Bulletin of the Atomic Scientists

ISSN: 0096-3402 (Print) 1938-3282 (Online) Journal homepage: https://www.tandfonline.com/loi/rbul20

An AI early warning system to monitor online disinformation, stop violence, and protect elections

Michael Yankoski, Tim Weninger & Walter Scheirer

To cite this article: Michael Yankoski, Tim Weninger & Walter Scheirer (2020) An AI early warning system to monitor online disinformation, stop violence, and protect elections, Bulletin of the Atomic Scientists, 76:2, 85-90, DOI: 10.1080/00963402.2020.1728976

To link to this article: https://doi.org/10.1080/00963402.2020.1728976

Published online: 19 Feb 2020.

BULLETIN OF THE ATOMIC SCIENTISTS
2020, VOL. 76, NO. 2, 85–90
https://doi.org/10.1080/00963402.2020.1728976

Routledge
Taylor & Francis Group

OTHER FEATURES

An AI early warning system to monitor online disinformation, stop violence, and protect elections

Michael Yankoski, Tim Weninger and Walter Scheirer

ABSTRACT
We're developing an AI early warning system to monitor how manipulated content online such as altered photos in memes leads, in some cases, to violent conflict and societal instability. It can also potentially interfere with democratic elections. Look no further than the 2019 Indonesian election to learn how online disinformation can have an unfortunate impact on the real world. Our system may prove useful to journalists, peacekeepers, election monitors, and others who need to understand how manipulated content is spreading online during elections and in other contexts.

KEYWORDS
Artificial intelligence; conflict; elections; manipulated content; disinformation; memes

Over half the world's population is now online (Internet World Stats 2019). More than 1.6 billion people use Facebook each day (Facebook 2019), while Twitter has another 145 million daily users (Twitter 2019). But our ability to connect online has far outpaced our capacity to distinguish between reliable and unreliable information. Intelligence agencies, political parties, and rogue actors are all taking advantage of this complex and exploitable web of interconnections and using social media to disseminate altered content, fake news, and propaganda to interfere in elections and even to incite direct violence.

As digital forensics researchers, we have been developing artificial intelligence (AI) capabilities that validate the integrity of media such as photos or videos for government, commercial, and humanitarian purposes. We believe one important use for this technology is to preserve peace and security by helping guarantee the integrity of democratic elections. Another use is to help provide real time forecasts of mass violence in volatile contexts. With these ends in mind, we are building an early warning system that will employ AI to observe social media for the coordinated dissemination of manipulated disinformation and other emerging malicious trends. This system will be able to warn human observers such as journalists or election monitors about potential threats in real time.

The broad reach of social media platforms – their incredible ability to sway opinions, shape perspectives, and incite action – means that attempts to exploit and manipulate the users of these platforms are likely to increase. But given the massive amount of data generated on the internet every minute, AI systems are the only tools capable of identifying and analyzing the trends and potential threats arising from disinformation in real time. Our AI early warning system will allow people to understand manipulative online content in order to limit its potential to disrupt elections or incite violence.

The problem posed by organized disinformation campaigns isn't limited to a few bad actors trying to cause chaos. A recent study at Oxford University discovered that 28 countries make use of coordinated social media influence campaigns for a variety of purposes (Bradshaw and Howard 2017). While authoritarian regimes all had social media campaigns targeting their own populations, democracies tended to direct their influence abroad. It's hard to say if one of these findings is more unnerving than the other.

But is manipulated media really something to worry about? A faked photograph of Elvis shaking hands with Kim Jong Un is more or less harmless, no (Waddell 2018)? What is the problem if someone uses cutting-edge AI technology to make a so-called deepfake video showing Barack Obama (Vincent 2018) saying something he didn't? Or if people make videos of Kim Kardashian or Mark Zuckerberg appearing to make references to a mysterious person, group, or thing they call "Spectre"? The videos go viral and a few million people laugh.

But not all fakes are intended for laughs. Imagine if a faked video appeared online showing a prominent world-leader or a significant national security official declaring a war or announcing a military attack. Would other countries respond? It's all too easy, for instance, to imagine a faked image or audio clip of a political figure being tortured causing protestors to take to the streets in demand of vengeance. The technology that produces fake images, video, and audio is becoming so convincing

CONTACT Walter Scheirer walter.scheirer@nd.edu

that it is plausible that it will be deployed in an effort to spark (or justify) a real war, or even a real genocide.

These concerns have led AI researchers such as ourselves to form new research collaborations with scholars in Peace Studies, an interdisciplinary field devoted to analyzing the causes of conflict and peace. The intersection between the two fields makes sense. For instance, a system designed to flag all the faked images, videos, or audio clips released online would have to scan literally every bit of posted content – no simple task given the terabytes of data generated every minute. Instead, we are developing systems with the specific goal of helping prevent violence due to malignant disinformation in particular places and at particular times – places where there is a strong likelihood of violence breaking out. By incorporating lessons from Peace Studies, AI developers can target an early warning system to where there's the highest probability of disinformation leading to election manipulation or violence.

The reasons why violence, genocide, and mass atrocities might be more or less likely in a given country or region are well understood (Straus 2016). Peace Studies scholars have provided a helpful distinction between so-called conflict epicenters and conflict episodes (Lederach 2003). Conflict epicenters are the complex and deep-seated tensions that exist among groups of people, whatever their form or origin. Historical violence, longstanding ethnic tensions, pervasive distrust and suspicion, ongoing entrenched forms of injustice and oppression, etc., are all examples of conflict epicenters, and they are the seedbeds of violent outbreaks. Conflict episodes, on the other hand, explode from conflict epicenters, and are distinct and identifiable moments of direct violence. This is not a rigid distinction, of course, but these definitions help bring into focus the historical fact that eruptions of violence among groups of people tend to emerge in the context of deep-seated tensions and incompatibilities (Varshney 2014).

By using insights from both artificial intelligence and Peace Studies research, we are building predictive computer models of the ways nefarious actors might manipulate conflict epicenters through social media in order to provoke conflict episodes. This includes building both new algorithms as well as dynamic, context-specific libraries of inflammatory language, images, and symbols. We intend to use these tools to provide information on situations that warrant further investigation to journalists, human rights experts, and government officials.

Recent US government investment in research and development efforts such as the Defense Advanced Research Project Agency's (or DARPA's) Media Forensics program (Turek 2019) has supported breakthroughs across the entire spectrum of digital image forensics. The program has aided the development of technologies ranging from low-level detectors of image manipulation to high-level analysis programs that can establish the provenance of media content. With respect to the former, software now exists to detect photoshopped (or altered) images (Farid 2016), deepfake videos (Rössler et al. 2019), and voice-swapped audio (Agarwal et al. 2019). With respect to the latter, novel algorithms can perform sophisticated data mining operations to identify related content and trace the order of its creation. Many of the related algorithms harness recent breakthroughs in artificial neural networks, supercomputing, and vast troves of data – the cutting edge of AI (Moreira et al. 2018).

The digital forensics community has started to turn its attention to manipulated media in the context of international elections. Unlike elections of the past, where voters have largely been consumers of professionally curated media content, anyone can now make and disseminate their own political messages to an audience of millions on the internet. A primary tool for doing this has been the meme (Schifman 2014). Memes are cultural artifacts that evolve as they spread among internet users. Frequently they consist of humorous images and text that adheres to a set formula acting as a guideline for the creation of new instances (see Figure 1). But the images are often more than just jokes. Memes have served as the impetus for political actions in movements as diverse as the Arab Spring (York 2012) and Occupy Wall Street (Know Your Meme 2020). And they are now a significant resource for monitoring the pulse of an election.

Figure 1. Examples of a typical humorous meme, in which the public remixes an original image many times on the internet.

The 2019 Indonesian general election

For a clear illustration of our concerns about the impact coordinated online disinformation can have on electoral politics, look no further than the 2019 Indonesian general election. In a rematch of the 2014 election, incumbent president Joko Widodo ran against challenger Prabowo Subianto in what turned out to be a rancorous campaign. Widodo, a left-leaning centrist, defeated the conservative populist Subianto with 55 percent of the vote. Subianto promptly contested the outcome of the election and mobilized his supporters to take to the streets of Jakarta. The ensuing mayhem left eight people dead and hundreds injured (Lipson 2019). Online information about the protests was rife with inaccuracies and, in some cases, outright hoaxes (The Jakarta Post 2019). In response, the Indonesian government blocked access to social media to prevent the spread of such content. What was the government so concerned about to take such a drastic measure (Singh and Russell 2019)?

Much can be learned about a country, its people, and its fears by studying the way images are manipulated in online forums. We monitored Indonesian social media as the election was unfolding using a prototype of our early warning system for violence. In our analysis of over two million meme-style images collected from Twitter and Instagram over a period of a year, we used pattern recognition algorithms to identify related political messages. Our AI system was designed to detect stylistic similarities between images from a targeted collection, allowing it to automatically group variations on the same message together. Users of the system can see the groupings and make further assessment of them. To ingest relevant content for analysis, we followed so-called buzzer accounts (fake accounts used to spread content in a way that makes the public believe that it is more popular than it really is) (Potkin and Beo Da Costa 2019) and hashtags known to be signifiers of questionable content. In total we found 1,500 different groups of images, many containing thousands of variations on a meme. We also discovered that supporters of both candidates deployed an astonishing array of deceptive content (see Figure 2).

Some of the memes spread by Widodo's supporters were designed to portray the opposition as being prone to violence. Pro-Subianto memes, on the other hand, sometimes depicted the desecration of objects sacred to Islam by secular politicians running for office. Tens of thousands of these images were detected as being intentionally manipulated to mislead the viewer. In many circumstances, the images were changed to incorporate political iconography in offensive contexts, e.g., a hammer and sickle superimposed on an Islamic prayer mat, meant to insinuate that the people in the image are crypto-communists. Another common genre of memes reflected suspicion of Chinese interference in Indonesia. This category included repurposed images of fair-skinned police officers being depicted as Chinese security forces as well as allegations that imported chili peppers were vectors for biological weapons (Chew and Barahamin 2019). Such images were meant to provoke targeted populations based on their concerns about Indonesia's troubled past and uncertain present (Reuters 2016).

More subtle forms of manipulation were equally alarming (see Figure 3). For instance, a composite image contained a bogus painted sign on a building that promoted a hashtag used to categorize content critical of Widodo. We also found instances of co-opted corporate logos. In one such example, the CNN International logo was added

Figure 2. Examples of memes that appeared during the 2019 Indonesian general election campaign. Left: Pro-Widodo meme. The peaceful political demonstration and text reading "when our leader lost" on the top is contrasted with an image of rioting in Jakarta and text reading "when your leader lost" on the bottom. This meme is intended to provoke supporters of Subianto. Center: Possible pro-Subianto meme mimicking the desecration of an Islamic grave. Right: Meme meant to invoke the controversial role of communism in Indonesia's past.

Figure 3. Examples of composite images designed to be used in disinformation campaigns. Left and Center: This anti-Widodo hashtag was added to the building through the use of a digital image-editing tool. The modification was detected by an algorithm that searches images for inconsistencies in the statistics of their pixels. Right: the CNN International logo has been added to a false news story about Widodo.

to a false news story about Widodo that circulated as a purported screen capture on Facebook. (CEKFAKTA 2019). That logo was an addition intended to make the image look credible. Most insidious of all were the changes to images that appeared entirely plausible. Without assistance from sophisticated AI tools, viewers would find it nearly impossible to detect subtle changes of this nature.

Implications for AI and peace studies research

While using AI to find groups of election-related memes is one thing, developing an automated means of understanding the "why" behind manipulations and misinformation will be much more difficult. Doing so requires considering questions such as: Who first disseminated a piece of content? When? What other content has the source created in the past? How has the content been manipulated? Who else has shared the manipulations? While a human might look at a meme and easily understand its implications, AI researchers are still developing these capabilities in machines. The ability to replicate the meaningful associations that people effortlessly make can further automate the process, facilitating a rapid response to emerging threats. Although computer scientists haven't achieved that goal yet, AI technology is already being evaluated by USAID for its initiatives that promote democracy and fair and transparent elections.

We uncovered a trove of disturbing material related to the Indonesian election. A key question for policy makers is how this information should be used. With respect to a policy agenda, there are several immediate use cases for an AI early warning system for violence that can benefit the international security community. Aided by real-time information about emerging threats of violence or escalating tensions, peacekeeping missions can react swiftly to defuse an incident as it unfolds. Similarly, election monitors can better gauge the fairness and legitimacy of an election if they have an understanding of what influence campaigns or intimidation tactics are being deployed on the internet. When used in a forensics context, there is the possibility of collecting and understanding image and video evidence at an unprecedented scale, changing the way investigations into human rights abuses and war crimes are conducted. In all cases, sufficient human resources must be committed to examine any warnings that appear most urgent. As AI evolves, it may be able to take on more of this task by itself.

And then there is the equally important question of access: Who should have access to such a potent violence early warning system? The ethical and political ramifications to this question are significant. We advocate strongly for a sliding-scale subscription fee access model. A structured committee should be responsible for granting access to the system. We argue that access must be guided by a set of principles such as a) empowering human rights monitors; b) strong support for civil-society actors working to secure human rights; c) transparency and education about how the AI technology functions, including its limitations; d) oversight by and accountability to a larger institutional structure such as an arm of the United Nations, or an electoral oversight

element of regional organizations like the European Union, African Union, or Organization of American States.

To combat misinformation effectively, we must not only be aware of the social media landscape, but also of our own actions when participating online. We have become the editors of our friends' news. It's important that social media users become more knowledgeable about the pitfalls and perils of the platforms. Much like how educational campaigns such as "Click it or Ticket" or "Stop Drop and Roll" are in place to encourage seat belt use and fire safety, educational memes, like, say, "Think Before You Share," or "Recognize Your Reaction," could help social media users become more thoughtful.

Together, AI tools like our early warning system and education campaigns can help combat the psychological warfare of targeted misinformation campaigns on the internet.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This article is based on research sponsored by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number [FA8750-16-2-0173]. Support was also given by USAID under agreement number [7200AA18CA00054].

Notes on contributors

Michael Yankoski is a doctoral candidate at the University of Notre Dame's Kroc Institute for International Peace Studies. His dissertation research explores the intersection of anthropogenic climate change, virtue theory, strategic peacebuilding, and human population displacement.

Tim Weninger is an associate professor in the Department of Computer Science and Engineering at the University of Notre Dame. His research is at the intersection of social media, artificial intelligence, and graphs.

Walter Scheirer is an assistant professor in the Department of Computer Science and Engineering at the University of Notre Dame. His research is in the area of artificial intelligence, with a focus on visual recognition, media forensics, and ethics.

References

Agarwal, S., H. Farid, G. Yuming, H. Mingming, K. Nagano, and L. Hao. 2019. "Protecting World Leaders Against Deep Fakes." Paper presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, 16–20.

Bradshaw, S., and P. Howard. 2017. "Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation." Vol. 2017.12. Oxford Internet Institute. https://ora.ox.ac.uk/objects/uuid:cef7e8d9-27bf-4ea5-9fd6-855209b3e1f6

CEKFAKTA. 2019. "[Salah] 'Jokowi Fake and Authoritary President of Rrc Communists,indonesia Branch'." *CEKFAKTA*, November 7. https://cekfakta.com/focus/2799

Chew, A., and A. Barahamin. 2019. "Chinese Indonesians in Jakarta Fear Attacks on the Community, as Anti-China Hoaxes Spread on Social Media." *South China Morning Post*, May 22. https://www.scmp.com/week-asia/politics/article/3011392/chinese-indonesians-jakarta-fear-attacks-community-anti-china

Facebook. 2019. "Facebook Reports Third Quarter 2019 Results." Facebook, October 30. https://s21.q4cdn.com/399680738/files/doc_financials/2019/q3/FB-Q3-2019-Earnings-Release.pdf

Farid, H. 2016. *Photo Forensics*. Cambridge, MA: MIT Press.

Internet World Stats. 2019. "INTERNET USAGE STATISTICS: The INTERNET Big Picture. World INTERNET Users and 2019 Population Stats." *Internet World Stats*, June 30. https://www.internetworldstats.com/stats.htm

Know Your Meme. 2020. "Occupy Wall Street." *Know Your Meme*, January (updated). https://knowyourmeme.com/memes/events/occupy-wall-street

Lederach, J. 2003. *The Little Book of Conflict Transformation: Clear Articulation of the Guiding Principles by a Pioneer in the Field*. Intercourse, PA: Good Books.

Lipson, D. 2019 "Indonesia's Worst Political Violence in Two Decades Brings Out the Comically Absurd." ABC News, May 24. https://www.abc.net.au/news/2019-05-25/indonesian-riots-bring-out-the-comically-absurd/11148770

Moreira, D., A. Bharati, J. Brogan, A. Pinto, M. Parowski, K.W. Bowyer, P.J. Flynn, et al. 2018. "Image Provenance Analysis at Scale." *IEEE Transactions on Image Processing* 27 (12): 6109–6123.

Potkin, F., and A. Beo Da Costa. 2019. "In Indonesia, Facebook and Twitter are 'Buzzer' Battlegrounds as Elections Loom." *Reuters*, March 12. https://www.reuters.com/article/us-indonesia-election-socialmedia-insigh/in-indonesia-facebook-and-twitter-are-buzzer-battlegrounds-as-elections-loom-idUSKBN1QU0AS

Reuters. 2016. "China Alarmed as Chili 'Conspiracy' Heats up Indonesians." *Reuters*, December 16. https://www.reuters.com/article/us-indonesia-china-chili-idUSKBN1451G4

Rössler, A., D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. 2019. "FaceForensics++: Learning to Detect Manipulated Facial Images." arXiv:1901.08971. Advance online publication.

Schifman, L. 2014. *Memes in Digital Culture*. Cambridge, MA: MIT Press.

Singh, M., and J. Russell. 2019. "Indonesia Restricts WhatsApp, Facebook and Instagram Usage following Deadly Riots." *TechCrunch*, May 22. https://techcrunch.com/2019/05/22/indonesia-restricts-whatsapp-and-instagram

Straus, S. 2016. *Fundamentals of Genocide and Mass Atrocity Prevention*. Washington D.C.: United States Holocaust Memorial Museum.

The Jakarta Post. 2019. "Six Dead, 200 Injured in Jakarta Riot: Jakarta Governor." *The Jakarta Post*, May 22. https://www.thejakartapost.com/news/2019/05/22/six-dead-200-injured-in-jakarta-riot-governor-anies.html

Turek, M. 2019. "Media Forensics (Medifor)." DARPA. Accessed 17 January. https://www.darpa.mil/program/media-forensics

Twitter. 2019. "Twitter Announces Third Quarter 2019 Results." Twitter, October 24. https://s22.q4cdn.com/826641620/files/doc_financials/2019/q3/Q3'19-Earnings-Press-Release-FINAL-(1).pdf

Varshney, A. 2014. *Battles Half Won: India's Improbable Democracy*. UK: Penguin Global.

Vincent, J. 2018 "TL;DR Watch Jordan Peele Use AI to Make Barack Obama Deliver a PSA about Fake News." *The Verge*, April 17. https://www.theverge.com/tldr/2018/4/17/17247334/ai-fake-news-video-barack-obama-jordan-peele-buzzfeed

Waddell, K. 2018. "The Impending War over Deepfakes." *Axios*, July 22. https://www.axios.com/the-impending-war-over-deepfakes-b3427757-2ed7-4fbc-9edb-45e461eb87ba.html

York, J. C. 2012. "Middle East Memes: A Guide." *The Guardian*, April 20. https://www.theguardian.com/commentisfree/2012/apr/20/middle-east-memes-guide