Undergraduate Thesis Prospectus

Learning to Automate the Factchecking of News Articles

(technical research project in Computer Science)

The Competition over Automation in News Media

(sociotechnical research project)

by

Christine Baca

November 2, 2020

technical project collaborators:

Sharon Bryant

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.
Christine Baca

Technical advisor: Bloomfield T. Advisor, Department of Computer Science

STS advisor: Peter Norton, Department of Engineering and Society

General Research Problem

How can machine learning best augment human communication in the area of journalism?

With machine learning (ML), engineers can apply artificial intelligence to communications media, introducing new advantages and new hazards. ML can promote the distribution and analysis of accurate and false information on social media (Span, 2020). ML algorithms can automate validation of online news, and it can even automate journalism itself, through algorithms that "automatically generate news" (Graefe, 2016). Within the news industry, machine learning enables automated journalism: "the use of algorithms to automatically generate news" (Graefe, 2016). Media companies, news organizations, journalists, tech companies, and readers promote, resist or otherwise influence applications of ML in news media. This competition will determine how ML's technical possibilities in news media will develop.

Learning to Automate the Factchecking of News Articles

How do journalists use machine learning to advance the news industry?

The spread of fake news damages the reputation of quality journalism. Jeff Hancock, a psychologist at Stanford University, says social medias like Facebook and Twitter deploy "mechanisms" of machine learning to surreptitiously promote fake news to the elderly, and, in response, many courses like MediaWise, a free "digital literacy course," exist to teach Americans to "detect and combat misinformation" (Span, 2020). To further help consumers of journalism identify fake news and to further understand the technical

relationship between machine learning, media, and language, Sharon Bryant and I propose building an application that uses machine learning to deduce the legitimacies and bias of web articles. We plan to complete the Capstone Project in the spring with Professor Aaron Bloomfield from the Computer Science department.

Project goals include designing a classifier that trains on large datasets of real and fake news. State of the art classifiers, such as Google AI's Bidirectional Encoder Representations from Transformers or BERT, utilize natural language processing and unsupervised learning (Devlin, 2019). By testing different modes of modeling and researching other state of the art classifiers like Google AI's BERT, we hope to design a machine learning algorithm that classifies and analyzes fake news and real news to better understand and improve upon the quality of information and journalism on social media platforms.

The Competition over Automation in News Media

How are journalists responding to efforts to automate aspects of journalism?

How are journalists adapting to machine learning systems in media? News organizations have applied machine learning (ML) algorithms to introduce automated journalism: written articles that emulate journalism by human reporters. To many journalists, such "robot journalism" threatens "journalism's value" (Kim, 2018), but many also agree that automated data collection, language processing, and analysis, properly applied, can augment the industry's speed, scale, and accuracy (Graefe, 2016). The *Los Angeles Times* has used its Quakebot application to analyze data from the U.S. Geological Survey and report local earthquakes since 2017 (Miller, 2019).

The *Washington Post*'s Heliograf is an "automated storytelling technology" that uses data and natural language processing for hyperlocal news (WashPostPR, 2017).

Researchers have investigated the dangers and benefits of automated journalism. Dörr (2015) contends that, because natural language generation (NLG) can "perform tasks of professional journalism at a technical level," it can revolutionize journalism. Lewis (2019) warns, however, that the emergence of "newswriting bots" introduces problems of legal responsibility, including "the complicated matter of determining fault in a case of algorithm-based libel." According to Galily (2018), automated journalism will continue to develop despite the "worry that automation will either cause or be used as an excuse for job cuts and dismissal of journalists." As Graefe (2016) notes, automation's advantages are compelling: algorithms can already "create thousands of news stories for a particular topic," and "do it more quickly, cheaply, and potentially with fewer errors than any human journalist." But to reporters, such advantages are also threats. Automation "has fueled journalists' fears that automated content production will eventually eliminate newsroom jobs" (Graefe, 2016).

Journalists disagree about the implications of automation for journalism. Kim (2018) found that many journalists resist robot journalism. Some see in it a threat to "their status in their organization"; some think "robots are likely to damage journalism's value" (Kim, 2018). Many executives in news organizations, however, welcome automated journalism. Lisa Gibbs, the director of news partnerships for the Associated Press, argues that automated journalism liberates journalists: "The work of journalism is creative, it's about curiosity, it's about storytelling, it's about digging and holding governments accountable, it's critical thinking, it's judgment —and that is where we

want our journalists spending their energy." Francesco Marconi, the *Wall Street Journal*'s head of research and development, predicts that "a lot of the tools in journalism will soon be powered by artificial intelligence" (Peiser, 2019). Tech companies that market journalism programs extol their products' advantages. The CEO of Syllabs, a company that "produced 150,000 web pages ... during France's 2015 election," told reporters: "Robots can't do what journalists do," but they "can do amazing things and it's a revolution for the media" (Radcliffe, 2016).

Some readers distrust automated journalism. Critics warn that "algorithmic authorship" complicates responsibility, and invites "discrepancies between the perceptions of authorship and crediting policy" (Montal, 2016). Some question the algorithms' reliability as "fair and accurate," and "free from subjectivity, error, or attempted influence" (Gillespie, 2014). Readers cannot reliably distinguish articles written by humans from articles written by software (Clerwall, 2014), and Graefe (2016) finds that readers even "rate automated news as more credible than human-written news." Graefe adds, however, that readers "do not particularly enjoy reading automated content." Robot journalists' limitations as writers will constrain their growth in journalism; it may also be symptomatic of their limits as interpreters of the news.

References

Clerwall, Christer (2014) Enter the Robot Journalist, *Journalism Practice*, *8*(5), 519-531, doi: 10.1080/17512786.2014.883116

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*

Dörr, K. N. (2015). Mapping the field of Algorithmic Journalism. *Digital Journalism, 4*(6), 700-722. doi:10.1080/21670811.2015.1096748

Galily, Y. (2018, August 1). Artificial intelligence and sports journalism: Is it a sweeping change?. *Technology In Society*, *54*, 47 – 51.

Gillespie, T., Boczkowski, P. J., & Foot, K. A. (2014). Media Technologies. *MIT Press Scholarship Online*. doi:10.7551/mitpress/9780262525374.001.0001

Graefe, A. (2017). Guide to Automated Journalism. *Columbia Journalism Review*. doi: https://doi.org/10.7916/D80G3XDJ

Kim, D., & Kim, S. (2018, May 1). Newspaper journalists' attitudes towards robot journalism. *Telematics and Informatics*, *35*(2), 340 - 357.

Miller, C. (2014). Quakebot. *Los Angeles Times.* https://www.latimes.com/people/quakebot

Montal, T., & Reich, Z. (2016). I, Robot. You, Journalist. Who is the Author? *Digital Journalism, 5*(7), 829-849. doi:10.1080/21670811.2016.1209083

Peiser, J. (2019, February 05). The Rise of the Robot Reporter. *New York Times.* https://www.nytimes.com/2019/02/05/business/media/artificial-intelligence-journalism-robots.html

Radcliffe, D. (2016, July 07). The Upsides (and Downsides) of Automated Robot Journalism. *MediaShift.* http://mediashift.org/2016/07/upsides-downsides-automated-robot-journalism/

Span, P. (2020, September 11). Getting Wise to Fake News. *New York Times.* https://www.nytimes.com/2020/09/11/health/misinformation-social-media-elderly.html

WashPostPR. (2019, February 28). The Washington Post leverages automated storytelling to cover high school football. *Washington Post.* https://www.washingtonpost.com/pr/wp/2017/09/01/the-washington-post-leverages-heliograf-to-cover-high-school-football/