Study reveals strong demand for open-access science
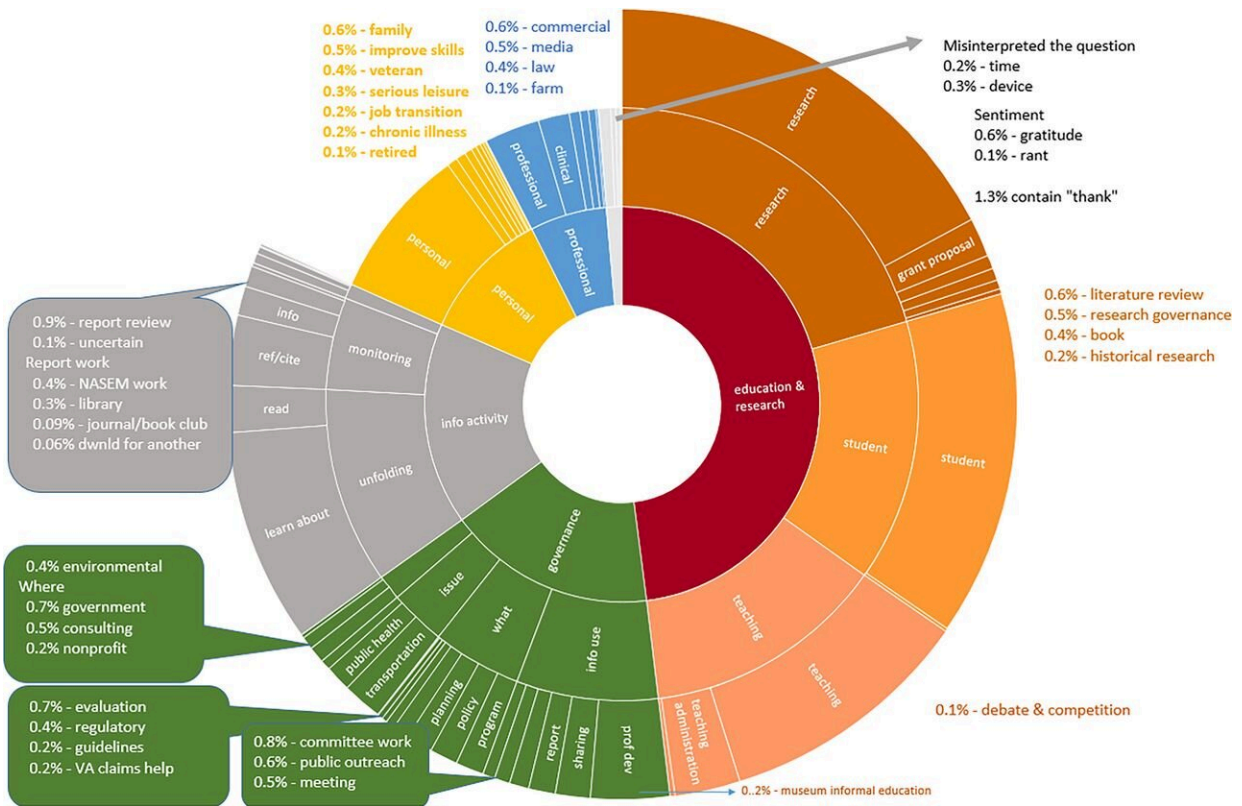
February 24 2022

Fig. 1. How were NASEM reports used? Classification into 64 categories of 1.6 million comments left by US downloaders of NASEM reports between 2011 and 2020. Downloaders were asked how they will use the report. BERT machine learning algorithm was used to classify. Credit: DOI: 10.1073/pnas.2107760119

While a dominant narrative of American life paints a bleak picture of

poorly informed internet partisans duking it out over a landscape denuded of anything resembling truth or reality, a new study from the Georgia Tech School of Public Policy offers a different take while also advancing the use of machine learning in the social sciences and an understanding of the importance of open-access, science-based information to everyday Americans.

The study, published Feb. 23, 2022, in the *Proceedings of the National Academy of Sciences* (*PNAS*), analyzed the reasons for 1.6 million downloads of National Academies of Sciences, Engineering, and Medicine (NASEM) consensus reports, considered among the highest credibility science-based literature.

The resulting analysis, which included U.S. downloads only, is the first to look at who is using such information and why. Professor Diana Hicks, Assistant Professor Omar I. Asensio, and Ph.D. students Matteo Zullo and Ameet Doshi, all of Georgia Tech's School of Public Policy, co-authored the study.

They found that while nearly half of the reports were downloaded for academic purposes, even more were accessed by people outside strictly educational settings, such as veterans, chaplains, and writers. The word "edification" appeared 3,700 times in the data set, signaling a strong desire for lifelong learning among users.

"This study shows strong demand among everyday Americans for the highest quality information to help improve at their jobs, to help their relatives, neighbors, and communities, and in some cases simply to learn for learning's sake," said Hicks. "We never hear these stories because everyone is focusing on all the misinformation that goes out over social media."

The study emphatically shows that open access to scientific information

matters to the average American, said co-author Ameet Doshi, a School of Public Policy Ph.D. student and the head of the Donald E. Stokes Library at Princeton University.

"This research will, hopefully, raise awareness about the positive returns that accrue to society from investments in institutions that democratize public access to high-quality research," said Doshi, who for his dissertation is analyzing similar data on downloads from Harvard University's open-access portal.

Machine learning growing into key tool for social scientists

The study examined 1.6 million comments left on 6.6 million downloads of NASEM consensus reports since 2011, when the Academies first began to offer them for free. The comments were left in response to a prompt asking users how they planned to use the reports.

The authors used a machine learning algorithm called BERT to analyze the comments—a further expansion into the social sciences of Asensio's use of machine learning techniques, which organize and make sense of unstructured data that is too time-consuming for people to analyze directly. Asensio's work has shown that such data can be of immense utility to researchers and policymakers when researchers properly teach algorithms how to do the hard work. Asensio' s Data Science and Policy Lab has used deep learning techniques in recent years to advance knowledge in energy efficiency, sustainable plastics, and electric vehicle charging infrastructure.

"When you get data at this scale, especially when you have unstructured data that grows in real-time, there are practical limitations as to why this kind of behavioral information was not previously known," Asensio said.

"We are showing in a number of research areas that experimental approaches to curate human-labeled training data can boost the performance of popular supervised ML algorithms, at a level that can match or even exceed human performance. This is expanding opportunities for data discovery in the social sciences. So, there was a compelling need to use these computational solutions to automatically classify behavioral evidence about public interest in scientific information."

The analysis found that academic users accounted for 48% of the downloads that had comments, an unsurprising result given the nature of the reports, which are densely scientific and primarily intended to meet the technical needs of federal agencies.

Learning for learning's sake

It is the other uses that most interested the researchers, including downloads from ham radio operators, amateur astronomers, lifelong learning providers, and retirees interested in keeping up.

About 150,000 downloads were categorized as having to do with "personal use," including topics such as cannabis, dying, genetically engineered crops, evolution versus creationism, and reducing gun violence. The analysis also revealed thousands of veterans planned to use NASEM reports as part of their disability claim to the U.S. Veteran's Administration, with NASEM's 20 reports on Agent Orange, the health effects of burn pits or high noise levels the most frequent downloads.

More than 25,000 doctors and nurses downloaded reports with plans to use the details to improve their clinical work.

One user downloaded 551 reports for "personal edification," according to the study.

The researchers also noted downloads by non-fiction authors, science fiction writers, and even visual artists. The reports were even the subject of book club discussion, according to the comments.

The algorithm the researchers used was able to identify the correct meaning about 84% of the time.

"There's a performative aspect to language, there's a descriptive aspect," Zullo said. "The fact that a machine learning tool can predict the meaning of it with that kind of accuracy is incredible."

Americans 'innately curious'

Overall, the findings point to broad and impactful diffusion of knowledge stemming from NASEM's decision to make the reports freely available, the authors wrote.

"The results reveal adults motivated to seek out the most credible sources, engage with challenging material, use it to improve the services they provide and learn more about the world they live in," they said. "The picture contrasts starkly with the dominant narrative of a misinformed and manipulated public targeted by social media."

That's not to say that social media misinformation isn't a problem, the authors note. Social media platforms are rife with millions of false and misleading posts, many of them posted by bots, that can help drive belief in conspiracy theories, misinformation, and even state-sponsored disinformation.

However, in this case, the study shows a public that—despite the widespread narrative lamenting the politicization of science and distrust in scientists—still turns to experts to help sort out a complicated, ever-changing world.

"A large part of the American public is innately curious and is willing to tackle the academic jargon to gain some insight," Doshi said. "That in itself is a comforting finding."

More information: Diana Hicks et al, Widespread use of National Academies consensus reports by the American public, *Proceedings of the National Academy of Sciences* (2022). DOI: 10.1073/pnas.2107760119

Provided by Georgia Institute of Technology