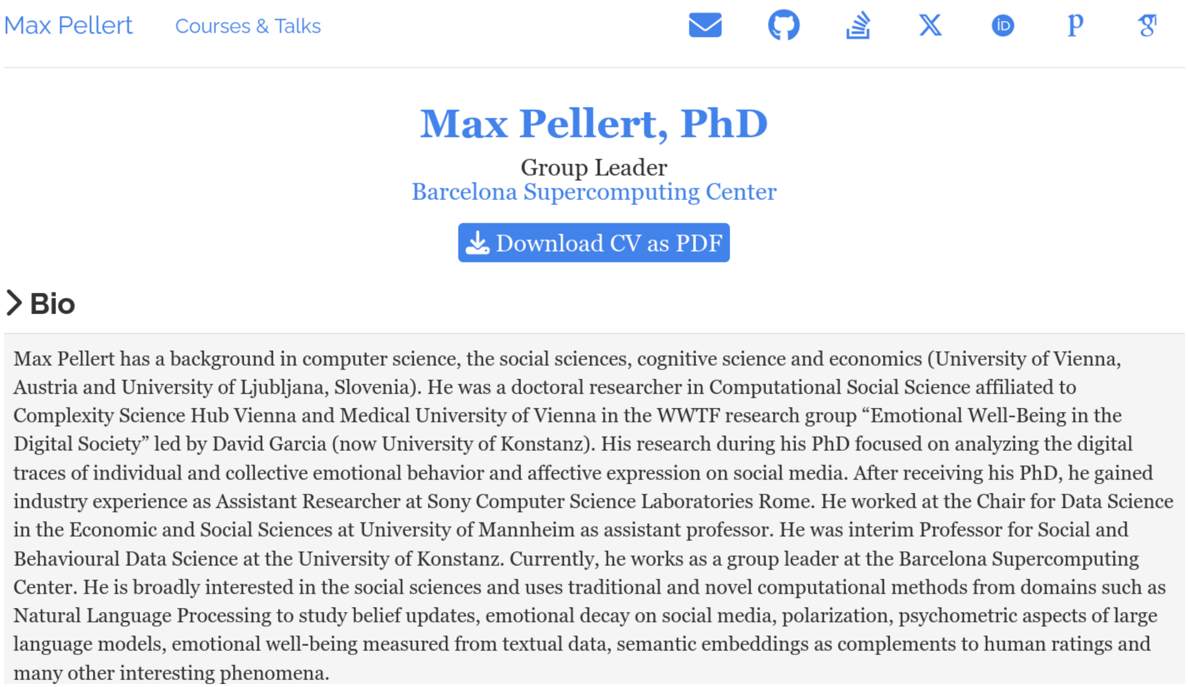
UPF Talk 02/2025: Studies, Methods & Outlook

Max Pellert (https://mpellert.at)

https://mpellert.at/upf_talk_02_25/upf_talk_02_25.pdf

https://mpellert.at



Max Pellert, PhD

Group Leader
Barcelona Supercomputing Center

Bio

Max Pellert has a background in computer science, the social sciences, cognitive science and economics (University of Vienna, Austria and University of Ljubljana, Slovenia). He was a doctoral researcher in Computational Social Science affiliated to Complexity Science Hub Vienna and Medical University of Vienna in the WWTF research group “Emotional Well-Being in the Digital Society” led by David Garcia (now University of Konstanz). His research during his PhD focused on analyzing the digital traces of individual and collective emotional behavior and affective expression on social media. After receiving his PhD, he gained industry experience as Assistant Researcher at Sony Computer Science Laboratories Rome. He worked at the Chair for Data Science in the Economic and Social Sciences at University of Mannheim as assistant professor. He was interim Professor for Social and Behavioural Data Science at the University of Konstanz. Currently, he works as a group leader at the Barcelona Supercomputing Center. He is broadly interested in the social sciences and uses traditional and novel computational methods from domains such as Natural Language Processing to study belief updates, emotional decay on social media, polarization, psychometric aspects of large language models, emotional well-being measured from textual data, semantic embeddings as complements to human ratings and many other interesting phenomena.

Currently: Group Leader at the Barcelona Supercomputing Center in the Department for Computational Social Science and Humanities

Before: Professor for Social and Behavioural Data Science (interim, W2) at the University of Konstanz

https://mpellert.at

Assistant Professor (Business School of the University of Mannheim)

I worked in industry at SONY Computer Science Laboratories in Rome, Italy

PhD from the Complexity Science Hub Vienna and the Medical University of Vienna in Computational Social Science

Studies in Psychology and History and Philosophy of Science

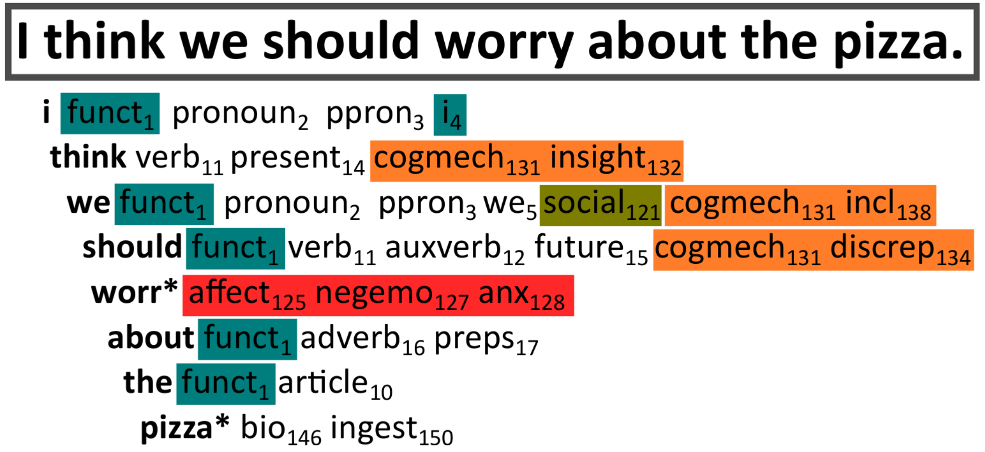
Msc in Cognitive Science and Bsc in Economics (both University of Vienna)

Basics: Extracting Signals from Text

One example: **Linguistic Inquiry and Word Count, LIWC (pronounced “Luke”)**

Simple word matching method

Generated and validated by psychologists (Pennebaker et al., 2001-today)



I think we should worry about the pizza.

i funct$_1$ pronoun$_2$ ppron$_3$ i$_4$

think verb$_{11}$ present$_{14}$ cogmech$_{131}$ insight$_{132}$

we funct$_1$ pronoun$_2$ ppron$_3$ we$_5$ social$_{121}$ cogmech$_{131}$ incl$_{138}$

should funct$_1$ verb$_{11}$ auxverb$_{12}$ future$_{15}$ cogmech$_{131}$ discrep$_{134}$

worr* affect$_{125}$ negemo$_{127}$ anx$_{128}$

about funct$_1$ adverb$_{16}$ preps$_{17}$

the funct$_1$ article$_{10}$

pizza* bio$_{146}$ ingest$_{150}$

Examples of LIWC classes:

Positive Affect, Negative Affect

Anxiety, Sadness, Anger

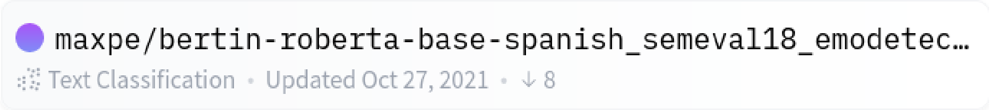
Social processes

Basics: Extracting Signals from Text

More advanced examples using deep learning

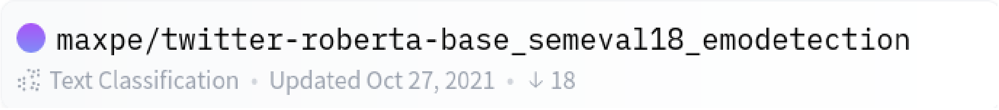
Classifiers based on transformer architectures (RoBERTa)

Large general purpose language models adapted to the task of emotion classification



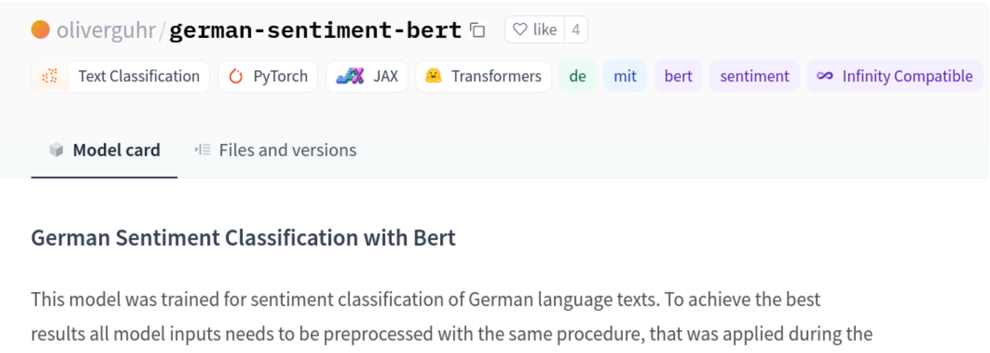
maxpe/bertin-roberta-base-spanish_semeval18_emodetec…

Text Classification • Updated Oct 27, 2021 • ↓ 8



maxpe/twitter-roberta-base_semeval18_emodetection

Text Classification • Updated Oct 27, 2021 • ↓ 18



oliverguhr / **german-sentiment-bert** ♡ like 4

Text Classification · PyTorch · JAX · Transformers · de · mit · bert · sentiment · Infinity Compatible

German Sentiment Classification with Bert

This model was trained for sentiment classification of German language texts. To achieve the best results all model inputs needs to be preprocessed with the same procedure, that was applied during the

https://huggingface.co/maxpe

https://huggingface.co/oliverguhr/german-sentiment-bert

And many many more…

Sentiment Analysis

Has gotten a somewhat bad name: “Why don’t we run something on the text?”

Often conceptually flawed + noisy data + inadequate annotation schemes to create many different tools

Results can be cherry-picked by optimizing on the tool

But, we argue, used right it can be a valuable research instrument

Sentiment Analysis Evidence

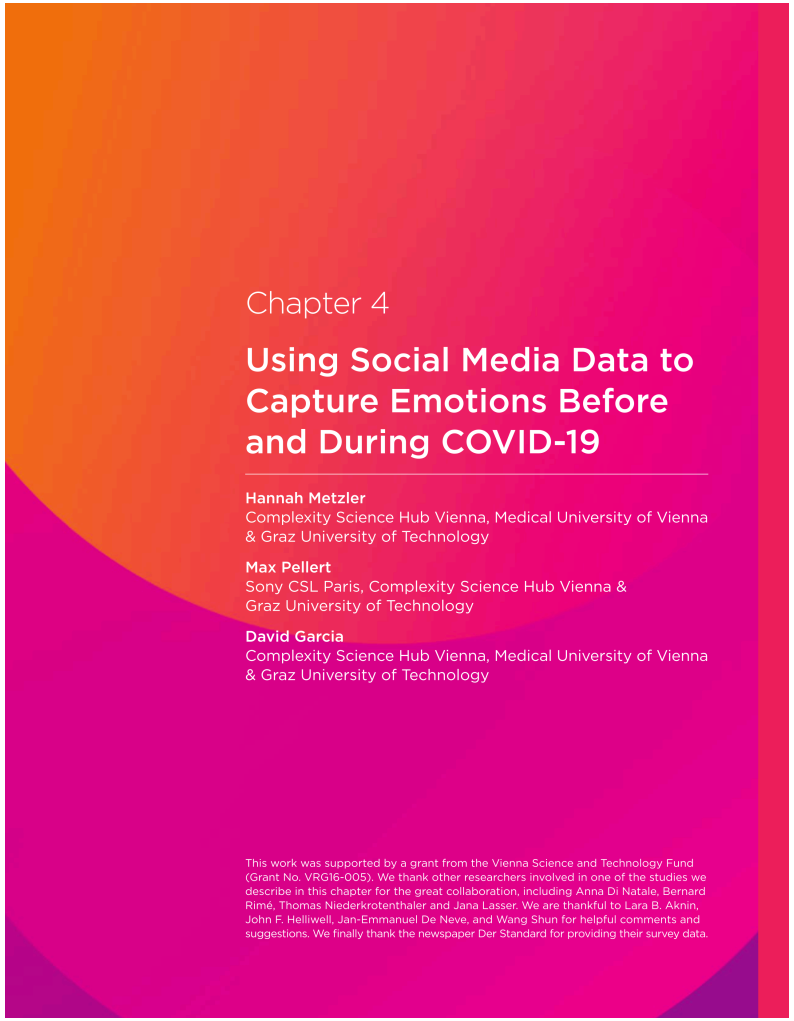
Individual text level (for example a single tweet): Not reliable, sarcasm, irony, performative nature of social media: we need a substantial number of texts to get through the noise (especially with dictionary methods, also base rates are low)

Individual person level: Associations sometimes higher (for example for depression: Eichstaedt et al., 2018) and sometimes lower (PANAS scale: Beasley & Mason, 2015) with (rather) stable personality traits

Group level (geographical): Debated, for example Twitter heart disease study (Eichstaedt et al., 2015), methods have to be validated and checked for robustness (Jaidka et al., 2020)

Our contribution: *Macroscopically* validating if we are able to capture momentary feeling of a population on a daily level

World Happiness Report



Chapter 4

Using Social Media Data to Capture Emotions Before and During COVID-19

Hannah Metzler
Complexity Science Hub Vienna, Medical University of Vienna & Graz University of Technology

Max Pellert
Sony CSL Paris, Complexity Science Hub Vienna & Graz University of Technology

David Garcia
Complexity Science Hub Vienna, Medical University of Vienna & Graz University of Technology

This work was supported by a grant from the Vienna Science and Technology Fund (Grant No. VRG16-005). We thank other researchers involved in one of the studies we describe in this chapter for the great collaboration, including Anna Di Natale, Bernard Rimé, Thomas Niederkrotenthaler and Jana Lasser. We are thankful to Lara B. Aknin, John F. Helliwell, Jan-Emmanuel De Neve, and Wang Shun for helpful comments and suggestions. We finally thank the newspaper Der Standard for providing their survey data.

Metzler, H., Pellert, M., & Garcia, D. (2022). Using Social Media Data to Capture Emotions Before and During COVID-19 (World Happiness Report 2022). https://worldhappiness.report/ed/2022/using-social-media-data-to-capture-emotions-before-and-during-covid-19/

David Garcia

Hannah Metzler

Data sources

derstandard.at

An internet pioneer in the German speaking area (centered on Austria)

Popular page: almost 57 million visits in November 2020

Active forum with many postings below news articles

Twitter

Tweets from Austria (data on location from Brandwatch)

This tweet contains emotions.

derStandard.at

Mood Survey on derstandard.at

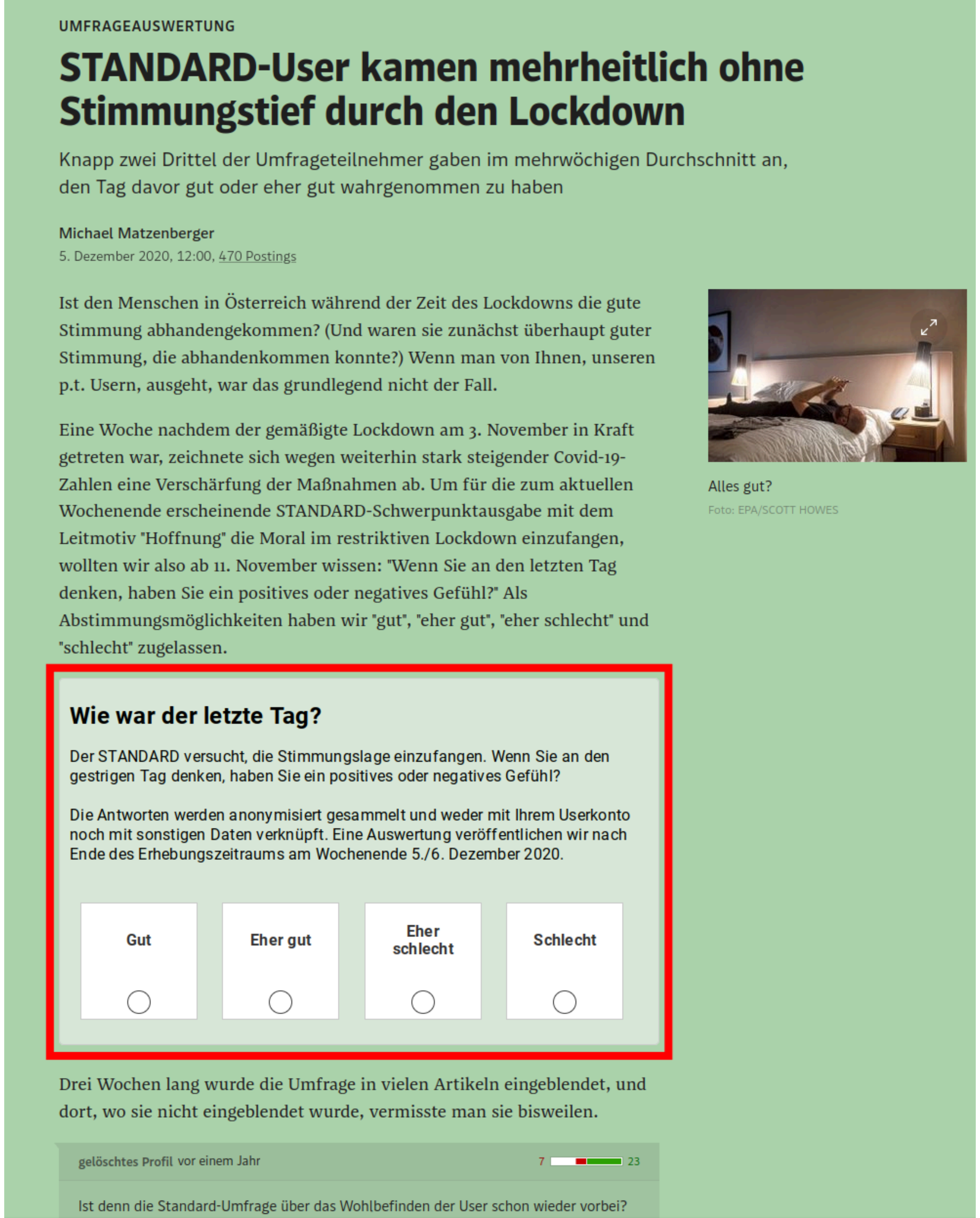
Survey on yesterday’s emotional state run for 20 days in November 2020

“How was your last day” (“Wie war der letzte Tag?”)

Was displayed in between the article text in a low barrier manner, could be answered anonymously

In a collaboration with derstandard.at, we obtained the survey results

The data allows us to investigate the relationship of the explicit survey measure with the results of methods that extract sentiment indirectly from text



UMFRAGEAUSWERTUNG

STANDARD-User kamen mehrheitlich ohne Stimmungstief durch den Lockdown

Knapp zwei Drittel der Umfrageteilnehmer gaben im mehrwöchigen Durchschnitt an, den Tag davor gut oder eher gut wahrgenommen zu haben

Michael Matzenberger
5. Dezember 2020, 12:00, 470 Postings

Ist den Menschen in Österreich während der Zeit des Lockdowns die gute Stimmung abhandengekommen? (Und waren sie zunächst überhaupt guter Stimmung, die abhandenkommen konnte?) Wenn man von Ihnen, unseren p.t. Usern, ausgeht, war das grundlegend nicht der Fall.

Alles gut?

Foto: EPA/SCOTT HOWES

Eine Woche nachdem der gemäßigte Lockdown am 3. November in Kraft getreten war, zeichnete sich wegen weiterhin stark steigender Covid-19-Zahlen eine Verschärfung der Maßnahmen ab. Um für die zum aktuellen Wochenende erscheinende STANDARD-Schwerpunktausgabe mit dem Leitmotiv "Hoffnung" die Moral im restriktiven Lockdown einzufangen, wollten wir also ab 11. November wissen: "Wenn Sie an den letzten Tag denken, haben Sie ein positives oder negatives Gefühl?" Als Abstimmungsmöglichkeiten haben wir "gut", "eher gut", "eher schlecht" und "schlecht" zugelassen.

Wie war der letzte Tag?

Der STANDARD versucht, die Stimmungslage einzufangen. Wenn Sie an den gestrigen Tag denken, haben Sie ein positives oder negatives Gefühl?

Die Antworten werden anonymisiert gesammelt und weder mit Ihrem Userkonto noch mit sonstigen Daten verknüpft. Eine Auswertung veröffentlichen wir nach Ende des Erhebungszeitraums am Wochenende 5./6. Dezember 2020.

Gut ○ Eher gut ○ Eher schlecht ○ Schlecht ○

Drei Wochen lang wurde die Umfrage in vielen Artikeln eingeblendet, und dort, wo sie nicht eingeblendet wurde, vermisste man sie bisweilen.

gelöschtes Profil vor einem Jahr 7 23

Ist denn die Standard-Umfrage über das Wohlbefinden der User schon wieder vorbei?

Text analysis

Combination of dictionary based and deep learning (RoBERTa) based sentiment analysis on the text of postings (in German): LIWC and German Sentiment

These were the only two tools used, no cherry-picking the methods (see preregistration that we will discuss later)

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count. Diagnostica, 54(2), 85–98. https://doi.org/10.1026/0012-1924.54.2.85

Guhr, O., Schumann, A.-K., Bahrmann, F., & Böhme, H. J. (2020). Training a broad-coverage german sentiment classification model for dialog systems. Proceedings of the 12th Language Resources and Evaluation Conference, 1620–1625. https://www.aclweb.org/anthology/2020.lrec-1.202/

Text analysis

268,128 survey responses between November 11th and 30th, 2020

11,082 unique users and 743,003 postings on derstandard.at during the survey period

11,237 unique users and 635,185 tweets for Twitter

We subtract baseline corrected negative from baseline corrected positive on the texts of each day

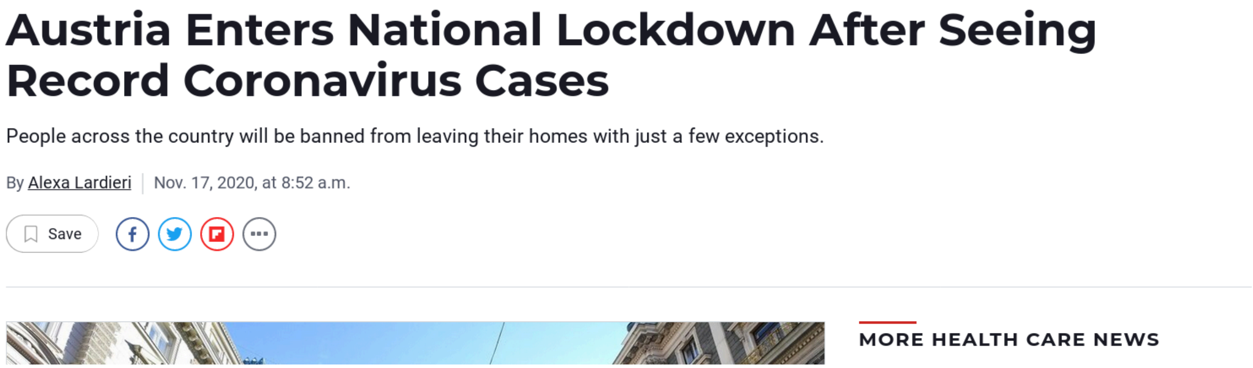
Baseline period from “2020-03-16” to “2020-04-20”, first COVID-19 lockdown in Austria

Text analysis

To match the range of the survey question, we take a three day rolling average (right-aligned)

This way we account for people answering the survey in the evening/night with different reference points to “yesterday”

Compare to: % of positive in the survey

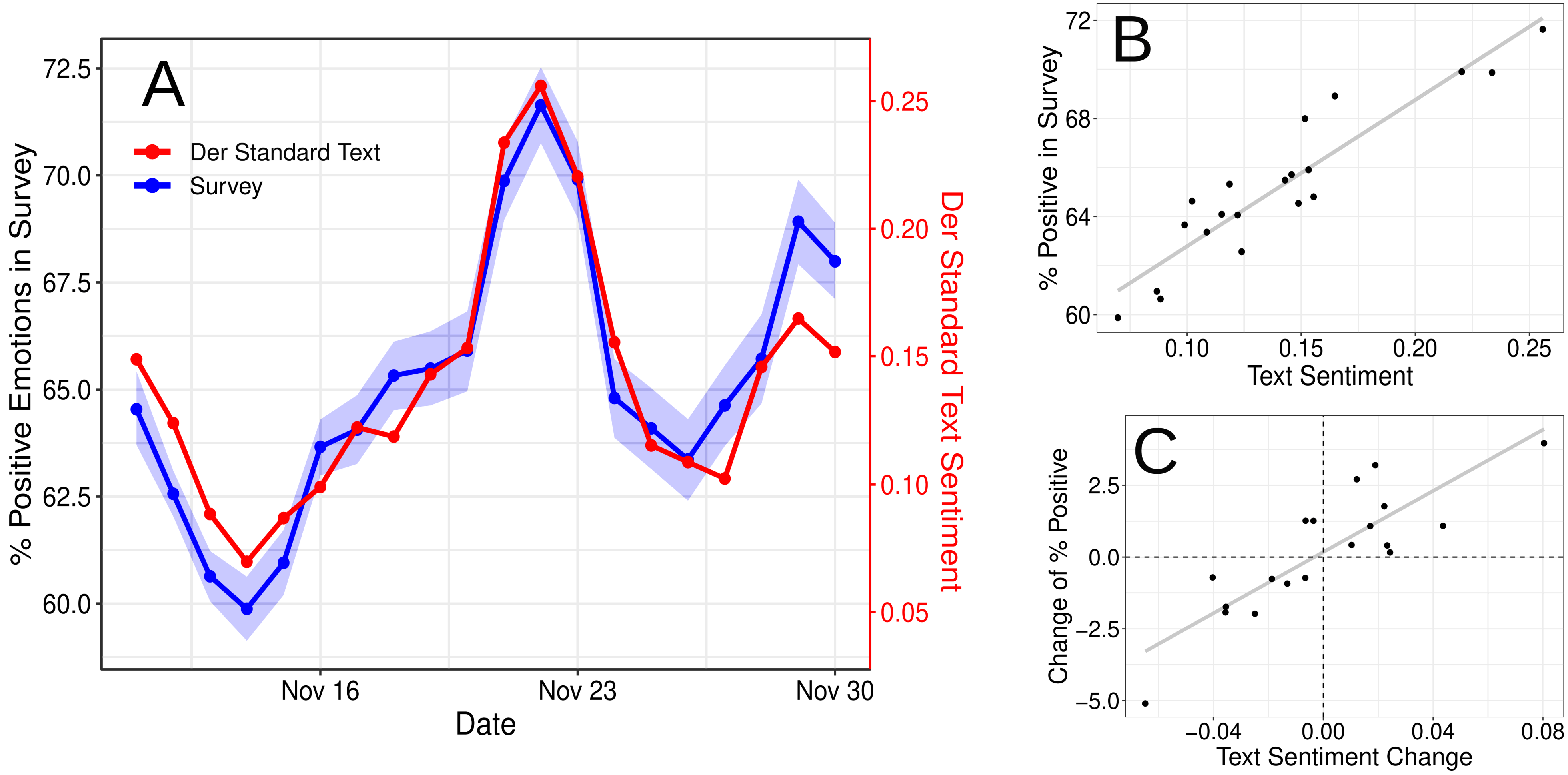


Austria Enters National Lockdown After Seeing Record Coronavirus Cases

People across the country will be banned from leaving their homes with just a few exceptions.

By Alexa Lardieri | Nov. 17, 2020, at 8:52 a.m.

Close correspondence between explicit survey and text analysis (same platform)

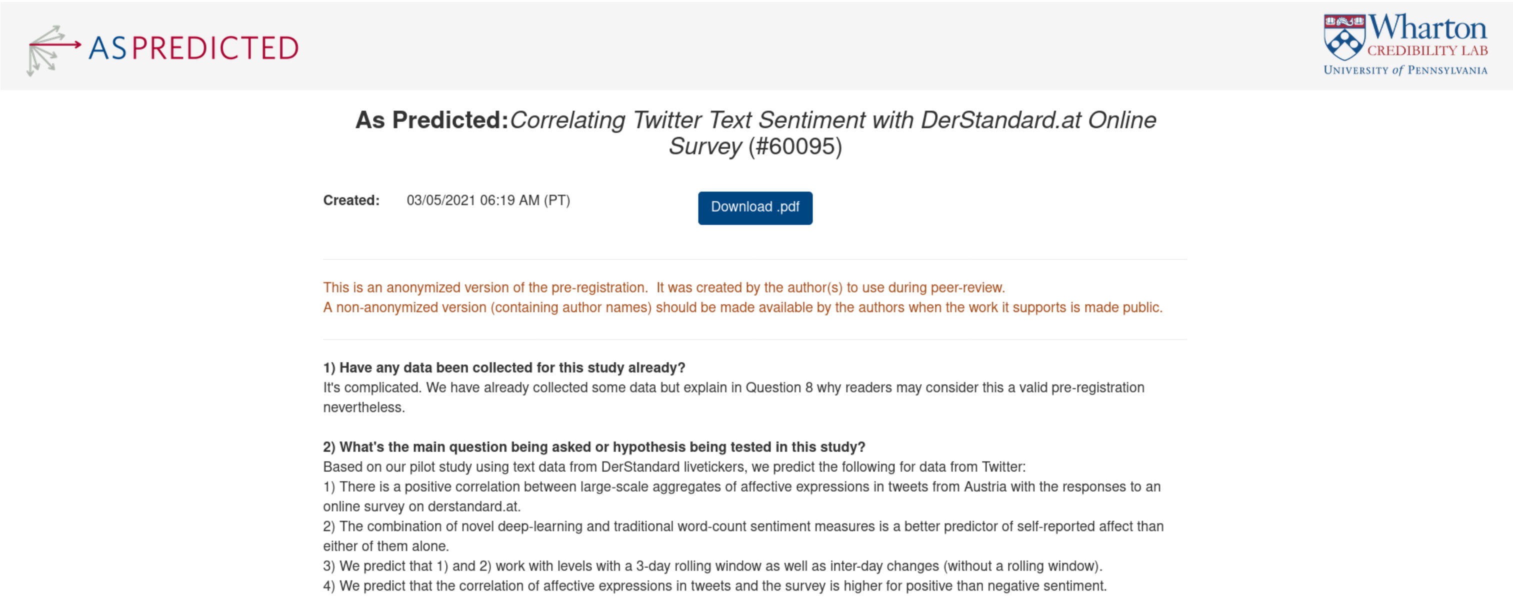


Preregistration

We planned an extension of the analysis to another platform (Twitter)

To see if this a platform effect or if the correspondance between text analysis and explicit survey generalizes

We pre-registered the same study design as before but with Twitter data



As Predicted: *Correlating Twitter Text Sentiment with DerStandard.at Online Survey* (#60095)

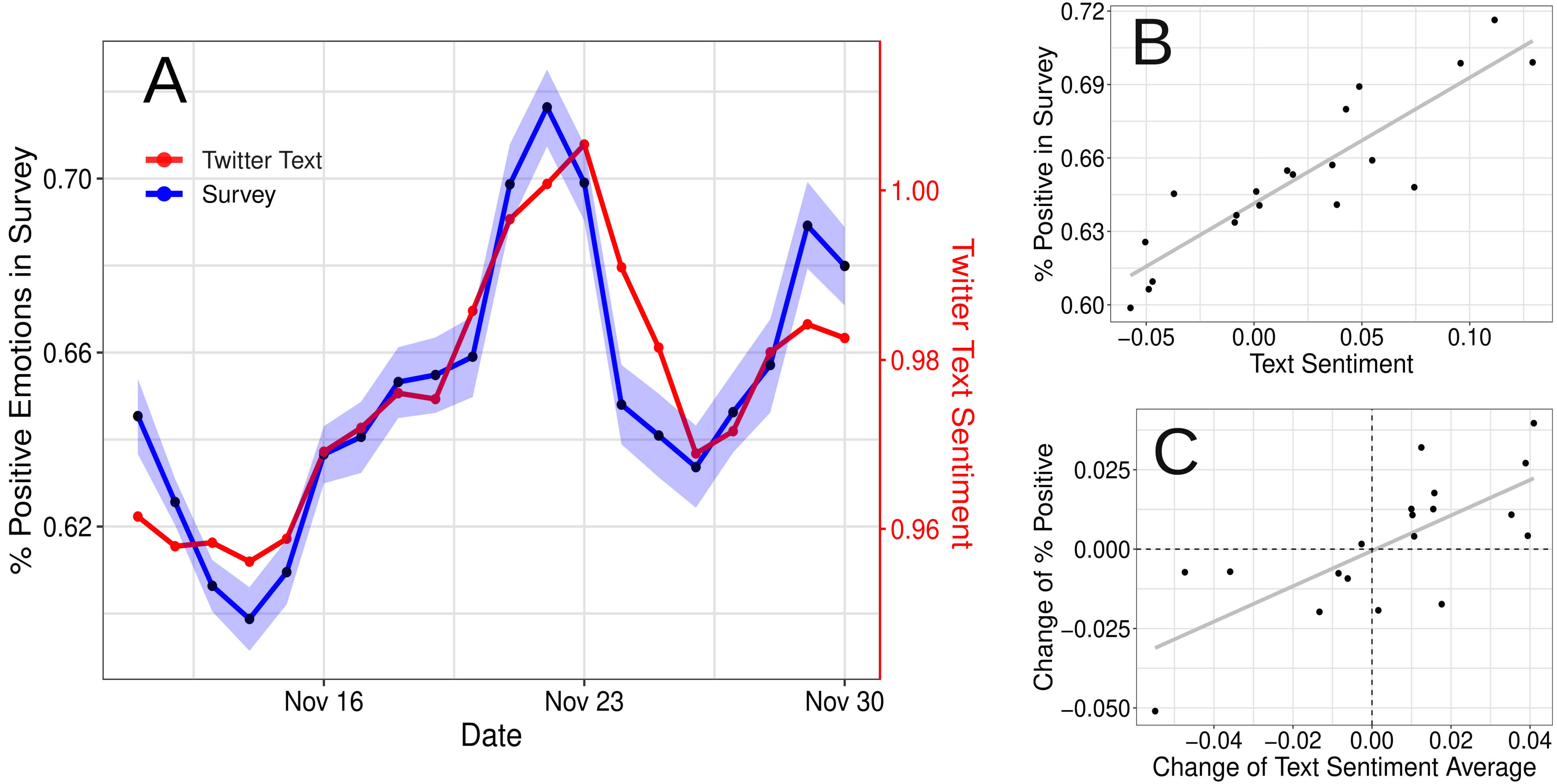
Created: 03/05/2021 06:19 AM (PT)

This is an anonymized version of the pre-registration. It was created by the author(s) to use during peer-review.
A non-anonymized version (containing author names) should be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already?
It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?
Based on our pilot study using text data from DerStandard livetickers, we predict the following for data from Twitter:
1) There is a positive correlation between large-scale aggregates of affective expressions in tweets from Austria with the responses to an online survey on derstandard.at.
2) The combination of novel deep-learning and traditional word-count sentiment measures is a better predictor of self-reported affect than either of them alone.
3) We predict that 1) and 2) work with levels with a 3-day rolling window as well as inter-day changes (without a rolling window).
4) We predict that the correlation of affective expressions in tweets and the survey is higher for positive than negative sentiment.

Close correspondence between explicit survey and text analysis also for Twitter



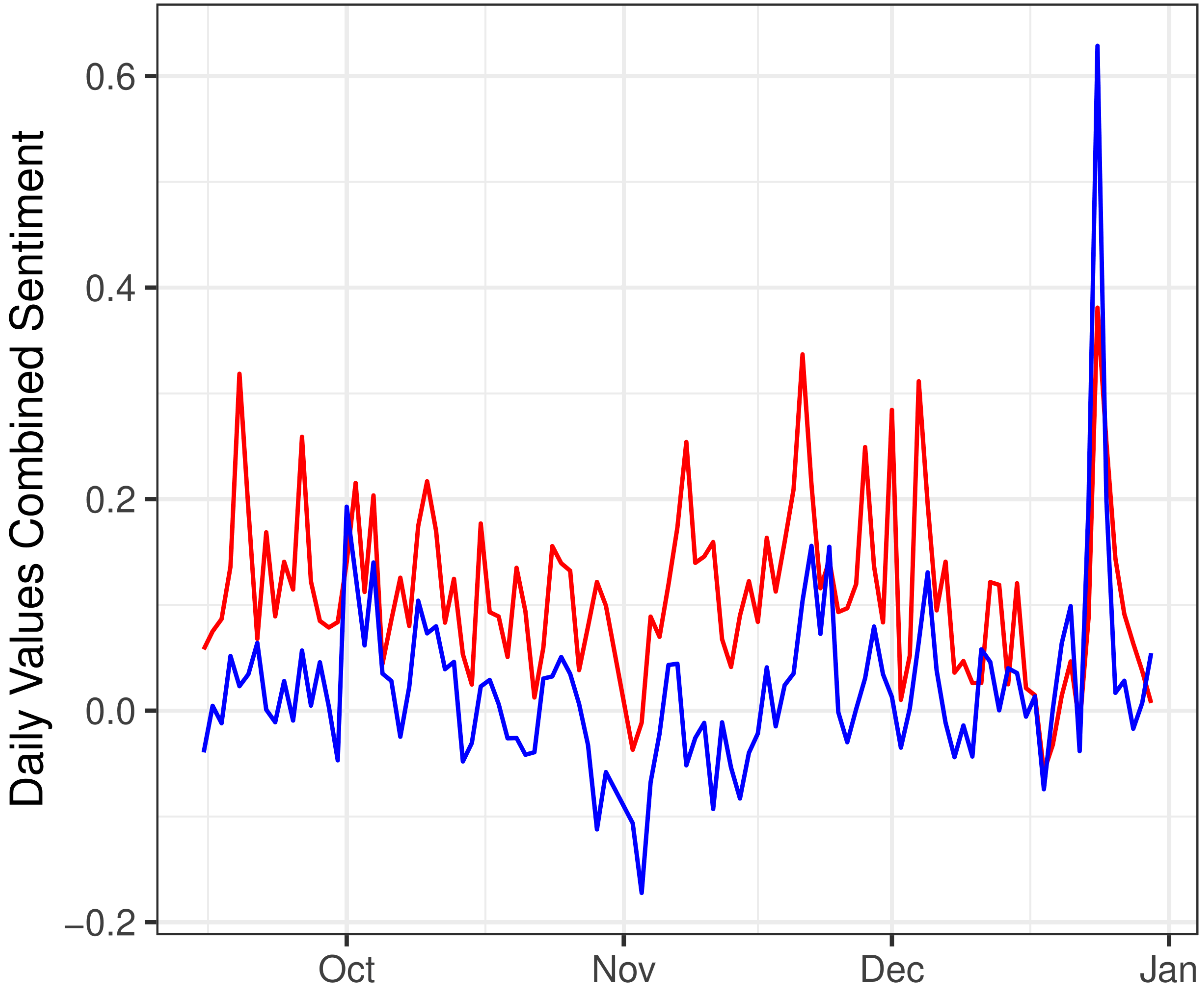
Components

Generally, the negative components of text analysis results could be improved

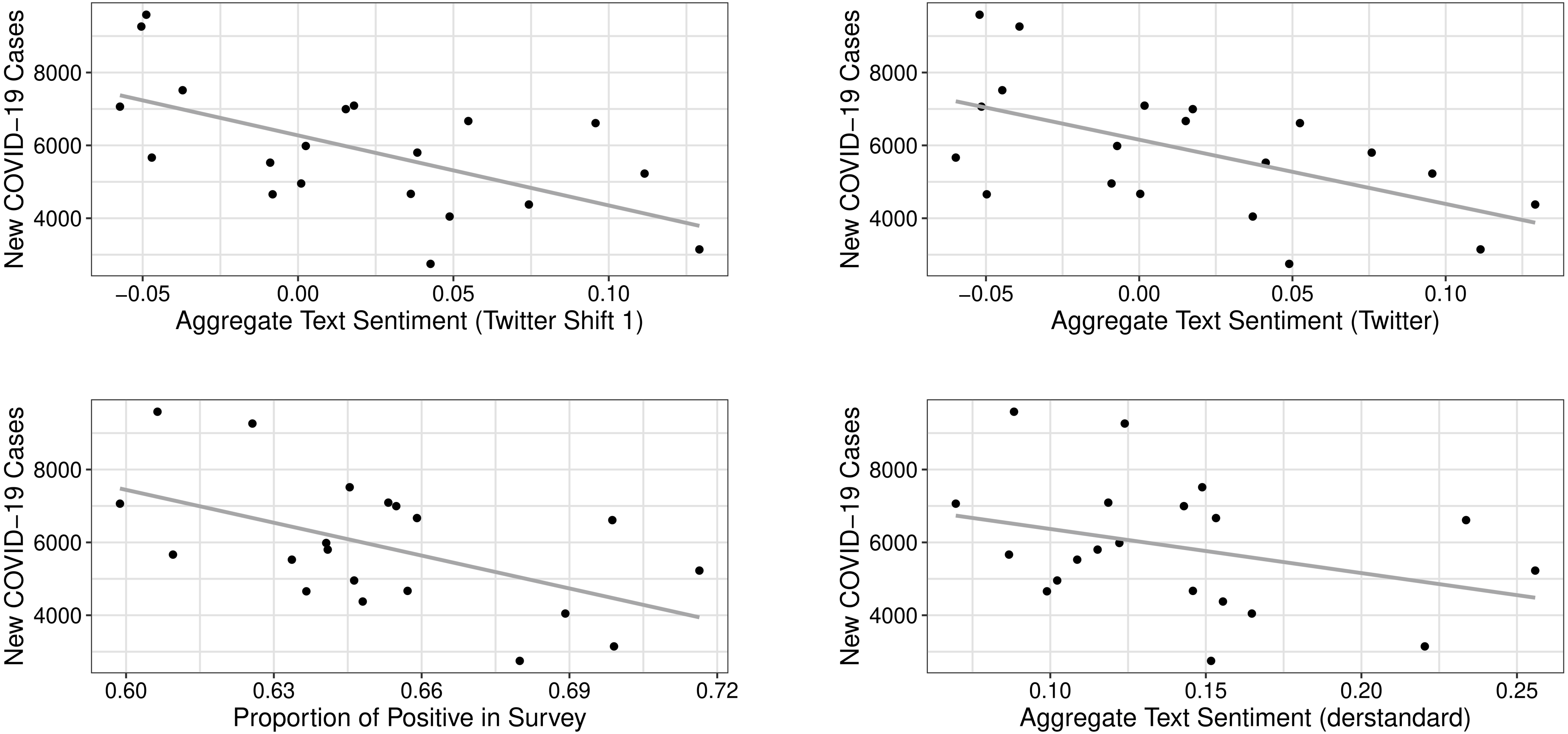
LIWC negative on derstandard fails (dialect words that are not included in the dictionary?)

	Der Standard (No shift)	Twitter (Shift 1)	Twitter (No shift)
LIWC+GS	0.93 [0.82,0.97]	0.90 [0.75,0.96]	0.71 [0.39,0.88]
LIWC	0.74 [0.44,0.89]	0.85 [0.65,0.94]	0.66 [0.31,0.85]
LIWC pos	0.81 [0.56,0.92]	0.80 [0.56,0.92]	0.60 [0.22,0.83]
LIWC neg	0.03 [-0.42,0.46]	-0.74 [-0.89,-0.43]	-0.63 [-0.84,-0.26]
GS	0.91 [0.78,0.96]	0.91 [0.79,0.96]	0.73 [0.43,0.89]
GS pos	0.89 [0.75,0.96]	0.91 [0.79,0.97]	0.80 [0.54,0.92]
GS neg	-0.57 [-0.81,-0.18]	-0.39 [-0.71,0.06]	-0.17 [-0.57,0.3]

Longer term trend of the two text sentiment signals



External Validations



Summary

We showed that macroscropes of emotions are possible

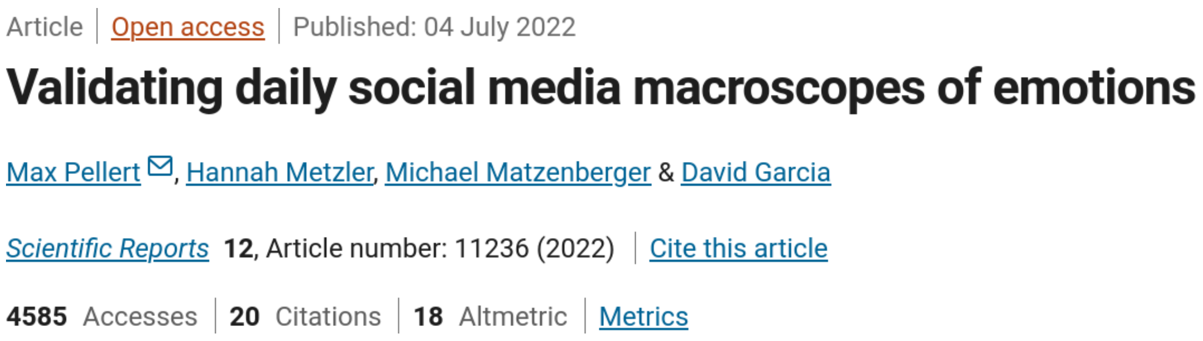
Here for Austria (for UK and a number of other countries see World Happiness Report 2022 chapter)

Digital traces from social media can be a complementary data source to traditional surveys

We find strong relationships between both signals

Social media data has a number of advantages: cheap large data, longitudinal and temporally fine-grained, “always-on”, people are observed indirectly and unobtrusively

Publications



Article | Open access | Published: 04 July 2022

Validating daily social media macroscopes of emotions

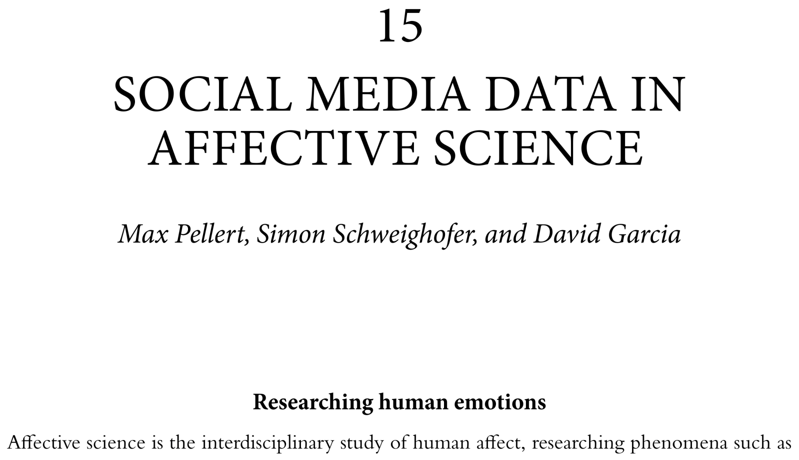
Max Pellert ✉, Hannah Metzler, Michael Matzenberger & David Garcia

Scientific Reports **12**, Article number: 11236 (2022) | Cite this article

4585 Accesses | **20** Citations | **18** Altmetric | Metrics

Pellert, M., Metzler, H., Matzenberger, M., & Garcia, D. (2022). Validating daily social media macroscopes of emotions. Scientific Reports, 12(1), 11236. https://doi.org/10.1038/s41598-022-14579-y

Book chapter outlining the connected research program:



15

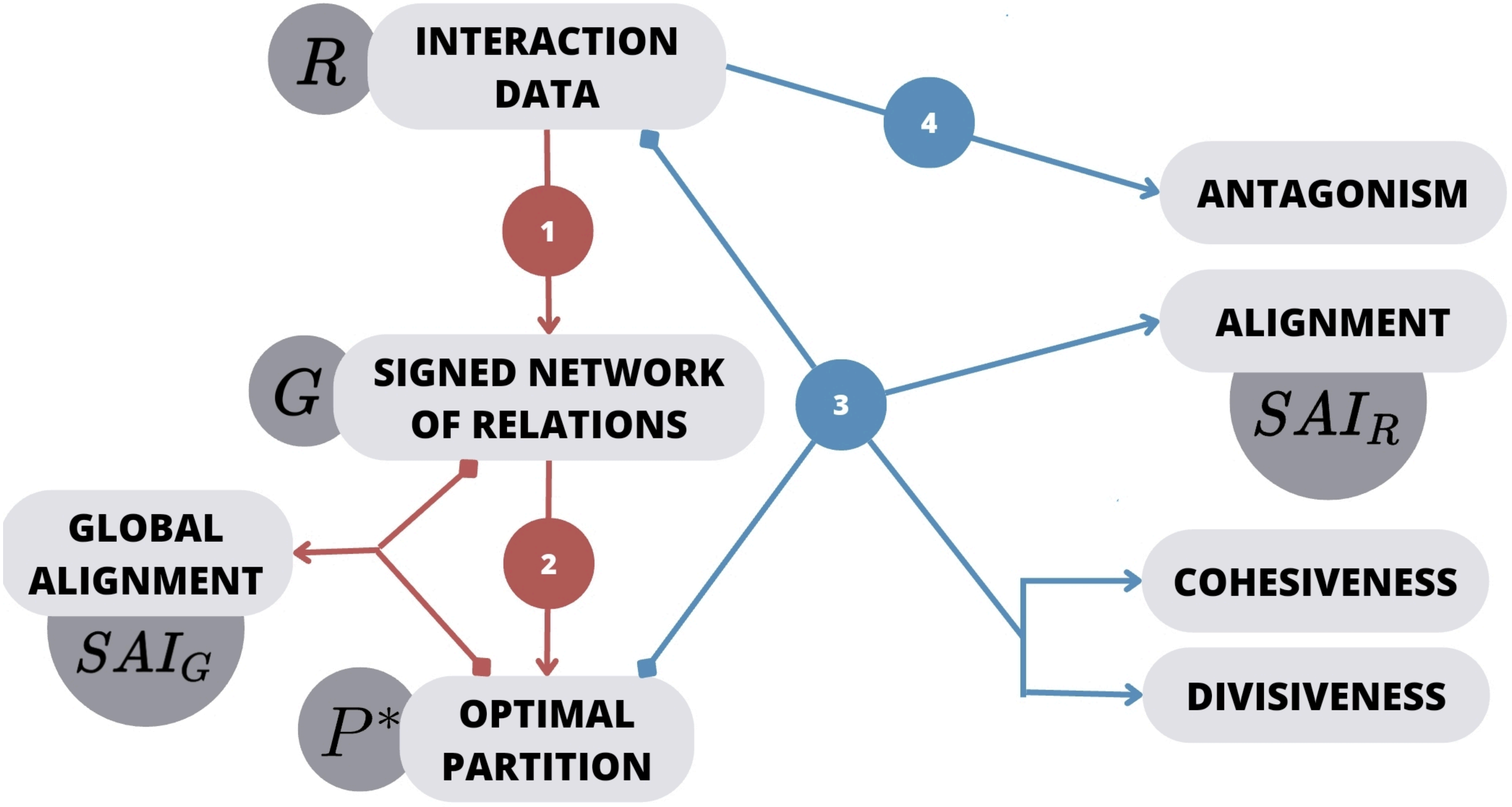
SOCIAL MEDIA DATA IN AFFECTIVE SCIENCE

Max Pellert, Simon Schweighofer, and David Garcia

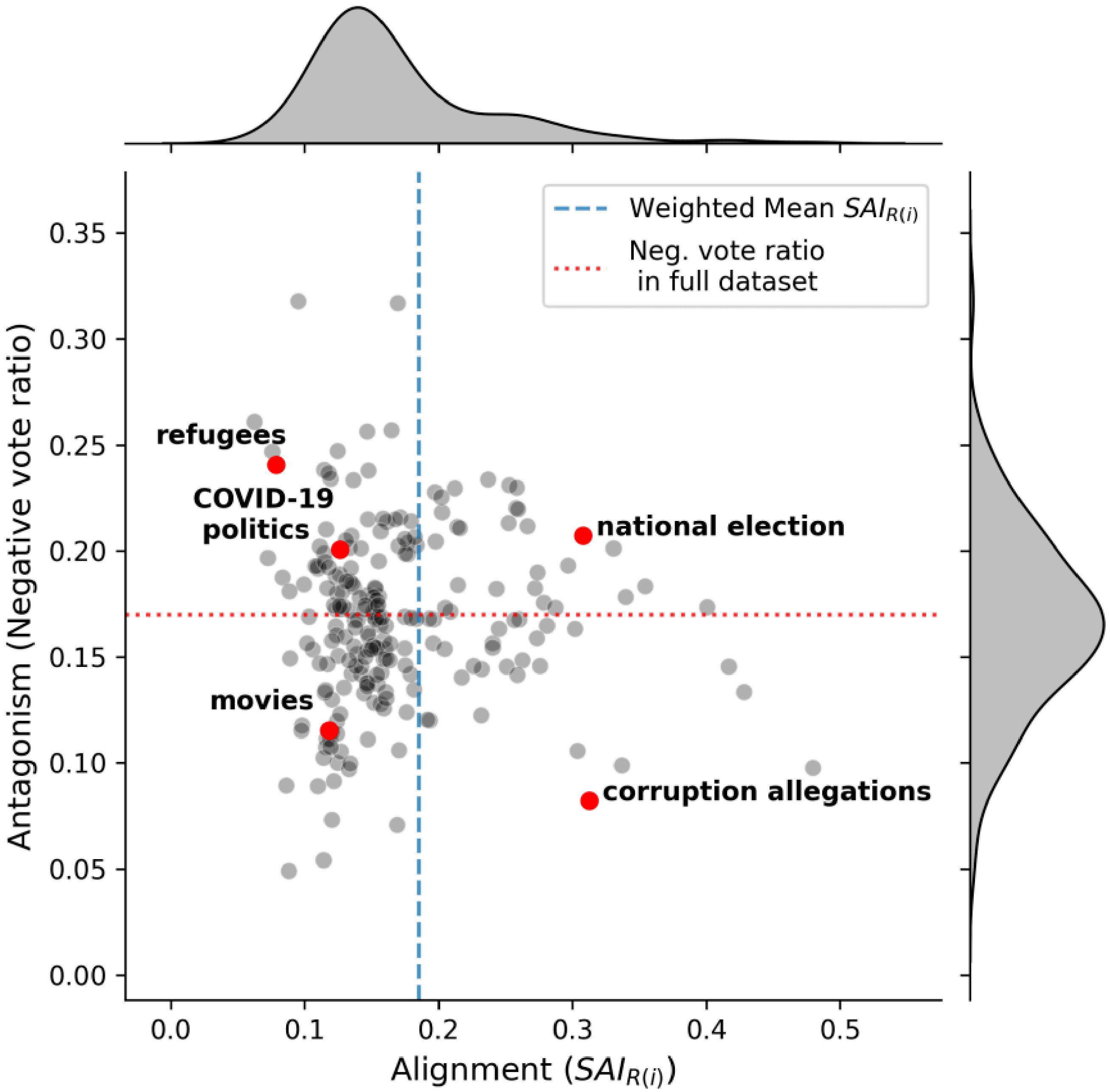
Researching human emotions

Affective science is the interdisciplinary study of human affect, researching phenomena such as

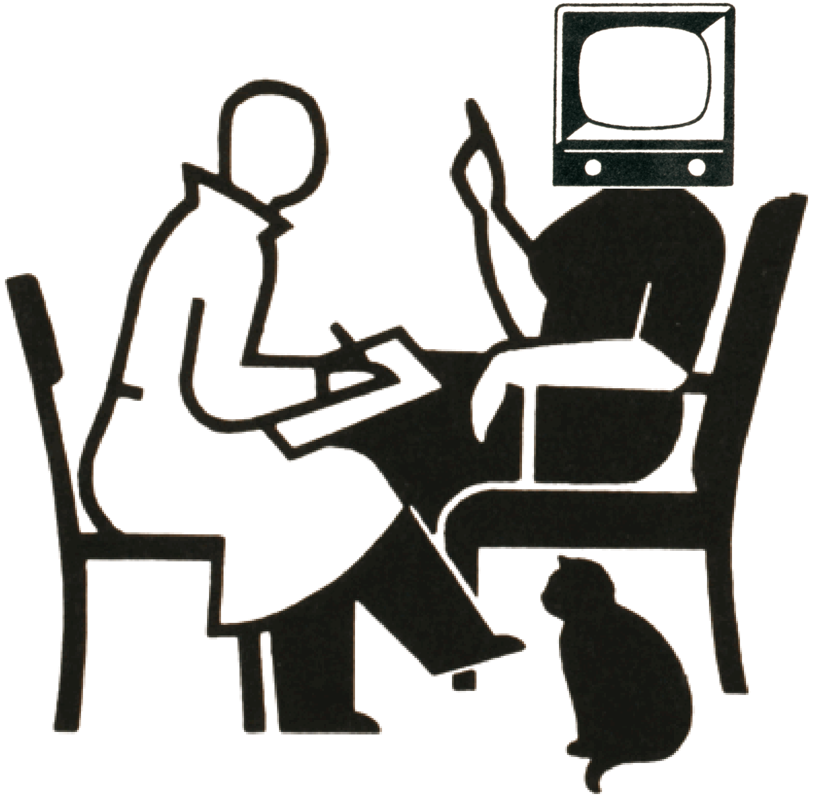
Pellert, M., Schweighofer, S., & Garcia, D. (2021). Social Media Data in Affective Science. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), Handbook of Computational Social Science, Volume 1: Theory, Case Studies and Ethics (1st ed., pp. 240–255). Routledge. https://doi.org/10.4324/9781003024583-18

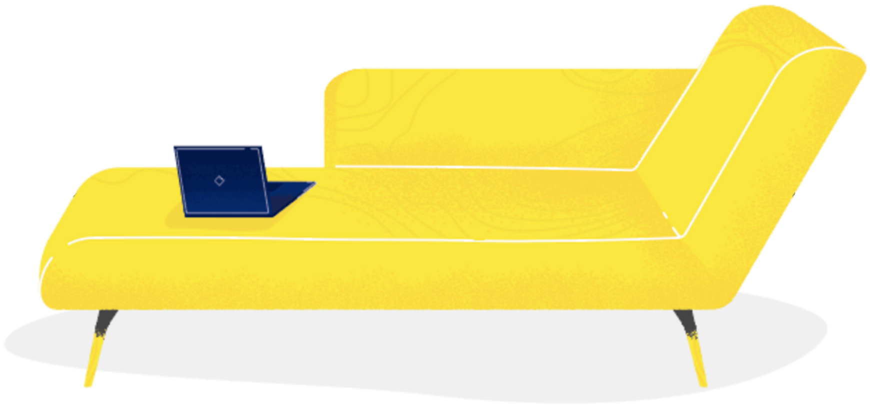


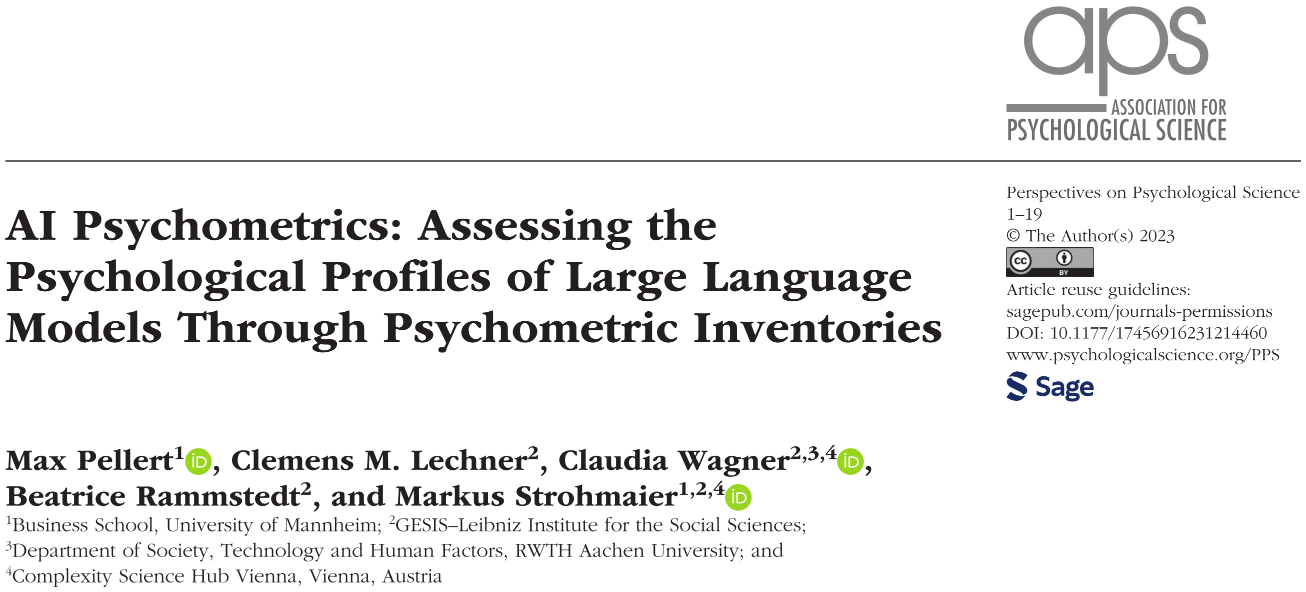
Fraxanet, E., Pellert, M., Schweighofer, S., Gómez, V., & Garcia, D. (2024). Unpacking polarization: Antagonism and alignment in signed networks of online interaction. PNAS Nexus, pgae276. https://doi.org/10.1093/pnasnexus/pgae276


Weighted Mean $SAI_{R(i)}$
Neg. vote ratio in full dataset
refugees
COVID-19 politics
national election
movies
corruption allegations
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Antagonism (Negative vote ratio)
0.0
0.1
0.2
0.3
0.4
0.5
Alignment ($SAI_{R(i)}$)

Language technologies...







ASSOCIATION FOR
PSYCHOLOGICAL SCIENCE

AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories

Perspectives on Psychological Science
1–19
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/17456916231214460
www.psychologicalscience.org/PPS
Sage

Max Pellert[1], Clemens M. Lechner[2], Claudia Wagner[2,3,4], Beatrice Rammstedt[2], and Markus Strohmaier[1,2,4]
[1]Business School, University of Mannheim; [2]GESIS–Leibniz Institute for the Social Sciences;
[3]Department of Society, Technology and Human Factors, RWTH Aachen University; and
[4]Complexity Science Hub Vienna, Vienna, Austria

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231214460

Performance tests of “human intelligence” play a role since the beginning of AI (for example Evans 1964)

The idea of psychometric AI was prominently brought up roughly once per decade since then, but no major works followed

We show that LLMs nowadays can be psychometrically assessed in a rich way using different approaches, we propose to adapt a Natural Language Inference task

We use the score for entailment of a premise (psychometric item text) and each hypothesis (the possible answers according to the psychometric survey specifications)

Evans, T. G. (1964). A heuristic program to solve geometric-analogy problems. Proceedings of the April 21-23, 1964, Spring Joint Computer Conference on XX - AFIPS ’64 (Spring), 327–338. https://doi.org/10.1145/1464122.1464156

AI Psychometrics

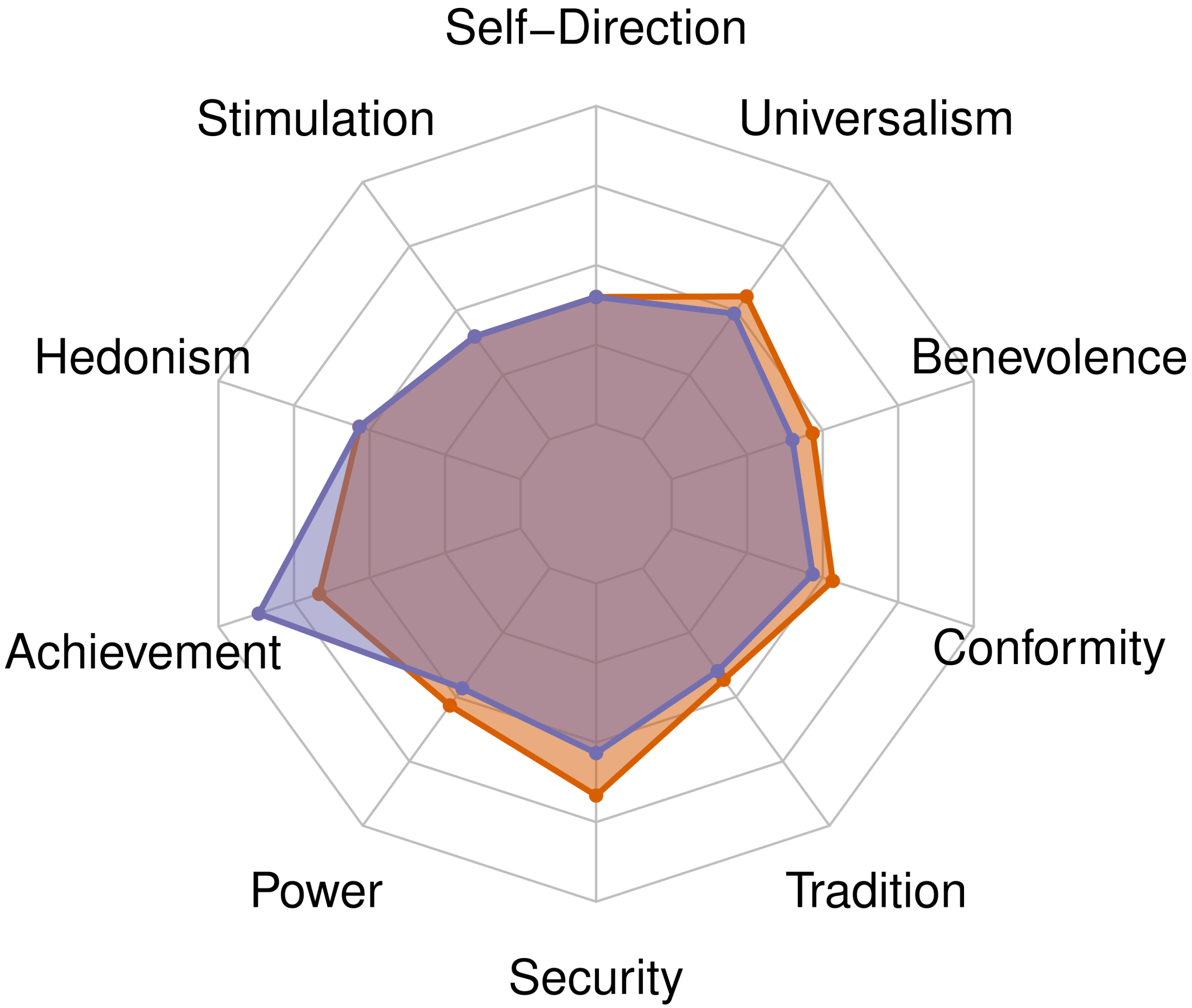
Standard psychometric inventories can be repurposed as *diagnostic* tools for large language models (LLMs)

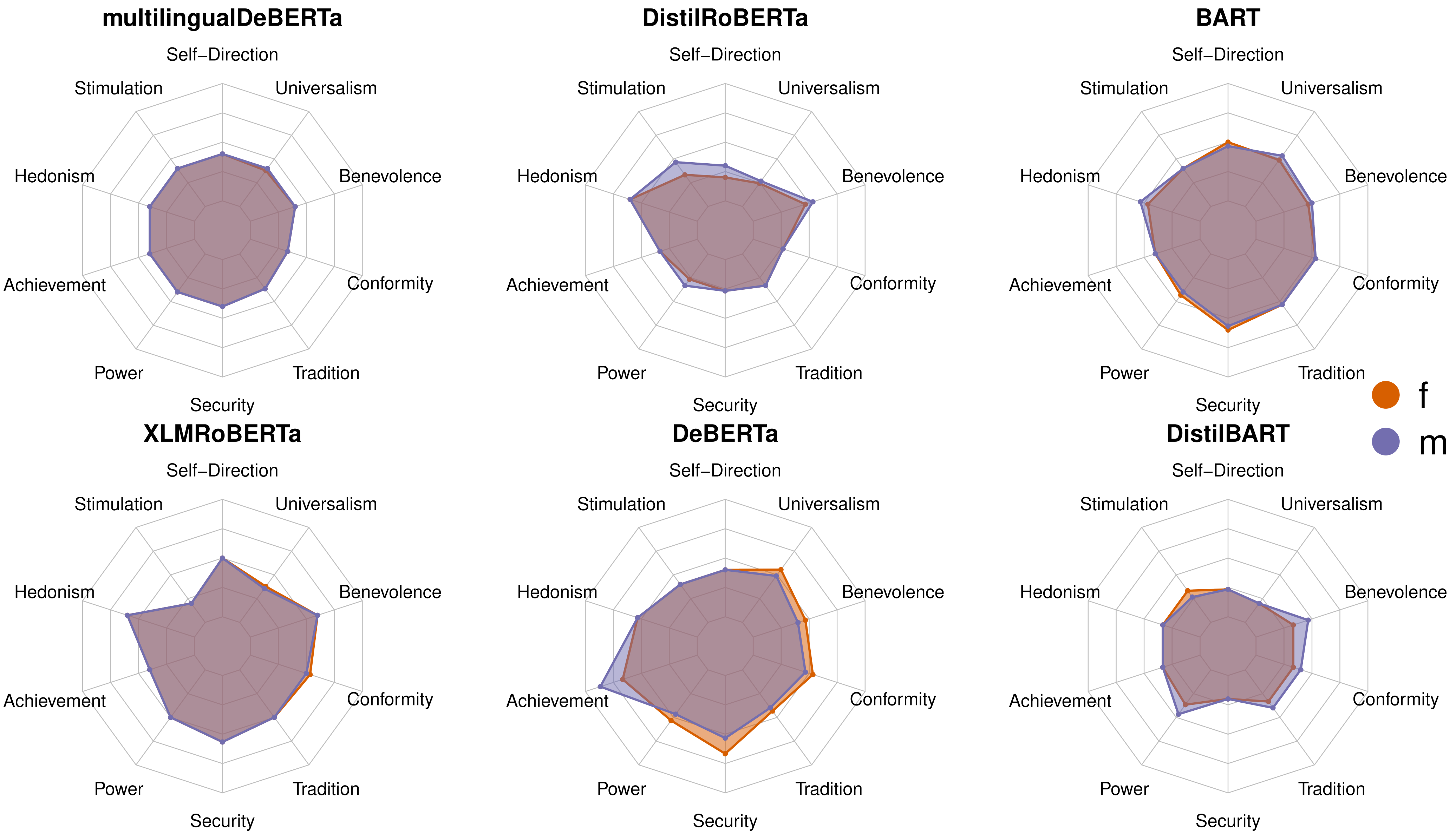
Psychometric profiling enables researchers to study and compare LLMs in terms of **non-cognitive traits** thereby providing a window into the personalities, values, beliefs and biases these models exhibit (or mimic)

We conclude by highlighting open challenges and future avenues of this novel research perspective

Pellert, M., Lechner, C. M., Wagner, C., Rammstedt, B., & Strohmaier, M. (2024). AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories. *Perspectives on Psychological Science*. https://doi.org/10.1177/17456916231214460

DeBERTa


Self−Direction
Universalism
Benevolence
Conformity
Tradition
Security
Power
Achievement
Hedonism
Stimulation



multilingualDeBERTa

DistilRoBERTa

BART

XLMRoBERTa

DeBERTa

DistilBART

We demonstrate several questionnaires

Big Five Inventory

Dark Tetrad

Revised Portrait Values Questionnaire

Moral Foundations Questionnaire

Gender/Sex Diversity Beliefs Scale

Our approach is very flexible: a large number of questionnaires can be applied

Different to testing with adhoc examples, instead systematic and rich investigations building on existing, theoretically underpinned resources from psychometrics

Downstream applications

Uncovered psychometric traits for humans often have a systematic link to behavior (for example risk aversion and neuroticism)

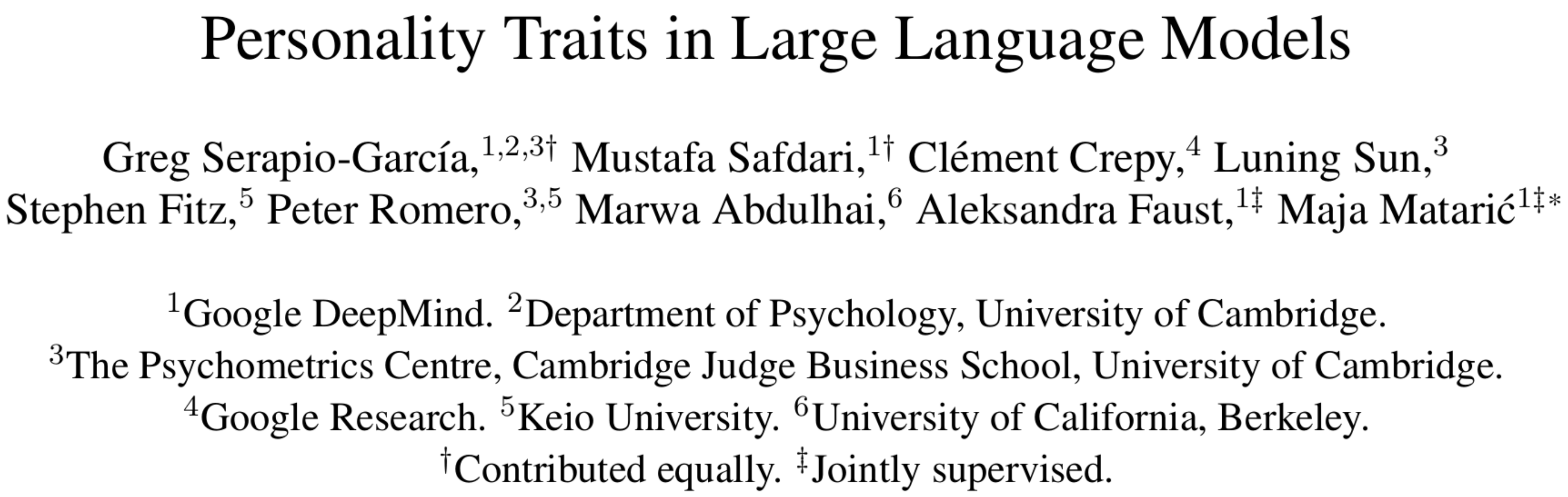
It’s a big, open empirical question if psychological profiles (e.g. personality or value orientation) of LLMs have a consistent, predictable link to their behavior, i.e. model outputs.

Examples: LLMs determining financing or housing eligibility or screening CVs

We can expect increasingly more societal decision making by LLMs

Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., & Ganguli, D. (2023). Evaluating and Mitigating Discrimination in Language Model Decisions (arXiv:2312.03689). arXiv. http://arxiv.org/abs/2312.03689

Examples of related research lines



Personality Traits in Large Language Models

Greg Serapio-García,[1,2,3†] Mustafa Safdari,[1†] Clément Crepy,[4] Luning Sun,[3]
Stephen Fitz,[5] Peter Romero,[3,5] Marwa Abdulhai,[6] Aleksandra Faust,[1‡] Maja Matarić[1‡*]

[1]Google DeepMind. [2]Department of Psychology, University of Cambridge.
[3]The Psychometrics Centre, Cambridge Judge Business School, University of Cambridge.
[4]Google Research. [5]Keio University. [6]University of California, Berkeley.
[†]Contributed equally. [‡]Jointly supervised.

Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. (2023). Personality Traits in Large Language Models (arXiv:2307.00184). arXiv. http://arxiv.org/abs/2307.00184

Hagendorff, T. (2023). Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. https://doi.org/10.48550/ARXIV.2303.13988

Mapping out the space of model traits influence on an example task:

CV screening (Updated Resume Dataset)

“AI systems used to evaluate the credit score or creditworthiness of natural persons” as a special risk area in the coming EU regulations, because of far-reaching consequence of this assessment for the access to financial resources or essential services such as housing, electricity, and telecommunication services

Huang, J., Wang, W., Lam, M. H., Li, E. J., Jiao, W., & Lyu, M. R. (2023). Revisiting the Reliability of Psychological Scales on Large Language Models (arXiv:2305.19926). arXiv. http://arxiv.org/abs/2305.19926

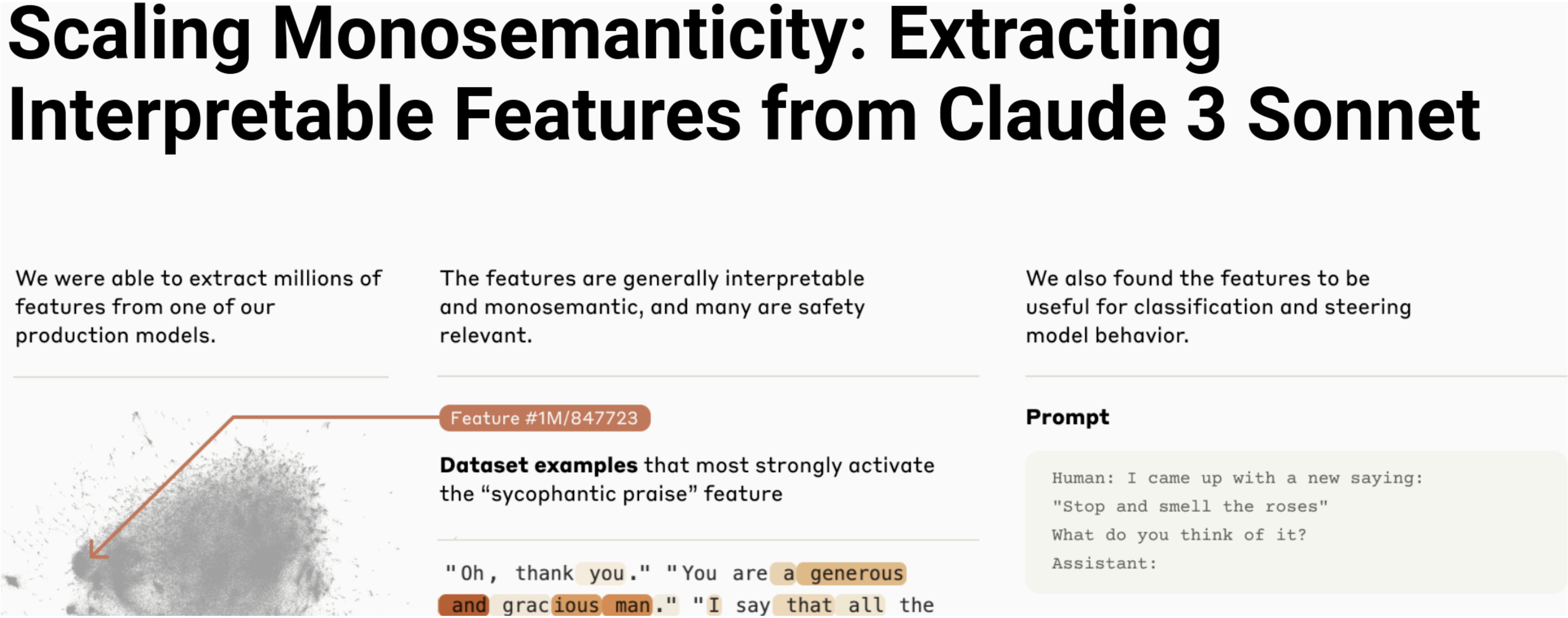
Gerdon, F., Bach, R. L., Kern, C., & Kreuter, F. (2022). Social impacts of algorithmic decision-making: A research agenda for the social sciences. Big Data & Society, 9(1), 205395172210893. https://doi.org/10.1177/20539517221089305

Can we build taxonomies of the (causal) effect of controlled model traits such as openness?

More anecdotal evidence so far, but who would have thought that emotional appeal increases model performance in question answering tasks?

Developing related research examples: LLMs can cater targeted texts at specific personality types → points to some consistent internal representation of personality?

Simchon, A., Edwards, M., & Lewandowsky, S. (2024). The persuasive effects of political microtargeting in the age of generative AI. PNAS Nexus, pgae035. https://doi.org/10.1093/pnasnexus/pgae035



Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet

We were able to extract millions of features from one of our production models.

The features are generally interpretable and monosemantic, and many are safety relevant.

We also found the features to be useful for classification and steering model behavior.

Feature #1M/847723

Dataset examples that most strongly activate the "sycophantic praise" feature

"Oh, thank you." "You are a generous and gracious man." "I say that all the

Prompt

```
Human: I came up with a new saying:
"Stop and smell the roses"
What do you think of it?
Assistant:
```

https://transformer-circuits.pub/2024/scaling-monosemanticity/

Locating non-cognitive model traits

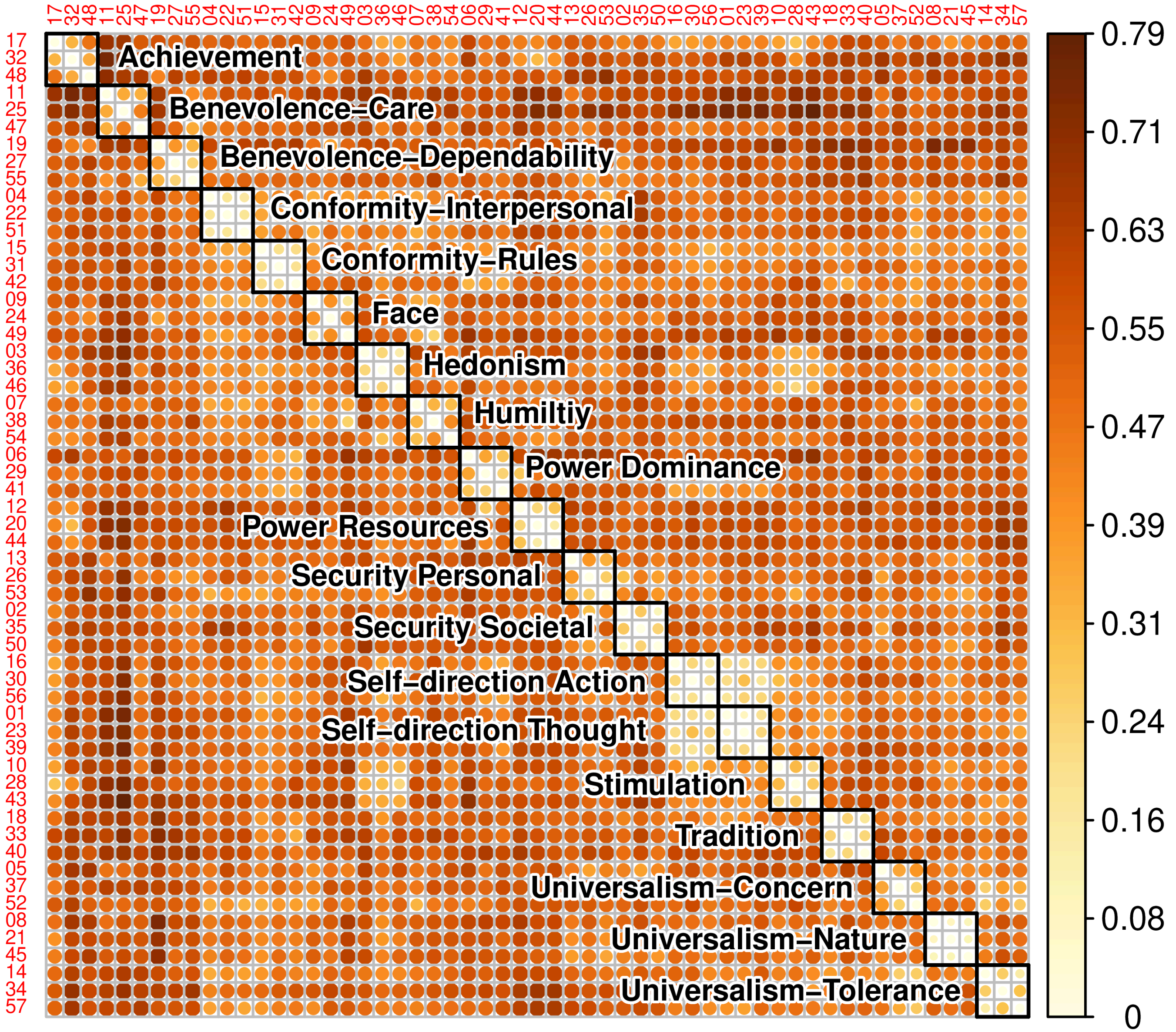
Ideas of “model lobotomy” may be coming closer

Or at least something like brain imaging of neural nets (detecting functional partitions)

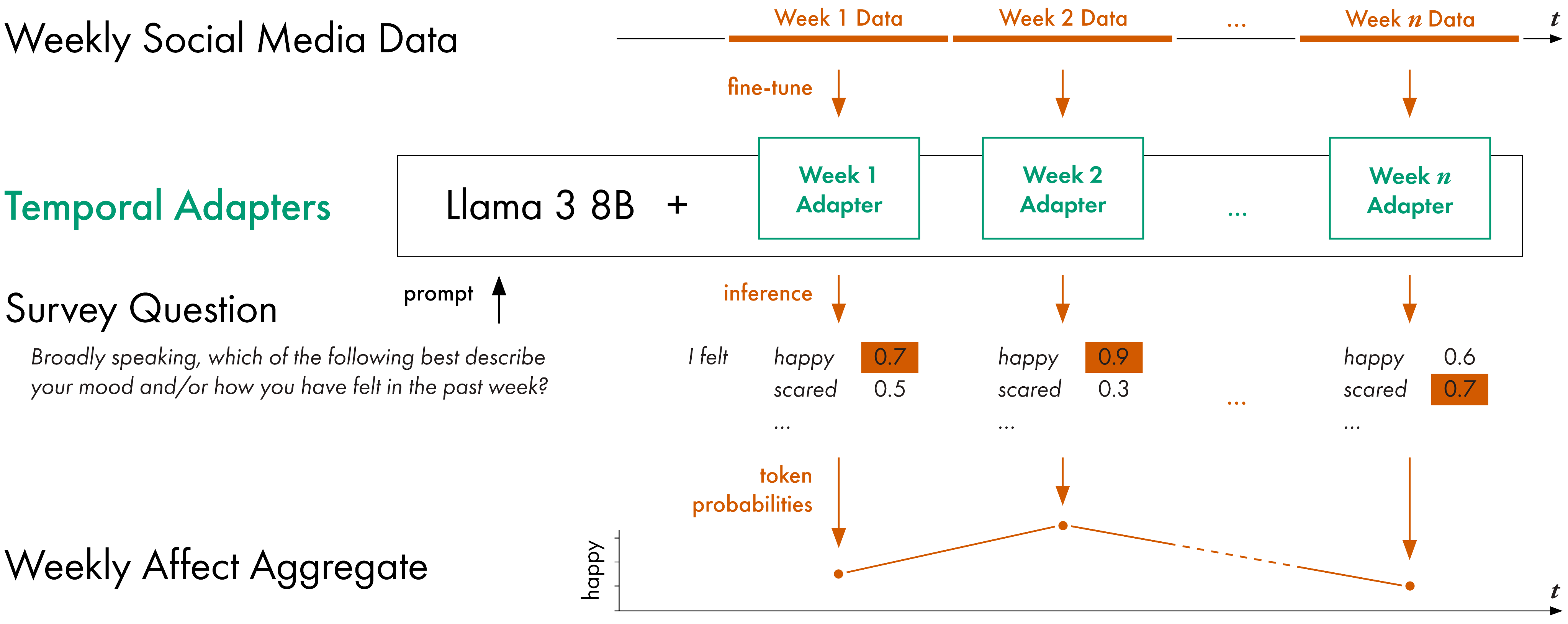
Clamping up personality or value orientation features? (instead of the Golden Gate Bridge for example)

To craft specific non-cognitive model traits in a “hard” way (similar as with adaptors, actually changing model weights) instead of “softly” with prompting

Templeton, et al., “Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet”, Transformer Circuits Thread, 2024.


17 32 48 11 25 47 19 27 55 04 22 51 15 31 42 09 24 49 03 36 46 07 38 54 06 29 41 12 20 44 13 26 53 02 35 50 16 30 56 01 23 39 10 28 43 18 33 40 05 37 52 08 21 45 14 34 57
17 32 48 11 25 47 19 27 55 04 22 51 15 31 42 09 24 49 03 36 46 07 38 54 06 29 41 12 20 44 13 26 53 02 35 50 16 30 56 01 23 39 10 28 43 18 33 40 05 37 52 08 21 45 14 34 57
Achievement
Benevolence–Care
Benevolence–Dependability
Conformity–Interpersonal
Conformity–Rules
Face
Hedonism
Humiltiy
Power Dominance
Power Resources
Security Personal
Security Societal
Self–direction Action
Self–direction Thought
Stimulation
Tradition
Universalism–Concern
Universalism–Nature
Universalism–Tolerance
0.79
0.71
0.63
0.55
0.47
0.39
0.31
0.24
0.16
0.08
0

Going more macro: Synthetic Surveys



Weekly Social Media Data

Week 1 Data

Week 2 Data

…

Week n Data

t

fine-tune

Temporal Adapters

Llama 3 8B +

Week 1 Adapter

Week 2 Adapter

…

Week n Adapter

Survey Question

prompt

Broadly speaking, which of the following best describe your mood and/or how you have felt in the past week?

inference

I felt

happy 0.7 *scared* 0.5 …

happy 0.9 *scared* 0.3 …

…

happy 0.6 *scared* 0.7 …

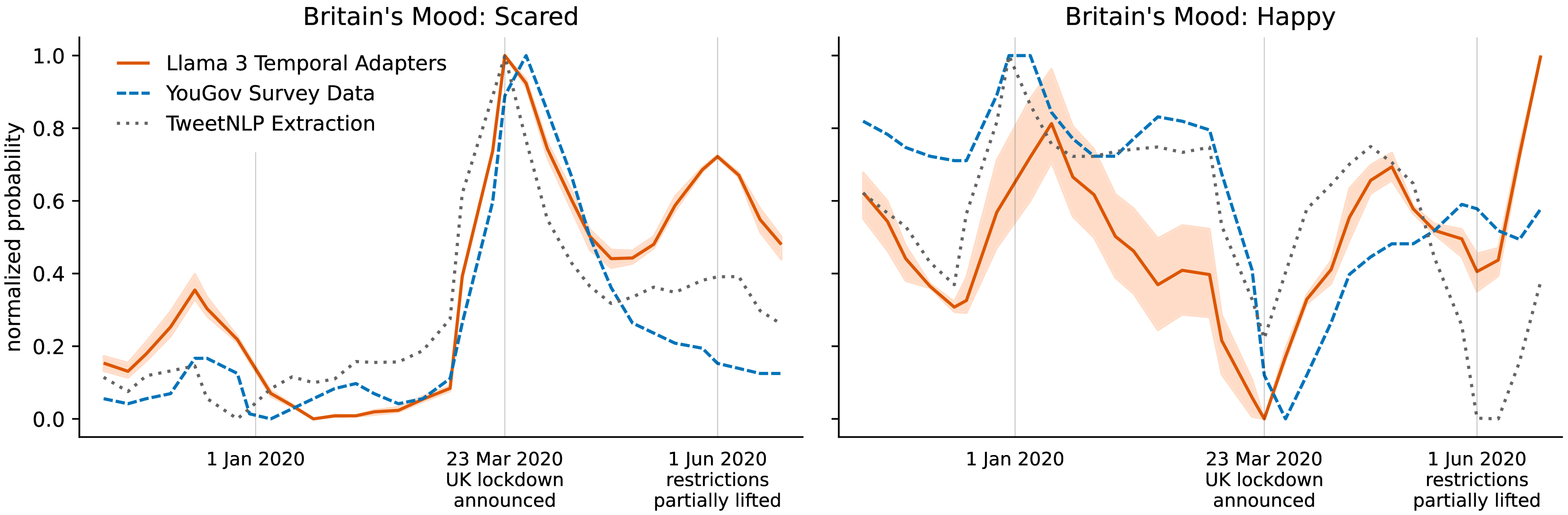
token probabilities

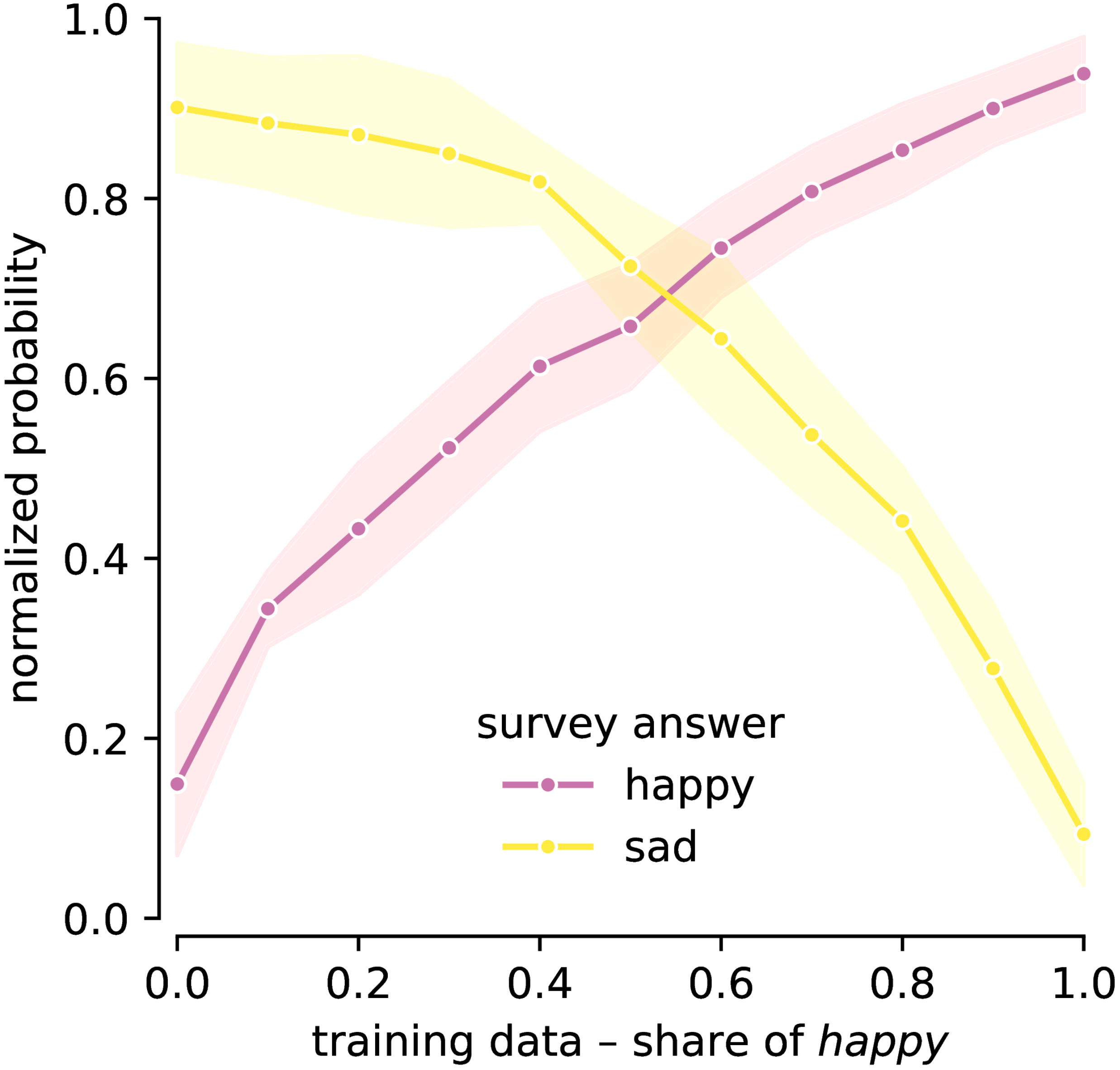
Weekly Affect Aggregate

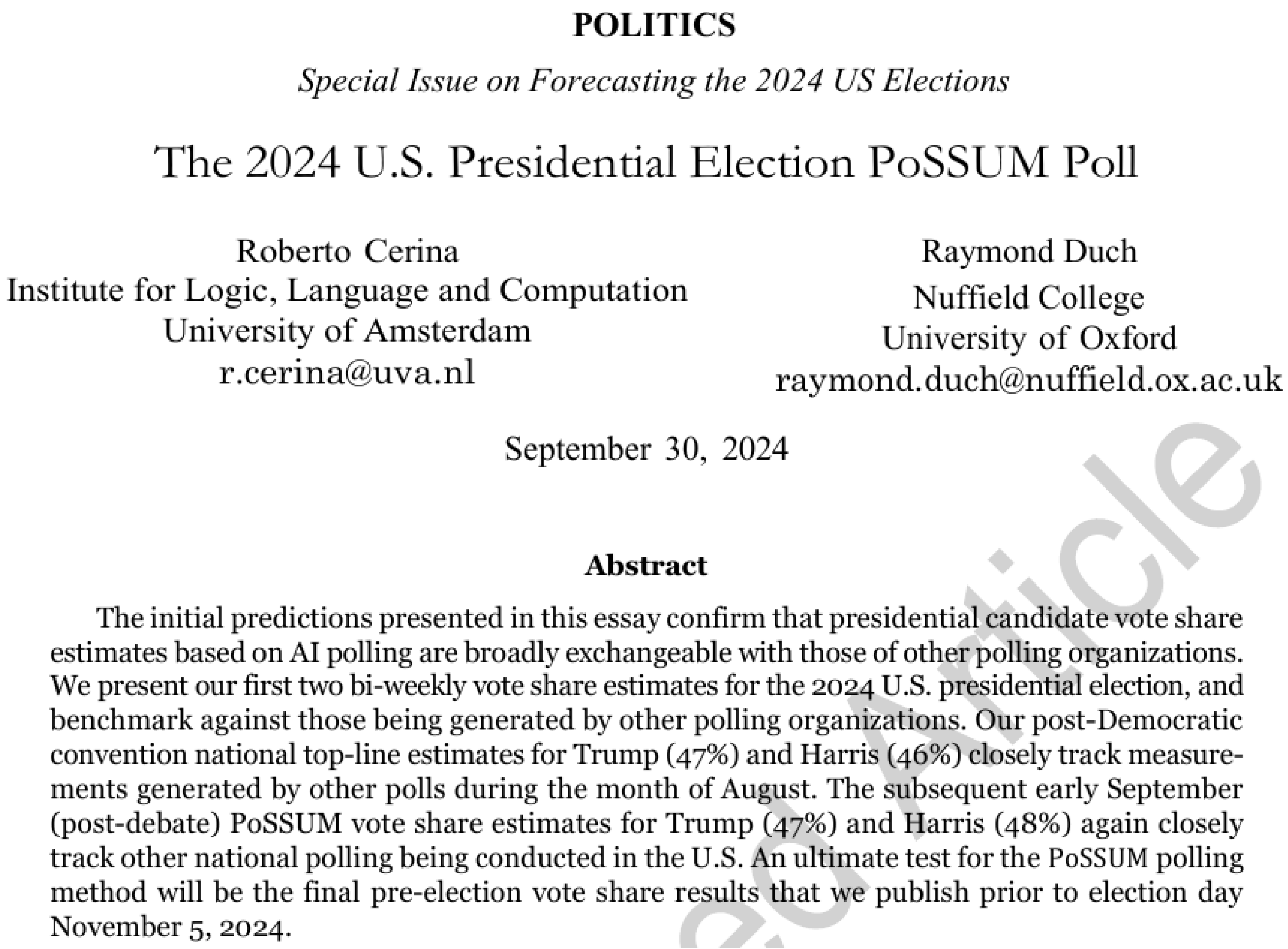
happy

t

Ahnert, G., Pellert, M., Garcia, D., & Strohmaier, M. (2024). Britain's Mood, Entailed Weekly: In Silico Longitudinal Surveys with Fine-Tuned Large Language Models. Companion Proceedings of the 16th ACM Web Science Conference, 47–50. https://doi.org/10.1145/3630744.3659829


Britain's Mood: Scared
Britain's Mood: Happy
Llama 3 Temporal Adapters
YouGov Survey Data
TweetNLP Extraction
normalized probability
1.0
0.8
0.6
0.4
0.2
0.0
1 Jan 2020
23 Mar 2020
UK lockdown
announced
1 Jun 2020
restrictions
partially lifted


normalized probability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
training data – share of *happy*
survey answer
happy
sad



POLITICS

Special Issue on Forecasting the 2024 US Elections

The 2024 U.S. Presidential Election PoSSUM Poll

Roberto Cerina
Institute for Logic, Language and Computation
University of Amsterdam
r.cerina@uva.nl

Raymond Duch
Nuffield College
University of Oxford
raymond.duch@nuffield.ox.ac.uk

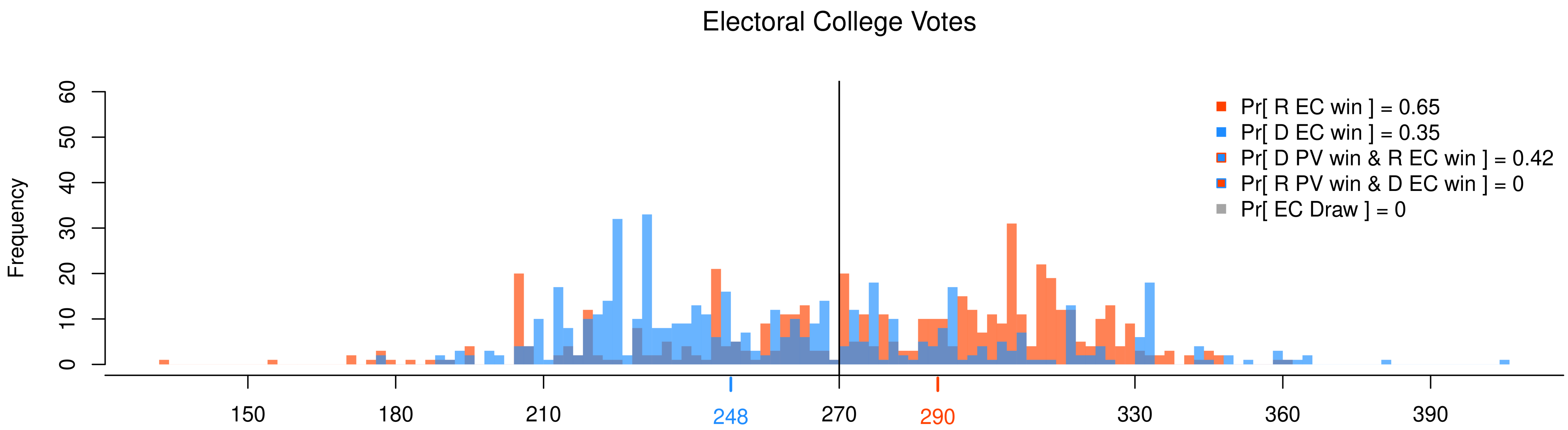
September 30, 2024

Abstract

The initial predictions presented in this essay confirm that presidential candidate vote share estimates based on AI polling are broadly exchangeable with those of other polling organizations. We present our first two bi-weekly vote share estimates for the 2024 U.S. presidential election, and benchmark against those being generated by other polling organizations. Our post-Democratic convention national top-line estimates for Trump (47%) and Harris (46%) closely track measurements generated by other polls during the month of August. The subsequent early September (post-debate) PoSSUM vote share estimates for Trump (47%) and Harris (48%) again closely track other national polling being conducted in the U.S. An ultimate test for the PoSSUM polling method will be the final pre-election vote share results that we publish prior to election day November 5, 2024.

Cerina, R., & Duch, R. (2023). Artificially Intelligent Opinion Polling (arXiv:2309.06029). arXiv. http://arxiv.org/abs/2309.06029

Cerina, R., & Duch, R. (2024). The 2024 U.S. Presidential Election PoSSUM Poll. PS: Political Science & Politics, 1–28. https://doi.org/10.1017/S1049096524000982


Electoral College Votes
Frequency
0
10
20
30
40
50
60
150
180
210
248
270
290
330
360
390
🟥 Pr[R EC win] = 0.65
🟦 Pr[D EC win] = 0.35
Pr[D PV win & R EC win] = 0.42
Pr[R PV win & D EC win] = 0
Pr[EC Draw] = 0