UNITED STATES OF AMERICA
Federal Trade Commission
WASHINGTON, D.C. 20580

Office of Commissioner
Alvaro Martín Bedoya

STATEMENT OF COMMISSIONER ALVARO M. BEDOYA

Regarding Report to Congress on Combatting Online Harms Through Innovation
Matter No. P214501

Federal Trade Commission Open Meeting
June 16, 2022

In the 2021 Appropriations Act, Congress asked the Commission to report on the uses of artificial intelligence to detect or address harmful online content including fake reviews, opioid sales, hate crimes, and election-related disinformation.

There are countless positive uses of machine learning in society today.[1] But Congress asked us a more specific question: Is it a good idea to rely on these tools to identify unlawful and other harmful conduct—in a way that might result in a subsequent law enforcement action?

Today's report offers this answer: Proceed with caution. I agree with that answer. I emphatically agree with it. I will not elaborate on the warnings in this report, because I think they were stated accurately and persuasively—and I'm grateful for staff's work in this regard.

Instead, I will focus on one specific issue: The danger of software trained predominantly on one form of one language to analyze text in another language or even another form of that same language.

Natural Language Processing, or NLP is a branch of artificial intelligence that allows software to analyze natural speech. You may not realize it, but chances are good that you encounter NLP on most days when you run a search online, or when a string of predictive text appears while you're typing.[2]

But like every other technology, NLP has its strengths and weaknesses. One weakness is the fact that most leading NLP programs have been predominantly trained on English. Most Internet users do not speak English.[3]

---

[1] *See* Tejal A. Patel et al., *Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods*, 123 Cancer 114-121 (Jan 1, 2017; first published online Aug. 29, 2016) available at https://acsjournals.onlinelibrary.wiley.com/doi/full/10.1002/cncr.30245; Jeff Grubb, *Google Duplex: A.I. Assistant Calls Local Businesses To Make Appointments*, YouTube (May 8, 2018), https://www.youtube.com/watch?v=D5VN56jQMWM.

[2] *See* Natasha Duarte, et al., *Mixed Messages? The Limits of Automated Social Media Content Analysis*, Center for Democracy & Technology (2017), https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf.

[3] *See* Duarte, *supra* note 2; Su Lin Blodgett and Brendan O'Connor, *Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English* (Presented at Workshop on Fairness, Accountability, and Transparency in Machine Learning, 2017) available at https://arxiv.org/pdf/1707.00061.pdf;

Those that are trained on languages other than English are often trained on languages spoken in countries with the resources to conduct intensive research in this area—languages like French, German, and Mandarin. And many of the data sets that supposedly contain data from these languages are instead filled with garbage—either text from other languages or text that isn't language at all.[4] As a result, there are currently few programs that effectively analyze languages such as Bengali, Hindi, Indonesian, Punjabi, Cebuano, and Swahili—collectively spoken and written by more than a billion people across the world.[5]

It's even more complicated than that, however. Even when you have an NLP program trained on English, research has shown that it can misfire when it is used to analyze English used by specific communities, or by all people in particular contexts.

This is because these models are trained on language taken from very specific contexts, like the comments section of Reddit or biography pages from Wikipedia, and then generalized to all English speakers.[6] This can result in poor performance when analyzing other dialects, casual speech, and slang.

For example, researchers found that one popular NLP tool would categorize African American Vernacular English as Danish—with a 99.9% confidence level.[7] In another study, researchers found that YouTube auto-captioning had a higher error rate for captioning female speakers than male speakers in videos.[8]

The relevance of this research to the report today is obvious. Before using NLP to support high-risk, dangerous decisions, policymakers and law enforcement officials must be aware of its shortcomings, particularly when it comes to processing language. And when presented with the question, "Should we use artificial intelligence to identify unlawful conduct?" —"no" has got to be an option.

Holly Young, *The Digital Language Divide: How does the language you speak shape your experience of the internet?*, The Guardian, http://labs.theguardian.com/digital-language-divide/; *see also* Maarten Sap, et al., *The Risk of Racial Bias in Hate Speech Detection*, Proc. of the 57th Ann. Meeting of the Ass'n for Computational Linguistics 1668 (2019), available at https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf; Thomas Davidson, et al., *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, Proc. of the Third Abusive Language Workshop at the Ann. Meeting for the Ass'n for Computational Linguistics 6 (Aug. 1–2, 2019), available at https://arxiv.org/pdf/1905.12516.pdf.

[4] Julia Kreutzer et al., *Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets*, 10 Transactions of the Association for Computational Linguistics 50 (2022) available at https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00447/109285/Quality-at-a-Glance-An-Audit-of-Web-Crawled

[5] *See* Duarte, *et al*., *supra* note 2.

[6] Anjalie Field et al., *Controlled Analyses of Social Biases in Wikipedia Bios*, WWW '22: Proceedings of the ACM Web Conference 2022 2624 (2022), available at https://arxiv.org/pdf/2101.00078.pdf.

[7] Blodgett & O'Connor *supra* note 3; Rachael Tatman, *Gender and Dialect Bias in YouTube's Automatic Cations*, 1 Proceedings of the First Association for Computational Linguistics Workshop on Ethics in Natural Language Processing 53, 53-59 (2017), available at http://www.aclweb.org/anthology/W/W17/W17-1606.

[8] *See* Blodgett & *O'Connor* supra note 3 at 1-2.