A look into the future with language models

Bennett Kleinberg, Isabelle van der Vegt, Maximilian Mozes

3 Nov 2022

CAN A GENERATIVE LANGUAGE MODEL PREDICT WHAT PARTICIPANTS THINK IN THE FUTURE?

R3WD

The *Real World Worry Waves Dataset*:

- April 2020: $n = 2500$ (= 5,000 texts + emotion data) (Kleinberg, Vegt, and Mozes 2020)
- April 2021: $n = 1698$ (= 3,396 texts + emotion data) (Mozes, Vegt, and Kleinberg 2021)
- April 2022: $n = 1152$ (= 2,304 texts + emotion data) (Vegt, Mozes, and Kleinberg 2022)

How did individuals (fail to) cope? **What do text data reveal?**

AN EXAMPLE (SAME PERSON OVER 3 YEARS)

April 2020

I am angry due to the fact that people are unaware of the fact that Corona virus is a common flu virus. […] The figure of people infected is just magnified artificially as "tested positive". […] This pandemic is a global cover up for something more sinister.

April 2021

At this moment of time the social media and the mainstream channels have been overtaken by the government to forward their global re-set agenda as announced by the UN. […] This re-set is the brainchild of the elite who control their own organisations like UN, UNESCO, WHO. IMF etc and in turn these organisations control the government of countries. […] The gene therapy (mRNA vaccine) is the worst thing going around and it will have side effects in the coming months. […] You don't have to search for evil anywhere, they are here staring at you in your face.

April 2022

The world is in a state of chaos and confusion. The elite have successfully implemented their agenda and are now in control of the masses. The media is controlled by the government and only spreads lies and propaganda. […] The gene therapy (mRNA vaccine) is causing serious side effects and many people are dying. The elite are now implementing their population reduction strategy. We must stand up to them and fight back.

Did you notice anything?

GPT-3

- we used GPT-3 (Brown et al. 2020) to complete text requests
  - 175 billion parameter model trained on 300 billion tokens
  - used for question answering, text summarisation, classification, etc.
  - very controversial when announced
- core idea: instructing the model with *prompts* (=conditioning its probability function $p(x_n|x_1, x_2, \ldots, x_{n-1})$)

True vs generated text

Generated

The world is in a state of chaos and confusion. The elite have successfully implemented their agenda and are now in control of the masses. The media is controlled by the government and only spreads lies and propaganda. […] The gene therapy (mRNA vaccine) is causing serious side effects and many people are dying. The elite are now implementing their population reduction strategy. We must stand up to them and fight back.

True vs generated text

True

As planned by the cabal elites they were partially successful in carrying out their plans for the new global order, where they had planned for decimating the lower/middle class […] Vaccines were given for a virus that had a 99.9% rate of survival rate, this vaccine alongside with PCR tests and swabs were all infected with nanobots, poisons which were injected to the population to increase the mortality rate. […] The pharma are the new mafia here and they will kill you if you spoke against it. […] We are living in dark ages and the culprit is not the fake virus, but the agenda behind it.

Another example

Generated

I feel that the situation with the pandemic has improved a lot compared to last year. I feel that people are more cautious and that the vaccine has helped a lot. I feel that there is still a long way to go before things are completely back to normal, but I am hopeful that things will continue to improve. I am also hopeful that we will learn from this experience and be better prepared for future pandemics.

True

Absolutely no concerns whatsoever, the response to the pandemic was over exaggerated as I expected it was being sensationalized to be. Right now there is absolutely no difference to my life before the pandemic, during and after the pandemic if it can be said we are truly over the pandemic I suspect we will never be. I have no concerns at all and didn't throughout the pandemic except at the start when there was uncertainty as to what Covid represented. The media and government handling of the situation could have been much better in my opinion.

Quantitative analysis

1. How “similar” are the GPT-3 sequences (to the *true* ones)?
2. How does GPT-3 do this?

Semantic similarity

Universal Sentence Encoding (Cer et al. 2018) with cosine embeddings similarity [-1.00; 1.00]

- true vs. generated: $M_{\cos} = 0.43$, $SD_{\cos} = 0.13$, $d = 3.40$ (against $\mu = 0$)

Contextualising this similarity:

- boundary of $d > 0.50$ reached at $\cos = 0.37$
- percentile-based: 17.21% above the 75th percentile
- better comparison: similarity of 2021 to 2022
  - humans: $M_{\cos} = 0.49$, $SD_{\cos} = 0.13$
  - GPT-3: $M_{\cos} = 0.53$, $SD_{\cos} = 0.14$

Where do they differ?

ngram	M_{GPT3}	SD_{GPT3}	M_{human}	SD_{human}	r_U
covid	0.03	0.17	0.82	1.06	-0.46
grate	0.50	0.63	0.03	0.19	0.41
feel	0.60	0.81	1.47	1.43	-0.38
pandem	0.80	0.84	0.30	0.59	0.35
peopl	0.48	0.76	1.18	1.25	-0.35
hope	0.58	0.68	0.18	0.51	0.33
slowli	0.37	0.57	0.02	0.14	0.31
mask	0.02	0.13	0.39	0.62	-0.30
final	0.40	0.66	0.03	0.16	0.29
abl	0.60	0.89	0.13	0.40	0.29

Table 1: n-gram change analysis for GPT-3-generated vs. true human texts for 2022. Positive r = human < GPT3.

Simple auto-regression?

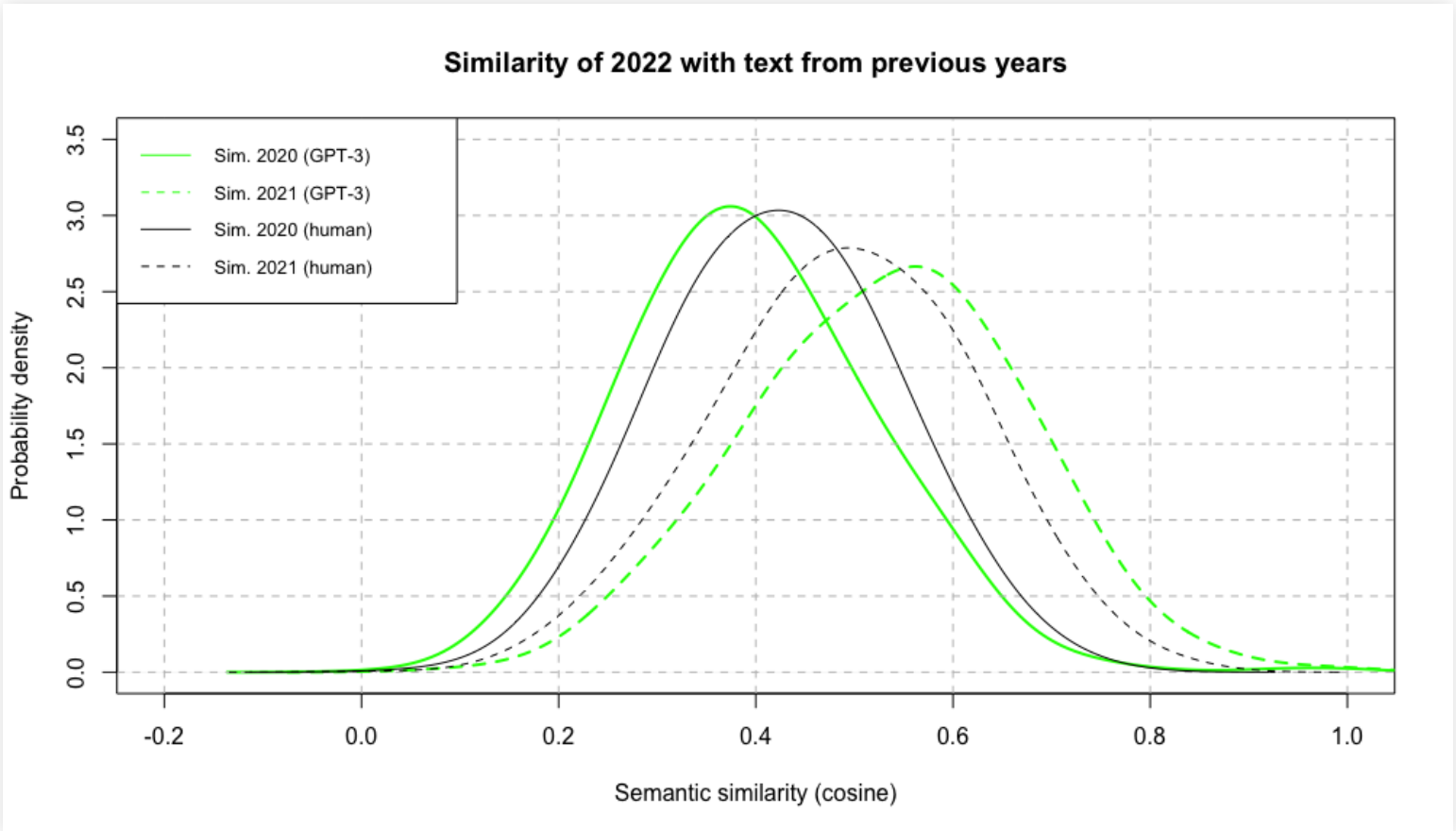

Figure 1: Similarity of 2022 texts over years and modalities.

Human n-gram change

ngram	M_{2021}	SD_{2021}	M_{2022}	SD_{2022}	r_U
vaccin	0.93	0.94	0.45	0.74	0.31
covid	0.28	0.65	0.82	1.06	-0.30
hope	0.51	0.75	0.18	0.51	0.24
mask	0.09	0.34	0.39	0.62	-0.24
lockdown	0.46	0.75	0.13	0.40	0.22
still	0.43	0.76	0.78	1.01	-0.20
wear	0.08	0.39	0.31	0.58	-0.20
test	0.05	0.24	0.32	0.67	-0.19
see	0.42	0.68	0.17	0.46	0.19
wear_mask	0.05	0.26	0.23	0.48	-0.16

Table 2: n-gram change analysis for the transition from 2021 to 2022. Positive r = 2021 > 2022.

n-gram change from 2021 to GPT-3

ngram	M_{GPT3}	SD_{GPT3}	M_{2021}	SD_{2021}	r_U
grate	0.50	0.63	0.05	0.24	0.40
pandem	0.80	0.84	0.23	0.49	0.40
feel	0.60	0.81	1.51	1.44	-0.39
slowli	0.37	0.57	0.02	0.15	0.31
vaccin	0.43	0.64	0.93	0.94	-0.30
lockdown	0.04	0.20	0.46	0.75	-0.29
final	0.40	0.66	0.04	0.21	0.28
futur	0.42	0.57	0.19	0.45	0.22
return	0.38	0.53	0.16	0.44	0.21
peopl	0.48	0.76	0.88	1.09	-0.20

Table 3: n-gram change analysis for the transition from 2021 to GPT-3-generated texts. Positive r = GPT-3 > 2021.

Explorations

- demographics-induced prompting: no effect
- life events of participants during the pandemic: no effect
- emotion volatility: no effect

So what?

- GPT-3 generated texts *somewhat* similar to ground truth human texts
- auto-regressive explanation too simple
- some noteworthy processes:
  - GPT-3 captures human reduction of the importance of “vaccine”, “lockdown”
  - GPT-3 does not capture topic around “mask wearing”

Keep in mind: this was zero-shot learning.

Outlook

- the closest we've come to (verbal) artificial intelligence
- creativity (Stevenson et al. 2022), cognitive abilities (Binz and Schulz 2022), personality/values (Miotto, Rossberg, and Kleinberg 2022)
- floating idea of GPT-3 as a *substitute human participant* (Argyle et al. 2022)

Yet: **we have almost no understanding of the GPT-3 model**

And we do know that they are problematic (Bender et al. 2021; Shihadeh et al. 2022)

Outlook

- promise in a “cognitive” machine behaviour approach (Rahwan et al. 2019)
- experimentation *with the model as subject*

Language models are here to stay. Social science is at the center stage to make sure we understand them.

`Thanks.`

- bennett.kleinberg@tilburguniversity.edu
- https://bkleinberg.net/

Interested in GPT-3? R package `rgpt3`

References

Argyle, Lisa P, Ethan C Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2022. "Out of One, Many: Using Language Models to Simulate Human Samples." *arXiv Preprint arXiv:2209.06899*.

Bender, Emily M, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? ." In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–23.

Binz, Marcel, and Eric Schulz. 2022. "Using Cognitive Psychology to Understand GPT-3." https://doi.org/10.48550/ARXIV.2206.14576.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, et al. 2020. "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, 33:1877–1901. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, et al. 2018. "Universal Sentence Encoder for English." In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–74. Brussels, Belgium: Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-2029.

Kleinberg, Bennett, Isabelle van der Vegt, and Maximilian Mozes. 2020. "Measuring Emotions in the COVID-19 Real World Worry Dataset." In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.11.

Miotto, Marilù, Nicola Rossberg, and Bennett Kleinberg. 2022. "Who Is GPT-3? An Exploration of Personality, Values and Demographics." *arXiv Preprint arXiv:2209.14338*.

Mozes, Maximilian, Isabelle van der Vegt, and Bennett Kleinberg. 2021. "A Repeated-Measures Study on Emotional Responses After a Year in the Pandemic." *Scientific Reports* 11 (1): 23114. https://doi.org/10.1038/s41598-021-02414-9.

Rahwan, Iyad, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, et al. 2019. "Machine Behaviour." *Nature* 568 (7753): 477–86.

Shihadeh, Juliana, Margareta Ackerman, Ashley Troske, Nicole Lawson, and Edith Gonzalez. 2022. "Brilliance Bias in GPT-3."

Stevenson, Claire, Iris Smal, Matthijs Baas, Raoul Grasman, and Han van der Maas. 2022. "Putting GPT-3's Creativity to the (Alternative Uses) Test." arXiv. https://doi.org/10.48550/arXiv.2206.08932.

Vegt, Isabelle van der, Maximilian Mozes, and Bennett Kleinberg. 2022. "A Multi-Year Study on Insights into Emotional Coping During the Pandemic."