Jan 17th, 12:00 AM

Exploring Audience's Attitudes Towards Machine Learning-based Automation in Comment Moderation

Kilian Müller
European Research Center for Information Systems, University of Münster, Germany,
kilian.mueller@ercis.uni-muenster.de

Holger Koelmann
European Research Center for Information Systems, University of Münster, Germany,
holger.koelmann@ercis.uni-muenster.de

Marco Niemann
European Research Center for Information Systems, University of Münster, Germany,
marco.niemann@ercis.uni-muenster.de

Ralf Plattfaut
Process Innovation & Automation Lab, South Westphalia University of Applied Sciences, Soest, Germany,
plattfaut.ralf@fh-swf.de

Jörg Becker
European Research Center for Information Systems, University of Münster, Germany, becker@ercis.uni-muenster.de

Exploring Audience's Attitudes Towards Machine Learning-based Automation in Comment Moderation

Kilian Müller[1], Holger Koelmann[1], Marco Niemann[1], Ralf Plattfaut[2], and Jörg Becker[1]

[1] European Research Center for Information Systems, University of Münster, Germany
{kilian.mueller, holger.koelmann, marco.niemann, becker}@ercis.uni-muenster.de
[2] Process Innovation & Automation Lab, South Westphalia University of Applied Sciences, Soest, Germany
plattfaut.ralf@fh-swf.de

Abstract. Digital technologies, particularly the internet, led to unprecedented opportunities to freely inform oneself, debate, and share thoughts. However, the reduced level of control through traditional gatekeepers such as journalists also led to a surge in problematic (e.g., fake news), straight-up abusive, and hateful content (e.g., hate speech). Being under ethical and often legal pressures, many operators of platforms respond to the onslaught of abusive user-generated content by introducing automated, machine learning-enabled moderation tools. Even though meant to protect online audiences, such systems have massive implications regarding free speech, algorithmic fairness, and algorithmic transparency. We set forth to present a large-scale survey experiment that aims at illuminating how the degree of transparency influences the commenter's acceptance of the machine-made decision, dependent on its outcome. With the presented study design, we seek to determine the necessary amount of transparency needed for automated comment moderation to be accepted by commenters.

Keywords: Community Management, Machine Learning, Content Moderation, Algorithmic Transparency, Freedom of Expression

1 Introduction and Motivation

Imagine posting a critical comment towards the latest political news report. It is late, the day at work was hard, and all your frustration seeks its way out through this post. A few hours later, you wake up surrounded by the police, taking you to prison. An autonomously working machine learning (ML) system classified your post as subversive and hateful, and an automatically informed follow-up system determined you to be too dangerous to be allowed to remain free.

So far, this story is a dystopian idea, threatening but not real for most people. However, the increasingly sophisticated censorship in China [1, 2] and recent developments such as the planned introduction of filtering regulations in India [3] or the EU [4] fuel such visions. Even though this sounds like a malevolent plot to restrict freedom rights, such actions are not unfounded. While the internet is an unprecedented opportunity for free communication and democratizes the voicing of individual opinions, it also gave rise to phenomena such as misinformation [5–7], cyberbullying [8, 9], and what is often

17th International Conference on Wirtschaftsinformatik
February 2022, Nürnberg, Germany

subsumed as hate speech (e.g., insults, racism, or sexism) [10–13]. As especially the latter is often prohibited by law, operators of such discussion and debate opportunities (e.g., online news discussion boards, comment sections in media websites) are obliged to filter user-generated content [14]. With most discussion opportunities being free of charge and moderation traditionally being a cumbersome manual task (and hence economically prohibitive), the majority of outlets is left with two options: Giving up these opportunities (e.g., closing the discussion section in an online newspaper) [15–17] or resorting to automated moderation [15, 18–20].

Both approaches have apparent implications for the right and the ability to exercise free speech. In the first case, the respective opportunity is lost entirely; however, in the second case of automated moderation, the exact implications are less clear and bring back the idea of the initial story[1]. Furthermore, if customers reject algorithmic comment moderation, newspapers which utilize such systems might be at risk of losing parts of their current and future audience and thus, both exposure and monetary benefits generated by them. While people acknowledge and welcome the assistance of automation in many areas of life [21], their attitude towards algorithms controlling their utterances and opinions is less clear [22]; this holds even true for less critical instances such as joke recommenders [23]. One common line of thought that is assumed to limit acceptance of algorithms in general and machine moderation, in particular, is their black-box nature that makes decisions opaque [24–26]. This perception is well-grounded in several existing biases that can be hidden by black-box models [27, 28] and has also been acknowledged by legislators through the introduction of legal action requiring operators of ML-based systems to provide insights into the decision-making [29–31]. Yet, there exist few studies (e.g., Brunk et al. [32]) that assess whether additional transparency mitigates negative attitudes to algorithmic decisions and the degree of transparency that would be required. Hence, the goal of this study-in-progress is to compare commenters' reactions to different degrees of transparency w.r.t. algorithmic moderation decisions, given either an acceptance or rejection decision of the posted comment by the machine moderator. To fulfill this research objective, we plan to conduct a large-scale survey experiment as detailed in the following.

2 Research Background

In practice, companies like *Facebook*, *Google*, or *Twitter* employ their own algorithms to moderate user-generated content [18]. However, these algorithms are often proprietary and not open source. Researchers, also mainly utilizing ML to detect hate speech, provide insights into their models [12, 33–39]. Nevertheless, even presupposing a reasonable good classification, the application of automated ML algorithms in content moderation leaves three political issues according to Gorwa et al. [18]: transparency, fairness, and depoliticization. In this study, we aim to evaluate the effects of increasing levels of transparency on the user's acceptance of an (automated) ML-based moderation system. Thus, we will not focus on the areas of fairness and depoliticization. To tackle the

[1] The factor of legally enforced automated content moderation leading to over-moderation and the introduction of more political biases further boosts such concerns [14, 18].

transparency problems, approaches like [40] try to make complex ML models (i.e., neural models) interpretable by including surrogate models such as Local Interpretable Model-agnostic Explanations [41] or SHapley Additive exPlanations [42]. These surrogate models try to open up the ML black-box not by explaining the underlying algorithms but by representing the importance of different input features (i.e., words). By understanding the input-output relationship, the ML process should become more transparent.

As recent legislation such as the GDPR [29,43,44] and rulings of the German Federal Supreme Court (demanding more transparency in comment moderation and the blocking of user accounts) [45] indicate, an increase in transparency seems especially needed. There already exists some research within the topics of moderation transparency. For example, Jhaver et al. [46], and Juneja et al. [47] investigate the effects of transparency during moderation feedback on future user behavior on the *Reddit* network. Furthermore, Wang [48] studies the changes in the opinion of users on news articles that are moderated automatically. Brunk et al. [32] conducted a first study linking increased transparency to an increase in trust and finally to a higher probability of acceptance. By utilizing the decision output of a functioning moderation system (`www.moderat.nrw`) instead of manual annotation, we will vary the degree of feedback transparency. Through this, we determine if an increasing amount of feedback also leads to increasing perceived transparency on the commenter's end and if this, in turn, leads to a higher acceptance of the decision made by the machine. In addition, we vary the made decision, to see, if the decision output, being acceptance or rejection of the comment, changes the user's acceptance of such a system. To deal with this more complex study design, we aim at recruiting a sufficiently large number of study participants from the German-speaking population, as the comments within the dataset are in German, via a Crowdsourcing-Platform.

3 Study Design

3.1 Hypotheses Building

Building upon this content creation and moderation background, we first need to evaluate, if the acceptance of an automated moderation system depends on the transparency of the system itself or the match between the commenter's opinion about the comment in question and the moderation results of this comment, meaning, we have to evaluate if potential issues with automated moderation lie within the *automated* or the *moderation* part of the process. Thus, we need to differentiate between the moderation decision itself and the potential automation. Besides that, research suggests that machine-made decision-making processes should be more transparent to increase the acceptance of the resulting decisions. Exemplary, Yeomans et al. [23] pointed out that to overcome the algorithm aversion, recommender systems need to be understood by their human counterparts. However, as stated by Burton et al. (2020) [49], increased transparency "often comes as a trade-off with the performance of the algorithm". Previous studies have already tried to show if transparency increases the acceptance of recommender system decisions [50] and automatic comment moderation [32] with mixed but promising first results and a call for more research into the field. Consequently, our first hypothesis is

H1: *The acceptance of an automated moderation decision depends on the fit between the expected moderation decision of the commenter and the actual moderation decision (**H1a**) and the perceived transparency of the moderation decision (**H1b**).* For those two influences on the decision acceptance, we have to look deeper into both, the fit of the result expected by the commenter and the actual result of the moderation decision, as well as the degree of transparency of decision presentation.

Therefore, the second hypothesis is meant to evaluate the commenter's tendencies towards the moderation in general. For that, we expect if a comment is more extreme (both in regards to being completely uncritical to completely critical) people might inherently understand potential moderation results. However, if their view differs from the actual moderation result this might change. Therefore, the second hypothesis is meant to evaluate their acceptance of moderation results both in matching and mismatching result situations. This leads lead us to the second hypothesis ***H2***: *The fit between the expected moderation result of the commenter and the actual moderation result is higher for more extreme comments (i.e., comments that are non-critical or highly critical).*

In addition, to get a more detailed understanding of how much transparency may be required, we aim at differentiating between different forms of algorithm feedback, an approach, which, e.g., has already been done for different types of transparency mechanisms for privacy decision making on smartphones [51]. We differentiate the degree of feedback into the three nuances, no explanation for the decision made, naming the classification result (e.g., "Threat" or "Insult"), and word highlighting within the user's comment. We expect naming the classification result to be perceived as more transparent than no explanation and word highlighting to be perceived as more transparent, both than no explanation and or only naming the classification result. This leads to our third hypothesis ***H3***: *The perceived transparency of the moderation decision is higher the more additional information is given.* The resulting research model is depicted in Fig. 1.

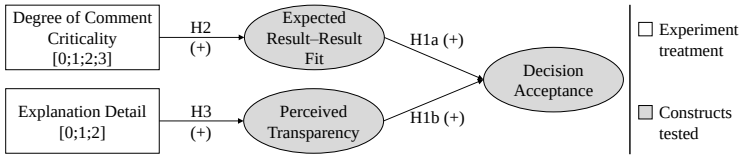

Figure 1. Resulting research model.

3.2 Survey Experiment

To assess our hypotheses, we plan to conduct an online experiment. Due to the most likely non-normality distribution of our data (especially w.r.t. the treatment), we will employ structural equation modeling (SEM) using partial least squares (PLS) for data analysis [52,53]. PLS-SEM is well accepted in IS research for these types of studies [54].

In our experiment, we will assess the hypotheses w.r.t. additional explanations in the case of ML-based content moderation. We will simulate a real-life situation with a news article and four existing comments (mentioning that comments are moderated with regards to the terms of use) and will present participants a specific comment (comments

can either be completely non-critical, borderline, slightly critical, and completely critical). To acquire a generally accepted view on the criticality of different comments, we rely on an existing dataset of comments with their corresponding criticality [55]. Next, participants are asked to report their expected result of the moderation. This expected moderation result will later be used to calculate the expected result-result fit. Afterwards, the participant will need to copy and paste the comment into a text field (i.e., simulating the comment process) and, as a result, will be presented with the moderation outcome. This outcome will then differ on three treatment options: (1) The text is shown as being accepted or rejected by the automated comment moderation system. (2) The text is shown as being accepted or rejected with an explanation that an algorithm has either detected a terms of use-violation and has been classified by the algorithm accordingly (naming the classification result) or as not having detected any issues. (3) The text is shown as being accepted or rejected with an explanation that an algorithm has either detected a terms of use-violation highlighting the critical words of the comment (word highlighting) or as not having detected any issues whilst still providing the word highlighting. After the treatment, we will conduct a manipulation check and assess the endogenous variables.

4 Concluding Discussion and Way Forward

For our study, we suggest a research model which shall be used to determine how increased transparency during ML-based moderation affects the acceptance of automated moderation systems. Further, we aim to show if this acceptance differs if the output of the automated moderation system varies from the user's own judgment. The results could further show how the utilization of ML-algorithms is generally perceived if the freedom of expression is concerned. We will continue this research project in the coming months and are looking forward to presenting the results.

As newspapers and similar outlets in which opinions are shared (e.g., social media, forums, etc.) are faced with an increased amount of work in order to keep their comment sections clean, one solution could be the use of ML-algorithms to reduce moderation effort. However, if their readers do not accept the automated moderation of their comments, the newspaper risks losing market share and thus profit. Therefore, this study could have numerous implications for practitioners. If the readers fully accept ML-algorithms, they could successfully be utilized to detect and moderate hate speech, cyberbullying, and other forms of digital aggression to ensure a free and civilized discourse. Our study could suggest the needed amount of transparency necessary to still utilize ML-algorithms without the risk of alienating the readers. However, it could lead to further implications if users tend to always reject automated moderation that differs from their own expectations, regardless of the level of transparency.

Acknowledgements

The research leading to these results received funding from the federal state of North Rhine-Westphalia and the European Regional Development Fund (EFRE.NRW 2014-2020), Project: MODERAT! (No. CM-2-2-036a).

References

1. Tai, Z.: Casting the Ubiquitous Net of Information Control. International Journal of Advanced Pervasive and Ubiquitous Computing 2(1), 53–70 (2010)
2. King, G., Pan, J., Roberts, M.E.: How Censorship in China Allows Government Criticism but Silences Collective Expression. American Political Science Review 107(2), 326–343 (2013)
3. Dash, S.: Mozilla, GitHub and Cloudflare fear 'automated censorship' in India's new internet laws (jan 2020), https://www.businessinsider.in/policy/news/mozilla-github-and-cloudflare-fear-automated-censorship-in-indias-new-internet-laws/articleshow/73136671.cms
4. Schiller, A., Weiskopf, T.: Automated censorship in the digital space (may 2019), https://www.thenewfederalist.eu/automated-censorship-in-the-digital-space
5. Lazer, D.M.J., Baum, M.A., Benkler, Y., Berinsky, A.J., Greenhill, K.M., Menczer, F., Metzger, M.J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S.A., Sunstein, C.R., Thorson, E.A., Watts, D.J., Zittrain, J.L.: The science of fake news. Science 359(6380), 1094–1096 (2018)
6. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science 359(6380), 1146–1151 (2018)
7. Meinert, J., Mirbabaie, M., Dungs, S., Aker, A.: Is it really fake?–towards an understanding of fake news in social media communication. In: International Conference on Social Computing and Social Media. pp. 484–497. Springer (2018)
8. Kowalski, R.M., Giumetti, G.W., Schroeder, A.N., Lattanner, M.R.: Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth. Psychological Bulletin 140(4), 1073–1137 (2014)
9. Whittaker, E., Kowalski, R.M.: Cyberbullying Via Social Media. Journal of School Violence 14(1), 11–29 (2015)
10. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.: Abusive Language Detection in Online User Content. In: Proceedings of the 25th International Conference on World Wide Web. pp. 145–153. WWW '16, ACM Press, Montreal, Canada (2016)
11. Schmidt, A., Wiegand, M.: A Survey on Hate Speech Detection using Natural Language Processing. In: Ku, L.W., Li, C.T. (eds.) Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media. pp. 1–10. SocialNLP 2017, Association for Computational Linguistics, Valencia, Spain (2017)
12. Mondal, M., Silva, L.A., Benevenuto, F.: A measurement study of hate speech in social media. In: Dolong, P., Vojtas, P. (eds.) Proceedings of the 28th ACM Conference on Hypertext and Social Media. pp. 85–94. HT 2017, ACM, Prague, Czech Republic (2017)
13. Fortuna, P., Nunes, S.: A Survey on Automatic Detection of Hate Speech in Text. ACM Computing Surveys 51(4), 1–30 (2018)
14. Bloch-Wehba, H.: Automation in Moderation. Cornell International Law Journal 53(1), 41–96 (2020)
15. Chen, Y., Zhou, Y., Zhu, S., Xu, H.: Detecting Offensive Language in Social Media to Protect Adolescent Online Safety. In: Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust. pp. 71–80. SOCIALCOM-PASSAT '12, IEEE, Amsterdam, Netherlands (2012)
16. Muddiman, A., Stroud, N.J.: News Values, Cognitive Biases, and Partisan Incivility in Comment Sections. Journal of Communication 67(4), 586–609 (2017)
17. Liu, J., McLeod, D.M.: Pathways to news commenting and the removal of the comment system on news websites. Journalism 22(4), 867–881 (2021)

18. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data and Society 7(1), 1–15 (2020)
19. Gillespie, T.: Content moderation, AI, and the question of scale. Big Data and Society 7(2), 1–5 (2020)
20. Laaksonen, S.M., Haapoja, J., Kinnunen, T., Nelimarkka, M., Pöyhtäri, R.: The Datafication of Hate: Expectations and Challenges in Automated Hate Speech Monitoring. Frontiers in Big Data 3, 1–16 (2020)
21. Bogert, E., Schecter, A., Watson, R.T.: Humans rely more on algorithms than social influence as a task becomes more difficult. Scientific Reports 11(1), 1–9 (2021)
22. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. Journal of Experimental Psychology: General 144(1), 114–126 (2015)
23. Yeomans, M., Shah, A., Mullainathan, S., Kleinberg, J.: Making sense of recommendations. Journal of Behavioral Decision Making 32(4), 403–414 (2019)
24. Adadi, A., Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160 (2018)
25. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1(5), 206–215 (2019)
26. Zednik, C.: Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. Philosophy and Technology 34(2), 265–288 (2021)
27. Garcia, M.: Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal 33(4), 111–117 (2016)
28. Lambrecht, A., Tucker, C.: Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. Management Science 65(7), 2966–2981 (2019)
29. Goodman, B., Flaxman, S.: European Union Regulations on Algorithmic Decision Making and a "Right to Explanation". AI Magazine 38(3), 50–57 (2017)
30. 116th Congress: Algorithmic Accountability Act of 2019 (2019), https://www.congress.gov/bill/116th-congress/house-bill/2231
31. Larsson, S., Heintz, F.: Transparency in artificial intelligence. Internet Policy Review 9(2), 1–16 (2020)
32. Brunk, J., Mattern, J., Riehle, D.M.: Effect of Transparency and Trust on Acceptance of Automatic Online Comment Moderation Systems. In: Proceedings of the 21st IEEE Conference on Business Informatics. pp. 429–435. IEEE, Moscow, Russia (2019)
33. Agarwal, S., Sureka, A.: Using knn and svm based one-class classifier for detecting online radicalization on twitter. In: International Conference on Distributed Computing and Internet Technology. pp. 431–442. Springer (2015)
34. Bartlett, J., Reffin, J., Rumball, N., Williamson, S.: Anti-social media. Demos 2014, 1–51 (2014)
35. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. International Journal of Multimedia and Ubiquitous Engineering 10(4), 215–230 (2015)
36. Ting, I.H., Chi, H.M., Wu, J.S., Wang, S.L.: An approach for hate groups detection in facebook. In: The 3rd International Workshop on Intelligent Data Analysis and Management. pp. 101–106. Springer (2013)
37. Warner, W., Hirschberg, J.: Detecting hate speech on the world wide web. In: Proceedings of the second workshop on language in social media. pp. 19–26 (2012)
38. Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., Bhamidipati, N.: Hate speech detection with comment embeddings. In: Proceedings of the 24th international conference on world wide web. pp. 29–30 (2015)

39. Badjatiya, P., Gupta, S., Gupta, M., Varma, V.: Deep learning for hate speech detection in tweets. In: Proceedings of the 26th international conference on World Wide Web companion. pp. 759–760 (2017)
40. Švec, A., Pikuliak, M., Šimko, M., Bieliková, M.: Improving Moderation of Online Discussions via Interpretable Neural Models. In: Fišer, D., Huang, R., Prabhakaran, V., Voigt, R., Waseem, Z., Wernimont, J. (eds.) Proceedings of the Second Workshop on Abusive Language Online. pp. 60–65. ALW2, Association for Computational Linguistics, Brussels, Belgium (2018)
41. Ribeiro, M.T., Singh, S., Guestrin, C.: " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
42. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems. pp. 4768–4777 (2017)
43. The European Parliament, The Council of the European Union: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Da. Official Journal of the European Union 119, 1–88 (2016)
44. Felzmann, H., Villaronga, E.F., Lutz, C., Tamò-Larrieux, A.: Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society 6(1), 1–14 (2019)
45. Pressestelle des Bundesgerichtshofs: Bundesgerichtshof zu Ansprüchen gegen die Anbieterin eines sozialen Netzwerks, die unter dem Vorwurf der "Hassrede" Beiträge gelöscht und Konten gesperrt hat. https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/20201/2021149 (2021), accessed 30.08.2021
46. Jhaver, S., Bruckman, A., Gilbert, E.: Does transparency in moderation really matter? user behavior after content removal explanations on reddit. Proceedings of the ACM on Human-Computer Interaction 3(CSCW), 1–27 (2019)
47. Juneja, P., Rama Subramanian, D., Mitra, T.: Through the looking glass: Study of transparency in reddit's moderation practices. Proceedings of the ACM on Human-Computer Interaction 4(GROUP), 1–35 (2020)
48. Wang, S.: Moderating uncivil user comments by humans or machines? the effects of moderation agent on perceptions of bias and credibility in news content. Digital Journalism 9(1), 64–83 (2021)
49. Burton, J.W., Stein, M.K., Jensen, T.B.: A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making 33(2), 220–239 (2020)
50. Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Stash, N., Aroyo, L., Wielinga, B.: The effects of transparency on trust in and acceptance of a content-based art recommender. User Modeling and User-adapted interaction 18(5), 455 (2008)
51. Betzing, J.H., Tietz, M., vom Brocke, J., Becker, J.: The impact of transparency on mobile privacy decision making. Electronic Markets 30(3), 607–625 (2020)
52. Hair, J.F., Hult, G.T.M., Ringle, C.M., Sarstedt, M.: A primer on partial least squares structural equation modeling (PLS-SEM). Sage, Los Angeles and London and New Delhi and Singapore and Washington DC and Melbourne, second edition edn. (2017)
53. Hair, J.F., Risher, J.J., Sarstedt, M., Ringle, C.M.: When to use and how to report the results of pls-sem. European Business Review 31(1), 2–24 (2019)
54. Petter, S.: "haters gonna hate": Pls and information systems research. SIGMIS Database 49(2), 10–13 (May 2018), https://doi.org/10.1145/3229335.3229337

55. Assenmacher, D., Niemann, M., Müller, K., Seiler, M.V., Riehle, D.M., Trautmann, H.: RP-Mod & RP-Crowd: Moderator- and Crowd-Annotated German News Comment Datasets (Aug 2021), https://doi.org/10.5281/zenodo.5291339