[409] N. McIntyre, R. Bradbury, and B. Perrigo, ‘Behind TikTok’s boom: A legion of traumatised, $10-a-day content moderators’, The Bureau of Investigative Journalism (en-GB). Dec 2023.

[410] F. Hibaq, ‘Diary of a TikTok moderator: “we are the people who sweep up the mess”’, The Guardian. Dec 2023.

[411] D. Kapellmann Zafra, R. Serabian, S. Riddell, and N. Brubaker, ‘How to understand and action Mandiant’s intelligence on information operations’, Mandiant. Oct 2022.

[412] Microsoft Incident Response and Microsoft Threat Intelligence, ‘Dev-0537 criminal actor targeting organizations for data exfiltration and destruction’, Microsoft Security Blog. Sep 2023.

[413] B. Nimmo, ‘Meta’s adversarial threat report, fourth quarter 2022’, Meta. Feb 2023.

[414] S. Srinivasan, ‘The global disinformation index’, GDI.

[415] I. Lapowsky, ‘Inside the research lab teaching Facebook about its trolls’, Wired. Aug 2018.

[416] A. Schiffrin, ‘sing AI to combat MIS/disinformation – an evolving story’, Tech Policy Press. Oct 2023.

[417] ‘How it works’, Graphika.

[418] A. Storey, ‘Our ongoing work to fight misinformation online’, Google. Oct 2023.

[419] OpenAI, ‘Using machine learning to reduce toxicity online’, Perspective.

[420] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. Raji, and T. Gebru, ‘Model Cards for Model Reporting’, Proceedings of the Conference on Fairness, Accountability, and Transparency. 2019.

[421] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, ‘The pushshift reddit dataset’, Proceedings of the international AAAI conference on web and social media. 2020.

[422] K. Yang, E. Ferrara, and F. Menczer, ‘Botometer 101: Social bot practicum for computational social scientists’, Journal of Computational Social Science. 2022.

[423] Democratics House Permanent Select Committee on Intelligence, ‘Exhibit B.’ May 2018.

[424] J. Habgood-Coote, “The term “fake news” is doing great harm,” The Conversation. July 2018.

[425] J. Jackson, “What are influence operations and why are we investigating them?”, The Bureau of Investigative Journalism. July 2023.

Appendix

1. Annotation Guide

Our guide for annotating papers follows below:

Target selection:
- What is the stated target of detection?
- What is the actual target of detection?
- Is the detection method *directly* performing fact-checking and misinformation detection, or is it detecting *proxy signals* instead (e.g., emotional language, topics associated with rumoring narratives)?

Dataset curation:
- Note dataset size, source, and temporality.
- Note *information scope*: what is the “operative unit” of checkable information? **Where this might become more complicated:** some multimodal methods, particularly network-based ones, usually accept some combination of user- and post-scoped data, *but only make a final decision about the veracity of one of those things.*
- Note *information domain*: does the data set coalesce around a topic / a set of topics? Are these pre-defined by the study authors?
- Fine to name multiple scopes if one info scope is a vital parameter / input to the model in its determination of the veracity of another; e.g., a model that performs network analysis of social media posts that share a link to a specific news article; the model might make a final determination about the truthfulness of the news article, but propagation patterns + social media posts were a crucial element of that decision.

Feature set selection:
- How were features selected?
- Are features explainable and accessible? Were they handcrafted? Were they generated by unsupervised models?
- Do positive signals indicate evidence of actually suspicious behavior?
- Do authors account for feature noise?

Test-train split:
- Note how/if this split was observed, and the percentage of the total data set assigned to test and training sets.
- Note any overlap between both data sets (ideally, none).
- Note any evidence of temporal leakage (future data used to make a prediction about past events).

Cross-validation performed:
- Note size and number of folds.
- Note presence of any development data sets (what percentage of total data set was this?).
- Note presence of any hyperparameter tuning during cross-validation process – are these parameters published in-text?

Out-of-sample testing:
- Note presence, if any.
- What data sets were used for this?
- Note any evidence of leakage during this process.

Third-party references:
- Particularly for claim-based studies that require some form of ground-truth reference for fact-checking, what is the ground truth reference employed by the study?
- Is this ground truth site reliable and up-to-date?
- What kind of taxonomy does the ground-truth reference employ for labeling true / false statements?
- What baked-in biases or slants might exist in the ground-truth reference’s data set (e.g., reference only labels political news sites; ground-truth reference is run by a known conservative organization)?

Model choice:

Evaluation:
- Note dataset(s) used for evaluating classifier performance.
- How does the evaluation dataset(s) compare to testing and training datasets in terms of size, sourcing, contents, and temporality?

- How do authors account for distribution shifts in rumoring topics or activity types? Is there a road map for iterative updates to their method?

2. Commercial fact-checking services

We include here a brief market survey of commercial and LLM-powered fact-checking and IO detection services. In general, these services fall into five categories: 1) media fact-checking organizations; 2) brand safety and suitability services; 3) trust & safety operations at large social media platforms, 4) threat detection operations, and 5) analytics organizations unaffiliated with a media outlet that offer research capacity to governments and businesses. We define each service category and (with the exception of the first category, which comprises human media workers and fact-checkers) discuss automated content moderation operations deployed by three prominent exemplars within each service category. In general, in instances where such information is made available, we observe that at-scale content moderation businesses *at least* employ human-labeled datasets to train classifiers, and some retain subject-area experts to adjudicate complex moderation decisions. On social media platforms, in particular, human moderators and automated systems appear to work hand-in-hand: automated systems surface potentially misinformative content that receives final verification from a human moderator. For IO detection, specialized knowledge (pertaining to specific geographies, languages, or political climates) is often invoked.

Brand safety and suitability companies. B2B companies that detect categories of potentially harmful speech on websites where ads might appear. Advertisers wishing to protect "brand safety" contract with these services to ensure that their ads do not appear alongside problematic content. The Global Alliance for Responsible Media (GARM) is the standards-setting body for brand safety and suitability companies [403].

- *Zefr*, a GARM member company, deploys AI to detect material that falls within predefined subcategories of problematic content (e.g., explicit content, misinformation, spam). In a press release for Zefr's acquisition of an AI-driven content moderation company (AdVerif.ai) from 2022, the company disclosed that AdVerif.ai is "powered by fact-checking data from more than 50 IFCN-certified organizations around the globe" [404]—that is, AdVerif.ai trains its models on labeled datasets produced by (human) IFCN affiliates.
- *DoubleVerify*, a GARM member company, "uses sophisticated approaches that rely on a combination of AI and comprehensive human review" [405]. According to the company's documentation, human assessors (a "semantic science team") evaluate site infrastructure and contents; AI is used to scale their assessments.
- *Integral Ad Science* (IAS), a GARM member company, deploys AI to detect low-quality sites via infrastructure features. The company's data sources, and deployment methodology were not immediately evident upon web search; IAS recently announced a new partnership with Meta for ad placement management on Facebook [406].

Trust & safety operations. In-house content moderation teams at large social media platforms.

- *Facebook* partners with IFCN affiliates to perform third-party manual checking of possibly misinformative content; first-line automated methods detect potentially harmful speech and surface near-duplicates of known problematic image (SimSearchNet++) and text content [339], [352], [407].
- *Twitter* has deployed a crowd-sourced annotations platform called Community Notes (formerly Birdwatch) since 2021 [408].
- *TikTok* employs thousands of content moderators across the globe who "work alongside automated moderation systems" [409], [410].

Threat intelligence services. At-scale detection of advanced persistent threats, foreign influence operations, and other cyberattacks oftentimes perpetrated by nation state actors.

- *Mandiant* strongly implies the use of hybrid detection methods, and disclaims that "defenders must constantly explore different techniques and leverage both subject matter expertise and technical capabilities to filter and uncover malicious activity") [411].
- *Microsoft Threat Intelligence* strongly implies the use of hybrid detection methods; in a report from September 2023, MTI cites the work of in-house "Microsoft Security teams" which are tracking an advanced social engineering attack [412]. Other details—including possible use of automated methods—are undisclosed.
- *Facebook Coordinated Inauthentic Behavior* reports share quarterly updates about Meta's takedown of coordinated activities across its platforms *and* others, including local news outlets. In a report from February 2023, Meta describes a CIB network in Serbia that used local news media to create the impression of grassroots support for the Serbian Progressive Party; while the nature of the detection methodology is unspecified, the complexity and geographic specificity of the CIB described suggest that specialists with country-level expertise were likely consulted [413].

Analytics firms. For- and non-profit organizations that offer checking services and research capacity to governments and businesses.

- *The Global Disinformation Index (GDI)* "reviews news domains based on various metadata and computational signals." Content, however, is manually reviewed by a "country expert," who analyzes a random sample of 10 articles from a news site to determine veracity [414].
- *DFRLabs (Digital Forensic Research Lab)* has disclosed that it employs human subject-area experts, and primarily addresses technology and policy issues pertaining to global and international affairs. In 2018, Facebook contracted its services to detect online trolls [415].

TABLE 3. Replication analysis of Baly et al.: Dropout(−) and feature importance(+) analyses of subsets of Baly et al.'s EMNLP18 dataset, stratified by political leaning and credibility. Most (secondmost) performant feature, as determined by its contribution to overall classifier accuracy on the full feature set, is highlighted in darker (lighter) hues. Fact and bias classification task performances are reported in the top and bottom halves of the table, respectively.

Dataset (size)	All features	articles		traffic		twitter		wikipedia		url	
		−	+	−	+	−	+	−	+	−	+
Full corpus (1066)	0.654	0.631	0.644	0.654	0.508	0.648	0.550	0.627	0.606	0.638	0.533
Med. corpus (400)	0.623	0.608	0.630	0.620	0.488	0.635	0.500	0.590	0.588	0.623	0.495
Small corpus (250)	0.636	0.632	0.596	0.632	0.524	0.608	0.512	0.588	0.536	0.624	0.516
Left bias (398)	0.691	0.683	0.671	0.688	0.668	0.686	0.628	0.678	0.683	0.678	0.636
Center (263)	0.913	0.810	0.890	0.913	0.700	0.924	0.741	0.920	0.776	0.890	0.635
Right bias (405)	0.279	0.267	0.252	0.279	0.173	0.286	0.230	0.274	0.205	0.272	0.121
Full corpus (1066)	0.569	0.523	0.595	0.569	0.399	0.580	0.440	0.552	0.538	0.577	0.373
Med. corpus (400)	0.563	0.517	0.580	0.560	0.420	0.578	0.478	0.585	0.545	0.568	0.360
Small corpus (250)	0.456	0.424	0.560	0.452	0.364	0.500	0.408	0.400	0.496	0.444	0.436
Low cred. (256)	0.590	0.516	0.633	0.590	0.641	0.629	0.445	0.609	0.633	0.590	0.473
Mixed cred. (268)	0.407	0.340	0.474	0.407	0.0522	0.414	0.258	0.362	0.276	0.414	0.198
High cred. (542)	0.349	0.336	0.408	0.349	0.255	0.369	0.271	0.341	0.316	0.351	0.218

- *Graphika Labs* leverages network analysis to identify influence operations online. On its own website and in the popular press, Graphika has disclosed that it uses AI to map online networks and trace information flows [416], [417].

LLM-driven detection. A few LLM-powered detection methods have been discussed in the popular press, including those advertised by Google [418] and OpenAI [355], but these deployments appear to be mostly experimental, or have required additional adjudication from human moderators. OpenAI in particular has advertised content moderation tools that address misinformation-adjacent tasks, such as toxic speech detection [419]. Misinformation and toxic speech detection are not equivalent tasks, however, and the latter is narrowly defined in the Perspective training data documentation as a four-way classification task (the four class labels are "profanity/obscenity," "identity-based negativity," "insults," and "threatening" language).