[429] David Snelling. *Xperia XZ4 release this month - five things every Sony fan should K...* Feb. 2019. URL: https://www.express.co.uk/life-style/science-technology/1084439/Sony-Xperia-XZ4-release-date-price-specs-mobile-world-congress.

[430] Charles F Bond and Bella M DePaulo. *Accuracy of deception judgments*. 2006. URL: https://pubmed.ncbi.nlm.nih.gov/16859438/.

[431] OpenAI. *ChatGPT*. URL: https://openai.com/product/chatgpt.

[432] Ina Fried. "OpenAI touts GPT-4 for content moderation". In: (2023).

[433] Matthew R DeVerna et al. "Artificial intelligence is ineffective and potentially harmful for fact checking". In: *arXiv preprint arXiv:2308.10800* (2023).

[434] James Vincent. *OpenAI isn't doing enough to make CHATGPT's limitations clear*. May 2023. URL: https://www.theverge.com/2023/5/30/23741996/openai-chatgpt-false-information-misinformation-responsibility.

[435] Shirin Ali. *Facebook's formula prioritized anger and ended up spreading misinformation*. Oct. 2021. URL: https://thehill.com/changing-america/enrichment/arts-culture/578724-5-points-for-anger-1-for-a-like-how-facebooks/.

[436] Wall Street Journal staff. *The Facebook Files*. Oct. 2021. URL: https://www.wsj.com/articles/the-facebook-files-11631713039.

[437] Devin Coldewey. *Deconstructing "the Twitter files"*. Jan. 2023. URL: https://techcrunch.com/2023/01/13/deconstructing-the-twitter-files/.

Appendix

A 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.

Media fact-checking. Human fact-checkers and content moderators affiliated with news outlets, or who work as freelance fact-checkers. *The International Fact-Checking Network (IFCN)* is a professional network of media workers and fact-checkers; IFCN is also the de facto standards setting body for media fact-checking, and maintains a fact-checking code of ethics [404]. In general, media fact-checking organizations with IFCN affiliations are established news organizations, non-profits, and watchdog organizations that employ human journalists and fact-checkers. Furthermore, (human) fact-checkers can receive IFCN compliance certificates after passing a qualifying exam.

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 [215].

- *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" [216]—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" [244]. 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 [241].

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 [236, 419, 436].
- *Twitter* has deployed a crowd-sourced annotations platform called Community Notes (formerly Birdwatch) since 2021 [247].
- *TikTok* employs thousands of content moderators across the globe who "work alongside automated moderation systems" [237, 246].

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") [218].
- *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 [240]. 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 [235].

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 [219].
- *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 [238].
- *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 [231, 233].

LLM-driven detection. A few LLM-powered detection methods have been discussed in the popular press, including those advertised by Google [230] and OpenAI [432], 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 [232]. 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).

Paper		① Target	② Dataset	③ Model	④ Features				⑤ Performance
Work	Scope				Textual	Network-based	Author-based	Infrastructural	Accuracy/AUROC
1. Ajao et al. [65]	C A	Sentiment (**A.i**)	PHEME [**pheme_2018**]	LSTM, DT, RF, SVM	●				0.86 (Acc.)
2. Abulldah-All-Tanvir et al. [66]	C A N	Content (**C.i**)	Twitter (API)	NB, RNN, LSTM, SVM, Logit	●	●			0.89 (Acc.)
3. Bhutani et al. [63]	C A	Content (**C.i**); sentiment (**A.i**)	Twitter (API), PolitiFact [55]	Naive Bayes, RF	●				0.60 (AUC)
4. Bozarth et al. [60]	C	Contents (**C.i**)	PolitiFact [55], Daily Dot, Zimdars, MBFC	LDA	●				n/a
5. Ciampaglia et al. [62]	C N	Shortest path search (**C.ii**)	DBpedia	kNN, RF	●				0.97 (AUC)
6. Cui et al. [101]	C A U	Content (**C.i**); sentiment (**A.i**); user response (**U.ii**)	PolitiFact[55], GossipCop[424]	KNN, SVM, CSI [111], RMSprop	●				0.82 (F1)
7. Debnath et al. [58]	C	Content (**C.i**)	LIAR [54]	CNN	●				0.27 (Acc.)
8. Dey et al. [332]	C A	Content (**C.i**); sentiment (**A.i**)	Twitter (API)	Clustering (kNN)	●				0.67 (Acc.)
9. Galitsky et al. [312]	C	Content (**C.i**)	Amazon reviews	Parse thicket	●				0.81 (Prec.)
10. Glockner et al. [92]	C	Content (**C.i**)	PolitiFact[55], Snopes[155], MultiFC	CNN, DNN	●				0.58 (Acc.)
11. Gordon et al. [268]	C A	Content (**C.i**); source rep (**U.i**)	Credibility-Factors2020	SVD	●	●			0.63 (Acc.)
12. Gupta et al. [100]	C A N	Content (**C.i**); stance (**A.iii**)	Twitter (API)	SVM	●	●			0.60 (Agreement)
13. Hassan et al. [88]	C	Content (**C.i**); checkability (**C.iii**)	NBA, weather datasets	Frequency	●				n/a
14. Jain et al. [381]	C	Content (**C.i**); sentiment (**A.i**)	Twitter (API)	Gensim/TextBlob	●				0.77 (Acc.)
15. Jiang et al. [188]	C	Content (**C.i**); ling./syntax (**C.ii**)	PolitiFact[55], Snopes[155]	SVM	●				0.81 (Acc.)
16. Karimi et al. [346]	C	Content (**C.i**)	LIAR[54]	LSTM, CNN	●				0.39 (Acc.)
17. Kartal et al. [93]	C	Content (**C.i**); checkability (**C.iii**)	Check That! dataset	Logit, SVM, RF	●				0.26 (MAP)
18. Kou et al. [90]	C	Content (**C.i**)	CoAID, CONSTRAINT	Knowledge graph	●				0.90 (Acc.)
19. Paudel et al. [99]	N A	Keyword detection (**A.ii**)	Abilov et al. dataset, Twitter (API)	AdaRank, ListNet, RF	●				0.79 (MAP)
20. Popat et al. [253]	C	Content (**C.i**)	PolitiFact[55], Snopes[155], NewsTrust	biLSTM, CNN	●				0.88 (AUC)
21. Shiralkar et al. [61]	C N	KG search (**C.ii**)	DBpedia	Knowledge graph	●				1.00 (AUC)
22. Shu et al. [64]	C U	Word encoding (**C.i**); user response (**U.ii**)	GossipCop [424], PolitiFact [55]	RNN/RMSprop, CSI [111], LSTM, CNN	●		●		0.93 (F1)
23. Tian et al. [97]	C U	Content (**C.i**); user response (**U.ii**)	Twitter15, Twitter16	CNN-biLSTM	●		●		0.82 (F1)
24. Zhang et al. [365]	C U	Content (**C.i**); user response (**U.ii**)	RumourEval, PHEME[**pheme_2018**]	biLSTM, Multitask, SVM, CNN,	●	●			0.89 (Acc.)
1. Afroz et al. [67]	A	Content (**C.i**); syntax (**A.i**)	Brennan-Greenstadt	SVM, J48 Decision Trees	●				0.97 (F1)
2. Ahmed et al. [68]	A	Syntax (**A.i**)	Twitter, Kaggle, Horne and Adali [69]	SVM	●				0.92 (Acc.)
3. Bourgonje et al. [70]	A	Stance (**A.ii**)	Fake News Challenge Data	Logit	●				0.90 (Acc.)
4. Brasoveanu et al. [71]	A C	Sentiment (**A.i**); keywords (**A.ii**)	LIAR[54]	CNN, LSTM, CN	●	●			0.64 (Acc.)
5. Della Vedova et al. [72]	A N	Content (**C.i**); virality (**N.iv**)	FakeNewsNet, Buzzfeed	Logit	●	●			0.82 (Acc.)
6. Horne et al. [69]	A	Syntax (**A.i**), headline (**A.ii**)	Buzzfeed, Burfoot & Baldwin	SVM	●				0.78 (Acc.)
7. Jabiyev et al. [274]	A W	Topic detection (**A.ii**); site cred. (**W.iv**)	Snopes[155], FactCheck, PolitiFact[55]	SVM, DT, RF	●		●		0.87 (Acc.)
8. Jadhav et al. [96]	A	Content (**C.i**); syntax (**A.i**)	LIAR [54]	DSSM/RNN	●				0.99 (Acc.)
9. Jin et al. [114]	A N U	(**A.ii**); (**N.i**); suspicious accounts (**U.i**)	Tweets; articles	n/a	●	●	●		0.87 (Prec.)
10. Kapusta et al. [252]	A	Sentiment & word freq. (**A.i**)	MBFC and custom	n/a	●				n/a
11. Kumar et al. [187]	A W U	(**A.i**); UI (**W.iii**); Author cred. (**U.i**)	20K Wiki Hoaxes	Random forest	●	●	●	●	0.87 (AUC)
12. Magdy et al. [98]	A	Content (**C.i**)	NYT Corpus [**nyt_corpus**], 100 Wikis	Pattern recog.	●				0.99 (Recall)
13. Monti et al. [139]	A N U	(**A.ii**); (**N.i**); suspicious accounts (**U.i**)	Tweets; articles	RNN/CNN	●	●	●		0.927 (AUC)
14. Nasir et al. [112]	A	Syntax (**A.i**)	ISOT [158]; FAKES [159]	RNN/CNN	●				0.99 (Acc.)
15. Perez-Rosas et al. [107]	A	Syntax (**A.i**)	FakeNewsAMT; Celebrity	SVM	●				0.74 (Acc.)
16. Potthast et al. [104]	A	Syntax, sentiment, readability (**A.i**)	Buzzfeed-Webis	Bag-of-words	●				0.46 (F1)
17. Reis et al. [167]	A	Syntax (**A.i**); contents (**C.i**); source rep. (**U.i**); timing (**N.ii**); URL (**W.i**)	BuzzFeed	GBM	●	●	●	●	0.85 (AUC)
18. Rubin et al. [382]	A	Syntax (**A.i**)	AMT	Clustering	●				0.67 (Agreement)
19. Ruchansky et al. [111]	A U	(**A.i**); Responses (**A.ii**); acct metadata (**U.i**)	Twitter/Weibo posts	RNN/LSTM	●		●		0.95 (Acc.)
20. Santos et al. [168]	A	Readability (**A.i**)	Fake.Br corpus	SVM	●				0.92 (Acc.)
21. Silva et al. [117]	A N	Topic detection (**A.ii**); propagation (**N.i**)	PolitiFact[55], GossipCop[424], CoAID	Clustering	●	●			0.88 (Acc.)
21. Singh et al. [304]	A	Syntax (**A.i**)	Kaggle Fake News	SVM	●				0.87 (Acc.)
22. Uppal et al. [263]	A	Discourse structure (**A.i**)	Buzzfeed, PolitiFact[55]	GRU, Dependency tree	●				0.74 (Acc.)
1. Cao et al. [73]	U N	Acct. cred. (**U.i**); prop. (**N.i**)	Tuenti social network	Louvain clustering		●	●		0.90+ (TP)
2. Danezis et al. [74]	U N	Acct. cred. (**U.i**); prop. (**N.i**)	LiveJournal data	Bayesian inf.		●	●		n/a*
3. Ezzeddine et al. [75]	U	Acct. behaviors (**U.ii**)	DATA	LSTM		●	●		0.91 (AUC)
4. Hamdi et al. [76]	U N	Account metadata (**U.i**); prop. (**N.i**)	CREDBANK	LDA, Bayes, Logit, SVM		●	●		0.99 (AUC)
5. Helmstetter et al. [77]	U N A	Acct metadata (**U.i**); post sharing data (**U.ii**)	Public site cred. lists	SVM, NB, DT, RF	●	●	●		0.936 (F1)
6. Jain et al. [381]	U A	Acct metadata (**U.i**); topic det. (**A.ii**)	Twitter (API)	Gensim, TextBlob	●		●		0.77 (Acc.)
7. Leonardi et al. [164]	U N A	Acct metadata (**U.i**); prop. (**N.i**)	CoAID	RF	●	●	●		0.81 (F1)
8. Saeed et al. [118]	U	User behavior (**U.ii**)	Reddit Pushshift; Reddit IRA trolls list	RF			●		0.98 (Acc.)
9. Sansonetti et al. [130]	U C	Acct metadata (**U.i**); acct activity (**U.ii**)	PolitiFact [55], Twitter (API)	LSTM-CNN, SVM, KNN	●				0.92 (Acc.)
10. Santia et al. [165]	U A	Source rep. (**U.i**); user response (**U.ii**); syntax (**A.i**)	BuzzFeed	SVM, RF, DT, NB	●	●	●		0.77 (Prec.)
11. Shu et al. [321]	U N A	User behavior (**U.ii**); prop. (**N.i**) syntax (**A.i**)	Buzzfeed, PolitiFact	Gibbs sampling	●	●	●		0.85+ (Acc.)
12. Vargas et al. [306]	N A	Prop. (**N.i**); topic det. (**A.ii**)	Twitter (API)	RF	●	●			0.98 (F1)
13. Wang et al. [169]	U N	User behavior (**U.ii**)	Renren data	SVM		●	●		0.99 (Acc.)
14. Yu et al. [170]	U N	User behavior (**U.i**); prop. (**N.i**)	LiveJournal, Friendster, DBLP accounts	Random route		●	●		n/a*
15. Yuan et al. [140]	U N	Acct metadata (**U.i**); timing (**N.ii**)	WeChat data	Clustering		●	●		0.90+ (Prec.)
16. Zhang et al. [129]	U N	User behavior (**U.ii**); prop. (**N.i**)	Twitter, Slashdot, Epinion	Graph cut		●	●		n/a
17. Zhou et al. [213]	U N	User suscept. (**U.i**); prop (**N.i**)	PolitiFact [55], BuzzFeed	SVM, KNN, NB, DT, RF		●	●		0.93 (Acc.)
1. Alizadeh et al. [78]	N A U	Propagation (**N.i**); syntax (**A.i**); acct metadata (**U.i**)	Twitter (API), Reddit IRA troll list	RF	●	●	●		0.70+ (F1)
2. Antoniadis et al. [79]	U A	Acct metadata (**U.i**); syntax (**A.i**)	Hurricane Sandy tweet dataset	J48, RF, KNN, Bayes	●			●	0.79 (Avg. Prec.)
3. Assenmacher et al. [80]	N A	Propagation (**N.i**); topic det. (**A.ii**)	Twitter (API)	Clustering	●	●			not reported
4. Buntain et al. [81]	N U A	Time (**N.ii**); acct metadata (**U.i**); sentiment (**A.i**)	CREDBANK, Buzzfeed	RF	●	●	●		0.65 (Acc.)
5. Castillo et al. [82]	U A	Syntax (**A.i**); user behavior (**U.ii**)	Twitter Monitor events	SVM, DT	●	●	●		0.874 (P)
6. Chen et al. [166]	N U A	Social graph (**N.iii**); syntax (**A.i**); user behavior (**U.ii**)	Weibo	RNN	●	●	●		0.92 (Acc.)
7. Guo et al. [367]	N C	Prop. (**N.i**); semantics (**C.i**)	Twitter, Weibo	LSTM	●	●			0.9 (Acc.)
8. Jin et al. [349]	N A	Propagation (**N.i**); stance (**A.iii**)	Sina Weibo posts	Clustering	●	●			0.84 (Acc.)
9. Liu et al. [143]	N	Propagation (**N.i**)	Weibo, Twitter15, Twitter16	RNN, CNN		●			0.897 (Acc.)
10. Ma et al. [144]	N	Propagation (**N.i**)	Kochina, Ma, Shu Twitter datasets	RNN/BiLSTM		●			0.75 (Acc.)
11. Magelinski et al. [171]	N U	Prop. (**N.i**); timing (**N.ii**); user behavior (**U.ii**)	Twitter (API)	–		●	●		– n/a
12. Nguyen et al. [210]	N A	Prop. (**N.i**); semantics (**A.i**)	Twitter, Weibo, PHEME [**pheme_2018**]	SVM, RNN, DT	●	●			0.970 (Acc.)
13. Pacheco et al. [131]	U N	Account metadata (**U.i**); timing (**N.i**)	Twitter (API)	Clustering		●	●		0.8+ (Prec.)
14. Ratkiewicz et al. [142]	N A U	Prop. (**N.i**); keywords (**A.ii**); user behavior (**U.ii**)	Twitter (API)	AdaBoost, SVM	●	●	●		0.96 (Acc.)
15. Sharma et al. [172]	N A U	Prop. (**N.i**); keywords (**A.ii**); user behavior (**U.ii**)	Twitter (API); Twitter IRA trolls list	GMM, Kmeans, NN	●	●	●		0.94 (AUC)
16. Tschiatschek et al. [324]	N U	Prop. (**N.i**); user rep. (**U.i**)	Facebook dataset	Bayes inf.		●	●		n/a
17. Zeng et al. [284]	N A	Prop. (**N.i**); stance (**A.iii**)	4.3K tweets (from API)	Logit, NB, RF	●	●			0.88 (Acc.)
1. Asr et al. [83]	W A	Source rep. (**W.i**); Syntax (**A.i**)	BuzzfeedUSE, Snopes, Rashkin, Rubin	CNN, SVM, NB	●	●		●	not reported
2. Baly et al. [5]	W A N	Source rep. (**W.i**); Site infra. (**W.ii**); (**A.i**)	MediaBiasFactCheck[84]	SVM	●			●	0.66 (Acc.)
3. Baly et al. [85]	W A N	Source rep. (**W.i**); Site infra. (**W.ii**); (**A.i**)	MediaBiasFactCheck[84]	SVM	●			●	0.7152 (Acc.)
4. Castelo et al. [86]	W A	Site infra. (**W.ii**); syntax (**A.i**)	Celebrity, US-Election2016	SVM, kNN, RF	●			●	0.86 (Acc.)
5. Chen et al. [87]	W A	Hosting infra. (URL) (**W.ii**); syntax (**A.i**)	PoliticalFakeNews	Clustering	●			●	0.97 (AUC)
6. Hounsel et al. [154]	W	Site infra. (**W.i**)	FactCheck, Snopes, PolitiFact, Buzzfeed	RF				●	0.98 (AUC)
7. Ribeiro et al. [132]	W U	Site infra. (**W.i**); User demog. (**U.i**)	Facebook API	Graph search		●	●	●	0.97 (PCC)

Table 4: **Focus corpus by scope and target.** Coding of a focus set of 87 papers, sorted by information scope. Values in parentheses in "Target" field correspond to highlighted subcategories presented in Section 4 (e.g., "C.i" denotes target (i) in "Claims," Section 4.1). If authors present evaluation results for multiple models, we underline the most performant model and record its corresponding performance score.

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

Table 5: 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.