Flagging Misinformation and Disinformation

Introduction

This paper surveys research in detecting misinformation and disinformation. These two terms are distinguished by the goal of the author or publisher: Misinformation, although demonstrably false, is not created with the intent to cause harm. Disinformation, by contrast, is false information created and disseminated with an explicit intent to inflict damage – reputational, emotional, physical, or other – on an individual, group, or organization. Some authorities[1] define a third category, mal-information, which uses information based on reality to inflict harm, often by taking such information out of its original context. This third category will not be considered here.

For convenience, this paper will use "MIDI" as shorthand for misinformation and disinformation. MIDI itself is nothing new: often-cited antecedents include the "yellow journalism of William Randolph Hearst and Joseph Pulitzer at the end of the 19th century, sensationalist tabloids like *Weekly World News*, *National Enquirer* and *Globe*, and the "Great Moon Hoax" of 1835[2]. Given MIDI's persistence throughout human history, it is worthwhile to quantify MIDI's impact before reactively investing in systems to detect and oppose it. The cost of COVID MIDI has been estimated at $50 to $300 million per day[3]. On the user side, one study found that individuals would be willing to pay over $9 per year for a virtual, public-run fact-checking system[4], with younger persons being willing to pay more. One may conclude there is a recognized and unsatisfied need.

The popularized but imprecise expression "fake news" can be said to encompass both misinformation (e.g., journalistic mistakes) and disinformation, but it also has been applied to other content never intended as news. Hangloo and Arora[5] propose four additional categories:

- Propaganda – content created and propagated by a political entity to influence political view, like that deployed in America during both World Wars[6].
- Rumors / Hoaxes – inaccurate or fabricated content claimed to have been verified by traditional news outlets.
- Parody / Satire – content that typically mimics mainstream news media for humorous or dramatic effect. Examples include *Saturday Night Live's* "Weekend Update", Jon Stewart's *The Daily Show*, and Orson Welles' *War of the Worlds* radio broadcast in 1938.
- Clickbait – sensational headlines intended to draw attention and direct users to a different website.

[1] "Mis-, Dis-, and Malinformation". *United States Cybersecurity & Infrastructure Security Agency*, 20 Oct. 2022, https://www.cisa.gov/mdm. See also "Misinformation, Disinformation and Mal-Information", *Media Defence*, retrieved 20 Oct. 2022, https://www.mediadefence.org/ereader/publications/introductory-modules-on-digital-rights-and-freedom-of-expression-online/module-8-false-news-misinformation-and-propaganda/misinformation-disinformation-and-mal-information/.

[2] Thornton, Brian. "The moon hoax: Debates about ethics in 1835 New York newspapers." *Journal of Mass Media Ethics 15.2* (2000): 89-100.

[3] Bruns, Richard, et al., "COVID-19 Vaccine Misinformation and Disinformation Costs an Estimated $50 to $300 Million Each Day", *Baltimore, MD: The Johns Hopkins Center for Health Security*, https://www.centerforhealthsecurity.org/our-work/publications/covid-19-vaccine-misinformation-and-disinformation-costs-an-estimated-50-to-300-million-each-da (2022)

[4] Jo, Hanseul, et al., "Estimating cost of fighting against fake news during catastrophic situations". *Telematics and Informatics* 66, (2022): 101734.

[5] Hangloo, Sakshini and Bhavna Arora, "Fake News Detection Tools and Methods". *arXiv preprint ArXiv:2112:11185* (2021).

[6] Krause, Nicole M., et al., "Fake News: A New Obsession with an Old Phenomenon", *Journalism and Truth in an Age of Social Media* (2019), 58-78.

Hangloo and Arora also suggest that the proliferation of all six forms of content has made "challenging for regular Internet users to distinguish between real and fake news content," although they provide no citations in support of this statement. Many other authors, however, have pointed out a crucial factor: Historically, users selected their news source before reading the articles therein, with at least some knowledge of the source's biases and limitations. Algorithmic news feeds – including but not limited to social media – mix articles from many different sources of varying reputation, typically without advertising the origin of each article. Deprived of this important indicator of reputability, curious users have no choice but to click on a headline.

Cognitive Bias of Users

Nevertheless, no user is forced to consume misinformation or disinformation. For their own reasons, individuals click on headlines, real and fake, that appear "relevant" to them. CS 410 lectures on information retrieval have emphasized, more than once, that users are the ultimate judges of relevance. If users can reliably detect MIDI for themselves, it is more difficult to question their choice to consume it. 84% of Americans believe they can detect fake news, and 39% are "extremely confident" of their ability to do so[7].

However, Moravec et al. found that only 17% of the 80 participants were better than random chance at judging credibility of story headlines in a social media setting[8]. Using time-frequency analysis of EEG data to measure cognitive activity, these researchers found headlines agreeing with test subjects' *a priori* opinions triggered the greatest amount of cognitive activity, and a high likelihood of accepting the headline as truth. Headlines disagreeing with subjects' existing opinions received substantially less consideration.

This confirmation bias results from how humans process information: Cognitive "System 1" continuously and unconsciously searches long-term memory, supplying in less than one second any matches with new incoming information. System 2 is deliberate, conscious, and is at work when humans make more considered evaluations of new information[9]. However System 2 requires more effort and is easily overwhelmed, so it is invoked only under conditions of significant uncertainty. Even when System 2 has been engaged, its conclusions are influenced by System 1 results that already are in working memory.

Therefore, a MIDI flag or marker by itself is unlikely to create sufficient cognitive dissonance to trigger System 2. To quote Moravec, "flagging articles as fake was not effective; it had no effect on users' beliefs". This clai is supported by the failure of Facebook's "disputed" flag, introduced in 2017 and withdrawn in 2018 for having little to no effect.

Spezzano[10] cites a number of studies asserting that automated MIDI detectors generally perform better than humans at separating truth from falsehood. The difference is less significant when both humans and computers have access to the same meta-data as well as text excerpts. For both humans and computers, performance is worst when only excerpt text is considered. For humans to assess meta-data, however, may require System 2.

[7] Barthel, Michael, Amy Mitchell, and Jesse Holcomb. "Many Americans believe fake news is sowing confusion", *Pew Research Center*, 2016, https://policycommons.net/artifacts/618138/many-americans-believe-fake-news-is-sowing-confusion/1599054/.

[8] Moravec, Patricia, Randall Minas, and Alan R. Dennis. "Fake news on social media: People believe what they want to believe when it makes no sense at all". Kelley School of Business research paper 18-87 (2018). Another study found that only 26% of Americans could correctly judge news veracity. See Amy Mitchell, Jeffrey Gottfriedd, Michael Barthel, and Nami Sumida, "Distinguishing Between Factual and Opinion Statements in the News", *Pew Research Center,* 2018.

[9] Kahneman, Daniel. Thinking, fast and slow. Macmillan, 2011.

[10] Spezzano, Francesca, Anu Shrestha, Jerry Alan Fails, and Brian W. Stone, "That's Fake News! Reliability of News When Provided Title, Image, Source Bias & Full Article." *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021): 1-19.

Similar to Moravec, Spezzano concludes that that labeling a news item as "disputed" is insufficient. Automated systems that call the user's attention to relevant credibility markers are most likely to receive consideration from System 1. Such credibility markers include: emotionality or neutrality of headline or title text, presence or absence of statistics and supporting evidence, presence or absence of quotes and cited sources, and relevant elements of images associated with the news item.

The above studies indicate that flagging MIDI can help users make more informed judgments, but only if human factors are considered in how flags are presented. An effective flag is one that can act on both System 1 and System 2 cognitive processes[11].

Detection Techniques

For this technical review, I read a number of different papers. The following subset is intended as a more or less representative cross-section of recent research. Each study is summarized and followed by a brief commentary.

[A] Kim and Ko[12] propose a two-stage method that, first, builds a context graph of all sentences within a news document where each node represents a sentence and each edge weight represents the "relational strength" between two connected sentences. Second, a summarization technique is deployed to identify subject information and rank sentences that contain the most information about the subject. Using a training data set of 6,452 documents and a test set of 134 documents, the authors claim a 3.72% performance improvements compared to a BERT model, and up to 12.68% compared to various other models.

Because this technique acts only upon the full text of news items, its ability to influence users in real-world settings is debatable. The scalability of this technique also is open to question, given the speed with which truthful and false news items can be generated in 2022.

[B] Researchers in Spain[13] have attempted to address the problem of COVID-19 MIDI by creating an authoritative source for validating information. using open-source intelligence (OSINT) tools. Specifically, this group created MedOSINT, a compendium of COVID-19 information from official bulletins such as those from the World Health Organization and governmental agencies. The compiled data are run through a Case-Based Reasoning (CBR) system that purports to explain why a particular news item is classified as true or false and presented through a web interface.

The explanatory capability of MedOSINT could be a real benefit, in terms of creating a more educated user community. However, users must navigate to MedOSINT to use it. Also, the amount of human effort to create and maintain MedOSINT appears to be considerable, likely making this technique unsuitable for highly dynamic news environments such as elections.

[C] Shim et al. recognize several limitations of a text-only approach, beyond the human factors mentioned above. First, the writing style of fake news has evolved to more closely resemble genuine news. Second, news documents (particularly those on social media) tend to be short and thus are less amenable to detailed text analysis, such as sentiment analysis and syntactic analysis. Third, automatic classifiers trained on previous news documents are less

[11] Moravec, Patricia, Antino Kim, and Alan Dennis. "Flagging fake news: System 1 vs. System 2." *39th International Conference on Information Systems*, San Francisco (2018).

[12] Kim, Gihwan, and Youngjoong Ko. "Effective fake news detection using graph and summarization techniques." *Pattern Recognition Letters 151* (2021): 135-139.

[13] Monterrubio, Sergio Mauricio Martinez, et al. "Coronavirus fake news detection via MedOSINT check in health care official bulletins with CBR explanation: The way to find the real information source through OSINT, the verifier tool for official journals." *Information Sciences 574* (2021): 210-237.

likely to work well with new documents, as news topics change constantly and frequently. Fourth, text-based models generally are language-dependent and cannot be readily applied to languages other than English, where most research has occurred.

For these reasons, context-based detection is receiving more attention. “Context” can include:

- User information, e.g., number of followers and posts by those who distribute news, both actual and fake.
- Network properties, e.g., density, clustering and diffusion within networks such as Twitter and Facebook
- Source link properties, e.g., is the source link on a “whitelist” of trusted authorities or a “blacklist” of known purveyors of MIDI[14]. Are there classes of links, such as those ending in “.gov”, that can be presumed to be trustworthy?

Shim and his colleagues propose an “self-supervised” deep learning model based on link information extracted from search results[15]. Starting from the “word2vec” method of clustering words according to the similarity of their contexts, Shim’s “link2vec” method vectorizes link information in search results. These features are run through different machine learning algorithms, including Artifical neural Network and Support Vector Machine. Using an English-language dataset with 14,000 fake news domains and 23,000 genuine news domains, link2vec combined with Support Vector Machine achieved 93.1% accuracy, much higher than text-only and whitelist models used for comparison. Using a comparable Korean-language dataset, link2vec combined with Artificial Neural Network achieved only 81.9% accuracy, which still was better than text-only and whitelist models. The difference between English and Korean results is attributed to Korean new organizations’ propensity to print articles without extensive fact-checking.

[D] Yang et al.[16] propose an unsupervised generative approach that uses Bayes rule to calculate conditional relationships among the veracity of news documents, the credibility of individual users, and users’ opinions about documents. They address the challenge of conflicting and unreliable information by dividing the user community into two groups: well-known verified users whose tweets are popular and receive a lot of attention, and unverified users whose tweets receive little attention. A Monte Carlo technique known as Gibbs sampling is used to generate conditional probability distributions.

The authors implicitly assume that verified users have higher credibility in differentiating between fake news and real news. Unverified users, by contrast, are assumed to have opinions influenced primarily by verified users and by news documents. No empirical basis for these assumptions is presented. The top five verified users in this study are: ABC7 (New York) reporter Amy Hollyfield, Politico, senior correspondent for PolitiFact Lou Jacobson, the Washington Examiner, and Fox News. Only three of these five sources are rated by Media Bias Fact Check as having high credibility. This difference of opinion, combined with the unsupported assumptions about verified users, raise questions about the validity and extensibility of this research.

[14] For example, blacklists pertaining to COVID-19 are available from https://www.newsguardtech.com/wp-content/uploads/2020/04/FB-Superspreaders-Spreadsheet-20200423-1.csv and https://counterhate.com/wp-content/uploads/2022/05/210324-The-Disinformation-Dozen.pdf.

[15] Shim, Jae-Seung, Yunju Lee, and Hyunchul Ahn. "A link2vec-based fake news detection model using web search results." *Expert Systems with Applications 184* (2021): 115491.

[16] Yang, Shuo, et al. "Unsupervised fake news detection on social media: A generative approach." P*roceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. No. 01. 2019.

[E] Finally, Gaozhao[17] conducted an experiment in which 717 participants were randomly assigned to one of three groups. Group 1 say news items with flags from a professional fact-checker, group 2 saw news items with flags that had been crowd-sourced, and group 3 (the control group) saw news items with no flags at all. Thestudy focused on two questions: First, do flags persuade users that information is real or fake? Second, are flags from experts or flags from peers more persuasive?

The study showed that fact-checking flags reduced subjects' uncertainty about whether a news items was true or false. Flags also were able to influence subjects' judgment of whether a news item was true or false – even when true news items were deliberately marked as false and vice versa. However, the analysis does not indicate that flags cause users to think more critically. Rather, it appears that users' System 1 cognition simply incorporated flags as another data point. Supporting this hypothesis, the study found insignificant differences between the perceived credibility of flags from professional fact-checkers and flags that were crowd-sourced.

Summary

This short paper touches only a small fraction of recent research. It is likely that any other existing studies, unexamined here, could shed additional light on the widespread issue of misinformation and disinformation. It does seem clear, however, that even the most sophisticated analytical and machine learning detection techniques will be valuable only if they can attract and engage users' attention.

[17] Gaozhao, Dongfang. "Flagging fake news on social media: An experimental study of media consumers' identification of fake news." *Government Information Quarterly 38.3* (2021): 101591.