All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation

Anonymous Author(s)
email

Abstract

1 Summary

With the recent progress in language models, it is now possible to generate text that is indistinguishable from human written text. Kreps et al. [6] investigate the consequences of this trend on attitudes about foreign policy. They test whether humans are able to decide if a generated news story is credible and if the size of the models influence this decision. Furthermore, they analyze if partisanship and disclaimers affect the perceived credibility of the articles. The experiments show that the larger the models, the harder the distinction (with diminishing returns), that partisanship influences the credibility (where politically-congenial stories were more likely to be rated credible), and that the value of disclaimers is only marginal.

2 Critique & Further Work

In my opinion, the authors do not provide any novel investigations or insights. It is already well established that humans are not able to distinguish texts that are created by sophisticated language models from human written texts [7, 5, 4]. The GPT-3 paper [2] also investigated the relationship between model size and credibility. Furthermore, other researchers already pointed out that partisan media can promote the endorsement of inaccurate beliefs [3] and the limited benefit of disclaimers [1]. So all in all, they mainly endorsed existing research in the area of language models and confirmed that the social science findings also hold for artificially generated text (which is to be expected, when they are not distinguishable by humans). Although they answered their research question (by confirming existing work), I think that there would be much more interesting questions to answer in the context of AI-generated text and media. For instance, they always used a relatively generic disclaimer in the second experiment. Would the result change with another disclaimer which for instance indicates that a text is automatically generated with a high probability? Would providing the probability (e.g., GROVER's [7] probabilities) or giving a reason for that judgement reduce the perceived credibility of the texts? There is existing research to automatically detect generated text [4, 5] and answers to these questions would help to decide how such systems could be deployed in practice.

Another interesting experiment would be the generation of misinformation that is tailored to each participant. For instance, one could deduce based on a questionnaire (e.g., automatically with Machine Learning algorithms) on which topics a person does not have a strong opinion about. This information could then be used to generate texts that influence the opinion of the participant. This would more realistically simulate "hyper-targeted synthetic misinformation" and could show the danger of voter targeting that was for instance employed by Cambridge Analytica. Such an experiment could also answer the question if their results only apply to controversial topics such as immigration or also to uncontentious ones.

Submitted as response essay for the *Sequencing Legal DNA: NLP for Law and Political Economy* course at ETH Zurich. Do not distribute.

Methodologically, I have some doubts on how well the population was represented in their experiments. By using a platform such as Mechanical Turk, it may be possible that certain demographics are under- or overrepresented. Judging from the basic information they give about the sample demographics in the appendix, it looks like people with a college degree (53%-57% of the participants, but only 33% of the US population) and democrats (47%-55% of the participants, but only 32% of the US population) are overrepresented. This may be because a lot of students participate in these surveys, but they do not provide enough data to analyze this issue in detail. As they asked a number of pre-treatment questions, I would be interested in analyses according to these characteristics. While they provide some in the appendix, it would also be interesting to analyze if the results depend on the age, technical literacy, or the socio-economic background of the participants.

I think additional visualizations would help the reader of the paper. For instance, a graph (in addition to the table) about experiment 1 could convey the main message visually. I would also use a bar plot (with 95% confidence intervals for every value) for the third experiment as extrapolating a continuous distribution based on 10 answer possibilities and 200 respondents seems questionable to me.

In the end of the paper, they argue that new AI-based natural language programs such as Blender, T5, or GPT-3 are unlikely to be many orders of magnitude more capable than the GPT-2 model with 1.5 billion parameters they tested. I do not think that such a conclusion is possible based on their experiment. The number of parameters is an oversimplification for the capability of a language model (there are other important aspects like the exact architecture, if the parameters are used for the word embeddings, attention blocks, context window, etc...) and other papers have demonstrated significant performance improvements for larger models [2]. Furthermore, they even talk about "natural language programs" in general and future novel architectures/algorithms might result in models with completely different performance characteristics, so this general extrapolation based on the results of three GPT-2 model sizes cannot be done in my opinion.

References

[1] Kevin Arceneaux, Martin Johnson, and Chad Murphy. “Polarized Political Communication, Oppositional Media Hostility, and Selective Exposure”. In: *The Journal of Politics* 74.1 (2012), pp. 174–186. ISSN: 0022-3816. DOI: 10.1017/s002238161100123x.

[2] Tom Brown et al. “Language Models Are Few-Shot Learners”. en. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[3] R Kelly Garrett, Jacob A Long, and Min Seon Jeong. “From Partisan Media to Misperception: Affective Polarization as Mediator”. In: *Journal of Communication* 69.5 (Oct. 2019), pp. 490–512. ISSN: 0021-9916. DOI: 10.1093/joc/jqz028.

[4] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. “GLTR: Statistical Detection and Visualization of Generated Text”. In: *arXiv:1906.04043 [cs]* (June 2019). arXiv: 1906.04043 [cs].

[5] Daphne Ippolito et al. “Automatic Detection of Generated Text Is Easiest When Humans Are Fooled”. In: *arXiv:1911.00650 [cs]* (May 2020). arXiv: 1911.00650 [cs].

[6] Sarah E. Kreps, Miles McCain, and Miles Brundage. *All the News That’s Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation*. en. SSRN Scholarly Paper ID 3525002. Rochester, NY: Social Science Research Network, Sept. 2020. DOI: 10.2139/ssrn.3525002.

[7] Rowan Zellers et al. “Defending against Neural Fake News”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 9054–9065.