7 Ethics considerations

Our author outreach survey design was approved by the [redacted] IRB. All statistics published from those communications are in aggregate, with no personal identifiers attached to individual author responses. Datasets used for this work were already publicly available or were obtained with permission from study authors.

8 Open science

Code and data used for our analyses, as well as our full paper corpus, are included in our online SI (available at https://anonymous.4open.science/r/sok_misinformation-41E8).

Acknowledgments

The authors would like to thank [redacted] for assisting with paper coding; [redacted] and [redacted] for productive conversations, and for reviewing multiple drafts of this manuscript; and [redacted] for retrieving data used for a replication study.

References

[1] Prateek Dewan and Ponnurangam Kumaraguru. "Towards automatic real time identification of malicious posts on Facebook". In: *2015 13th Annual Conference on Privacy, Security and Trust (PST)*. IEEE. 2015, pp. 85–92.

[2] Firoj Alam et al. "A Survey on Multimodal Disinformation Detection". en. In: *arXiv:2103.12541 [cs]* (Mar. 2021). URL: http://arxiv.org/abs/2103.12541 (visited on 03/29/2021).

[3] Darko Androcec. "Machine learning methods for toxic comment classification: a systematic review". In: *Acta Universitatis Sapientiae, Informatica* 12.2 (2020), pp. 205–216.

[4] Hee-Eun Lee et al. "Detecting child sexual abuse material: A comprehensive survey". In: *Forensic Science International: Digital Investigation* 34 (2020), p. 301022.

[5] Ramy Baly et al. "Predicting Factuality of Reporting and Bias of News Media Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (2018), pp. 3528–3539. URL: https://aclanthology.org/D18-1389/.

[6] Numa Dhamani et al. "Using Deep Networks and Transfer Learning to Address Disinformation". en. In: *arXiv:1905.10412 [cs]* (May 2019). URL: http://arxiv.org/abs/1905.10412 (visited on 06/03/2019).

[7] *Generalizing scaled misinformation detection*. 2023.

[8] Kelvin Chan, Barbara Ortutay, and Nicholas Riccardi. "Meta eliminates fact-checking in latest bow to Trump". In: *Associated Press* (2025).

[9] Daniel Arp et al. "Dos and Don'ts of Machine Learning in Computer Security". In: *31st USENIX Security Symposium (USENIX Security 22)*. Boston, MA: USENIX Association, Aug. 2022, pp. 3971–3988. ISBN: 978-1-939133-31-1. URL: https://www.usenix.org/conference/usenixsecurity22/presentation/arp.

[10] A. S. Jacobs et al. "AI/ML and Network Security: The Emperor has no Clothes". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. CCS '22. Los Angeles, CA, USA: Association for Computing Machinery, 2022.

[11] Sadegh M Milajerdi et al. "Holmes: real-time apt detection through correlation of suspicious information flows". In: *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2019, pp. 1137–1152.

[12] Samuel T King and Peter M Chen. "SubVirt: Implementing malware with virtual machines". In: *2006 IEEE Symposium on Security and Privacy (S&P'06)*. IEEE. 2006, 14–pp.

[13] Brett Stone-Gross et al. "Your botnet is my botnet: analysis of a botnet takeover". In: *Proceedings of the 16th ACM conference on Computer and communications security*. 2009, pp. 635–647.

[14] Shujaat Mirza et al. "Tactics, threats & targets: Modeling disinformation and its mitigation". In: *ISOC Network and Distributed Systems Security Symposium (NDSS)*. 2023.

[15] Xinyi Zhou and Reza Zafarani. "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities". In: *ACM Computing Surveys* 53.5 (Sept. 2020), 109:1–109:40. ISSN: 0360-0300. DOI: 10.1145/3395046. URL: https://doi.org/10.1145/3395046 (visited on 03/16/2023).

[16] Ray Oshikawa, Jing Qian, and William Yang Wang. "A Survey on Natural Language Processing for Fake News Detection". In: *arXiv:1811.00770 [cs]* (Nov. 2018). URL: http://arxiv.org/abs/1811.00770 (visited on 06/04/2019).

[17] Liang Wu et al. “Misinformation in Social Media: Definition, Manipulation, and Detection”. en. In: *ACM SIGKDD Explorations Newsletter* 21.2 (Nov. 2019), pp. 80–90. ISSN: 1931-0145, 1931-0153. DOI: 10.1145/3373464.3373475. URL: https://dl.acm.org/doi/10.1145/3373464.3373475 (visited on 03/16/2023).

[18] Axel Gelfert. “Fake News: A Definition”. en. In: *Informal Logic* 38.1 (Mar. 2018), pp. 84–117. ISSN: 0824-2577, 0824-2577. DOI: 10.22329/il.v38i1.5068. URL: https://ojs.uwindsor.ca/index.php/informal_logic/article/view/5068 (visited on 02/01/2019).

[19] Leonie Haiden and Jente Althuis. “The Definitional Challenges of Fake News”. In: *SBP-BRiMS 18*. June 2018. URL: http://sbp-brims.org/2018/proceedings/papers/challenge_papers/SBP-BRiMS_2018_paper_116.pdf (visited on 02/01/2019).

[20] David Klein and Joshua Wueller. *Fake News: A Legal Perspective*. en. SSRN Scholarly Paper ID 2958790. Rochester, NY: Social Science Research Network, Mar. 2017. URL: https://papers.ssrn.com/abstract=2958790 (visited on 02/01/2019).

[21] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. “Defining “Fake News”: A typology of scholarly definitions”. en. In: *Digital Journalism* 6.2 (Feb. 2018), pp. 137–153. ISSN: 2167-0811, 2167-082X. DOI: 10.1080/21670811.2017.1360143. URL: https://www.tandfonline.com/doi/full/10.1080/21670811.2017.1360143 (visited on 11/05/2018).

[22] Joshua Habgood-Coote. ““The term “fake news” is doing great harm””. In: *The Conversation* (July 2018). DOI: https://theconversation.com/the-term-fake-news-is-doing-great-harm-100406.

[23] Jasper Jackson. ““The term “fake news” is doing great harm””. In: *The Bureau of Investigative Journalism* (July 2023). DOI: https://www.thebureauinvestigates.com/stories/2023-07-27/what-are-influence-operations-and-why-are-we-investigating-them/.

[24] Sille Obelitz Søe. “Algorithmic detection of misinformation and disinformation: Gricean perspectives”. In: *Journal of Documentation* 74.2 (2017), pp. 309–332.

[25] Aldert Vrij. *Detecting lies and deceit: Pitfalls and opportunities*. John Wiley & Sons, 2008.

[26] Barbara G Amado, Ramón Arce, and Francisca Fariña. “Undeutsch hypothesis and Criteria Based Content Analysis: A meta-analytic review”. In: *The European Journal of Psychology Applied to Legal Context* 7.1 (2015), pp. 3–12.

[27] Carlos Carrasco-Farré. “The fingerprints of misinformation: how deceptive content differs from reliable sources in terms of cognitive effort and appeal to emotions”. In: *Humanities and Social Sciences Communications* 9.1 (2022), pp. 1–18.

[28] Junaed Younus Khan et al. “A benchmark study of machine learning models for online fake news detection”. en. In: *Machine Learning with Applications* 4 (June 2021), p. 100032. ISSN: 2666-8270. DOI: 10.1016/j.mlwa.2021.100032. URL: https://www.sciencedirect.com/science/article/pii/S266682702100013X (visited on 03/16/2023).

[29] Don Fallis. “A Functional Analysis of Disinformation”. In: *iConference 2014 Proceedings* (Mar. 2014). Publisher: iSchools. DOI: 10.9776/14278. URL: https://hdl.handle.net/2142/47258 (visited on 03/16/2023).

[30] Nadia K. Conroy, Victoria L. Rubin, and Yimin Chen. “Automatic deception detection: Methods for finding fake news”. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010082. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082 (visited on 03/16/2023).

[31] Karishma Sharma et al. “Combating Fake News: A Survey on Identification and Mitigation Techniques”. In: *ACM Transactions on Intelligent Systems and Technology* 10.3 (Apr. 2019), 21:1–21:42. ISSN: 2157-6904. DOI: 10.1145/3305260. URL: https://doi.org/10.1145/3305260 (visited on 03/16/2023).

[32] Md Rafiqul Islam et al. “Deep learning for misinformation detection on online social networks: a survey and new perspectives”. en. In: *Social Network Analysis and Mining* 10.1 (Sept. 2020), p. 82. ISSN: 1869-5469. DOI: 10.1007/s13278-020-00696-x. URL: https://doi.org/10.1007/s13278-020-00696-x (visited on 03/16/2023).

[33] Alim Al Ayub Ahmed et al. “Detecting Fake News using Machine Learning: A Systematic Literature Review”. In: *Psychology (Savannah, Ga.)* 58 (Jan. 2021), pp. 1932–1939. DOI: 10.17762/pae.v58i1.1046.

[34] Kai Shu et al. “Fake News Detection on Social Media: A Data Mining Perspective”. en. In: (), p. 15.

[35] Ammara Habib et al. “False information detection in online content and its role in decision making: a systematic literature review”. en. In: *Social Network Analysis and Mining* 9.1 (Sept. 2019), p. 50. ISSN: 1869-5469. DOI: 10.1007/s13278-019-0595-5. URL: https://doi.org/10.1007/s13278-019-0595-5 (visited on 03/16/2023).

[36] Jiawei Zhang et al. “Fake News Detection with Deep Diffusive Network Model”. In: *arXiv:1805.08751 [cs, stat]* (May 2018). URL: http://arxiv.org/abs/1805.08751 (visited on 06/04/2019).

[37] Nicollas R. de Oliveira et al. “Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges”. en. In: *Information* 12.1 (Jan. 2021). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 38. ISSN: 2078-2489. DOI: 10.3390/info12010038. URL: https://www.mdpi.com/2078-2489/12/1/38 (visited on 03/16/2023).

[38] Kai Shu et al. “Mining Disinformation and Fake News: Concepts, Methods, and Recent Advancements”. en. In: *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Ed. by Kai Shu et al. Lecture Notes in Social Networks. Cham: Springer International Publishing, 2020, pp. 1–19. ISBN: 978-3-030-42699-6. DOI: 10.1007/978-3-030-42699-6_1. URL: https://doi.org/10.1007/978-3-030-42699-6_1 (visited on 03/16/2023).

[39] Fernando Miró-Llinares and Jesús C. Aguerri. “Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a ‘threat’”. en. In: *European Journal of Criminology* (Apr. 2021), p. 1477370821994059. ISSN: 1477-3708. DOI: 10.1177/1477370821994059. URL: https://doi.org/10.1177/1477370821994059 (visited on 05/04/2021).

[40] Yimin Chen, Nadia K. Conroy, and Victoria L. Rubin. “News in an online world: The need for an “automatic crap detector””. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015), pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010081. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010081 (visited on 03/16/2023).

[41] Miriam Fernandez and Harith Alani. “Online Misinformation: Challenges and Future Directions”. en. In: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW ’18*. Lyon, France: ACM Press, 2018, pp. 595–602. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3188730. URL: http://dl.acm.org/citation.cfm?doid=3184558.3188730 (visited on 03/16/2023).

[42] Diego A Martin, Jacob N Shapiro, and Michelle Nedashkovskaya. “Recent Trends in Online Foreign Influence Efforts”. en. In: (), p. 34.

[43] Sushila Shelke and Vahida Attar. “Source detection of rumor in social network – A review”. en. In: *Online Social Networks and Media* 9 (Jan. 2019), pp. 30–42. ISSN: 2468-6964. DOI: 10.1016/j.osnem.2018.12.001. URL: https://www.sciencedirect.com/science/article/pii/S2468696418300934 (visited on 03/16/2023).

[44] Bin Guo et al. *The Future of Misinformation Detection: New Perspectives and Trends*. arXiv:1909.03654 [cs]. Sept. 2019. DOI: 10.48550/arXiv.1909.03654. URL: http://arxiv.org/abs/1909.03654 (visited on 03/16/2023).

[45] James Thorne and Andreas Vlachos. “Automated Fact Checking: Task Formulations, Methods and Future Directions”. en. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, NM, USA, Aug. 2018, pp. 3346–3359.

[46] Arkaitz Zubiaga et al. “Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads”. en. In: *PLOS ONE* 11.3 (Mar. 2016). Publisher: Public Library of Science, e0150989. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0150989. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0150989 (visited on 03/16/2023).

[47] Victoria L. Rubin, Yimin Chen, and Nadia K. Conroy. “Deception detection for news: Three types of fakes”. en. In: *Proceedings of the Association for Information Science and Technology* 52.1 (2015). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.1450520 pp. 1–4. ISSN: 2373-9231. DOI: 10.1002/pra2.2015.145052010083. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010083 (visited on 03/16/2023).

[48] L Padgham et al. ““CORE Rankings.”” In: (). DOI: https://www.core.edu.au/conference-portal.

[49] Miranda Wei et al. “‘SoK (or SoLK?): On the Quantitative Study of Sociodemographic Factors and Computer Security Behaviors’”. In: *USENIX Security* (2024). DOI: https://www.usenix.org/system/files/usenixsecurity24-wei-miranda-solk.pdf.

[50] Sarah Scheffler and Jonathan Mayer. “SoK: Content Moderation for End-to-End Encryption”. In: *Proceedings on Privacy Enhancing Technologies* (2023). DOI: https://petsymposium.org/popets/2023/popets-2023-0060.pdf.

[51] Noel Warford et al. "SoK: A Framework for Unifying At-Risk User Research". In: *IEEE S&P* (2021). DOI: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9833643.

[52] Steve Mollman. *OpenAI is getting trolled for its name after refusing to be open about its A.I.* Mar. 2023. URL: https://fortune.com/2023/03/17/sam-altman-rivals-rip-openai-name-not-open-artificial-intelligence-gpt-4/.

[53] Matteo Wong. *There was never such a thing as "open" ai*. Jan. 2024. URL: https://www.theatlantic.com/technology/archive/2024/01/ai-transparency-meta-microsoft/677022/.

[54] William Yang Wang. ""Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection". en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 422–426. DOI: 10.18653/v1/P17-2067. URL: http://aclweb.org/anthology/P17-2067 (visited on 05/29/2019).

[55] "PolitiFact". In: (). URL: https://www.politifact.com/.

[56] Yangqian Wang et al. "Learning Contextual Features with Multi-head Self-attention for Fake News Detection". en. In: *Cognitive Computing – ICCC 2019*. Ed. by Ruifeng Xu, Jianzong Wang, and Liang-Jie Zhang. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 132–142. ISBN: 978-3-030-23407-2.

[57] Michael Soprano et al. "The Many Dimensions of Truthfulness: Crowdsourcing Misinformation Assessments on a Multidimensional Scale". en. In: *arXiv:2108.01222 [cs]* (Aug. 2021). URL: http://arxiv.org/abs/2108.01222 (visited on 08/11/2021).

[58] Amar Debnath et al. "A Hierarchical Learning Model for Claim Validation". en. In: *Proceedings of International Joint Conference on Computational Intelligence*. Ed. by Mohammad Shorif Uddin and Jagdish Chand Bansal. Algorithms for Intelligent Systems. Springer Singapore, 2020, pp. 431–441. ISBN: 9789811375644.

[59] Tayyaba Rasool et al. "Multi-Label Fake News Detection using Multi-layered Supervised Learning". In: *Proceedings of the 2019 11th International Conference on Computer and Automation Engineering*. ICCAE 2019. New York, NY, USA: Association for Computing Machinery, Feb. 2019, pp. 73–77. ISBN: 978-1-4503-6287-0. DOI: 10.1145/3313991.3314008. URL: https://doi.org/10.1145/3313991.3314008 (visited on 03/16/2023).

[60] Lia Bozarth, Aparajita Saraf, and Ceren Budak. "Higher Ground? How Groundtruth Labeling Impacts Our Understanding of Fake News about the 2016 U.S. Presidential Nominees". In: *Proceedings of the International AAAI Conference on Web and Social Media* 46.2 (May 2020), pp. 48–59. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7278.

[61] Prashant Shiralkar et al. "Finding Streams in Knowledge Graphs to Support Fact Checking". In: Nov. 2017, pp. 859–864. DOI: 10.1109/ICDM.2017.105.

[62] Ciampaglia GL et al. "Computational Fact Checking from Knowledge Networks". In: *PLoS ONE* 10.6 (June 2015). DOI: 10.1371/journal.pone.0128193.

[63] Bhavika Bhutani et al. "Fake News Detection Using Sentiment Analysis". In: *2019 Twelfth International Conference on Contemporary Computing (IC3)*. ISSN: 2572-6129. Aug. 2019, pp. 1–5. DOI: 10.1109/IC3.2019.8844880.

[64] Kai Shu et al. "dEFEND: Explainable Fake News Detection". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 395–405. ISBN: 978-1-4503-6201-6. DOI: 10.1145/3292500.3330935. URL: https://dl.acm.org/doi/10.1145/3292500.3330935 (visited on 03/16/2023).

[65] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. "Sentiment Aware Fake News Detection on Online Social Networks". In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. ISSN: 2379-190X. May 2019, pp. 2507–2511. DOI: 10.1109/ICASSP.2019.8683170.

[66] Abdullah-All-Tanvir et al. "Detecting Fake News using Machine Learning and Deep Learning Algorithms". In: *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. June 2019, pp. 1–5. DOI: 10.1109/ICSCC.2019.8843612.

[67] Sadia Afroz, Michael Brennan, and Rachel Greenstadt. "Detecting Hoaxes, Frauds, and Deception in Writing Style Online". In: *2012 IEEE Symposium on Security and Privacy*. ISSN: 2375-1207. May 2012, pp. 461–475. DOI: 10.1109/SP.2012.34.

[68] Hadeer Ahmed, Issa Traore, and Sherif Saad. "Detecting opinion spams and fake news using text classification". en. In: *SECURITY AND PRIVACY* 1.1 (2018). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/spy2.9,

e9. ISSN: 2475-6725. DOI: 10.1002/spy2.9. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/spy2.9 (visited on 03/16/2023).

[69] Benjamin D. Horne and Sibel Adali. "This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News". In: *AAAI CWSM'17*. Mar. 2017. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772 (visited on 01/07/2019).

[70] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. "From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles". en. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*. Copenhagen, Denmark: Association for Computational Linguistics, 2017, pp. 84–89. DOI: 10.18653/v1/W17-4215. URL: http://aclweb.org/anthology/W17-4215 (visited on 06/04/2019).

[71] Adrian M. P. Braşoveanu and Răzvan Andonie. "Semantic Fake News Detection: A Machine Learning Perspective". en. In: *Advances in Computational Intelligence*. Ed. by Ignacio Rojas, Gonzalo Joya, and Andreu Catala. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 656–667. ISBN: 978-3-030-20521-8.

[72] Marco Della Vedova et al. "Automatic Online Fake News Detection Combining Content and Social Signals". In: May 2018. DOI: 10.23919/FRUCT.2018.8468301.

[73] Qiang Cao et al. "'Aiding the detection of fake accounts in large scale social online services'". In: *USENIX NSDI* (2012). DOI: https://www.usenix.org/system/files/conference/nsdi12/nsdi12-final42_2.pdf.

[74] George Danezis and Prateek Mittal. "Sybilinfer: Detecting sybil nodes using social networks." In: *Ndss*. San Diego, CA. 2009, pp. 1–15.

[75] Fatima Ezzeddine et al. "Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls". In: *EPJ Data Science* 12.1 (2023), p. 46.

[76] Tarek Hamdi et al. "A Hybrid Approach for Fake News Detection in Twitter Based on User Features and Graph Embedding". en. In: *Distributed Computing and Internet Technology*. Ed. by Dang Van Hung and Meenakshi D´Souza. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 266–280. ISBN: 978-3-030-36987-3. DOI: 10.1007/978-3-030-36987-3_17.

[77] Stefan Helmstetter and Heiko Paulheim. "Weakly Supervised Learning for Fake News Detection on Twitter". In: *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (Aug. 2018). DOI: https://ieeexplore.ieee.org/document/8508520.

[78] M. Alizadeh et al. "Content-based features predict social media influence operations". In: *Science Advances* (July 2020). DOI: 10.1126/sciadv.abb5824.

[79] Sotirios Antoniadis, Iouliana Litou, and Vana Kalogeraki. "A Model for Identifying Misinformation in Online Social Networks". en. In: *On the Move to Meaningful Internet Systems: OTM 2015 Conferences*. Ed. by Christophe Debruyne et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2015, pp. 473–482. ISBN: 978-3-319-26148-5. DOI: 10.1007/978-3-319-26148-5_32.

[80] Dennis Assenmacher et al. "A two-phase framework for detecting manipulation campaigns in social media". In: *Social Computing and Social Media. Design, Ethics, User Behavior, and Social Network Analysis: 12th International Conference, SCSM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part I 22*. Springer. 2020, pp. 201–214.

[81] Cody Buntain and Jennifer Golbeck. "Automatically Identifying Fake News in Popular Twitter Threads". In: *2017 IEEE International Conference on Smart Cloud (SmartCloud)* (Nov. 2017), pp. 208–215. DOI: 10.1109/SmartCloud.2017.40. URL: http://arxiv.org/abs/1705.01613 (visited on 06/04/2019).

[82] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. "Information credibility on twitter". In: *Proceedings of the 20th international conference on World wide web*. WWW '11. New York, NY, USA: Association for Computing Machinery, Mar. 2011, pp. 675–684. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963500. URL: https://doi.org/10.1145/1963405.1963500 (visited on 03/16/2023).

[83] Fatemeh Torabi Asr and Maite Taboada. "The Data Challenge in Misinformation Detection: Source Reputation vs. Content Veracity". In: *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 10–15. DOI: 10.18653/v1/W18-5502. URL: https://aclanthology.org/W18-5502 (visited on 03/16/2023).

[84] *Media Bias/Fact Check News*. July 2021. URL: https://mediabiasfactcheck.com/.

[85] Ramy Baly et al. "What Was Written vs. Who Read It: News Media Profiling Using Text Analysis and Social Media Context". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 3364–3374. DOI: 10.18653/v1/2020.acl-main.308. URL: https://aclanthology.org/2020.acl-main.308.

[86] Sonia Castelo et al. *A Topic-Agnostic Approach for Identifying Fake News Pages*. San Francisco, USA, 2019. DOI: 10.1145/3308560.3316739. URL: http://doi.acm.org/10.1145/3308560.3316739.

[87] Zhouhan Chen and Juliana Freire. "Proactive Discovery of Fake News Domains from Real-Time Social Media Feeds". In: *Companion Proceedings of the Web Conference 2020*. WWW '20. Taipei, Taiwan: Association for Computing Machinery, Apr. 2020, pp. 584–592. ISBN: 978-1-4503-7024-0. DOI: 10.1145/3366424.3385772. URL: https://doi.org/10.1145/3366424.3385772 (visited on 07/31/2020).

[88] Naeemul Hassan et al. "Data in, fact out: automated monitoring of facts by FactWatcher". In: *Proceedings of the VLDB Endowment* 7.13 (Aug. 2014), pp. 1557–1560. ISSN: 2150-8097. DOI: 10.14778/2733004.2733029. URL: https://doi.org/10.14778/2733004.2733029 (visited on 11/06/2022).

[89] Dimitrios Katsaros, George Stavropoulos, and Dimitrios Papakostas. "Which machine learning paradigm for fake news detection?" In: *2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*. Oct. 2019, pp. 383–387.

[90] Ziyi Kou et al. "HC-COVID: A Hierarchical Crowdsource Knowledge Graph Approach to Explainable COVID-19 Misinformation Detection". In: *Proceedings of the ACM on Human-Computer Interaction* 6.GROUP (Jan. 2022), 36:1–36:25. DOI: 10.1145/3492855. URL: https://doi.org/10.1145/3492855 (visited on 01/19/2022).

[91] Kevin Roitero et al. "Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19". en. In: *arXiv:2107.11755 [cs]* (July 2021). URL: http://arxiv.org/abs/2107.11755 (visited on 08/03/2021).

[92] Max Glockner, Yufang Hou, and Iryna Gurevych. *Missing Counter-Evidence Renders NLP Fact-Checking Unrealistic for Misinformation*. arXiv:2210.13865 [cs]. Oct. 2022. DOI: 10.48550/arXiv.2210.13865. URL: http://arxiv.org/abs/2210.13865 (visited on 03/16/2023).

[93] Yavuz Selim Kartal, Busra Guvenen, and Mucahid Kutlu. "Too Many Claims to Fact-Check: Prioritizing Political Claims Based on Check-Worthiness". en. In: *arXiv:2004.08166 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.08166 (visited on 04/24/2020).

[94] Tanik Saikh et al. "A Novel Approach Towards Fake News Detection: Deep Learning Augmented with Textual Entailment Features". en. In: *Natural Language Processing and Information Systems*. Ed. by Elisabeth Métais et al. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 345–358. ISBN: 978-3-030-23281-8.

[95] Luís Borges, Bruno Martins, and Pável Calado. "Combining Similarity Features and Deep Representation Learning for Stance Detection in the Context of Checking Fake News". In: *arXiv:1811.00706 [cs, stat]* (Nov. 2018). URL: http://arxiv.org/abs/1811.00706 (visited on 06/04/2019).

[96] Shrutika S. Jadhav and Sudeep D. Thepade. "Fake News Identification and Classification Using DSSM and Improved Recurrent Neural Network Classifier". In: *Applied Artificial Intelligence* 33.12 (Oct. 2019). Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08839514.2019.1661579, pp. 1058–1068. ISSN: 0883-9514. DOI: 10.1080/08839514.2019.1661579. URL: https://doi.org/10.1080/08839514.2019.1661579 (visited on 03/16/2023).

[97] Lin Tian et al. "Early Detection of Rumours on Twitter via Stance Transfer Learning". en. In: *Advances in Information Retrieval*. Ed. by Joemon M. Jose et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 575–588. ISBN: 978-3-030-45439-5. DOI: 10.1007/978-3-030-45439-5_38.

[98] Amr Magdy and Nayer Wanas. "Web-Based Statistical Fact Checking of Textual Documents". In: *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*. SMUC '10. Toronto, ON, Canada: Association for Computing Machinery, 2010, pp. 103–110. ISBN: 9781450303866. DOI: 10.1145/1871985.1872002. URL: https://doi.org/10.1145/1871985.1872002.

[99] Pujan Paudel et al. "'LAMBRETTA: Learning to Rank for Twitter Soft Moderation'". In: *IEEE Security & Privacy* (2023).

[100] Aditi Gupta et al. *TweetCred: Real-Time Credibility Assessment of Content on Twitter*. arXiv:1405.5490 [physics]. Jan. 2015. DOI: 10.48550/arXiv.1405.5490. URL: http://arxiv.org/abs/1405.5490 (visited on 03/16/2023).

[101] Limeng Cui, Suhang Wang, and Dongwon Lee. "SAME: Sentiment-Aware Multi-Modal Embedding for Detecting Fake News". In: *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ISSN: 2473-991X. Aug. 2019, pp. 41–48. DOI: 10.1145/3341161.3342894.

[102] Naeemul Hassan et al. "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster". en. In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. Halifax, NS, Canada: ACM Press, 2017, pp. 1803–1812. ISBN: 978-1-4503-4887-4. DOI: 10.1145/3097983.3098131. URL: http://dl.acm.org/citation.cfm?doid=3097983.3098131 (visited on 06/04/2019).

[103] James Fairbanks et al. "Credibility assessment in the news: do we need to read". In: *Proc. of the MIS2 Workshop held in conjuction with 11th Int'l Conf. on Web Search and Data Mining*. ACM. 2018, pp. 799–800.

[104] Martin Potthast et al. "A Stylometric Inquiry into Hyperpartisan and Fake News". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 231–240. DOI: 10.18653/v1/P18-1022. URL: https://aclanthology.org/P18-1022.

[105] Rada Mihalcea and Carlo Strapparava. "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language". In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, Aug. 2009, pp. 309–312. URL: https://aclanthology.org/P09-2078.

[106] Kai Shu, Suhang Wang, and Huan Liu. "Beyond News Contents: The Role of Social Context for Fake News Detection". en. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining - WSDM '19*. Melbourne VIC, Australia: ACM Press, 2019, pp. 312–320. ISBN: 978-1-4503-5940-5. DOI: 10.1145/3289600.3290994. URL: http://dl.acm.org/citation.cfm?doid=3289600.3290994 (visited on 05/29/2019).

[107] Verónica Pérez-Rosas et al. "Automatic Detection of Fake News". In: *arXiv:1708.07104 [cs]* (Aug. 2017). URL: http://arxiv.org/abs/1708.07104 (visited on 06/04/2019).

[108] Rada Mihalcea and Carlo Strapparava. "The lie detector: explorations in the automatic recognition of deceptive language". en. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers on - ACL-IJCNLP '09*. Suntec, Singapore: Association for Computational Linguistics, 2009, p. 309. DOI: 10.3115/1667583.1667679. URL: http://portal.acm.org/citation.cfm?doid=1667583.1667679 (visited on 02/13/2019).

[109] Victoria Rubin et al. "Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News". In: *Proceedings of the Second Workshop on Computational Approaches to Deception Detection*. San Diego, California: Association for Computational Linguistics, June 2016, pp. 7–17. DOI: 10.18653/v1/W16-0802. URL: https://aclanthology.org/W16-0802 (visited on 03/16/2023).

[110] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. "Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles". en. In: (), p. 8.

[111] Natali Ruchansky, Sungyong Seo, and Yan Liu. "CSI: A Hybrid Deep Model for Fake News Detection". In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. CIKM '17. New York, NY, USA: ACM, 2017, pp. 797–806. ISBN: 978-1-4503-4918-5. DOI: 10.1145/3132847.3132877. URL: http://doi.acm.org/10.1145/3132847.3132877 (visited on 05/29/2019).

[112] Jamal Abdul Nasir, Osama Subhani Khan, and Iraklis Varlamis. "Fake news detection: A hybrid CNN-RNN based deep learning approach". en. In: *International Journal of Information Management Data Insights* 1.1 (Apr. 2021), p. 100007. ISSN: 2667-0968. DOI: 10.1016/j.jjimei.2020.100007. URL: https://www.sciencedirect.com/science/article/pii/S2667096820300070 (visited on 03/16/2023).

[113] "Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing". en. In: *iConference 2014 Proceedings*. iSchools, Mar. 2014. ISBN: 978-0-9884900-1-7. DOI: 10.9776/14308. URL: https://www.ideals.illinois.edu/handle/2142/47257 (visited on 03/16/2023).

[114] Zhiwei Jin et al. "Detection and Analysis of 2016 US Presidential Election Related Rumors on Twitter". en. In: *Social, Cultural, and Behavioral Modeling*. Ed. by Dongwon Lee et al. Lecture Notes in Computer

Science. Cham: Springer International Publishing, 2017, pp. 14–24. ISBN: 978-3-319-60240-0. DOI: 10.1007/978-3-319-60240-0_2.

[115] Kai Shu et al. “Detecting fake news with weak social supervision”. In: *IEEE Intelligent Systems* 36.4 (2020), pp. 96–103.

[116] Chaowei Zhang et al. “Detecting Fake News for Reducing Misinformation Risks Using Analytics Approaches”. In: *European Journal of Operational Research* (June 2019). ISSN: 0377-2217. DOI: 10.1016/j.ejor.2019.06.022. URL: http://www.sciencedirect.com/science/article/pii/S0377221719304977 (visited on 06/18/2019).

[117] Amila Silva et al. “Embracing Domain Differences in Fake News: Cross-domain Fake News Detection using Multimodal Data”. en. In: *arXiv:2102.06314 [cs]* (Feb. 2021). URL: http://arxiv.org/abs/2102.06314 (visited on 02/22/2021).

[118] Mohammad Hammas Saeed et al. “Trollmagnifier: Detecting state-sponsored troll accounts on reddit”. In: *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2022, pp. 2161–2175.

[119] Savvas Zannettou et al. “Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web”. In: *Companion proceedings of the 2019 world wide web conference*. 2019, pp. 218–226.

[120] Josephine Lukito. “Coordinating a multi-platform disinformation campaign: Internet Research Agency Activity on three US Social Media Platforms, 2015 to 2017”. In: *Political Communication* 37.2 (2020), pp. 238–255.

[121] Franziska B Keller et al. “Political astroturfing on twitter: How to coordinate a disinformation campaign”. In: *Political communication* 37.2 (2020), pp. 256–280.

[122] Stefano Cresci. “A Decade of Social Bot Detection”. In: *Commun. ACM* 63.10 (Sept. 2020), pp. 72–83. ISSN: 0001-0782. DOI: 10.1145/3409116. URL: https://doi.org/10.1145/3409116.

[123] Zi Chu et al. “Detecting automation of twitter accounts: Are you a human, bot, or cyborg?” In: *IEEE Transactions on dependable and secure computing* 9.6 (2012), pp. 811–824.

[124] Jinxue Zhang et al. “The rise of social botnets: Attacks and countermeasures”. In: *IEEE Transactions on Dependable and Secure Computing* 15.6 (2016), pp. 1068–1082.

[125] Craig Silverman et al. *Facebook groups topped 10,000 daily attacks on election before Jan. 6, analysis shows*. Jan. 2022. URL: https://www.washingtonpost.com/technology/2022/01/04/facebook-election-misinformation-capitol-riot/.

[126] Darren L Linvill and Patrick L Warren. “Troll factories: Manufacturing specialized disinformation on Twitter”. In: *Political Communication* 37.4 (2020), pp. 447–467.

[127] URL: https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf.

[128] Gang Wang et al. “Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers”. In: *23rd USENIX Security Symposium (USENIX Security 14)*. 2014, pp. 239–254.

[129] Huiling Zhang et al. “Misinformation in Online Social Networks: Detect Them All with a Limited Budget”. In: *ACM Transactions on Information Systems* 34.3 (Apr. 2016), 18:1–18:24. ISSN: 1046-8188. DOI: 10.1145/2885494. URL: https://doi.org/10.1145/2885494 (visited on 03/16/2023).

[130] Giuseppe Sansonetti et al. “‘Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection’”. In: *IEEE Access* (2020). DOI: 2020.

[131] Diogo Pacheco et al. “Uncovering Coordinated Networks on Social Media”. en. In: *arXiv:2001.05658 [physics]* (Jan. 2020). URL: http://arxiv.org/abs/2001.05658 (visited on 01/22/2020).

[132] Filipe Ribeiro et al. “Media Bias Monitor: Quantifying Biases of Social Media News Outlets at Large-Scale”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 12.1 (June 2018). Number: 1. ISSN: 2334-0770. DOI: 10.1609/icwsm.v12i1.15025. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/15025 (visited on 03/16/2023).

[133] Deen Freelon et al. “Black Trolls Matter: Racial and Ideological Asymmetries in Social Media Disinformation”. en. In: *Social Science Computer Review* (Apr. 2020), p. 0894439320914853. ISSN: 0894-4393. DOI: 10.1177/0894439320914853. URL: https://doi.org/10.1177/0894439320914853 (visited on 04/12/2020).

[134] Aseel Addawood et al. “Linguistic Cues to Deception: Identifying Political Trolls on Social Media”. en. In: *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media*. 2019, p. 11.

[135] Common Thread. *Four truths about bots*. Sept. 2021. URL: https://blog.twitter.com/common-thread/en/topics/stories/2021/four-truths-about-bots.

[136] Christian Grimme, Dennis Assenmacher, and Lena Adam. "Changing perspectives: Is it sufficient to detect social bots?" In: *Social Computing and Social Media. User Experience and Behavior: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings, Part I 10*. Springer. 2018, pp. 445–461.

[137] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (Mar. 2018), pp. 1146–1151. DOI: 10.1126/science.aap9559.

[138] Marcella Tambuscio et al. "Fact-checking Effect on Viral Hoaxes: A Model of Misinformation Spread in Social Networks". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15 Companion. New York, NY, USA: Association for Computing Machinery, May 2015, pp. 977–982. ISBN: 978-1-4503-3473-0. DOI: 10.1145/2740908.2742572. URL: https://doi.org/10.1145/2740908.2742572 (visited on 03/16/2023).

[139] Federico Monti et al. *Fake News Detection on Social Media using Geometric Deep Learning*. 2019. arXiv: 1902.06673 [cs.SI].

[140] Dong Yuan et al. "Detecting fake accounts in online social networks at the time of registrations". In: *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2019, pp. 1423–1438.

[141] *Developer policy – twitter developers | twitter developer platform*. URL: https://developer.twitter.com/en/developer-terms/policy#4-e.

[142] Jacob Ratkiewicz et al. "Detecting and tracking political abuse in social media". In: *Proceedings of the International AAAI Conference on Web and social media*. Vol. 5. 1. 2011, pp. 297–304.

[143] Yang Liu and Yi-Fang Brook Wu. "Early Detection of Fake News on Social Media Through Propagation Path Classification with Recurrent and Convolutional Networks". en. In: (), p. 8.

[144] Jing Ma, Wei Gao, and Kam-Fai Wong. "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning". en. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 708–717. DOI: 10.18653/v1/P17-1066. URL: http://aclweb.org/anthology/P17-1066 (visited on 02/13/2019).

[145] Jing Ma et al. "Detecting rumors from microblogs with recurrent neural networks". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016). DOI: https://www.ijcai.org/Proceedings/16/Papers/537.pdf.

[146] Xiamo Liu et al. "Real-time Rumor Debunking on Twitter". In: *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management (CIKM '15)* (2015). DOI: https://doi.org/10.1145/2806416.2806651.

[147] Daniel Williams. *Misinformation is the symptom, not the disease: Daniel Williams*. Dec. 2023. URL: https://iai.tv/articles/misinformation-is-the-symptom-not-the-disease-daniel-walliams-auid-2690.

[148] Li Qian Tay et al. "Thinking clearly about misinformation". In: *Communications Psychology* 2.1 (2024), p. 4.

[149] Fang Jin et al. "Epidemiological modeling of news and rumors on Twitter". en. In: *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. Chicago Illinois: ACM, Aug. 2013, pp. 1–9. ISBN: 978-1-4503-2330-7. DOI: 10.1145/2501025.2501027. URL: https://dl.acm.org/doi/10.1145/2501025.2501027 (visited on 01/05/2023).

[150] Sander van der Linden. *Misinformation: Susceptibility, spread, and interventions to immunize the public*. Mar. 2022. URL: https://www.nature.com/articles/s41591-022-01713-6.

[151] Nam P. Nguyen et al. "Containment of misinformation spread in online social networks". In: *Proceedings of the 4th Annual ACM Web Science Conference*. WebSci '12. New York, NY, USA: Association for Computing Machinery, June 2012, pp. 213–222. ISBN: 978-1-4503-1228-8. DOI: 10.1145/2380718.2380746. URL: https://doi.org/10.1145/2380718.2380746 (visited on 03/16/2023).

[152] John R Douceur. "The sybil attack". In: *International workshop on peer-to-peer systems*. Springer. 2002, pp. 251–260.

[153] Emilio Ferrara et al. "The rise of social bots". In: *Communications of the ACM* 59.7 (2016), pp. 96–104.

[154] Austin Hounsel et al. "Identifying Disinformation Websites Using Infrastructure Features". In: *FOCI* (2020). URL: https://www.usenix.org/system/files/foci20-paper-hounsel.pdf.

[155] *Snopes*. URL: https://www.snopes.com/.

[156] *FactCheck.org*. URL: https://www.factcheck.org/.

[157] Manolis Chalkiadakis et al. "The Rise and Fall of Fake News sites: A Traffic Analysis". en. In: *arXiv:2103.09258 [cs]* (Mar. 2021). URL: http://arxiv.org/abs/2103.09258 (visited on 03/23/2021).

[158] *Fake news detection datasets*. URL: https://onlineacademiccommunity.uvic.ca/isot/2022/11/27/fake-news-detection-datasets/.

[159] Fatima K. Abu Salem et al. *FA-Kes: A fake news dataset around the Syrian War*. Jan. 2019. URL: https://zenodo.org/record/2607278.

[160] Jason Baumgartner et al. "The pushshift reddit dataset". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14. 2020, pp. 830–839.

[161] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. "Botometer 101: Social bot practicum for computational social scientists". In: *Journal of Computational Social Science* 5.2 (2022), pp. 1511–1528.

[162] Connie Moon Sehat et al. "Misinformation as a harm: structured approaches for fact-checking prioritization". In: *Proceedings of the ACM on Human-Computer Interaction* 8.CSCW1 (2024), pp. 1–36.

[163] Thodoris Lykouris and Wentao Weng. "Learning to defer in content moderation: The human-ai interplay". In: *arXiv preprint arXiv:2402.12237* (2024).

[164] Simone Leonardi, Giuseppe Rizzo, and Maurizio Morisio. "'Automated Classification of Fake News Spreaders to Break the Misinformation Chain'". In: *Information* (2021). DOI: https://doi.org/10.3390/info12060248.

[165] Giovanni Santia, Munif Mujib, and Jake Williams. "'Detecting Social Bots on Facebook in an Information Veracity Context'". In: *ICWSM* (2019). DOI: https://doi.org/10.1609/icwsm.v13i01.3244.

[166] Weiling Chen et al. "'Unsupervised rumor detection based on users' behaviors using neural networks'". In: *Pattern Recognition Letters* (2018). DOI: https://doi.org/10.1016/j.patrec.2017.10.014.

[167] Julio CS Reis et al. "Supervised learning for fake news detection". In: *IEEE Intelligent Systems* 34.2 (2019), pp. 76–81.

[168] Roney Santos et al. "Measuring the impact of readability features in fake news detection". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 1404–1413.

[169] Gang Wang et al. "You are how you click: Clickstream analysis for sybil detection". In: *22nd USENIX security symposium (USENIX Security 13)*. 2013, pp. 241–256.

[170] Haifeng Yu et al. "Sybillimit: A near-optimal social network defense against sybil attacks". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 3–17.

[171] Thomas Magelinski, Lynnette Hui Xian Ng, and Kathleen M Carley. "A synchronized action framework for responsible detection of coordination on social media". In: *arXiv preprint arXiv:2105.07454* (2021).

[172] Karishma Sharma et al. "Identifying coordinated accounts on social media through hidden influence and group behaviours". In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1441–1451.

[173] Margaret Mitchell et al. "Model Cards for Model Reporting". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019). DOI: https://doi.org/10.1145/3287560.3287596.

[174] Jim Maddock et al. "Characterizing Online Rumoring Behavior Using Multi-Dimensional Signatures". In: *CSCW* (Mar. 2015). DOI: 10.1145/2675133.2675280.

[175] Adrien Friggeri et al. "Rumor Cascades". In: *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014* 8 (May 2014), pp. 101–110. DOI: 10.1609/icwsm.v8i1.14559.

[176] James Thorne et al. "FEVER: a Large-scale Dataset for Fact Extraction and VERification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 809–819. DOI: 10.18653/v1/N18-1074. URL: https://aclanthology.org/N18-1074.

[177] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". In: *Proceedings of the 25th International Conference on World Wide Web*. WWW '16. Montréal, Québec, Canada: International World Wide Web Conferences Steering Committee, 2016, pp. 591–602. ISBN: 9781450341431. DOI: 10.1145/2872427.2883085. URL: https://doi.org/10.1145/2872427.2883085.

[178] Vahed Qazvinian et al. "Rumor has it: Identifying Misinformation in Microblogs". In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Edinburgh, Scotland, UK.: Association for Computational Linguistics, July 2011, pp. 1589–1599. URL: https://aclanthology.org/D11-1147.

[179] Pew Research Center. "Social Media Fact Sheet". In: *Pew Research Center* (Sept. 2022). URL: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/.

[180] Press Association. "Blue ticks for all: Twitter allows users to apply to be verified". In: *The Guardian* (July 2016). URL: https://www.theguardian.com/technology/2016/jul/19/blue-ticks-for-all-twitter-allows-all-users-to-be-verified.

[181] Hannah Rashkin et al. "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sept. 2017, pp. 2931–2937. DOI: 10.18653/v1/D17-1317. URL: https://aclanthology.org/D17-1317.

[182] Justin Cheng et al. "Can Cascades Be Predicted?" In: *Proceedings of the 23rd International Conference on World Wide Web*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 925–936. ISBN: 9781450327442. DOI: 10.1145/2566486.2567997. URL: https://doi.org/10.1145/2566486.2567997.

[183] Tal Schuster et al. "Limitations of stylometry for detecting machine-generated fake news". In: *Computational Linguistics* 46 (June 2020), pp. 499–510. URL: https://direct.mit.edu/coli/article/46/2/499/93369/The-Limitations-of-Stylometry-for-Detecting.

[184] Craig Silverman and Jeff Kao. *Infamous Russian troll farm appears to be source of Anti-Ukraine propaganda*. Mar. 2022. URL: https://www.propublica.org/article/infamous-russian-troll-farm-appears-to-be-source-of-anti-ukraine-propaganda.

[185] Andreas Pitsillidis et al. "Botnet Judo: Fighting Spam with Itself." In: *NDSS*. 2010.

[186] Fan Yang et al. "Automatic detection of rumor on Sina Weibo". en. In: *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics - MDS '12*. Beijing, China: ACM Press, 2012, pp. 1–7. ISBN: 978-1-4503-1546-3. DOI: 10.1145/2350190.2350203. URL: http://dl.acm.org/citation.cfm?doid=2350190.2350203 (visited on 02/13/2019).

[187] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". en. In: *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. Montr&#233;al, Qu&#233;bec, Canada: ACM Press, 2016, pp. 591–602. ISBN: 978-1-4503-4143-1. DOI: 10.1145/2872427.2883085. URL: http://dl.acm.org/citation.cfm?doid=2872427.2883085 (visited on 02/13/2019).

[188] Shan Jiang and Christo Wilson. "Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media". en. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (Nov. 2018), pp. 1–23. ISSN: 25730142. DOI: 10.1145/3274351. URL: http://dl.acm.org/citation.cfm?doid=3290265.3274351 (visited on 03/28/2019).

[189] Jeppe Norregaard, Benjamin D. Horne, and Sibel Adali. "NELA-GT-2018: A Large Multi-Labelled News Dataset for The Study of Misinformation in News Articles". en. In: *arXiv:1904.01546 [cs]* (Apr. 2019). URL: http://arxiv.org/abs/1904.01546 (visited on 04/04/2019).

[190] Jooyeon Kim et al. "Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation". en. In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining - WSDM '18*. Marina Del Rey, CA, USA: ACM Press, 2018, pp. 324–332. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159734. URL: http://dl.acm.org/citation.cfm?doid=3159652.3159734 (visited on 05/29/2019).

[191] Kai Shu, Suhang Wang, and Huan Liu. "Understanding User Profiles on Social Media for Fake News Detection". en. In: *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*. Miami, FL: IEEE, Apr. 2018, pp. 430–435. ISBN: 978-1-5386-1857-8. DOI: 10.1109/MIPR.2018.00092. URL: https://ieeexplore.ieee.org/document/8397048/ (visited on 05/29/2019).

[192] Feng Qian et al. "Neural User Response Generator: Fake News Detection with Collective User Intelligence". en. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. Stockholm, Sweden: International Joint Conferences on Artificial Intelligence Organization, July 2018, pp. 3834–3840. ISBN: 978-0-9992411-2-7. DOI: 10.24963/ijcai.2018/533. URL: https://www.ijcai.org/proceedings/2018/533 (visited on 05/29/2019).

[193] Yimin Chen, Niall J. Conroy, and Victoria L. Rubin. “Misleading Online Content: Recognizing Clickbait As "False News"”. In: *Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection*. WMDD ’15. New York, NY, USA: ACM, 2015, pp. 15–19. ISBN: 978-1-4503-3987-2. DOI: 10.1145/2823465.2823467. URL: http://doi.acm.org/10.1145/2823465.2823467 (visited on 05/29/2019).

[194] Xinyi Zhou and Reza Zafarani. “Fake News: A Survey of Research, Detection Methods, and Opportunities”. In: *arXiv:1812.00315 [cs]* (Dec. 2018). URL: http://arxiv.org/abs/1812.00315 (visited on 06/04/2019).

[195] Julio CS Reis et al. “Explainable machine learning for fake news detection”. In: *Proceedings of the 10th ACM conference on web science*. 2019, pp. 17–26.

[196] Diego Sáez-Trumper. “Fake tweet buster: a webtool to identify users promoting fake news ontwitter”. In: *HT*. 2014. DOI: 10.1145/2631775.2631786.

[197] Nguyen Vo and Kyumin Lee. “The Rise of Guardians: Fact-checking URL Recommendation to Combat Fake News”. en. In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval - SIGIR ’18*. Ann Arbor, MI, USA: ACM Press, 2018, pp. 275–284. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210037. URL: http://dl.acm.org/citation.cfm?doid=3209978.3210037 (visited on 06/04/2019).

[198] Julio Amador, Axel Oehmichen, and Miguel Molina-Solana. “Characterizing Political Fake News in Twitter by its Meta-Data”. In: *arXiv:1712.05999 [cs, stat]* (Dec. 2017). URL: http://arxiv.org/abs/1712.05999 (visited on 06/05/2019).

[199] Yang Yang et al. “TI-CNN: Convolutional Neural Networks for Fake News Detection”. In: *arXiv:1806.00749 [cs]* (June 2018). URL: http://arxiv.org/abs/1806.00749 (visited on 06/04/2019).

[200] Melanie Tosik, Antonio Mallia, and Kedar Gangopadhyay. “Debunking Fake News One Feature at a Time”. In: *arXiv:1808.02831 [cs]* (Aug. 2018). URL: http://arxiv.org/abs/1808.02831 (visited on 06/04/2019).

[201] Xinyi Zhou et al. “Fake News Early Detection: A Theory-driven Model”. In: *arXiv:1904.11679 [cs]* (Apr. 2019). URL: http://arxiv.org/abs/1904.11679 (visited on 06/04/2019).

[202] Shuo Yang et al. “Unsupervised Fake News Detection on Social Media: A Generative Approach”. In: Feb. 2019.

[203] Kashyap Popat et al. “DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning”. en. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: ACL, Nov. 2018, pp. 22–32.

[204] Eugenio Tacchini et al. “Some Like it Hoax: Automated Fake News Detection in Social Networks”. In: *arXiv:1704.07506 [cs]* (Apr. 2017). URL: http://arxiv.org/abs/1704.07506 (visited on 06/04/2019).

[205] Eugenio Tacchini et al. “Automated Fake News Detection in Social Networks”. en. In: (2017), p. 15.

[206] S. Krishnan and M. Chen. “Cloud-Based System for Fake Tweet Identification”. In: *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*. Apr. 2019, pp. 720–721.

[207] Svitlana Volkova et al. “Explaining Multimodal Deceptive News Prediction Models”. en. In: 2019, p. 4.

[208] Shamoz Shah and Madhu Goyal. “Anomaly Detection in Social Media Using Recurrent Neural Network”. en. In: *Computational Science – ICCS 2019*. Ed. by João M. F. Rodrigues et al. Lecture Notes in Computer Science. Springer International Publishing, 2019, pp. 74–83. ISBN: 978-3-030-22747-0.

[209] Sung Soo Park and Kun Chang Lee. “A Comparative Study of Text analysis and Network embedding Methods for Effective Fake News Detection”. ko. In: *Journal of Digital Convergence* 17.5 (May 2019), pp. 137–143. DOI: 10.14400/JDC.2019.17.5.137. URL: https://doi.org/10.14400/JDC.2019.17.5.137 (visited on 06/12/2019).

[210] Duc Minh Nguyen et al. “Fake News Detection using Deep Markov Random Fields”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1391–1400. URL: https://www.aclweb.org/anthology/N19-1141 (visited on 06/12/2019).

[211] Laura Burbach et al. “Who Shares Fake News in Online Social Networks?” en. In: *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization - UMAP ’19*. Larnaca, Cyprus: ACM Press, 2019, pp. 234–242. ISBN: 978-1-4503-6021-0. DOI: 10.1145/3320435.3320456. URL: http://dl.acm.org/citation.cfm?doid=3320435.3320456 (visited on 06/17/2019).

[212] Maryam Ramezani et al. “News Labeling as Early as Possible: Real or Fake?” en. In: *arXiv:1906.03423 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.03423 (visited on 06/17/2019).

[213] Xinyi Zhou and Reza Zafarani. “Network-based Fake News Detection: A Pattern-driven Approach”. en. In: *arXiv:1906.04210 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.04210 (visited on 06/17/2019).

[214] Adrian Rauchfleisch and Jonas Kaiser. “The false positive problem of automatic bot detection in social science research”. In: *PloS one* 15.10 (2020), e0241045.

[215] *GARM Brand Safety Floor and Suitability Framework*. URL: https://wfanet.org/l/library/download/urn:uuid:7d484745-41cd-4cce-a1b9-a1b4e30928ea/garm+brand+safety+floor+suitability+framework+23+sept.pdf.

[216] Zefr. July 2022. URL: https://www.prnewswire.com/news-releases/zefr-acquires-israeli-ai-firm-adverifai-bolstering-technology-led-approach-to-identifying-and-defunding-misinformation-301590009.html.

[217] Saleem Alhabash et al. “Pathways to virality: Psychophysiological responses preceding likes, shares, comments, and status updates on Facebook”. In: *Media Psychology* 22.2 (2019), pp. 196–216.

[218] Daniel Kapellmann Zafra et al. *How to understand and action Mandiant’s intelligence on information operations*. Oct. 2022. URL: https://www.mandiant.com/resources/blog/understand-action-intelligence-information-operations.

[219] Santhosh Srinivasan. *The global disinformation index*. URL: https://www.disinformationindex.org/.

[220] Kevin Aslett et al. “News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions”. In: *Science advances* 8.18 (2022), eabl3844.

[221] Mingkun Gao et al. “To label or not to label: The effect of stance and credibility labels on readers’ selection and perception of news articles”. In: *Proceedings of the ACM on Human-Computer Interaction* 2.CSCW (2018), pp. 1–16.

[222] Mark Stencel, Erica Ryan, and Joel Luther. *Misinformation spreads, but fact-checking has leveled off*. June 2023. URL: https://reporterslab.org/misinformation-spreads-but-fact-checking-has-leveled-off/.

[223] Xishuang Dong et al. “Deep Two-path Semi-supervised Learning for Fake News Detection”. en. In: *arXiv:1906.05659 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.05659 (visited on 06/18/2019).

[224] Javier Sánchez-Junquera et al. “Unmasking Bias in News”. en. In: *arXiv:1906.04836 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.04836 (visited on 06/18/2019).

[225] Anoop K, Deepak P, and Lajish V. L. “Emotion Cognizance Improves Fake News Identification”. en. In: *arXiv:1906.10365 [cs]* (June 2019). URL: http://arxiv.org/abs/1906.10365 (visited on 07/02/2019).

[226] Munyeong Lim and Sungbum Park. “A Study on the Preemptive Measure for Fake News Eradication Using Data Mining Algorithms : Focused on the M Online Community Postings”. kor. In: *Journal of Information Technology Services* 18.1 (2019), pp. 219–234. ISSN: 1975-4256. DOI: 10.9716/KITS.2019.18.1.219. URL: http://www.koreascience.or.kr/article/JAKO201915561990199.page (visited on 07/02/2019).

[227] Atif Ahmad et al. “Strategically-Motivated Advanced Persistent Threat: Definition, Process, Tactics and a Disinformation Model of Counterattack”. In: *Computers & Security* (July 2019). ISSN: 0167-4048. DOI: 10.1016/j.cose.2019.07.001. URL: http://www.sciencedirect.com/science/article/pii/S0167404818310988 (visited on 07/11/2019).

[228] Michelle Seref and Onur Seref. “Rhetoric Mining for Fake News: Identifying Moves of Persuasion and Disinformation”. In: *AMCIS 2019 Proceedings* (July 2019). URL: https://aisel.aisnet.org/amcis2019/rhetoric_social_media_disinformation/rhetoric_social_media_disinformation/1.

[229] Nicolas Papernot et al. “SoK: Security and Privacy in Machine Learning”. In: *IEEE Explore* (2018). DOI: 10.1109/EuroSP.2018.00035.

[230] Amanda Storey. *Our ongoing work to fight misinformation online*. Oct. 2023. URL: https://blog.google/around-the-globe/google-europe/our-ongoing-work-to-fight-misinformation-online/.

[231] URL: https://graphika.com/how-it-works.

[232] OpenAI. *Using machine learning to reduce toxicity online*. URL: https://perspectiveapi.com/.

[233] Anya Schiffrin. *Using AI to combat MIS/disinformation – an evolving story*. Oct. 2023. URL: https://www.techpolicy.press/using-ai-to-combat-mis-disinformation-an-evolving-story/.

[234] Kai Shu and Huan Liu. "Detecting Fake News on Social Media". In: *Synthesis Lectures on Data Mining and Knowledge Discovery* 11.3 (July 2019), pp. 1–129. ISSN: 2151-0067. DOI: 10.2200/S00926ED1V01Y201906DMK018. URL: https://www.morganclaypool.com/doi/abs/10.2200/S00926ED1V01Y201906DMK018 (visited on 07/11/2019).

[235] Ben Nimmo. *Meta's adversarial threat report, fourth quarter 2022*. Feb. 2023. URL: https://about.fb.com/news/2023/02/metas-adversarial-threat-report-q4-2022/.

[236] Nov. 2020. URL: https://ai.meta.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

[237] Hibaq Farah. *Diary of a TikTok moderator: "we are the people who sweep up the mess"*. Dec. 2023. URL: https://www.theguardian.com/technology/2023/dec/21/diary-of-a-tiktok-moderator-we-are-the-people-who-sweep-up-the-mess#:~:text=TikTok%20says%20it%20has%20more,in%20more%20than%2070%20languages..

[238] Issie Lapowsky. *Inside the research lab teaching Facebook about its trolls*. Aug. 2018. URL: https://www.wired.com/story/facebook-enlists-dfrlab-track-trolls/.

[239] URL: https://www.reuters.com/technology/musk-owned-xs-content-moderation-shift-complicated-effort-win-back-brands-former-2023-09-07/.

[240] Microsoft Threat Intelligence Microsoft Incident Response. *Dev-0537 criminal actor targeting organizations for data exfiltration and destruction*. Sept. 2023. URL: https://www.microsoft.com/en-us/security/blog/2022/03/22/dev-0537-criminal-actor-targeting-organizations-for-data-exfiltration-and-destruction/.

[241] Inc. Integral Ad Science. *IAS expands AI-driven brand safety and suitability measurement to Meta*. Feb. 2024. URL: https://www.prnewswire.com/news-releases/ias-expands-ai-driven-brand-safety-and-suitability-measurement-to-meta-302052737.html.

[242] URL: https://zefr.com/misinformation.

[243] Issie Lapowsky. *After sending content moderators home, YouTube doubled its video removals*. Aug. 2020. URL: https://www.protocol.com/youtube-content-moderation-covid-19.

[244] Frank Pasquine. *What advertisers need to know about surges in online hate speech*. Feb. 2022. URL: https://doubleverify.com/online-hate-speech-surges-amid-protests/.

[245] The YouTube Team. *Responsible policy enforcement during COVID-19*. Aug. 2020. URL: https://blog.youtube/inside-youtube/responsible-policy-enforcement-during-covid-19/.

[246] Niamh McIntyre, Rosie Bradbury, and Billy Perrigo. *Behind TikTok's boom: A legion of traumatised, $10-a-day content moderators*. Dec. 2023. URL: https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatised-10-a-day-content-moderators.

[247] Stefan Wojcik et al. "Birdwatch: Crowd wisdom and bridging algorithms can inform understanding and reduce the spread of misinformation". In: *arXiv preprint arXiv:2210.15723* (2022).

[248] URL: https://www.facebook.com/business/help/866941431052802?id=2520940424820218.

[249] Jesus Reyes and Leon Palafox. "Detection of Fake News based on readability". en. In: (), p. 6.

[250] Michael A. Stefanone, Matthew Vollmer, and Jessica M. Covert. "In News We Trust?: Examining Credibility and Sharing Behaviors of Fake News". In: *Proceedings of the 10th International Conference on Social Media and Society*. SMSociety '19. New York, NY, USA: ACM, 2019, pp. 136–147. ISBN: 978-1-4503-6651-9. DOI: 10.1145/3328529.3328554. URL: http://doi.acm.org/10.1145/3328529.3328554 (visited on 07/11/2019).

[251] Aarash Heydari et al. "YouTube Chatter: Understanding Online Comments Discourse on Misinformative and Political YouTube Videos". en. In: (), p. 32.

[252] Jozef Kapusta, Ľubomír Benko, and Michal Munk. "Fake News Identification Based on Sentiment and Frequency Analysis". en. In: *Innovation in Information Systems and Technologies to Support Learning Research*. Ed. by Mohammed Serrhini, Carla Silva, and Sultan Aljahdali. Learning and Analytics in Intelligent Systems. Cham: Springer International Publishing, 2020, pp. 400–409. ISBN: 978-3-030-36778-7. DOI: 10.1007/978-3-030-36778-7_44.

[253] Kashyap Popat. “"Credibility Analysis of Textual Claims with Explainable Evidence"”. en. In: (), p. 134.

[254] Leonardo Nizzoli et al. "Coordinated behavior on social media in 2019 UK general election". In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15. 2021, pp. 443–454.

[255] Onur Varol et al. "Early detection of promoted campaigns on social media". In: *EPJ data science* 6 (2017), pp. 1–19.

[256] Rohit Kumar Kaliyar et al. "FNDNet- A Deep Convolutional Neural Network for Fake News Detection". en. In: *Cognitive Systems Research* (Jan. 2020). ISSN: 1389-0417. DOI: 10.1016/j.cogsys.2019.12.005. URL: http://www.sciencedirect.com/science/article/pii/S1389041720300085 (visited on 01/25/2020).

[257] Ning Xin Nyow and Hui Na Chua. "Detecting Fake News with Tweets' Properties". In: *2019 IEEE Conference on Application, Information and Network Security (AINS)*. Nov. 2019, pp. 24–29. DOI: 10.1109/AINS47559.2019.8968706.

[258] Francesco Pierri, Carlo Piccardi, and Stefano Ceri. "Topology comparison of Twitter diffusion networks effectively reveals misleading information". en. In: *Scientific Reports* 10.1 (Jan. 2020), pp. 1–9. ISSN: 2045-2322. DOI: 10.1038/s41598-020-58166-5. URL: https://www.nature.com/articles/s41598-020-58166-5 (visited on 02/04/2020).

[259] Harita Reddy et al. "Text-mining-based Fake News Detection Using Ensemble Methods". en. In: *International Journal of Automation and Computing* (Feb. 2020). ISSN: 1751-8520. DOI: 10.1007/s11633-019-1216-5. URL: https://doi.org/10.1007/s11633-019-1216-5 (visited on 02/22/2020).

[260] Juan Cao et al. "Exploring the Role of Visual Content in Fake News Detection". en. In: *arXiv:2003.05096 [cs]* (Mar. 2020). DOI: 10.1007/978-3-030-42699-6. URL: http://arxiv.org/abs/2003.05096 (visited on 03/17/2020).

[261] Kai Shu et al. "Leveraging Multi-Source Weak Social Supervision for Early Detection of Fake News". en. In: *arXiv:2004.01732 [cs, stat]* (Apr. 2020). URL: http://arxiv.org/abs/2004.01732 (visited on 04/12/2020).

[262] Harika Kudarvalli and Jinan Fiaidhi. "Detecting Fake News using Machine Learning Algorithms". en. In: (Apr. 2020). ISSN: doi:10.36227/techrxiv.12089133.v1. DOI: 10.36227/techrxiv.12089133.v1. URL: https://www.techrxiv.org/articles/Detecting_Fake_News_using_Machine_Learning_Algorithms/12089133 (visited on 04/12/2020).

[263] Anmol Uppal, Vipul Sachdeva, and Seema Sharma. "Fake news detection using discourse segment structure analysis". In: *2020 10th International Conference on Cloud Computing, Data Science Engineering (Confluence)*. Jan. 2020, pp. 751–756. DOI: 10.1109/Confluence47617.2020.9058106.

[264] Marialaura Previti et al. "Fake News Detection Using Time Series and User Features Classification". en. In: *Applications of Evolutionary Computation*. Ed. by Pedro A. Castillo, Juan Luis Jiménez Laredo, and Francisco Fernández de Vega. Vol. 12104. Cham: Springer International Publishing, 2020, pp. 339–353. ISBN: 978-3-030-43721-3 978-3-030-43722-0. DOI: 10.1007/978-3-030-43722-0_22. URL: http://link.springer.com/10.1007/978-3-030-43722-0_22 (visited on 04/15/2020).

[265] Yue Zhou, Yan Zhang, and JingTao Yao. "Satirical News Detection with Semantic Feature Extraction and Game-theoretic Rough Sets". en. In: *arXiv:2004.03788 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.03788 (visited on 04/15/2020).

[266] Pepa Atanasova et al. "Generating Fact Checking Explanations". en. In: *arXiv:2004.05773 [cs]* (Apr. 2020). URL: http://arxiv.org/abs/2004.05773 (visited on 04/17/2020).

[267] Joma George, Shintu Mariam Skariah, and T Aleena Xavier. "Role of Contextual Features in Fake News Detection: A Review". In: *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*. Feb. 2020, pp. 1–6. DOI: 10.1109/ICITIIT49094.2020.9071524.

[268] Mitchell L Gordon et al. "The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality". en. In: (2021), p. 14.

[269] L. Kurasinski and R.-C. Mihailescu. "Towards Machine Learning Explainability in Text Classification for Fake News Detection". In: *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Dec. 2020, pp. 775–781. DOI: 10.1109/ICMLA51294.2020.00127. URL: https://ieeexplore.ieee.org/abstract/document/9356374.

[270] Paula Carvalho et al. *Assessing News Credibility: Misinformation Content Indicators*. en. preprint. In Review, Mar. 2021. DOI: 10.21203/rs.3.rs-173067/v1. URL: https://www.researchsquare.com/article/rs-173067/v1 (visited on 03/08/2021).

[271] Benjamin Krämer. “Stop studying “fake news” (we can still fight against disinformation in the media)”. In: *Studies in Communication and Media* 10.1 (2021), pp. 6–30. ISSN: 2192-4007. DOI: 10.5771/2192-4007-2021-1-6. URL: https://www.nomos-elibrary.de/index.php?doi=10.5771/2192-4007-2021-1-6 (visited on 04/06/2021).

[272] Nayeon Lee et al. “On Unifying Misinformation Detection”. en. In: *arXiv:2104.05243 [cs]* (Apr. 2021). URL: http://arxiv.org/abs/2104.05243 (visited on 04/16/2021).

[273] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. “The Surprising Performance of Simple Baselines for Misinformation Detection”. en. In: *arXiv:2104.06952 [cs]* (Apr. 2021). URL: http://arxiv.org/abs/2104.06952 (visited on 04/21/2021).

[274] Bahruz Jabiyev et al. “FADE: Detecting Fake News Articles on the Web”. en. In: (2021), p. 10.

[275] Jakub Simko et al. “Towards Continuous Automatic Audits of Social Media Adaptive Behavior and its Role in Misinformation Spreading”. en. In: *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. Utrecht Netherlands: ACM, June 2021, pp. 411–414. ISBN: 978-1-4503-8367-7. DOI: 10.1145/3450614.3463353. URL: https://dl.acm.org/doi/10.1145/3450614.3463353 (visited on 07/15/2021).

[276] Yunkang Yang, Trevor Davis, and Matthew Hindman. “Visual Misinformation on Facebook”. en. In: (), p. 8.

[277] Kevin Roitero et al. “The COVID-19 Infodemic: Can the Crowd Judge Recent Misinformation Objectively?” en. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Oct. 2020), pp. 1305–1314. DOI: 10.1145/3340531.3412048. URL: http://arxiv.org/abs/2008.05701 (visited on 08/03/2021).

[278] Antino Kim, Patricia Moravec, and Alan Dennis. “Behind the Stars: The Effects of News Source Ratings on Fake News in Social Media”. In: *SSRN Electronic Journal* (Jan. 2017). DOI: 10.2139/ssrn.3090355.

[279] Paul Resnick et al. “Informed Crowds Can Effectively Identify Misinformation”. en. In: *arXiv:2108.07898 [cs]* (Aug. 2021). URL: http://arxiv.org/abs/2108.07898 (visited on 08/25/2021).

[280] Prerna Juneja and Tanushree Mitra. “Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation”. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 186. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 1–27. ISBN: 978-1-4503-8096-6. URL: https://doi.org/10.1145/3411764.3445250 (visited on 09/17/2021).

[281] Maria Janicka, Maria Pszona, and Aleksander Wawer. “Cross-Domain Failures of Fake News Detection”. en. In: *Computación y Sistemas* 23.3 (Oct. 2019). ISSN: 2007-9737, 1405-5546. DOI: 10.13053/cys-23-3-3281. URL: https://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/3281 (visited on 10/29/2021).

[282] Ran Wang et al. “RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media”. In: *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI EA ’22. New York, NY, USA: Association for Computing Machinery, Apr. 2022, pp. 1–7. ISBN: 978-1-4503-9156-6. DOI: 10.1145/3491101.3519712. URL: https://doi.org/10.1145/3491101.3519712 (visited on 11/06/2022).

[283] James Thorne and Andreas Vlachos. “Automated Fact Checking: Task Formulations, Methods and Future Directions”. en. In: (), p. 14.

[284] Li Zeng, Kate Starbird, and Emma Spiro. “#Unconfirmed: Classifying Rumor Stance in Crisis-Related Social Media Messages”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 10.1 (2016). Number: 1, pp. 747–750. ISSN: 2334-0770. DOI: 10.1609/icwsm.v10i1.14788. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14788 (visited on 01/05/2023).

[285] Gregory Eady et al. “Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior”. en. In: *Nature Communications* 14.1 (Jan. 2023). Number: 1 Publisher: Nature Publishing Group, p. 62. ISSN: 2041-1723. DOI: 10.1038/s41467-022-35576-9. URL: https://www.nature.com/articles/s41467-022-35576-9 (visited on 01/09/2023).

[286] Abdullah Talha Kabakuş and Mehmet Şimşek. “An Analysis of the Characteristics of Verified Twitter Users”. en. In: *Sakarya University Journal of Computer and Information Sciences* 2.3 (Dec. 2019), pp. 180–186. ISSN: 2636-8129. DOI: 10.35377/saucis.02.03.649708. URL: https://dergipark.org.tr/en/doi/10.35377/saucis.02.03.649708 (visited on 01/12/2023).

[287] Alison Hearn. “Verified: Self-presentation, identity management, and selfhood in the age of big data”. In: *Popular Communication* 15.2 (Apr. 2017). Publisher: Routledge _eprint: https://doi.org/10.1080/15405702.2016.1269909,

pp. 62–77. ISSN: 1540-5702. DOI: 10.1080/15405702.2016.1269909. URL: https://doi.org/10.1080/15405702.2016.1269909 (visited on 01/12/2023).

[288] *Twitter Verification requirements - how to get the blue check*. en. URL: https://help.twitter.com/en/managing-your-account/about-twitter-verified-accounts (visited on 01/12/2023).

[289] *Prolific · Quickly find research participants you can trust*. en. URL: https://www.prolific.co/ (visited on 01/12/2023).

[290] Ahmer Arif et al. "How Information Snowballs: Exploring the Role of Exposure in Online Rumor Propagation". In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. CSCW '16. New York, NY, USA: Association for Computing Machinery, Feb. 2016, pp. 466–477. ISBN: 978-1-4503-3592-8. DOI: 10.1145/2818048.2819964. URL: https://doi.org/10.1145/2818048.2819964 (visited on 01/12/2023).

[291] Sander van der Linden. "Misinformation: susceptibility, spread, and interventions to immunize the public". en. In: *Nature Medicine* 28.3 (Mar. 2022). Number: 3 Publisher: Nature Publishing Group, pp. 460–467. ISSN: 1546-170X. DOI: 10.1038/s41591-022-01713-6. URL: https://www.nature.com/articles/s41591-022-01713-6 (visited on 01/12/2023).

[292] *RumorLens: Interactive Analysis and Validation of Suspected Rumors on Social Media | Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*. URL: https://dl.acm.org/doi/10.1145/3491101.3519712 (visited on 03/16/2023).

[293] Ala Mughaid et al. "An intelligent cybersecurity system for detecting fake news in social media websites". en. In: *Soft Computing* 26.12 (June 2022), pp. 5577–5591. ISSN: 1433-7479. DOI: 10.1007/s00500-022-07080-1. URL: https://doi.org/10.1007/s00500-022-07080-1 (visited on 03/16/2023).

[294] Firoj Alam et al. *A Survey on Multimodal Disinformation Detection*. arXiv:2103.12541 [cs]. Sept. 2022. DOI: 10.48550/arXiv.2103.12541. URL: http://arxiv.org/abs/2103.12541 (visited on 03/16/2023).

[295] Fernando Miró-Llinares and Jesús C. Aguerri. "Misinformation about fake news: A systematic critical review of empirical studies on the phenomenon and its status as a 'threat'". en. In: *European Journal of Criminology* 20.1 (Jan. 2023). Publisher: SAGE Publications, pp. 356–374. ISSN: 1477-3708. DOI: 10.1177/1477370821994059. URL: https://doi.org/10.1177/1477370821994059 (visited on 03/16/2023).

[296] Paul Resnick et al. *Informed Crowds Can Effectively Identify Misinformation*. arXiv:2108.07898 [cs]. Feb. 2022. DOI: 10.48550/arXiv.2108.07898. URL: http://arxiv.org/abs/2108.07898 (visited on 03/16/2023).

[297] *Assessing News Credibility: Misinformation Content Indicators*. en. Mar. 2021. DOI: 10.21203/rs.3.rs-173067/v1. URL: https://www.researchsquare.com (visited on 03/16/2023).

[298] Kevin Roitero et al. "Can the Crowd Judge Truthfulness? A Longitudinal Study on Recent Misinformation about COVID-19". In: *Personal and Ubiquitous Computing* 27.1 (Feb. 2023). arXiv:2107.11755 [cs], pp. 59–89. ISSN: 1617-4909, 1617-4917. DOI: 10.1007/s00779-021-01604-6. URL: http://arxiv.org/abs/2107.11755 (visited on 03/16/2023).

[299] Feng Yu et al. "A Convolutional Approach for Misinformation Identification". In: (2017), pp. 3901–3907. URL: https://www.ijcai.org/proceedings/2017/545 (visited on 03/16/2023).

[300] Arjun Roy et al. "A Deep Ensemble Framework for Fake News Detection and Multi-Class Classification of Short Political Statements". In: *Proceedings of the 16th International Conference on Natural Language Processing*. International Institute of Information Technology, Hyderabad, India: NLP Association of India, Dec. 2019, pp. 9–17. URL: https://aclanthology.org/2019.icon-1.2 (visited on 03/16/2023).

[301] Bashar Al Asaad and Madalina Erascu. "A Tool for Fake News Detection". In: *2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*. Sept. 2018, pp. 379–386. DOI: 10.1109/SYNASC.2018.00064.

[302] A. Conrad Nied et al. "Alternative Narratives of Crisis Events: Communities and Social Botnets Engaged on Social Media". en. In: *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Portland Oregon USA: ACM, Feb. 2017, pp. 263–266. ISBN: 978-1-4503-4688-7. DOI: 10.1145/3022198.3026307. URL: https://dl.acm.org/doi/10.1145/3022198.3026307 (visited on 03/16/2023).

[303] Xichen Zhang and Ali A. Ghorbani. "An overview of online fake news: Characterization, detection, and discussion". en. In: *Information Processing & Management* 57.2 (Mar. 2020), p. 102025. ISSN: 0306-4573. DOI: 10.1016/j.ipm.2019.03.004. URL: https://www.sciencedirect.com/science/article/pii/S0306457318306794 (visited on 03/16/2023).

[304] Vivek Singh et al. “Automated Fake News Detection Using Linguistic Analy- sis and Machine Learning”. en. In: ().

[305] Sawinder Kaur, Parteek Kumar, and Ponnurangam Kumaraguru. “Automating fake news detection system using multi-level voting model”. en. In: *Soft Computing* 24.12 (June 2020), pp. 9049–9069. ISSN: 1433-7479. DOI: 10.1007/s00500-019-04436-y. URL: https://doi.org/10.1007/s00500-019-04436-y (visited on 03/16/2023).

[306] Luis Vargas, Patrick Emami, and Patrick Traynor. “On the detection of disinformation campaign activity with network analysis”. In: *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*. 2020, pp. 133–146.

[307] Karishma Sharma et al. “Identifying coordinated accounts on social media through hidden influence and group behaviours”. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2021, pp. 1441–1451.

[308] Tanushree Mitra and Eric Gilbert. “CREDBANK: A Large-Scale Social Media Corpus With Associated Credibility Annotations”. en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 9.1 (2015). Number: 1, pp. 258–267. ISSN: 2334-0770. DOI: 10.1609/icwsm.v9i1.14625. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14625 (visited on 03/16/2023).

[309] Christopher G. Harris. “Detecting Deceptive Opinion Spam Using Human Computation”. In: July 2012. URL: https://www.semanticscholar.org/paper/Detecting-Deceptive-Opinion-Spam-Using-Human-Harris/471dd1fd942e3fc83e0201f7e51415c4310cdfa1 (visited on 03/16/2023).

[310] Monther Aldwairi and Ali Alwahedi. “Detecting Fake News in Social Media Networks”. en. In: *Procedia Computer Science*. The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN-2018) / The 8th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2018) / Affiliated Workshops 141 (Jan. 2018), pp. 215–222. ISSN: 1877-0509. DOI: 10.1016/j.procs.2018.10.171. URL: https://www.sciencedirect.com/science/article/pii/S1877050918318210 (visited on 03/16/2023).

[311] Supanya Aphiwongsophon and Prabhas Chongstitvatana. “Detecting Fake News with Machine Learning Method”. In: *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*. July 2018, pp. 528–531. DOI: 10.1109/ECTICon.2018.8620051.

[312] Boris A. Galitsky. “Detecting Rumor and Disinformation by Web Mining”. In: Mar. 2015. URL: https://www.semanticscholar.org/paper/Detecting-Rumor-and-Disinformation-by-Web-Mining-Galitsky/449790743c7de01916182e96369ebaedb5ebe1a6 (visited on 03/16/2023).

[313] Arkaitz Zubiaga et al. “Detection and Resolution of Rumours in Social Media: A Survey”. In: *ACM Computing Surveys* 51.2 (Feb. 2018), 32:1–32:36. ISSN: 0360-0300. DOI: 10.1145/3161603. URL: https://dl.acm.org/doi/10.1145/3161603 (visited on 03/16/2023).

[314] Hadeer Ahmed, Issa Traore, and Sherif Saad. “Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques”. en. In: *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*. Ed. by Issa Traore, Isaac Woungang, and Ahmed Awad. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2017, pp. 127–138. ISBN: 978-3-319-69155-8. DOI: 10.1007/978-3-319-69155-8_9.

[315] Suman Kalyan Maity et al. “Detection of Sockpuppets in Social Media”. In: *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. CSCW ’17 Companion. New York, NY, USA: Association for Computing Machinery, Feb. 2017, pp. 243–246. ISBN: 978-1-4503-4688-7. DOI: 10.1145/3022198.3026360. URL: https://doi.org/10.1145/3022198.3026360 (visited on 03/16/2023).

[316] Peter Hernon. “Disinformation and misinformation through the internet: Findings of an exploratory study”. en. In: *Government Information Quarterly* 12.2 (Jan. 1995), pp. 133–139. ISSN: 0740-624X. DOI: 10.1016/0740-624X(95)90052-7. URL: https://www.sciencedirect.com/science/article/pii/0740624X95900527 (visited on 03/16/2023).

[317] Lennart van de Guchte et al. “Near Real-Time Detection of Misinformation on Online Social Networks”. en. In: *Disinformation in Open Online Media*. Ed. by Max van Duijn et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020, pp. 246–260. ISBN: 978-3-030-61841-4. DOI: 10.1007/978-3-030-61841-4_17.

[318] Manqing Dong et al. “DUAL: A Deep Unified Attention Model with Latent Relation Representations for Fake News Detection: 19th International Conference, Dubai, United Arab Emirates, November 12-15,

2018, Proceedings, Part I". In: Nov. 2018, pp. 199–209. ISBN: 978-3-030-02921-0. DOI: 10.1007/978-3-030-02922-7_14.

[319] Maria Konte, Nick Feamster, and Jaeyeon Jung. "Dynamics of Online Scam Hosting Infrastructure". en. In: *Passive and Active Network Measurement*. Ed. by Sue B. Moon, Renata Teixeira, and Steve Uhlig. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2009, pp. 219–228. ISBN: 978-3-642-00975-4. DOI: 10.1007/978-3-642-00975-4_22.

[320] Kate Starbird. "Examining the Alternative Media Ecosystem Through the Production of Alternative Narratives of Mass Shooting Events on Twitter". en. In: *Proceedings of the International AAAI Conference on Web and Social Media* 11.1 (May 2017). Number: 1, pp. 230–239. ISSN: 2334-0770. DOI: 10.1609/icwsm.v11i1.14878. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14878 (visited on 03/16/2023).

[321] Kai Shu, Suhang Wang, and Huan Liu. "Exploiting Tri-Relationship for Fake News Detection". In: (Dec. 2017).

[322] Todor Mihaylov et al. *Exposing Paid Opinion Manipulation Trolls*. arXiv:2109.13726 [cs]. Sept. 2021. DOI: 10.48550/arXiv.2109.13726. URL: http://arxiv.org/abs/2109.13726 (visited on 03/16/2023).

[323] Chandra Mouli Madhav Kotteti et al. "Fake news detection enhancement with data imputation". In: *Computer Information Systems Faculty Publications* (Oct. 2018). URL: https://digitalcommons.pvamu.edu/computer-information-facpubs/46.

[324] Sebastian Tschiatschek et al. "Fake News Detection in Social Networks via Crowd Signals". en. In: *Companion of the The Web Conference 2018 on The Web Conference 2018 - WWW '18*. Lyon, France: ACM Press, 2018, pp. 517–524. ISBN: 978-1-4503-5640-4. DOI: 10.1145/3184558.3188722. URL: http://dl.acm.org/citation.cfm?doid=3184558.3188722 (visited on 03/16/2023).

[325] Ankit Kesarwani, Sudakar Singh Chauhan, and Anil Ramachandran Nair. "Fake News Detection on Social Media using K-Nearest Neighbor Classifier". In: *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*. June 2020, pp. 1–4. DOI: 10.1109/ICACCE49060.2020.9154997.

[326] Yunfei Long et al. "Fake News Detection Through Multi-Perspective Speaker Profiles". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 252–256. URL: https://aclanthology.org/I17-2043 (visited on 03/16/2023).

[327] Sachin Kumar et al. "Fake news detection using deep learning models: A novel approach". en. In: *Transactions on Emerging Telecommunications Technologies* 31.2 (2020), e3767. ISSN: 2161-3915. DOI: 10.1002/ett.3767. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/ett.3767 (visited on 03/16/2023).

[328] Iftikhar Ahmad et al. "Fake News Detection Using Machine Learning Ensemble Methods". en. In: *Complexity* 2020 (Oct. 2020). Publisher: Hindawi, e8885861. ISSN: 1076-2787. DOI: 10.1155/2020/8885861. URL: https://www.hindawi.com/journals/complexity/2020/8885861/ (visited on 03/16/2023).

[329] Mykhailo Granik and Volodymyr Mesyura. "Fake news detection using naive Bayes classifier". In: *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. May 2017, pp. 900–903. DOI: 10.1109/UKRCON.2017.8100379.

[330] Aswini Thota et al. "Fake News Detection: A Deep Learning Approach". en. In: 1.3 (2018).

[331] Mehrdad Farajtabar et al. "Fake News Mitigation via Point Process Based Intervention". en. In: *Proceedings of the 34th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, July 2017, pp. 1097–1106. URL: https://proceedings.mlr.press/v70/farajtabar17a.html (visited on 03/16/2023).

[332] Amitabha Dey et al. "Fake News Pattern Recognition using Linguistic Analysis". In: *2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*. June 2018, pp. 305–309. DOI: 10.1109/ICIEV.2018.8641018.

[333] Lianwei Wu et al. "False Information Detection on Social Media via a Hybrid Deep Model". en. In: *Social Informatics*. Ed. by Steffen Staab, Olessia Koltsova, and Dmitry I. Ignatov. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018, pp. 323–333. ISBN: 978-3-030-01159-8. DOI: 10.1007/978-3-030-01159-8_31.

[334] Srijan Kumar and Neil Shah. "False Information on Web and Social Media: A Survey". In: (Apr. 2018).

[335] Ke Wu, Song Yang, and Kenny Q. Zhu. "False rumors detection on Sina Weibo by propagation structures". In: *2015 IEEE 31st International Conference on Data Engineering*. ISSN: 2375-026X. Apr. 2015, pp. 651–662. DOI: 10.1109/ICDE.2015.7113322.

[336] *FANG | Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. URL: https://dl.acm.org/doi/abs/10.1145/3340531.3412046 (visited on 03/16/2023).

[337] Niraj Sitaula et al. "Credibility-Based Fake News Detection". en. In: *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*. Ed. by Kai Shu et al. Lecture Notes in Social Networks. Cham: Springer International Publishing, 2020, pp. 163–182. ISBN: 978-3-030-42699-6. DOI: 10.1007/978-3-030-42699-6_9. URL: https://doi.org/10.1007/978-3-030-42699-6_9 (visited on 03/16/2023).

[338] Huiling Zhang et al. "Fight Under Uncertainty: Restraining Misinformation and Pushing out the Truth". In: *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. ISSN: 2473-991X. Aug. 2018, pp. 266–273. DOI: 10.1109/ASONAM.2018.8508402.

[339] Todor Mihaylov, Georgi Georgiev, and Preslav Nakov. "Finding Opinion Manipulation Trolls in News Community Forums". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Beijing, China: Association for Computational Linguistics, July 2015, pp. 310–314. DOI: 10.18653/v1/K15-1032. URL: https://aclanthology.org/K15-1032 (visited on 03/16/2023).

[340] Yang Liu and Yi-Fang Brook Wu. "FNED: A Deep Network for Fake News Early Detection on Social Media". In: *ACM Transactions on Information Systems* 38.3 (May 2020), 25:1–25:33. ISSN: 1046-8188. DOI: 10.1145/3386253. URL: https://doi.org/10.1145/3386253 (visited on 03/16/2023).

[341] Nitesh Chawla and Wei Wang, eds. *Proceedings of the 2017 SIAM International Conference on Data Mining*. en. Philadelphia, PA: Society for Industrial and Applied Mathematics, June 2017. ISBN: 978-1-61197-497-3. DOI: 10.1137/1.9781611974973. URL: https://epubs.siam.org/doi/book/10.1137/1.9781611974973 (visited on 03/16/2023).

[342] Liang Wu et al. "Gleaning Wisdom from the Past: Early Detection of Emerging Rumors in Social Media". In: June 2017, pp. 99–107. ISBN: 978-1-61197-497-3. DOI: 10.1137/1.9781611974973.12.

[343] Bo Ni et al. *Improving Generalizability of Fake News Detection Methods using Propensity Score Matching*. arXiv:2002.00838 [cs, stat]. Jan. 2020. DOI: 10.48550/arXiv.2002.00838. URL: http://arxiv.org/abs/2002.00838 (visited on 03/16/2023).

[344] Qiang Liu et al. "Mining Significant Microblogs for Misinformation Identification: An Attention-Based Approach". In: *ACM Transactions on Intelligent Systems and Technology* 9.5 (Apr. 2018), 50:1–50:20. ISSN: 2157-6904. DOI: 10.1145/3173458. URL: https://doi.org/10.1145/3173458 (visited on 03/16/2023).

[345] Steven Lloyd Wilson and Charles Wiysonge. "Social media and vaccine hesitancy". In: *BMJ global health* 5.10 (2020).

[346] Hamid Karimi et al. "Multi-Source Multi-Class Fake News Detection". In: *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 1546–1557. URL: https://aclanthology.org/C18-1131 (visited on 03/16/2023).

[347] Zhiwei Jin et al. "News Credibility Evaluation on Microblog with a Hierarchical Propagation Model". In: *2014 IEEE International Conference on Data Mining*. ISSN: 2374-8486. Dec. 2014, pp. 230–239. DOI: 10.1109/ICDM.2014.91.

[348] Dariusz Jemielniak and Yaroslav Krempovych. "An analysis of AstraZeneca COVID-19 vaccine misinformation and fear mongering on Twitter". In: *Public Health* 200 (2021), pp. 4–6.

[349] Zhiwei Jin et al. "News Verification by Exploiting Conflicting Social Viewpoints in Microblogs". en. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 30.1 (Mar. 2016). Number: 1. ISSN: 2374-3468. DOI: 10.1609/aaai.v30i1.10382. URL: https://ojs.aaai.org/index.php/AAAI/article/view/10382 (visited on 03/16/2023).

[350] Thomas Magelinski, Lynnette Ng, and Kathleen Carley. "A synchronized action framework for detection of coordination on social media". In: *Journal of Online Trust and Safety* 1.2 (2022).

[351] Sarthak Jindal et al. "NewsBag: A Multimodal Benchmark Dataset for Fake News Detection". en. In: ().

[352] Zhiwei Jin et al. "Novel Visual and Statistical Image Features for Microblogs News Verification". In: *IEEE Transactions on Multimedia* 19.3 (Mar. 2017). Conference Name: IEEE Transactions on Multimedia, pp. 598–608. ISSN: 1941-0077. DOI: 10.1109/TMM.2016.2617078.

[353] Michela Del Vicario et al. “Polarization and Fake News: Early Warning of Potential Misinformation Targets”. In: *ACM Transactions on the Web* 13.2 (Mar. 2019), 10:1–10:22. ISSN: 1559-1131. DOI: 10.1145/3316809. URL: https://doi.org/10.1145/3316809 (visited on 03/16/2023).

[354] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. “Limiting the spread of misinformation in social networks”. In: *Proceedings of the 20th international conference on World wide web*. WWW ’11. New York, NY, USA: Association for Computing Machinery, Mar. 2011, pp. 665–674. ISBN: 978-1-4503-0632-4. DOI: 10.1145/1963405.1963499. URL: https://doi.org/10.1145/1963405.1963499 (visited on 03/16/2023).

[355] *Web-based statistical fact checking of textual documents | Proceedings of the 2nd international workshop on Search and mining user-generated contents*. URL: https://dl.acm.org/doi/10.1145/1871985.1872002 (visited on 03/16/2023).

[356] Sejeong Kwon et al. “Prominent Features of Rumor Propagation in Online Social Media”. In: *2013 IEEE 13th International Conference on Data Mining*. ISSN: 2374-8486. Dec. 2013, pp. 1103–1108. DOI: 10.1109/ICDM.2013.61.

[357] Chris Grier et al. “@ spam: the underground on 140 characters or less”. In: *Proceedings of the 17th ACM conference on Computer and communications security*. 2010, pp. 27–37.

[358] Shirin Nilizadeh et al. “Poised: Spotting twitter spam off the beaten paths”. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 1159–1174.

[359] Taeri Kim et al. “Phishing URL Detection: A Network-based Approach Robust to Evasion”. In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 1769–1782.

[360] Hongyu Gao et al. “Detecting and characterizing social spam campaigns”. In: *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010, pp. 35–47.

[361] Kurt Thomas et al. “Design and evaluation of a real-time url spam filtering service”. In: *2011 IEEE symposium on security and privacy*. IEEE. 2011, pp. 447–462.

[362] Colin Whittaker, Brian Ryner, and Marria Nazif. “Large-scale automatic classification of phishing pages”. In: (2010).

[363] Mihai Christodorescu et al. “Semantics-aware malware detection”. In: *2005 IEEE symposium on security and privacy (S&P’05)*. IEEE. 2005, pp. 32–46.

[364] Fabio Giglietto et al. “It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 Italian elections”. In: *Information, Communication & Society* 23.6 (2020), pp. 867–891.

[365] Qiang Zhang et al. “Reply-Aided Detection of Misinformation via Bayesian Deep Learning”. In: *The World Wide Web Conference*. WWW ’19. New York, NY, USA: Association for Computing Machinery, May 2019, pp. 2333–2343. ISBN: 978-1-4503-6674-8. DOI: 10.1145/3308558.3313718. URL: https://doi.org/10.1145/3308558.3313718 (visited on 03/16/2023).

[366] Sejeong Kwon, Meeyoung Cha, and Kyomin Jung. “Rumor Detection over Varying Time Windows”. en. In: *PLOS ONE* 12.1 (Jan. 2017). Publisher: Public Library of Science, e0168344. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0168344. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0168344 (visited on 03/16/2023).

[367] Han Guo et al. “Rumor Detection with Hierarchical Social Attention Network”. In: *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. CIKM ’18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 943–951. ISBN: 978-1-4503-6014-2. DOI: 10.1145/3269206.3271709. URL: https://doi.org/10.1145/3269206.3271709 (visited on 03/16/2023).

[368] Xiang Lin et al. “Rumor Detection with Hierarchical Recurrent Convolutional Neural Network”. en. In: *Natural Language Processing and Chinese Computing*. Ed. by Jie Tang et al. Vol. 11839. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019, pp. 338–348. ISBN: 978-3-030-32235-9 978-3-030-32236-6. DOI: 10.1007/978-3-030-32236-6_30. URL: http://link.springer.com/10.1007/978-3-030-32236-6_30 (visited on 03/16/2023).

[369] Li Zeng, Kate Starbird, and Emma S. Spiro. “Rumors at the Speed of Light? Modeling the Rate of Rumor Transmission During Crisis”. In: *2016 49th Hawaii International Conference on System Sciences (HICSS)*. ISSN: 1530-1605. Jan. 2016, pp. 1969–1978. DOI: 10.1109/HICSS.2016.248.

[370] Kai-Cheng Yang et al. “Scalable and generalizable social bot detection through data selection”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 01. 2020, pp. 1096–1103.

[371] Laijun Zhao et al. “SIR rumor spreading model in the new media age”. en. In: *Physica A: Statistical Mechanics and its Applications* 392.4 (Feb. 2013), pp. 995–1003. ISSN: 0378-4371. DOI: 10.1016/j.physa.2012.09.030. URL: https://www.sciencedirect.com/science/article/pii/S037843711200934X (visited on 03/16/2023).

[372] Shivangi Singhal et al. “SpotFake: A Multi-modal Framework for Fake News Detection”. In: *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. Sept. 2019, pp. 39–47. DOI: 10.1109/BigMM.2019.00-44.

[373] Arjun Mukherjee, Bing Liu, and Natalie Glance. “Spotting fake reviewer groups in consumer reviews”. en. In: *Proceedings of the 21st international conference on World Wide Web*. Lyon France: ACM, Apr. 2012, pp. 191–200. ISBN: 978-1-4503-1229-5. DOI: 10.1145/2187836.2187863. URL: https://dl.acm.org/doi/10.1145/2187836.2187863 (visited on 03/16/2023).

[374] Daron Acemoglu, Asuman Ozdaglar, and Ali ParandehGheibi. “Spread of (mis)information in social networks”. en. In: *Games and Economic Behavior* 70.2 (Nov. 2010), pp. 194–227. ISSN: 0899-8256. DOI: 10.1016/j.geb.2010.01.005. URL: https://www.sciencedirect.com/science/article/pii/S0899825610000217 (visited on 03/16/2023).

[375] Inggrid Yanuar Risca Pratiwi, Rosa Andrie Asmara, and Faisal Rahutomo. “Study of hoax news detection using naïve bayes classifier in Indonesian language”. In: *2017 11th International Conference on Information & Communication Technology and System (ICTS)*. ISSN: 2338-185X. Oct. 2017, pp. 73–78. DOI: 10.1109/ICTS.2017.8265649.

[376] Kai Shu, H. Russell Bernard, and Huan Liu. *Studying Fake News via Network Analysis: Detection and Mitigation*. arXiv:1804.10233 [cs]. Apr. 2018. DOI: 10.48550/arXiv.1804.10233. URL: http://arxiv.org/abs/1804.10233 (visited on 03/16/2023).

[377] Julio C. S. Reis et al. “Supervised Learning for Fake News Detection”. en. In: *IEEE Intelligent Systems* 34.2 (Mar. 2019), pp. 76–81. ISSN: 1541-1672, 1941-1294. DOI: 10.1109/MIS.2019.2899143. URL: https://ieeexplore.ieee.org/document/8709925/ (visited on 03/16/2023).

[378] Kai Shu et al. “The role of user profiles for fake news detection”. In: *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*. ASONAM ’19. New York, NY, USA: Association for Computing Machinery, Jan. 2020, pp. 436–439. ISBN: 978-1-4503-6868-1. DOI: 10.1145/3341161.3342927. URL: https://dl.acm.org/doi/10.1145/3341161.3342927 (visited on 03/16/2023).

[379] Chengcheng Shao et al. “The spread of low-credibility content by social bots”. en. In: *Nature Communications* 9.1 (Nov. 2018). Number: 1 Publisher: Nature Publishing Group, p. 4787. ISSN: 2041-1723. DOI: 10.1038/s41467-018-06930-7. URL: https://www.nature.com/articles/s41467-018-06930-7 (visited on 03/16/2023).

[380] Jiwei Li et al. “Towards a General Rule for Identifying Deceptive Opinion Spam”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014, pp. 1566–1576. DOI: 10.3115/v1/P14-1147. URL: https://aclanthology.org/P14-1147 (visited on 03/16/2023).

[381] Suchita Jain, Vanya Sharma, and Rishabh Kaushal. “Towards automated real-time detection of misinformation on Twitter”. In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. Sept. 2016, pp. 2015–2020. DOI: 10.1109/ICACCI.2016.7732347.

[382] Victoria L Rubin, Niall J Conroy, and Yimin Chen. “Towards News Verification: Deception Detection Methods for News Discourse”. en. In: *HICSS2015* (2015).

[383] Victoria L. Rubin and Tatiana Lukoianova. “Truth and deception at the rhetorical structure level: Truth and Deception at the Rhetorical Structure Level”. en. In: *Journal of the Association for Information Science and Technology* 66.5 (May 2015), pp. 905–917. ISSN: 23301635. DOI: 10.1002/asi.23216. URL: https://onlinelibrary.wiley.com/doi/10.1002/asi.23216 (visited on 03/16/2023).

[384] Qiang Cao et al. “Aiding the detection of fake accounts in large scale social online services”. In: *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*. 2012, pp. 197–210.

[385] Savvas Zannettou et al. “Who let the trolls out? towards understanding state-sponsored trolls”. In: *Proceedings of the 10th acm conference on web science*. 2019, pp. 353–362.

[386] Gang Wang et al. “You are how you click: Clickstream analysis for sybil detection”. In: *22nd USENIX Security Symposium (USENIX Security 13)*. 2013, pp. 241–256.

[387] Haifeng Yu et al. “Sybillimit: A near-optimal social network defense against sybil attacks”. In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE. 2008, pp. 3–17.

[388] Zhi Yang, Christo Wilson, and Xiao Wang. "Uncovering Social Network Sybils in the Wild". en. In: ().

[389] Kareem Darwish et al. *Unsupervised User Stance Detection on Twitter*. arXiv:1904.02000 [cs]. May 2020. DOI: 10.48550/arXiv.1904.02000. URL: http://arxiv.org/abs/1904.02000 (visited on 03/16/2023).

[390] Brendan Nyhan and Jason Reifler. "When Corrections Fail: The Persistence of Political Misperceptions". en. In: *Political Behavior* 32.2 (June 2010), pp. 303–330. ISSN: 1573-6687. DOI: 10.1007/s11109-010-9112-2. URL: https://doi.org/10.1007/s11109-010-9112-2 (visited on 03/16/2023).

[391] Matthew Hindman and Vlad Barash. *Disinformation, 'Fake News' and Influence Campaigns on Twitter*. Oct. 2018.

[392] Pujan Paudel et al. "Lambretta: learning to rank for Twitter soft moderation". In: *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2023, pp. 311–326.

[393] Mary Ellen Zurko. "Disinformation and reflections from usable security". In: *IEEE Security & Privacy* 20.3 (2022), pp. 4–7.

[394] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives* 31.2 (May 2017), pp. 211–36. DOI: 10.1257/jep.31.2.211. URL: https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211.

[395] Tavish Vaidya et al. "Does Being Verified Make You More Credible? Account Verification's Effect on Tweet Credibility". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702. DOI: 10.1145/3290605.3300755. URL: https://doi.org/10.1145/3290605.3300755.

[396] Government News. "The Weibo rumor-refuting and co-governance platform is launched, giving government accounts the right to directly refute rumors". In: (Nov. 2018). URL: https://weibo.com/ttarticle/p/show?id=2309404301992787852969.

[397] Ivan Mehta and Manish Singh. *Twitter to end free access to its API in Elon Musk's latest monetization push*. Feb. 2023. URL: https://techcrunch.com/2023/02/01/twitter-to-end-free-access-to-its-api..

[398] Jon Porter. *Twitter announces new API pricing, posing a challenge for small developers*. Mar. 2023. URL: https://www.theverge.com/2023/3/30/23662832/twitter-api-tiers-free-bot-novelty-accounts-basic-enterprice-monthly-price.

[399] Simone McCarthy. *China's promotion of Russian disinformation indicates where its loyalties lie*. Mar. 2022. URL: https://www.cnn.com/2022/03/10/china/china-russia-disinformation-campaign-ukraine-intl-dst-hnk/index.html.

[400] Archive Team. *The Twitter Stream Grab*. URL: https://archive.org/details/twitterstream.

[401] Kai Shu et al. *FakeNewsNet: A data repository with news content, social context and spatialtemporal information for studying fake news on social media*. Mar. 2019. URL: https://arxiv.org/abs/1809.01286.

[402] Rafael Evangelista and Fernanda Bruno. *WhatsApp and political instability in Brazil: Targeted messages and political radicalisation*. Dec. 2019. URL: https://policyreview.info/articles/analysis/whatsapp-and-political-instability-brazil-targeted-messages-and-political.

[403] Kiran Garimella and Dean Eckles. *Images and misinformation in political groups: Evidence from WhatsApp in India: HKS Misinformation Review*. July 2022. URL: https://misinforeview.hks.harvard.edu/article/images-and-misinformation-in-political-groups-evidence-from-whatsapp-in-india/.

[404] Slick. *Commit to transparency - sign up for the international fact-checking network's code of Principles*. URL: https://ifcncodeofprinciples.poynter.org/.

[405] Sayash Kapoor and Arvind Narayanan. *Leakage and the Reproducibility Crisis in ML-based Science*. 2022. DOI: 10.48550/ARXIV.2207.07048. URL: https://arxiv.org/abs/2207.07048.

[406] Shachar Kaufman et al. "Leakage in Data Mining: Formulation, Detection, and Avoidance". In: *ACM Trans. Knowl. Discov. Data* 6.4 (Dec. 2012). ISSN: 1556-4681. DOI: 10.1145/2382577.2382579. URL: https://doi.org/10.1145/2382577.2382579.

[407] R. Nisbet, J. Elder, and G. Miner. *Handbook of Statistical Analysis and Data Mining Applications*. Association for Computing Machinery, 2009. ISBN: 978-0-12-374765-5.

[408] Song Feng, Ritwik Banerjee, and Yejin Choi. "Syntactic Stylometry for Deception Detection". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Jeju Island, Korea: Association for Computational Linguistics, July 2012, pp. 171–175. URL: https://aclanthology.org/P12-2034.

[409] Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. "Box of Lies: Multimodal Deception Detection in Dialogues". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 1768–1777. DOI: 10.18653/v1/N19-1175. URL: https://aclanthology.org/N19-1175.

[410] Joseph B. Bak-Coleman et al. *Combining interventions to reduce the spread of viral misinformation*. June 2022. URL: https://www.nature.com/articles/s41562-022-01388-6.

[411] Cristiano Lima. *A whistleblower's power: Key Takeaways from the Facebook papers*. Mar. 2022. URL: https://www.washingtonpost.com/technology/2021/10/25/what-are-the-facebook-papers/.

[412] *Content fact-checkers prioritize*. URL: https://transparency.fb.com/en-gb/features/content-fact-checkers-prioritize/.

[413] Fabiana Zollo and Walter Quattrociocchi. "Misinformation Spreading on Facebook". In: *Complex Spreading Phenomena in Social Systems: Influence and Contagion in Real-World Social Networks*. Ed. by Sune Lehmann and Yong-Yeol Ahn. Cham: Springer International Publishing, 2018, pp. 177–196. ISBN: 978-3-319-77332-2. DOI: 10.1007/978-3-319-77332-2_10. URL: https://doi.org/10.1007/978-3-319-77332-2_10.

[414] Alexandre Bovet and Hernán A. Makse. *Influence of fake news in Twitter during the 2016 US presidential election*. Jan. 2019. URL: https://www.nature.com/articles/s41467-018-07761-2#citeas.

[415] Sunday Oluwafemi Oyeyemi, Elia Gabarron, and Rolf Wynn. "Ebola, Twitter, and misinformation: a dangerous combination?" In: *Bmj* 349 (2014).

[416] Wasim Ahmed et al. "COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data". In: *Journal of medical internet research* 22.5 (2020), e19458.

[417] Darsh J Shah, Tal Schuster, and Regina Barzilay. "Automatic Fact-guided Sentence Modification". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. 2020. URL: https://arxiv.org/abs/1909.13838.

[418] Robert Geirhos et al. "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2.11 (Nov. 2020), pp. 665–673. DOI: 10.1038/s42256-020-00257-z. URL: https://doi.org/10.1038%2Fs42256-020-00257-z.

[419] Meta. *How Meta's third-party fact-checking program works*. URL: https://www.facebook.com/formedia/blog/third-party-fact-checking-how-it-works.

[420] Amber Jamieson and Olivia Solon. *Facebook to begin flagging fake news in response to mounting criticism*. Dec. 2016. URL: https://www.theguardian.com/technology/2016/dec/15/facebook-flag-fake-news-fact-check.

[421] Colin Crowell. *Our approach to bots and misinformation*. URL: https://blog.twitter.com/official/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html.

[422] Nur Ibrahim. *Why facebook won't be fact-checking Trump now that he's announced candidacy*. Nov. 2022. URL: https://www.snopes.com/news/2022/11/17/facebook-not-fact-checking-trump/.

[423] Ben Kaiser and Jonathan Mayer. *'It's the Algorithm: A Large-Scale Comparative Field Study of Misinformation Interventions'*. 2023.

[424] *GossipCop*. URL: https://web.archive.org/web/20221102102621/https://gossipcop.com/.

[425] Oliver Darcy. *BuzzFeed News will shut down | CNN business*. Apr. 2023. URL: https://www.cnn.com/2023/04/20/media/buzzfeed-news-shuts-down/index.html.

[426] Tiffany Hsu. *As covid-19 continues to spread, so does misinformation about it*. Dec. 2022. URL: https://www.nytimes.com/2022/12/28/technology/covid-misinformation-online.html.

[427] Emily Bell. *The fact-check industry*. Dec. 2019. URL: https://www.cjr.org/special_report/fact-check-industry-twitter.php.

[428] Ryan Woo Liangping Gao. *China's factory activity falls faster than expected as recovery stumbles*. May 2023. URL: https://www.reuters.com/markets/asia/chinas-factory-activity-falls-faster-than-expected-weak-demand-pmi-2023-05-31/.

[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].