Apps, AI, and Automated Fake News Detection

LIBRARIANS SHOULD KNOW WHAT AUTOMATED DETECTION TOOLS CAN AND CANNOT DO TO FLAG MISINFORMATION, AND THEY SHOULD PROMOTE AND TEACH INFORMATION LITERACY WHENEVER POSSIBLE.

BY DARCY GERVASIO, MLIS

In January 2017, when I teamed up with journalism faculty at SUNY Purchase College for our first fake news "teach-in," we strove to give students concrete strategies for fighting the spread of false information. At the time, journalists and academic librarians were focused on teaching users to *identify* fake news (and making so many libguides!). I wanted to empower disenchanted undergrads to take small, proactive *actions*, like flagging fake articles on social media, donating to legitimate news organizations, and installing fake news detection browser extensions.

These extensions (apps) seemed cutting-edge and popular with students, and it's easy to see why. A plug-in that automatically fact-checks search results and news feeds relieves the mental load of having to critically examine thousands of posts each day—plus, users don't have to stop what they're doing to deliberately visit a third-party website like Snopes or Politifact.

Apps seemed like a modern, proactive solution, but the more I recommended them, the more I questioned how they work and what level of human intervention is involved. In this article, I dive into the literature to give librarians a primer on the current state of fake news detection technology—and reveal how (un)automated many apps actually are.

It's All About the Apps

Since 2016, fake news apps have proliferated, with newsrooms (e.g., ThisIsFake by Slate), nonprofit centers for journalism (e.g., CrossCheck by First Draft), for-profit cybersecurity startups (e.g., CheckThis by Metacert), college students (e.g., Project FiB from a hackathon at Princeton), and concerned-citizen-coders (e.g., B.S. Detector and Fake News Detector) getting in on the action. Tech giants like Google, Microsoft, and Facebook have announced partnerships with journalists and programmers and filed patents for tools to address the fake news crisis (Lee 2019; Jackson 2016; Newton 2016).

Yet, despite many small tweaks,[1] we are still waiting for comprehensive solutions. In August 2018, Microsoft launched NewsGuard, the first fake news plug-in to come standard with the Edge browser on all Android OS devices (Lapowsky 2018; Warren 2019). NewsGuard flags news within search results and social media and provides a "nutrition label" indicating how trustworthy or biased the website is.

But here's the secret: NewsGuard, along with most "automatic" fake news detection apps, is barely automated at all. Rather than using AI to examine the actual content of posts or the complex ways they spread, most detection apps on the market today rely on simple keyword matching to check domains against a human-curated blacklist of

DARCY GERVASIO is coordinator of reference services (associate librarian) at Purchase College, State University of New York. She can be reached at darcy.gervasion@purchase.edu.

Fake News Detection Tools by Type

Web platforms for crowd-sourced fact-checking/flagging:

- CrossCheck
- ClimateFeedback.org
- Fiskkit
- Hypothes.is

Browser extensions that rely on curated blacklists:

- B.S. Detector
- Fake News Detector (hybrid)
- NewsGuard
- Project FiB (uses domain and text keyword-matching to "verify" posts against other online sources; does not use a learning algorithm)
- ThisIsFake (defunct)

Browser extensions that rely on learning algorithms (computational prediction):

- CheckThis
- Factmata
- Fake News Detector (hybrid)
- Hoaxy (web platform, not an extension)

"fake" or "suspicious" websites. In fact, the only commercially available browser plug-ins I found that use learning algorithms to analyze characteristics of fake news, rather than simply matching articles against blacklisted domains, were Factmata and CheckThis. Fake News Detector, a free Chrome extension by a Brazilian coder, uses a hybrid of crowd-sourced fact checking, plus a "baby bot" algorithm that learns from each flagged post. (Fake News Detector is more transparent than the former about how its algorithms work.)

What's wrong with expert-curated blacklists? Nothing, in theory. Even the most basic plug-in serves as a useful alarm bell. But every librarian knows that determining whether a given article is trustworthy goes beyond checking the source website. Apps that rely on blacklists—even ones like B.S. Detector that code for satire and political bias, or NewsGuard, which touts the transparency of its rubric—put a lot of faith and power in their human list makers.

Beyond ethical debates about media gatekeeping and the authority of list makers, relying on human-curated blacklists is simply not scalable. ThisIsFake, an ambitious plug-in from Slate that flagged individual articles and linked directly to debunking sources, shut down after a year (Oremus 2016). No explanation was given, but it's fair to assume Slate's fact checkers couldn't keep up with the onslaught of false stories.

A room of expert fact checkers—or even an international crowdsourced network like the CrossCheck or Fiskkit platforms—cannot keep pace with the creation of new hoax sites and fake posts. Shao et. al. discovered a lag of 10-20 hours before a false claim is fact-checked by journalists, plenty of time for a post to go viral (Shao et al. 2016, 1-2). Meanwhile, recent exposés on the poor labor conditions and long-term mental health consequences faced by social media "content moderators" reveal the human toll of large-scale fact checking (Chen 2014; Newton 2019). To stop fake news before it goes viral, automation must play a bigger role.

Three Methods of Detecting Fake News

What's the difference between truly automated detection and extensions like NewsGuard? First, consider that fake news is detected by three methods: (1) expert fact checking, (2) crowd-sourced flagging, and (3) computational prediction, also known as automatic detection (Shu et al. 2017). The first two have driven the solutions offered by journalists, whereas computational prediction has been the focus of computer scientists.

The scientific literature indicates an unfortunate communication gap between these two groups. Journalist-led initiatives have produced more user-friendly tools, in the form of crowd annotation/flagging web platforms (e.g., CrossCheck, Fiskkit, ClimateFeedback.org, and Hypothes.is) and browser extensions running on human-curated "blacklists" (e.g., ThisIsFake, B.S. Detector, and NewsGuard). In contrast, computer scientists are developing learning algorithms[2] that can spot fake news without human intervention. These researchers have focused mainly on testing their algorithms for accuracy, but have yet to create functional, publicly available apps. Several promising tools are now in beta, such as the University of Indiana's Hoaxy (Shao et al. 2016) and the Google-backed Factmata from University College London (Jackson 2016).

Three Types of Fake News Algorithms

Automated fake news detection involves three types of learning algorithms: (1) textual/content analysis, (2) user behavior/engagement analysis, and (3) diffusion analysis (tracking the spread of fake stories across networks).

Textual analysis alone can be quite challenging; it's hard to program algorithms to account for satire, bias, and intent (Papadopoulou et al. 2017; Edell 2018). Natural language processing algorithms that incorporate emotional affect and psycholinguistics look promising, since affective language appears

> It will always be important for librarians to host workshops, make libguides, and teach information literacy in all its messy glory.

more often in "clickbait" and contributes to its proliferation (Pérez-Rosas et al. 2017). Meanwhile, user behavior analysis suggest that *who* engages with a post can tell us nearly as much about its "fakeness" as the text itself (Tacchini et al. 2017; Shu, Wang, and Liu 2017). Finally, there's evidence that fake and real stories spread across networks differently (Shu et al. 2017; Zhao et al. 2018).

Successful fake news detection will likely require a combination of all three types of algorithms, or a hybrid approach that incorporates computational prediction as well as crowdsourcing and expert fact checking (Figueira and Oliveira 2017; Ruchansky, Seo, and Liu 2017; Wang 2017).

Facebook is an example of the hybrid approach (Figueira and Oliveira 2017). From what the company has shared publicly, Facebook uses crowdsourcing to flag fake news and other offensive content (users tap somewhat-hidden buttons to "report [an] ad" or "give feedback on this post"). A user behavior algorithm gives flaggers a "reliability score" to indicate how consistently they properly flag fake stories. Reliability scores are likely used to calculate the probability that a specific post is fake and rank the "worst" offenders (Newton 2016; Kozlowska 2017; Figueira and Oliveira 2017). Similar user behavior algorithms have also been used to suppress spam accounts, trolls, and bots (Adewole et al. 2017). Finally, posts identified as "fake" are sent to human "content moderators" Facebook hires through third-party companies, often overseas.

While we don't know the exact process for false news, this is how Facebook handles "offensive" content, including pornography, hate speech, and conspiracy theories that violate its "Community Standards" (Newton 2019; Chen 2014). The exploitative labor conditions and mental health risks for content moderators—many develop PTSD or come to believe conspiracy theories to which they are repeatedly exposed—pose ethical concerns, as detailed in a Verge article (Newton 2019) and upcoming book (Roberts 2019).

Not There Yet

If learning algorithms can be perfected into apps, we could (theoretically) rely less on reactive, costly, ethically problematic human content moderation. A truly automated tool could detect a false story before millions of people have been exposed to it just by analyzing its textual attributes, who has shared it, and how it spreads across social networks. An app could alert users to such stories or even suppress them, in a fraction of the time it takes humans to debunk them. While this is the lofty goal of browser extensions like Fake News Detector, CheckThis, and Factmata, the technology is not there yet.

Obviously, there are dangers in letting algorithms police the news. Most scientific literature ignores the ethical and free speech concerns posed by automation, though Figueira and Oliveira warn against giving machines "total control to decide which information is displayed" (Figueira and Oliveira 2017, 822).

I am somewhat reassured that human experts are still needed to create datasets for training algorithms. Indeed, much of the literature focuses on new sources of datasets (Pérez-Rosas et al. 2017; Shu et al. 2017; Wang 2017), whether crowdsourced fact checking is as reliable as expert fact checking (Tschiatschek et al. claim that it is), or how well algorithms perform compared with control groups of expert fact checkers (Tacchini et al. 2017). In short, expert fact checkers will always be integral to developing algorithms.

Likewise, I won't be giving up fake news teach-ins anytime soon. It will always be important for librarians to host workshops, make libguides, and teach information literacy in all its messy glory. An app can help users spot fake news quickly, but it's still up to readers to interpret the results of any automated solution. As librarians teach critical evaluation and media literacy, it can only help us to have a nuanced understanding of what automatic detection tools can—and cannot—do to stop fake news. **SLA**

NOTES

1 In 2017, Google tweaked its search algorithm to "surface more authoritative pages and demote low-quality content" (Gomes 2017). In 2018, Facebook shrank the size of fake posts and made factual "related" articles appear beside them in users' newsfeeds (Kozlowska 2017). In 2019, YouTube used a combination of AI and "real people" to keep conspiracy theories from popping up as recommended videos (YouTube Team 2019).

2 Sometimes called AI or artificial intelligence, learning algorithms make predictions or calculate probabilities based on existing datasets and become better at making predictions about new content over time, as the dataset grows.

REFERENCES

Adewole, Kayode Sakariyah, Nor Badrul Anuar, Amirrudin Kamsin, Kasturi Dewi Varathan, and Syed Abdul Razak. 2017. "Malicious Accounts: Dark of the Social Networks." *Journal of Network and Computer Applications*, 79(February): 41–67.

Chen, Adrian. 2014. "The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed." Wired, October 23.

Edell, Aaron. 2018. "I Trained Fake News Detection AI with >95% Accuracy, and Almost Went Crazy." Towards Data Science. January 11.

Figueira, Álvaro, and Luciana Oliveira. 2017. "The Current State of Fake News: Challenges and Opportunities." Procedia Computer Science, CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017, 121(January): 817–825.

Gomes, Ben. 2017. "Our Latest Quality Improvements for Search." Google, April 25.

Jackson, Jasper. 2016. "Fake News Clampdown: Google Gives 150,000 to Fact-Checking Projects." *The Guardian*, November 17.

Kozlowska, Hanna. 2017. "Facebook Is Ditching Its Own Solution to Fake News Because It Didn't Work." Quartz, December 22.

Lapowsky, Issie. 2018. "NewsGuard Wants to Fight Fake News with Humans, Not Algorithms." *Wired*, August 23.

Lee, Edmund. 2019. "Veterans of the News Business Are Now Fighting Fakes." *The New York Times*, January 17, Business section.

Newton, Casey. 2019. "The Trauma Floor: The Secret Lives of Facebook Moderators in America." The Verge, February 25.

_____. 2016. "Facebook Is Patenting a Tool That Could Help Automate Removal of Fake News." The Verge, December 7.

Oremus, Will. 2016. "It's Not Enough to Know What News Is Fake. Help Stop Its Spread with Slate's New Tool." *Slate Magazine*. December 13.

Papadopoulou, Olga, Markos Zampoglou, Symeon Papadopoulos, and Ioannis Kompatsiaris. 2017. "A Two-Level Classification Approach for Detecting Clickbait Posts Using Text-Based Features." Working Paper. Clickbait Challenge 2017. arXiv.org.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2017. "Automatic Detection of Fake News." https://doi.org/arXiv:1708.07104 [cs.CL].

Roberts, Sarah T. 2019. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven, Conn.: Yale University Press.

Ruchansky, Natali, Sungyong Seo, and Yan Liu. 2017. "CSI: A Hybrid Deep Model for Fake News Detection." In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 797–806. Singapore, Singapore: ACM.

Shao, Chengcheng, Giovanni Luca Ciampaglia, Alessandro Flammini, and Filippo Menczer. 2016. "Hoaxy: A Platform for Tracking Online Misinformation." In Proceedings of the 25th International Conference Companion on World Wide Web, 745–50. Montreal, Quebec, Canada.

Shu, Kai, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. "Fake News Detection on Social Media: A Data Mining Perspective." ACM SIGKDD Explorations Newsletter 19(1): 22–36.

Shu, Kai, Suhang Wang, and Huan Liu. 2017. "Exploiting Tri-Relationship for Fake News Detection." In Proceedings of 12th ACM International Conference on Web Search and Data Mining (WSDM 2019). Melbourne, Australia.

Tacchini, Eugenio, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. "Some Like It Hoax: Automated Fake News Detection in Social Networks." In Proceedings of the Second Workshop on Data Science for Social Good (SoGood), 1960:1–11. Skopje, Macedonia: CEUR Workshop Proceedings. http://arxiv.org/abs/1704.07506.

Tschiatschek, Sebastian, Adish Singla, Manuel Gomez Rodriguez, Arpit Merchant, and Andreas Krause. 2018. "Fake News Detection in Social Networks via Crowd Signals." In Companion Proceedings of the The Web Conference 2018, 517–524. WWW '18. Lyon, France: International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3184558.3188722.

Wang, William Yang. 2017. "'Liar, Liar Pants on Fire': A New Benchmark Dataset for Fake News Detection." In ACL 2017.

Warren, Tom. 2019. "Microsoft Is Trying to Fight Fake News with Its Edge Mobile Browser." The Verge. January 23.

YouTube Team. 2019. "Continuing Our Work to Improve Recommendations on YouTube." YouTube. Official YouTube Blog (blog). January 25.

Zhao, Zilong, Jichang Zhao, Yukie Sano, Orr Levy, Hideki Takayasu, Misako Takayasu, Daqing Li, and Shlomo Havlin. 2018. "Fake News Propagate Differently from Real News Even at Early Stages of Spreading." arXiv.org.