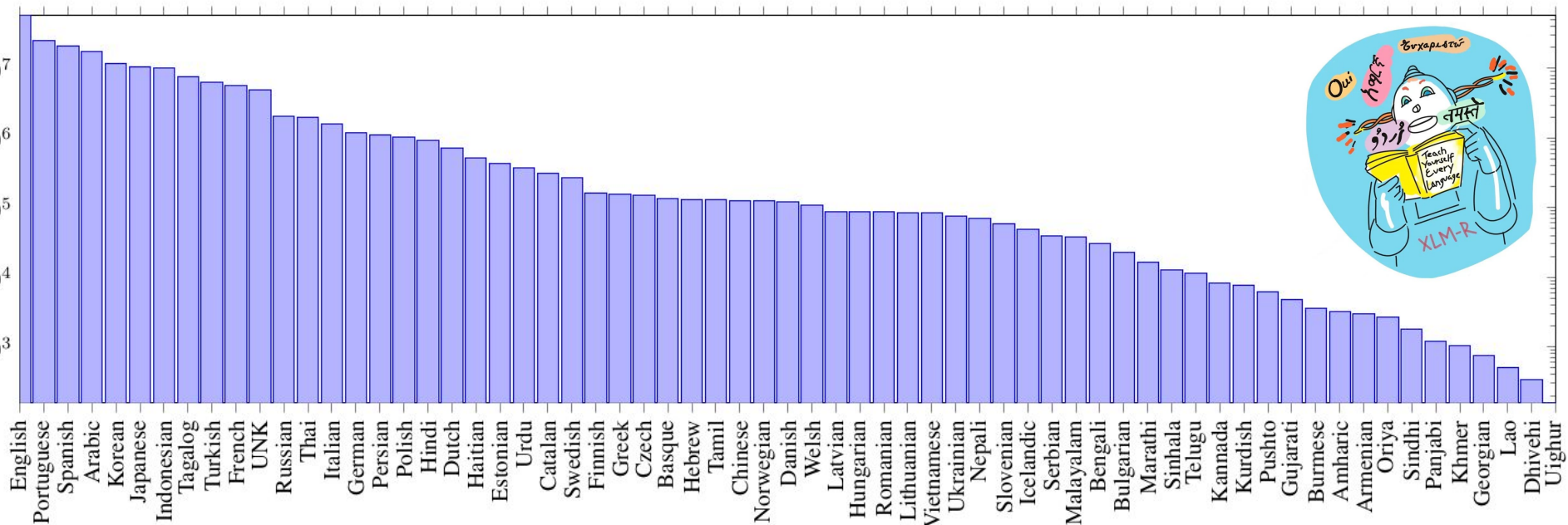
XLM-T

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

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

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

Pseudo-perplexity over time
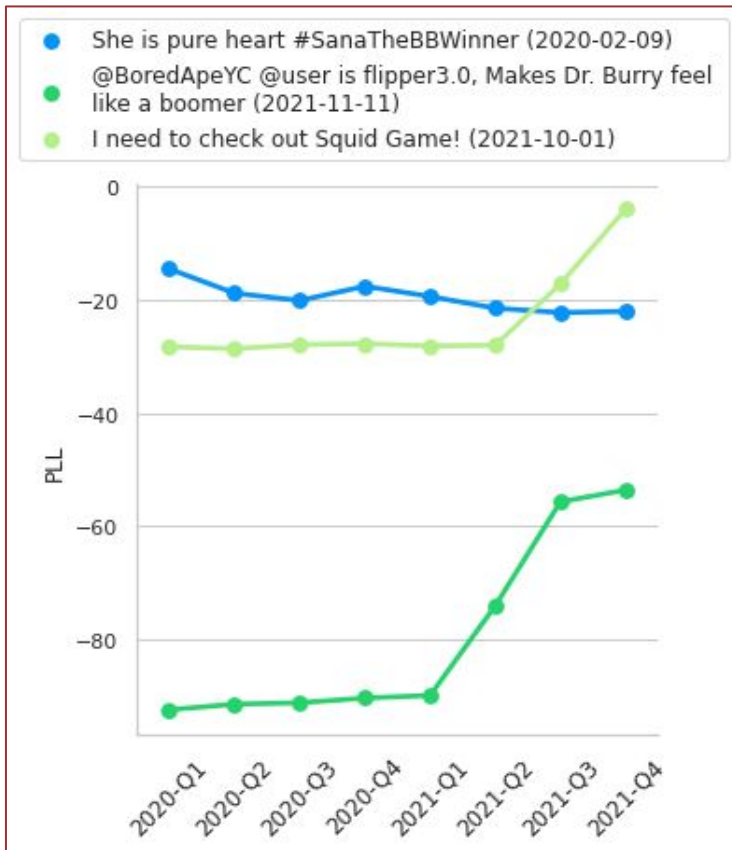
TimeLMs models

Always updated at github.com/cardiffnlp/timelms

And in Hugging Face 🤗. For instance, sentiment:

huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest
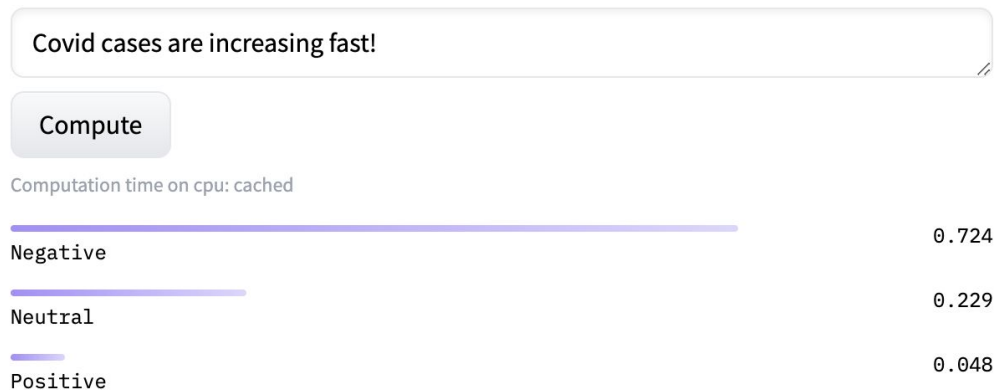
EMNLP-22 EvoNLP Workshop

The First Workshop on Ever Evolving NLP.

A forum to discuss the challenges posed by the **dynamic** nature of language in the specific context of the current NLP paradigm, dominated by language models.

Co-organised with **industry** partners of DeepMind and Snap Inc.
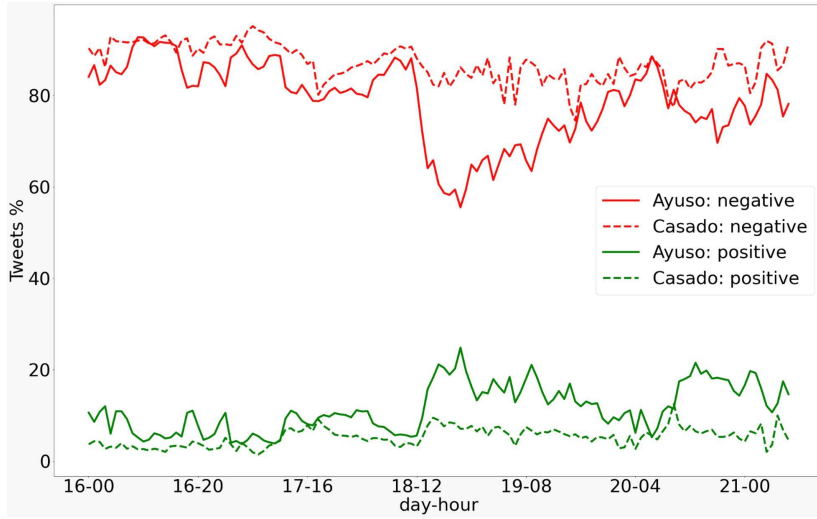
Invited

Speakers

Eunsol Choi

University of Austin, Texas

Jacob Eisenstein

Google AI

Adam Jatowt

University of Innsbruck, Austria

Ozan Sener

Intel Labs

Website: https://sites.google.com/view/evonlp

Applications of social media models

EL PAÍS
EL PERIÓDICO GLOBAL EN ESPAÑOL

TWITTER >
Un análisis de 1,3 millones de tuits sobre Casado y Ayuso da una ligera ventaja a la presidenta madrileña

Dos investigadores españoles han creado un modelo capaz de analizar el sentimiento en Twitter, que se ha convertido en uno de los más usados del mundo

https://elpais.com/tecnologia/2022-02-21/un-analisis-de-13-millones-de-tuits-sobre-casado-y-ayuso-da-una-ligera-ventaja-a-la-presidenta-madrilena.html

Politics, sentiment and virality

Dimosthenis Antypas

Alun Preece

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYⱭ

Crime and Security
Research Institute

Sefydliad Ymchwil
Trosedd a Diogelwch

Preprint (2022): https://arxiv.org/abs/2202.00396

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)
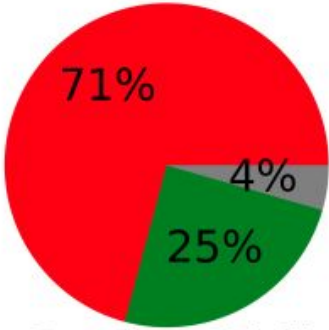
Politics, sentiment and virality

Collected a corpus of Twitter messages from MPs in **Greece, Spain and UK** (focus on 2021, 400K tweets)
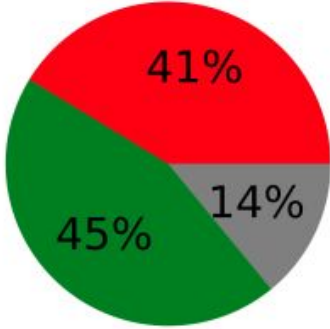
Analysed the relation between **sentiment** (as provided by our Twitter-based models) and **virality** (measured by number of retweets)

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

Sentiment of Spanish MP tweets

71%   4%
25%

**Popular tweets
(top 5%)**

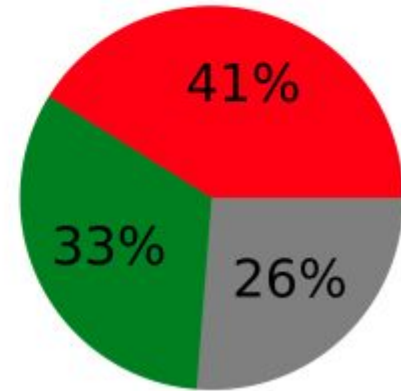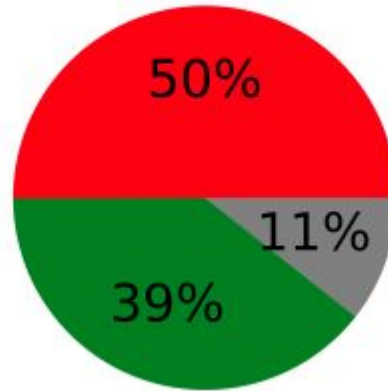41%
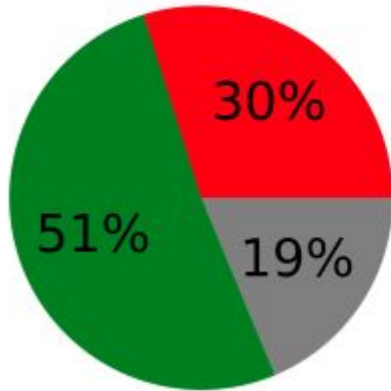45%   14%

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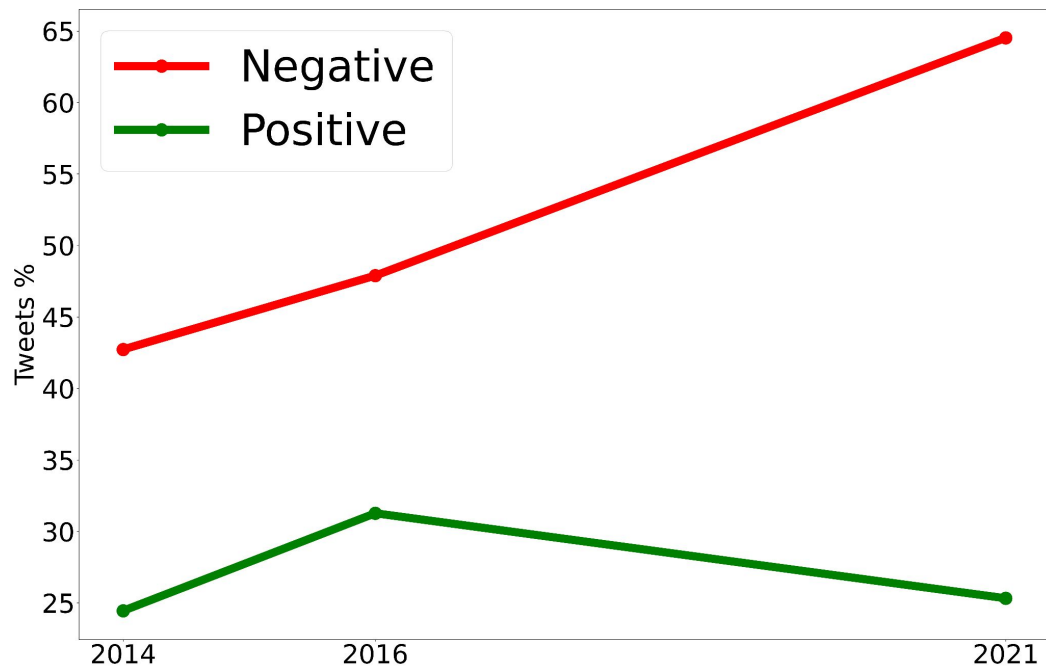
Most popular tweets

71% are negative!

■ Positive

■ Neutral
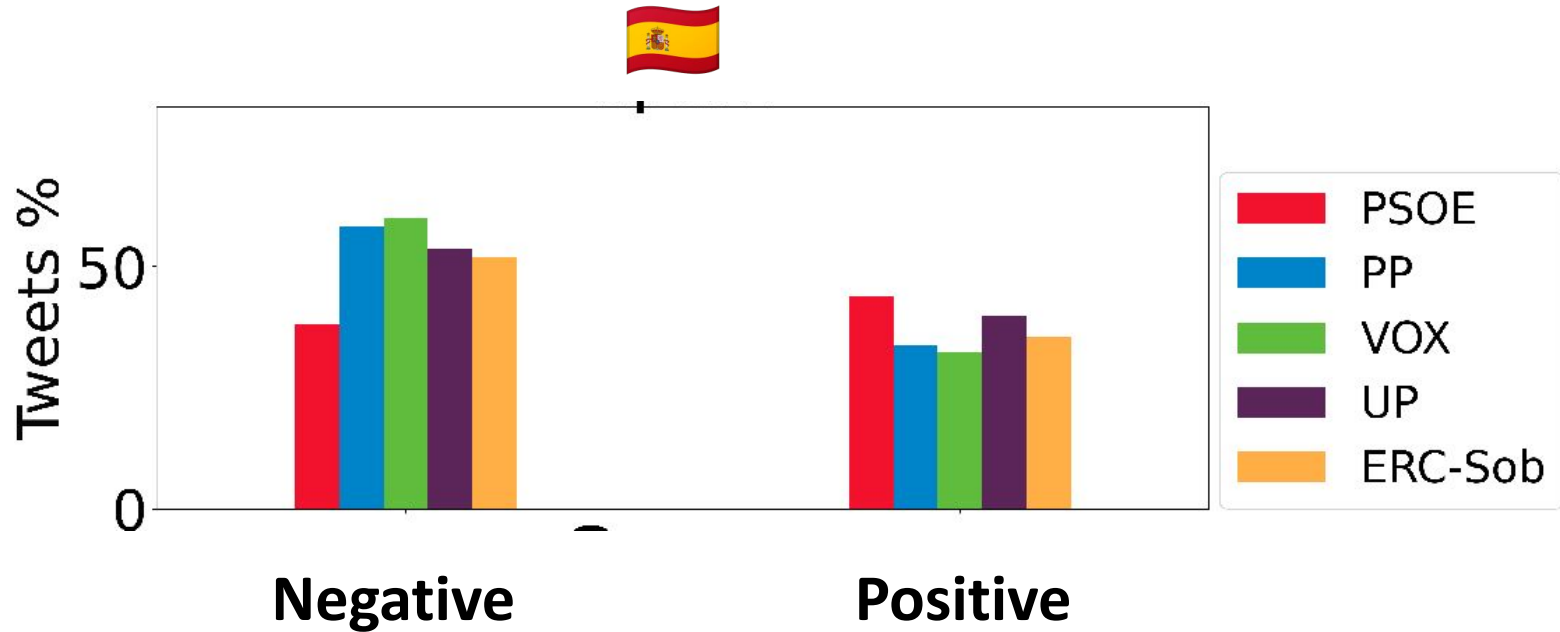
■ Negative

Sentiment of MPs' tweets

Spanish MPs more negative overall

Sentiment over time

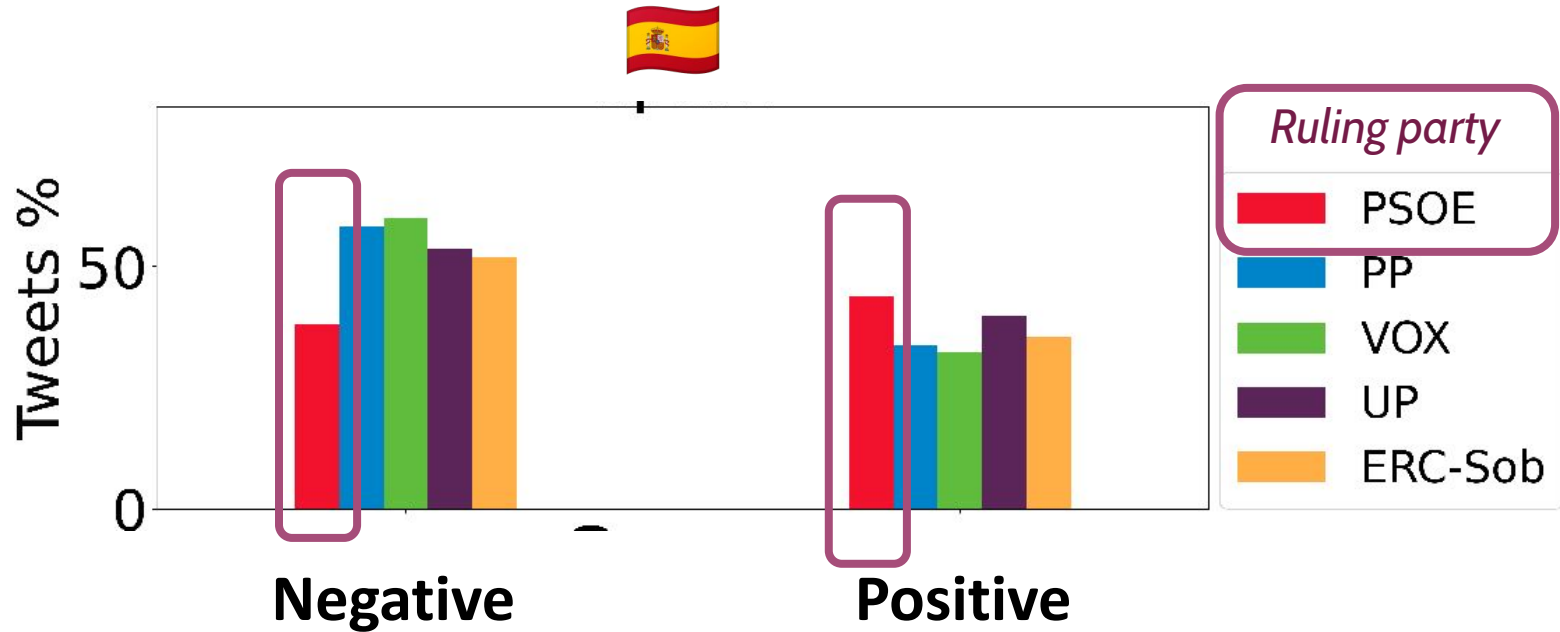Tweets by MPs are becoming more negative over time (UK)

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 (Spain)

13%   vs.   63%

Negative
tweets

Pedro Sánchez
Prime Minister

Pablo Casado
Leader of the
opposition

Disinformation: *Work in progress…*

Focusing on the **textual content** of social media posts.

Analysing various features using previous models: sentiment, emotion, hate speech, etc.

Study the **correlation** between these features and disinformation content/accounts.

Conclusion

Social media entails many **challenges.**

Specialized language models can help, but not the only solution.

Temporal adaptation is needed, but can be partially solved with newer models.

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

Summary of resources

➢ **TweetEval**: https://github.com/cardiffnlp/tweeteval

➢ **XLM-T**: https://github.com/cardiffnlp/xlm-t

➢ **TimeLMs**: https://github.com/cardiffnlp/timelms

All models available in the Hugging Face hub:

https://huggingface.co/cardiffnlp 🤗

Summary of resources

➢ **TweetEval**: https://github.com/cardiffnlp/tweeteval

➢ **XLM-T**: https://github.com/cardiffnlp/xlm-t

➢ **TimeLMs**: https://github.com/cardiffnlp/timelms

All models available in the Hugging Face hub:

https://huggingface.co/cardiffnlp 🤗

Thank you!