REFERENCES

[1] "Number of smartphone users worldwide," https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

[2] L. V. G. Carreno and K. Winbladh, "Analysis of user comments: an approach for software requirements evolution," in *2013 35th international conference on software engineering (ICSE)*. IEEE, 2013, pp. 582–591.

[3] E. Guzman and W. Maalej, "How do users like this feature? a fine grained sentiment analysis of app reviews," in *2014 IEEE 22nd international requirements engineering conference (RE)*. Ieee, 2014, pp. 153–162.

[4] D. Pagano and B. Bruegge, "User involvement in software evolution practice: a case study," in *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, 2013, pp. 953–962.

[5] M. Harman, Y. Jia, and Y. Zhang, "App store mining and analysis: Msr for app stores," in *2012 9th IEEE working conference on mining software repositories (MSR)*. IEEE, 2012, pp. 108–111.

[6] N. Chen, J. Lin, S. C. Hoi, X. Xiao, and B. Zhang, "Ar-miner: mining informative reviews for developers from mobile app marketplace," in *Proceedings of the 36th international conference on software engineering*, 2014, pp. 767–778.

[7] F. Palomba, M. Linares-Vásquez, G. Bavota, R. Oliveto, M. Di Penta, D. Poshyvanyk, and A. De Lucia, "Crowdsourcing user reviews to support the evolution of mobile apps," *Journal of Systems and Software*, vol. 137, pp. 143–162, 2018.

[8] E. Guzman, M. El-Haliby, and B. Bruegge, "Ensemble methods for app review classification: An approach for software evolution (n)," in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2015, pp. 771–776.

[9] S. Panichella, A. Di Sorbo, E. Guzman, C. A. Visaggio, G. Canfora, and H. C. Gall, "How can i improve my app? classifying user reviews for software maintenance and evolution," in *2015 IEEE international conference on software maintenance and evolution (ICSME)*. IEEE, 2015, pp. 281–290.

[10] E. Guzman, M. Ibrahim, and M. Glinz, "A little bird told me: Mining tweets for requirements and software evolution," 09 2017.

[11] N. Alturaief, H. Aljamaan, and M. Baslyman, "Aware: Aspect-based sentiment analysis dataset of apps reviews for requirements elicitation," in *2021 36th*

IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW). IEEE, 2021, pp. 211–218.

[12] D. Pagano and W. Maalej, "User feedback in the appstore: An empirical study," in *2013 21st IEEE international requirements engineering conference (RE)*. IEEE, 2013, pp. 125–134.

[13] H. Li, L. Zhang, L. Zhang, and J. Shen, "A user satisfaction analysis approach for software evolution," vol. 2, pp. 1093–1097, 2010.

[14] W. Maalej, M. Nayebi, T. Johann, and G. Ruhe, "Toward data-driven requirements engineering," *IEEE software*, vol. 33, no. 1, pp. 48–54, 2015.

[15] M. V. Phong, T. T. Nguyen, H. V. Pham, and T. T. Nguyen, "Mining user opinions in mobile app reviews: A keyword-based approach (t)," pp. 749–759, 2015.

[16] B. Fu, J. Lin, L. Li, C. Faloutsos, J. Hong, and N. Sadeh, "Why people hate your app: Making sense of user feedback in a mobile app store," pp. 1276–1284, 2013.

[17] R. T. Anchiêta and R. S. Moura, "Exploring unsupervised learning towards extractive summarization of user reviews," in *Proceedings of the 23rd Brazillian Symposium on Multimedia and the Web*, 2017, pp. 217–220.

[18] M. Gomez, R. Rouvoy, M. Monperrus, and L. Seinturier, "A recommender system of buggy app checkers for app store moderators," pp. 1–11, 2015.

[19] W. Maalej, Z. Kurtanović, H. Nabil, and C. Stanik, "On the automatic classification of app reviews," *Requirements Engineering*, vol. 21, no. 3, pp. 311–331, 2016.

[20] X. Gu and S. Kim, ""' what parts of your apps are loved by users?"(t)," in *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2015, pp. 760–770.

[21] V. T. Dhinakaran, R. Pulle, N. Ajmeri, and P. K. Murukannaiah, "App review analysis via active learning: reducing supervision effort without compromising classification accuracy," in *2018 IEEE 26th international requirements engineering conference (RE)*. IEEE, 2018, pp. 170–181.

[22] H. Guo and M. P. Singh, "Caspar: Extracting and synthesizing user stories of problems from app reviews," 2020, p. 628–640.

[23] C. Stanik, M. Haering, and W. Maalej, "Classifying multilingual user feedback using traditional machine learning and deep learning," in *2019 IEEE 27th international requirements engineering conference workshops (REW)*. IEEE, 2019, pp. 220–226.

[24] N. Aslam, W. Y. Ramay, K. Xia, and N. Sarwar, "Convolutional neural network based classification of app reviews," *IEEE Access*, vol. 8, pp. 185 619–185 628, 2020.

[25] M. A. Hadi and F. H. Fard, "Evaluating pre-trained models for user feedback analysis in software engineering: A study on classification of app-reviews," 2021.

[26] P. R. Henao, J. Fischbach, D. Spies, J. Frattini, and A. Vogelsang, "Transfer learning for mining feature requests and bug reports from tweets and app store reviews," in *2021 IEEE 29th International Requirements Engineering Conference Workshops (REW)*. IEEE, 2021, pp. 80–86.

[27] J. Verma and A. Patel, "Evaluation of unsupervised learning based extractive text summarization technique for large scale review and feedback data," *Indian Journal of Science and Technology*, vol. 10, pp. 1–6, 05 2017.

[28] S. Wang, Y. Liu, Y. Xu, C. Zhu, and M. Zeng, "Want to reduce labeling cost? gpt-3 can help," *arXiv preprint arXiv:2108.13487*, 2021.

[29] X. He, Z. Lin, Y. Gong, A. Jin, H. Zhang, C. Lin, J. Jiao, S. M. Yiu, N. Duan, W. Chen *et al.*, "Annollm: Making large language models to be better crowdsourced annotators," *arXiv preprint arXiv:2303.16854*, 2023.

[30] R. Zhang, Y. Li, Y. Ma, M. Zhou, and L. Zou, "Llmaaa: Making large language models as active annotators," *arXiv preprint arXiv:2310.19596*, 2023.

[31] J. Zhou, W. Du, M. O. F. Rokon, Z. Wang, J. Xu, I. Shah, K.-c. Lee, and M. Wen, "Enhanced e-commerce attribute extraction: Innovating with decorative relation correction and llama 2.0-based annotation," *arXiv preprint arXiv:2312.06684*, 2023.

[32] Z. He, C.-Y. Huang, C.-K. C. Ding, S. Rohatgi, and T.-H. K. Huang, "If in a crowdsourced data annotation pipeline, a gpt-4," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–25.

[33] Y. Tang, C.-M. Chang, and X. Yang, "Pdfchatannotator: A human-llm collaborative multi-modal data annotation tool for pdf-format catalogs," in *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 2024, pp. 419–430.

[34] D. Yu, L. Li, H. Su, and M. Fuoli, "Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology," *International Journal of Corpus Linguistics*, 2024.

[35] M. Imamovic, S. Deilen, D. Glynn, and E. Lapshinova-Koltunski, "Using chatgpt for annotation of attitude within the appraisal theory: Lessons learned," in *Proceedings of The 18th Linguistic Annotation Workshop (LAW-XVIII)*, 2024, pp. 112–123.

[36] A. Wang, J. Morgenstern, and J. P. Dickerson, "Large language models cannot replace human participants because they cannot portray identity groups," *arXiv preprint arXiv:2402.01908*, 2024.

[37] N. Pangakis, S. Wolken, and N. Fasching, "Automated annotation with generative ai requires validation," *arXiv preprint arXiv:2306.00176*, 2023.

[38] J. Tan, A. Zhang, X. Zhang, C. Xiao, Z. Ding, Y. Peng, C. Wu, X. Zhu, J. Zhou, and X. Huang, "Large language models for data annotation: A survey," *arXiv preprint arXiv:2402.13446*, 2024.

[39] S. Gunathilaka and N. De Silva, "Aspect-based sentiment analysis on mobile application reviews," in *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*. IEEE, 2022, pp. 183–188.

[40] S. Kobayashi, "Contextual augmentation: Data augmentation by words with paradigmatic relations," 2018.

[41] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[42] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[43] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez *et al.*, "Vicuna: An open-source chatbot impressing gpt-4 with 90

[44] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, E. Goffinet, D. Heslow, J. Launay, Q. Malartic *et al.*, "Falcon-40b: an open large language model with state-of-the-art performance," 2023.

[45] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[46] L. Reiter, "Zephyr," *Journal of Business Finance Librarianship*, vol. 18, no. 3, pp. 259–263, 2013.

[47] G. Colavito, F. Lanubile, N. Novielli, and L. Quaranta, "Leveraging gpt-like llms to automate issue labeling," in *2024 IEEE/ACM 21st International Conference on Mining Software Repositories (MSR)*. IEEE, 2024, pp. 469–480.

[48] W. Maalej and H. Nabil, "Bug report, feature request, or simply praise? on automatically classifying app reviews," in *2015 IEEE 23rd international requirements engineering conference (RE)*. IEEE, 2015, pp. 116–125.

[49] D. Yu, L. Li, H. Su, and M. Fuoli, "Assessing the potential of llm-assisted annotation for corpus-based pragmatics and discourse analysis."

[50] K. Hamilton, L. Longo, and B. Bozic, "Gpt assisted annotation of rhetorical and linguistic features for interpretable propaganda technique detection in news text." in *Companion Proceedings of the ACM on Web Conference 2024*, 2024, pp. 1431–1440.

[51] T. Zhang, I. C. Irsan, F. Thung, and D. Lo, "Revisiting sentiment analysis for software engineering in the era of large language models," *arXiv preprint arXiv:2310.11113*, 2023.

[52] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi, "Self-instruct: Aligning language models with self-generated instructions," *arXiv preprint arXiv:2212.10560*, 2022.

[53] R. R. Mekala, Y. Razeghi, and S. Singh, "Echoprompt: Instructing the model to rephrase queries for improved in-context learning," 2024. [Online]. Available: https://arxiv.org/abs/2309.10687

[54] S. Schulhoff, M. Ilie, N. Balepur, K. Kahadze, A. Liu, C. Si, Y. Li, A. Gupta, H. Han, S. Schulhoff, P. S. Dulepet, S. Vidyadhara, D. Ki, S. Agrawal, C. Pham, G. Kroiz, F. Li, H. Tao, A. Srivastava, H. D. Costa, S. Gupta, M. L. Rogers, I. Goncearenco, G. Sarli, I. Galynker, D. Peskoff, M. Carpuat, J. White, S. Anadkat, A. Hoyle, and P. Resnik, "The prompt report: A systematic survey of prompting techniques," 2024.

[55] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," 2023. [Online]. Available: https://arxiv.org/abs/2305.14314

[56] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "Peft: State-of-the-art parameter-efficient fine-tuning methods," https://github.com/huggingface/peft, 2022.

[57] R. V. Krejcie and D. W. Morgan, "Determining sample size for research activities," *Educational and psychological measurement*, vol. 30, no. 3, pp. 607–610, 1970.

[58] J. Cohen, "Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[59] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[60] L. Chen, M. Zaharia, and J. Zou, "How is chatgpt's behavior changing over time?" *arXiv preprint arXiv:2307.09009*, 2023.