REFERENCES

[1] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. 2007. DBpedia: A Nucleus for a Web of Open Data. In *ISWC'07/ASWC'07* (Busan, Korea). 722–735.

[2] L. Bernardi, J. Kamps, J. Kiseleva, and M. JI Müller. 2015. The continuous cold start problem in e-commerce recommender systems. *arXiv:1508.01177* (2015).

[3] A. Beutel, J. Chen, T. Doshi, H. Qian, L. Wei, Y. Wu, L. Heldt, Z. Zhao, L. Hong, E. H. Chi, and C. Goodrow. 2019. Fairness in Recommendation Ranking through Pairwise Comparisons. *KDD* (2019), 2212–2220.

[4] A. J. Biega, K. P. Gummadi, and G. Weikum. 2018. Equity of Attention: Amortizing Individual Fairness in Rankings. In *SIGIR*. ACM, 405–414.

[5] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe, and M. Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *CHI*.

[6] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S Liang. 2019. Unlabeled data improves adversarial robustness. In *NeurIPS*. 11192–11203.

[7] K. Christakopoulou and A. Banerjee. 2019. Adversarial Attacks on an Oblivious Recommender. In *RecSys* (Copenhagen, Denmark). 322–330.

[8] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918* (2019).

[9] Y. Deldjoo, V. W. Anelli, H. Zamani, A. Bellogin, and T. Di Noia. 2019. Recommender Systems Fairness Evaluation via Generalized Cross Entropy. arXiv:1908.06708

[10] Y. Deldjoo, T. Di Noia, E. Di Sciascio, and F. A. Merra. 2020. How Dataset Characteristics Affect the Robustness of Collaborative Recommendation Models. In *SIGIR* (Virtual Event, China). 951–960.

[11] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

[12] Y. Ding, J. Liu, and D. Wang. 2018. Deep Feature Fusion over Multi-Field Categorical Data for Rating Prediction. In *AICCC* (Tokyo, Japan). 16–22.

[13] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[14] E. D. Gennatas, J. H. Friedman, L. H. Ungar, R. Pirracchio, E. Eaton, L. G. Reichmann, Y. Interian, J. M. Luna, C. B. Simone, A. Auerbach, E. Delgado, M. J. van der Laan, T. D. Solberg, and G. Valdes. 2020. Expert-augmented machine learning. *PNAS* 117, 9 (2020), 4571–4577.

[15] A. Ghazimatin, O. Balalau, R. Saha Roy, and G. Weikum. 2020. PRINCE: Provider-Side Interpretability with Counterfactual Explanations in Recommender Systems. In *WSDM* (Houston, TX, USA). 196–204.

[16] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. A. Mann, and P. Kohli. 2018. On the Effectiveness of Interval Bound Propagation for Training Verifiably Robust Models. (2018). arXiv:1810.12715

[17] A. Gunawardana and C. Meek. 2009. A unified approach to building hybrid recommender systems. In *RecSys*. 117–124.

[18] G. Hadash, O. S. Shalom, and R. Osadchy. 2018. Rank and rate: multi-task learning for recommender systems. In *RecSys*. 451–454.

[19] F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4, Article 19 (Dec. 2015), 19 pages.

[20] Taha Hassan. 2019. Trust and trustworthiness in social recommender systems. In *Companion Proceedings of The 2019 World Wide Web Conference*. 529–532.

[21] X. He, Z. He, X. Du, and T.S. Chua. 2018. Adversarial Personalized Ranking for Recommendation. In *SIGIR*. 355–364.

[22] G. Hinton, O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[23] Y.L. Hsieh, M. Cheng, D.C. Juan, W. Wei, W.L. Hsu, and C.J. Hsieh. 2019. On the Robustness of Self-Attentive Models. In *ACL*. Florence, Italy, 1520–1529.

[24] N. J. Hurley. 2011. Robustness of Recommender Systems. In *RecSys*. 9–10.

[25] A Ilyas, S Santurkar, D Tsipras, L Engstrom, B Tran, and A Madry. 2019. Adversarial Examples Are Not Bugs, They Are Features. In *NeurIPS*, Vol. 32. 125–136.

[26] R. Jia, A. Raghunathan, K. Göksel, and P. Liang. 2019. Certified Robustness to Adversarial Word Substitutions. *EMNLP-IJCNLP* (2019).

[27] W.C. Kang, D. Z. Cheng, T. Chen, X. Yi, D. Lin, L. Hong, and E. H. Chi. 2020. Learning Multi-Granular Quantized Embeddings for Large-Vocab Categorical Features in Recommender Systems. In *WWW* (Taipei, Taiwan). 562–566.

[28] H Kannan, A Kurakin, and I Goodfellow. 2018. Adversarial logit pairing. *arXiv:1803.06373* (2018).

[29] D. Kaushik, E. Hovy, and Z. C. Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv:1909.12434* (2019).

[30] B. P. Knijnenburg, N. J.M. Reijmer, and M. C. Willemsen. 2011. Each to His Own: How Different Users Call for Different Interaction Methods in Recommender Systems. In *RecSys* (Chicago, Illinois, USA). 141–148.

[31] Igor Kononenko. 1993. Inductive and Bayesian learning in medical diagnosis. *Applied Artificial Intelligence* 7 (1993), 317–337.

[32] S. Krichene, M. Gartrell, and C. Calauzènes. 2019. Embedding models for recommendation under contextual constraints. *arXiv* abs/1907.01637 (2019).

[33] D. G. Lee, K. S. Ryu, M. Bashir, J.W. Bae, and K. H. Ryu. 2013. Discovering Medical Knowledge using Association Rule Mining in Young Adults with Acute Myocardial Infarction. *Journal of Medical Systems* 37, 2 (15 Jan 2013), 9896.

[34] Z. C. Lipton. 2016. The Mythos of Model Interpretability. (2016). arXiv:1606.03490

[35] P. Massa and P. Avesani. [n.d.]. Trust-aware recommender systems. In *RecSys*.

[36] J. McAuley, R. Pandey, and J. Leskovec. 2015. Inferring networks of substitutable and complementary products. In *KDD*. 785–794.

[37] K. Morik, P. Brockhausen, and T. Joachims. 1999. Combining Statistical Learning with a Knowledge-Based Approach - A Case Study in Intensive Care Monitoring. In *ICML*. 268–277.

[38] R.K Mothilal, A Sharma, and C Tan. 2019. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *CoRR* abs/1905.07697 (2019).

[39] M P O'Mahony, N J Hurley, and G Silvestre. 2005. Recommender systems: Attack types and strategies. In *AAAI*. 334–339.

[40] World Health Organization. 1978. *International classification of diseases*.

[41] World Health Organization. 2003. Anatomical Therapeutic Chemical (ATC) Classification System with Defined Daily Doses. (2003).

[42] M O'Mahony, N Hurley, N Kushmerick, and G Silvestre. 2004. Collaborative Recommendation: A Robustness Analysis. *ACM Trans. Internet Technol.* (2004).

[43] M Pawelczyk, K Broelemann, and G Kasneci. 2020. Learning Model-Agnostic Counterfactual Explanations for Tabular Data *(WWW '20)*.

[44] Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect* (1st ed.). Basic Books, Inc., USA.

[45] C Qin, J Martens, S Gowal, D Krishnan, K Dvijotham, A Fawzi, S De, R Stanforth, and P Kohli. 2019. Adversarial robustness through local linearization. In *NeurIPS*.

[46] T Qin, T Liu, J Xu, and H Li. 2010. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval. *Inf. Retr.* 13, 4 (2010).

[47] Y Qin, X Wang, A Beutel, and E.H Chi. 2020. Improving Uncertainty Estimates through the Relationship with Adversarial Robustness. *ArXiv* 2006.16375 (2020).

[48] B Rastegarpanah, K P Gummadi, and M Crovella. 2019. Fighting fire with fire: Using antidote data to improve polarization and fairness of recommender systems. In *WWW*. 231–239.

[49] M.T Ribeiro, S Singh, and C Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016).

[50] J Shang, T Ma, C Xiao, and J Sun. 2019. Pre-training of Graph Augmented Transformers for Medication Recommendation. *CoRR* abs/1906.00346 (2019).

[51] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*, Yike Guo and Faisal Farooq (Eds.). ACM, 2219–2228.

[52] A Sinha, D. F. Gleich, and K Ramani. 2017. Deconvolving Feedback Loops in Recommender Systems. *CoRR* abs/1703.01049 (2017).

[53] Harald Steck. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. *CoRR* abs/1905.03375 (2019). arXiv:1905.03375 http://arxiv.org/abs/1905.03375

[54] B Whitman T Bertin-Mahieux, D. P.W. Ellis and P Lamere. 2011. The Million Song Dataset. *ISMIR* (2011).

[55] A Taha and A Hadi. 2019. Anomaly Detection Methods for Categorical Data: A Review. *ACM Comput. Surv.* 52, 2, Article 38 (May 2019).

[56] Brendon Towle and Clark N. Quinn. 2000. Knowledge Based Recommender Systems Using Explicit User Models.

[57] P Velickovic, G Cucurull, A Casanova, A Romero, P Liò, and Y Bengio. 2018. Graph Attention Networks. *ArXiv* abs/1710.10903 (2018).

[58] J Villena-Román, S Collada-Pérez, S Lana-Serrano, and J.C González-Cristóbal. 2011. Hybrid approach combining machine learning and a rule-based expert system for text categorization. In *Twenty-Fourth International FLAIRS Conference*.

[59] J Wang and J Caverlee. 2019. Recurrent recommendation with local coherence *(WSDM)*.

[60] Z Wang, Z Jiang, Z Ren, J Tang, and D Yin. 2018. A path-constrained framework for discriminating substitutable and complementary products in e-commerce. In *WWW*.

[61] Filzmoser P Gyimesi M. Weisser A, Endel G. 2008. ATC -> ICD – evaluating the reliability of prognoses for ICD-10 diagnoses derived from the ATC-Code of prescriptions. *BMC Health Serv Res. 2008* (2008).

[62] Michael Wellman and Max Henrion. 1993. Explaining 'explaining away'. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 15 (04 1993), 287–292.

[63] A. B. Wilcox and G. Hripcsak. 2003. The role of domain knowledge in automating medical text report classification. *JAMIA* (2003).

[64] C Xie, Y Wu, L Maaten, A Yuille, and K He. 2019. Feature denoising for improving adversarial robustness. In *ICCV*.

[65] D Xin, N Mayoraz, H Pham, K Lakshmanan, and J Anderson. 2017. Folding: Why Good Models Sometimes Make Spurious Recommendations *(RecSys '17)*.

[66] H Zhang, Y Yu, J Jiao, E Xing, L Ghaoui, and M Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. CoRR/1901.08573 (2019).

[67] Z Zhang, C Xie, J Wang, L Xie, and A Yuille. 2018. Deepvoting: A robust and explainable deep network for semantic part detection under partial occlusion. In *ICCV, 2018*. 1372–1380.

[68] Q Zhao, J Chen, M Chen, S Jain, A Beutel, F Belletti, and E. Chi. 2018. Categorical-Attributes-Based Item Classification for Recommender Systems *(RecSys '18)*.

[69] G Zhou, X Zhu, C Song, Y Fan, H Zhu, X Ma, Y Yan, J Jin, H Li, and K Gai. 2018. Deep Interest Network for Click-Through Rate Prediction *(ACM SIGKDD '18)*.

[70] R Zmigrod, S Mielke, H Wallach, and R Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571* (2019).