Countering Digital Misinformation with Automatic Fact-Checkers

Sophia Campbell

Misinformation:

“Information that is not true” (Rulis 2024; Au 2022)

- Must be incorrect “based on the best available evidence from relevant experts at the time” (Vraga and Bode 2020)

- **Not** necessarily created or propagated as an intentional falsehood (Vasist 2023; Bradshaw 2024)

Challenges posed by Misinformation

- Positive correlation with societal ideological polarization and verbal/material conflict between citizens and governments (Vasist 2023; Rulis 2024)
  - Several international incidents of political violence attributed to widespread: misinformation: hate crimes in UK following Brexit, Jan 6th insurrection in US, anti-Muslim riots in Myanmar (Rulis 2024; Wardle and Derakhshan 2017).
- In cases of conspiracy theories, stronger belief of misinformation correlates with higher likelihood of supporting political violence (Enders et al. 2022)

Automatic Fact-Checkers (AFCs):

Digital tools that determine whether a claim is true (Guo, Schlichtkrull and Vachos 2021)

- Generally privately developed (Lee et al. 2013)
- Developed by use of training data, from which AFCs “learn” to recognize digital patterns common in pieces of misinformation (Akhtar 2023)
- May have web-scraping abilities to compare claims with data from other digital sources
- Can require degree of human oversight (Graves 2018)

Advantages

AFCs can:

- Accurately verify pieces of information faster and cheaper than human fact checkers (Quelle and Bovet 2024; Lim and Perrault 2023; Pathak, Shaikh, and Srihari 2020);

- Effectively impact users' belief of claims both the short and long term (Neilson and Graves 2020);

- Verify non-textual media (Akhtar et al. 2023)

Potential Negative Effects

- Could be impartial in detecting misinformation
  - Impartialities could result from biases in training data, selection of sources used to verify claims, or ideological biases in AFC creators (Unver 2023)

- Could misidentify as a result of attacks: “Planting” attacks—in which hackers tamper with verifying sources or flood the internet with false claims--can cause AFCs to incorrectly verify misinformation (Abdelnabi and Fritz 2023).

- Could cause Backfire Effect, in which correcting misinformation causes users to “double down” in believing it (Swire-Thompson et al. 2023).

Recommendations

- Increased Transparency: Privately-developed, publicly-used AFCs should disclose the external sources used to verify claims to the user. This would:
  - Show any selection bias in sources used to verify claims;
  - Identify sources that have been tampered with in case of planting attack;
  - (Maybe) assuage the backfire effect by providing evidence.

Sources (1/2)

- Abdelnabi, Sahar, and Mario Fritz. "Fact-Saboteurs: A Taxonomy of Evidence Manipulation Attacks Against Fact-Verification Systems." Proceedings of the 32nd USENIX Conference on Security Symposium (SEC '23), 2023, 6719–6736. https://www.usenix.org/conference/sec23/presentation/abdelnabi.
- Akhtar, Mubashara, et al. "Multimodal Automated Fact-Checking: A Survey." Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, 2023, pp. 5430–5448. https://aclanthology.org/2023.findings-emnlp.361/.
- Au, C.H., K.K.W. Ho, and D.K. Chiu. "The Role of Online Misinformation and Fake News in Ideological Polarization: Barriers, Catalysts, and Implications." Information Systems Frontiers 24 (2022): 1331–1354. https://doi.org/10.1007/s10796-021-10133-9.
- Bradshaw, Samantha. Disinformation and Identity-Based Violence. Stanley Center for Peace and Security, 2024. https://stanleycenter.org/wp-content/uploads/2024/10/Disinformation-and-Identity-Based-Violence-Bradshaw.pdf.
- Enders, Adam M., et al. "The Relationship Between Conspiracy Theory Beliefs and Political Violence." Misinformation Review, 2022. https://misinforeview.hks.harvard.edu/article/the-relationship-between-conspiracy-theory-beliefs-and-political-violence/
- Graves, Lucas. "Understanding the Promise and Limits of Automated Fact-Checking." Reuters Institute for the Study of Journalism, 2018. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2018-02/graves_factsheet_180226%20FINAL.pdf.
- Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos. "A Survey on Automated Fact-Checking." Transactions of the Association for Computational Linguistics 9 (2021): 1–21. https://neurips.cc/virtual/2024/poster/95091.
- Lee, Sian, Aiping Xiong, Haeseung Seo, and Dongwon Lee. "Fact-Checking Fact Checkers: A Data-Driven Approach." Harvard Kennedy School Misinformation Review 4, no. 5 (2023): 1–20. https://misinforeview.hks.harvard.edu/wp-content/uploads/2023/10/lee_fact-checking_fact_checkers_20231026.pdf.

Sources (2/2)

- Lewandowsky, Stephan, et al. "Misinformation, Disinformation, and Violent Conflict: From Iraq and the 'War on Terror' to Future Threats to Peace." American Psychologist 68, no. 7 (2013): 487–501. https://doi.org/10.1037/a0034515.
- Lim, Gionnieve, and Simon T. Perrault. "XAI in Automated Fact-Checking? The Benefits Are Modest and There's No One-Explanation-Fits-All." arXiv preprint, 2023. https://arxiv.org/pdf/2308.03372.
- Nielsen, Rasmus Kleis, and Lucas Graves. "News You Don't Believe: Audience Perspectives on Fake News." Reuters Institute for the Study of Journalism, 2020. https://reutersinstitute.politics.ox.ac.uk/our-research/news-you-dont-believe-audience-perspectives-fake-news.
- Pathak, Archita, Mohammad Abuzar Shaikh, and Rohini Srihari. "Self-Supervised Claim Identification for Automated Fact-Checking." In Proceedings of the 17th International Conference on Natural Language Processing (ICON), 213–227. Indian Institute of Technology Patna, 2020. https://aclanthology.org/2020.icon-main.28.
- Quelle, Dorian, and Alexandre Bovet. "The Perils and Promises of Fact-Checking with Large Language Models." Frontiers in Artificial Intelligence 7 (2024): 1341697. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2024.1341697/full.
- Rulis, M. "The Influences of Misinformation on Incidences of Politically Motivated Violence in Europe." The International Journal of Press/Politics 0 (2024). https://doi.org/10.1177/19401612241257873.
- Swire-Thompson, Bridget, Nathaniel Miklaucic, John P. Wihbey, David Lazer, and Julie DeGutis. "The Backfire Effect after Correcting Misinformation Is Strongly Associated with Reliability." Journal of Experimental Psychology: General 151, no. 7 (2022): 1655–1665. https://doi.org/10.1037/xge0001
- Unver, H. Akin. "Emerging Technologies and Automated Fact-Checking." SSRN Electronic Journal, 2023. https://edam.org.tr/Uploads/Yukleme_Resim/pdf-28-08-2023-23-40-14.pdf.
- Vasist, Pramukh Nanjundaswamy, et al. "The Polarizing Impact of Political Disinformation and Hate Speech: A Cross-country Configural Narrative." Information Systems Frontiers 26 (2023): 663–688. https://doi.org/10.1007/s10796-023-10390-w.
- Vraga, Emily K., and Leticia Bode. "Defining Misinformation and Understanding Its Bounded Nature: Using Expertise and Evidence for Describing Misinformation." Political Communication 37, no. 1 (2020): 136–144.
- Wardle, Claire, and Hossein Derakhshan. Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making. Council of Europe Report, 2017. https://firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf.