References:

[1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36]

[1] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, “Human-in-the-loop machine learning: a state of the art,” *Artif. Intell. Rev.*, vol. 56, no. 4, pp. 3005–3054, Apr. 2023, doi: 10.1007/s10462-022-10246-w.

[2] D. Xin, L. Ma, J. Liu, S. Macke, S. Song, and A. Parameswaran, “Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities,” in *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, Houston TX USA: ACM, June 2018, pp. 1–4. doi: 10.1145/3209889.3209897.

[3] C. Chai and G. Li, “Human-in-the-loop Techniques in Machine Learning”.

[4] O. Gómez-Carmona, D. Casado-Mansilla, D. López-de-Ipiña, and J. García-Zubia, “Human-in-the-loop machine learning: Reconceptualizing the role of the user in interactive approaches,” *Internet Things*, vol. 25, p. 101048, Apr. 2024, doi: 10.1016/j.iot.2023.101048.

[5] S. Rahman and E. Kandogan, “Characterizing Practices, Limitations, and Opportunities Related to Text Information Extraction Workflows: A Human-in-the-loop Perspective,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–15. doi: 10.1145/3491102.3502068.

[6] Y. Cui *et al.*, “Understanding the Relationship between Interactions and Outcomes in Human-in-the-Loop Machine Learning,” in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, Montreal, Canada: International Joint Conferences on Artificial Intelligence Organization, Aug. 2021, pp. 4382–4391. doi: 10.24963/ijcai.2021/599.

[7] Rahul, H. Kajla, J. Hooda, and G. Saini, “Classification of Online Toxic Comments Using Machine Learning Algorithms,” in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2020, pp. 1119–1123. doi: 10.1109/ICICCS48265.2020.9120939.

[8] M. O. Olayiwola, “23 PUBLICATIONS 57 CITATIONS SEE PROFILE”.

[9] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, “Measuring and Mitigating Unintended Bias in Text Classification,” in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, New Orleans LA USA: ACM, Dec. 2018, pp. 67–73. doi: 10.1145/3278721.3278729.

[10] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, “Nuanced Metrics for Measuring Unintended Bias with Real Data for Text Classification,” in *Companion Proceedings of The 2019 World Wide Web Conference*, San Francisco USA: ACM, May 2019, pp. 491–500. doi: 10.1145/3308560.3317593.

[11] D. S. Vijayarani and J. Ilamathi, “Preprocessing Techniques for Text Mining - An Overview,” vol. 5.

[12] R. Dzisevič and D. Šešok, “Text Classification using Different Feature Extraction Approaches,” in *2019 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, Apr. 2019, pp. 1–4. doi: 10.1109/eStream.2019.8732167.

[13] J. Wang, B. Guo, and L. Chen, “Human-in-the-loop Machine Learning: A Macro-Micro Perspective,” Feb. 21, 2022, *arXiv*: arXiv:2202.10564. doi: 10.48550/arXiv.2202.10564.

[14] J. L. Huan, A. A. Sekh, C. Quek, and D. K. Prasad, “Emotionally charged text classification with deep learning and sentiment semantic,” *Neural Comput. Appl.*, vol. 34, no. 3, pp. 2341–2351, Feb. 2022, doi: 10.1007/s00521-021-06542-1.

[15] A. Bandhakavi, N. Wiratunga, D. Padmanabhan, and S. Massie, “Lexicon based feature extraction for emotion text classification,” *Pattern Recognit. Lett.*, vol. 93, pp. 133–142, July 2017, doi: 10.1016/j.patrec.2016.12.009.

[16] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, “The Risk of Racial Bias in Hate Speech Detection,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1668–1678. doi: 10.18653/v1/P19-1163.

[17] H. Al Kuwatly, M. Wich, and G. Groh, “Identifying and Measuring Annotator Bias Based on Annotators’ Demographic Characteristics,” in *Proceedings of the Fourth Workshop on Online Abuse and Harms*, S. Akiwowo, B. Vidgen, V. Prabhakaran, and Z. Waseem, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 184–190. doi: 10.18653/v1/2020.alw-1.21.

[18] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, “‘It’s Reducing a Human Being to a Percentage’: Perceptions of Justice in Algorithmic Decisions,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, in CHI ’18. New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–14. doi: 10.1145/3173574.3173951.

[19] S. Pradha, M. N. Halgamuge, and N. Tran Quoc Vinh, “Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data,” in *2019 11th International Conference on Knowledge and Systems Engineering (KSE)*, Oct. 2019, pp. 1–8. doi: 10.1109/KSE.2019.8919368.

[20] A. Guo and T. Yang, “Research and improvement of feature words weight based on TFIDF algorithm,” in *2016 IEEE Information Technology, Networking, Electronic and Automation Control Conference*, May 2016, pp. 415–419. doi: 10.1109/ITNEC.2016.7560393.

[21] B. van Aken, J. Risch, R. Krestel, and A. Löser, “Challenges for Toxic Comment Classification: An In-Depth Error Analysis,” Sept. 20, 2018, *arXiv*: arXiv:1809.07572. doi: 10.48550/arXiv.1809.07572.

[22] M. Ibrahim, M. Torki, and N. El-Makky, “Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning,” in *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2018, pp. 875–878. doi: 10.1109/ICMLA.2018.00141.

[23] K. Matsumoto, F. Ren, M. Matsuoka, M. Yoshida, and K. Kita, “Slang feature extraction by analysing topic change on social media,” *CAAI Trans. Intell. Technol.*, vol. 4, no. 1, pp. 64–71, 2019, doi: 10.1049/trit.2018.1060.

[24] M. Shrivastava and S. Kumar, “A pragmatic and intelligent model for sarcasm detection in social media text,” *Technol. Soc.*, vol. 64, p. 101489, Feb. 2021, doi: 10.1016/j.techsoc.2020.101489.

[25] R. Misra and P. Arora, “Sarcasm detection using news headlines dataset,” *AI Open*, vol. 4, pp. 13–18, Jan. 2023, doi: 10.1016/j.aiopen.2023.01.001.

[26] X. Zhang, R. Mao, and E. Cambria, “Multilingual Emotion Recognition: Discovering the Variations of Lexical Semantics between Languages,” in *2024 International Joint Conference on Neural Networks (IJCNN)*, June 2024, pp. 1–9. doi: 10.1109/IJCNN60899.2024.10651409.

[27] B. Liu *et al.*, “Context-aware social media user sentiment analysis,” *Tsinghua Sci. Technol.*, vol. 25, no. 4, pp. 528–541, Aug. 2020, doi: 10.26599/TST.2019.9010021.

[28] X. Jiang, B. Ren, Q. Wu, W. Wang, and H. Li, “DCASAM: advancing aspect-based sentiment analysis through a deep context-aware sentiment analysis model,” *Complex Intell. Syst.*, vol. 10, no. 6, pp. 7907–7926, Dec. 2024, doi: 10.1007/s40747-024-01570-5.

[29] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi, and S. K. Shahzad, “Emotion Detection of Contextual Text using Deep learning,” in *2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, Oct. 2020, pp. 1–5. doi: 10.1109/ISMSIT50672.2020.9255279.

[30] J. Guo, “Deep learning approach to text analysis for human emotion detection from big data,” *J. Intell. Syst.*, vol. 31, no. 1, pp. 113–126, Jan. 2022, doi: 10.1515/jisys-2022-0001.

[31] M. Chiodo, D. Müller, P. Siewert, J.-L. Wetherall, Z. Yasmine, and J. Burden, “Formalising Human-in-the-Loop: Computational Reductions, Failure Modes, and Legal-Moral Responsibility,” May 15, 2025, *arXiv*: arXiv:2505.10426. doi: 10.48550/arXiv.2505.10426.

[32] M. Huelser, H. Mueller, N. Díaz-Rodríguez, and A. Holzinger, “On the disagreement problem in Human-in-the-Loop federated machine learning,” *J. Ind. Inf. Integr.*, vol. 45, p. 100827, May 2025, doi: 10.1016/j.jii.2025.100827.

[33] S. Wilson *et al.*, “Analyzing Privacy Policies at Scale: From Crowdsourcing to Automated Annotations,” *ACM Trans Web*, vol. 13, no. 1, p. 1:1-1:29, Dec. 2018, doi: 10.1145/3230665.

[34] A. Alabduljabbar, A. Abusnaina, Ü. Meteriz-Yildiran, and D. Mohaisen, “TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights,” in *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, in WPES ’21. New York, NY, USA: Association for Computing Machinery, Nov. 2021, pp. 103–118. doi: 10.1145/3463676.3485608.

[35] Z. Lin *et al.*, “ToxicChat: Unveiling Hidden Challenges of Toxicity Detection in Real-World User-AI Conversation,” Oct. 26, 2023, *arXiv*: arXiv:2310.17389. doi: 10.48550/arXiv.2310.17389.

[36] B. Settles, M. Craven, and L. Friedland, “Active Learning with Real Annotation Costs”.