Elicitation of Requirements for a NLP-Model Store for Abusive Language Detection

Kilian Müller(✉)

Department for Information Systems, University of Muenster, Leonardo-Campus 3,
48149 Münster, Germany
kilian.mueller@ercis.uni-muenster.de

Abstract. While in social media users most commonly interact with each other without a guiding entity, the discussion space of newspaper platforms is moderated by community managers, who invest considerable amounts of time and effort in keeping comment sections clean. To reduce this effort and allow community managers again to more freely interact with their users, automated comment moderation systems (ACMS) be be utilized. However, most newspapers do not have the expertise to create, update, and maintain machine learning (ML)-models. Thus, they are forced to rely on proprietary off-the-shelf solutions. However, if they want to keep their sovereignty over their data and systems, they would need access to models which could be integrated within their current moderation or content management systems. One option could be a platform for newspapers and data scientists where the data scientists could sell their pre-trained models and where newspapers could hire data scientists to create tailor-made models for them. In order to identify the requirements, community managers have for such systems, we conducted a series of semi-structured interviews with community managers of newspapers of varying size (from local to national). Furthermore, the information was enriched by the participation in multiple workshops on content moderation. We were able to elicit five major technical requirements necessary to create the described design artifact.

Keywords: Abusive language · Comment moderation · Model store

1 Introduction and Motivation

Hate and abusive comments are spreading through social media and other discussion spaces [16,18]. As this could pose a serious threat to online discussions [21] and is already being targeted by lawmakers (e.g., the Netzwerkdurchsetzungsgesetz (NetzDG) in Germany[1] [13], or other countries [6]), newspapers have

[1] The NetzDG forces the providers of comment spaces to remove abusive language from their sites within 24 h after being alerted.

C. Stephanidis et al. (Eds.): HCII 2022, CCIS 1582, pp. 581–588, 2022.
https://doi.org/10.1007/978-3-031-06391-6_72

to employ community managers in order to manually moderate user generated content.

These community managers used to fulfill the role of gatekeepers, interacting with their readers and fostering a healthy discussion space. However, in recent years, they are forced to act as content moderators, deleting or blocking user generated content [7]. This pushes the community managers away from their original job description [14]. Consequently, many newspapers have resorted to close their comment sections entirely (around 45% in Germany [17]), as the benefit of having a political discourse does not measure up to their necessary economic investment.

To support community managers in their work and allow newspapers to keep their comment sections, automated comment moderation systems (ACMS) utilizing different forms of machine learning (ML) are a promising option [10,20]. These ACMS utilize natural language processing (NLP) methods in order to automatically detect abusive content. Afterwards, this content can be automatically blocked from publication or be reviewed by community managers; resulting in semi-automated moderation.

In contrast to large newspapers, small- or medium-sized newspapers often times do not have the in-house competences to build such ACMS themselves [26]. Therefore, often times, they have to rely on proprietary solutions, e.g., Perspective[2] by Alphabet [25]. This, however, comes at the price of losing their sovereignty over their own and their users data, as the comments are transferred to the provider of the proprietary system. Therefore, a possible solution could be the utilization of publicly available research insights.

Favorably for newspapers, there already exist multiple, accessible research endeavors in which researchers investigate models suitable for abusive language detection (c.f. [2,4,5,8,9,16,22,24]). However, these insights are still not on a usable level for small- to medium-sized newspapers. The described approaches need to be applied to the newspaper's data and the resulting models need to be delivered in an easy to use format so that newspapers are able to seamlessly integrate them into their existing IT-architecture [26]. Thus, on the one hand, even though there exists available research in the domain, newspapers are still in need for external expertise to create proficient and usable models which can then be integrated into their ACMS. On the other hand, there exists the need for labeled training data, especially in non-English speaking communities, for the creation of refined models [3]. Therefore, apart from possible monetary benefits, data scientist could also benefit from receiving data from the different newspapers.

Apparently, both sides could benefit from either the provided data or the external expertise. Therefore, the goal of this paper is the elicitation of requirements for a NLP-model store, especially for the detection of abusive language which can be utilized to connect newspapers with data scientists/researchers. In the paper all five requirements are explained w.r.t. their implications on the community managers daily work. Additionally, the implications for the practical implementations of the desired artifact are detailed.

[2] https://www.perspectiveapi.com/.

The paper is structured as follows: Sect. 2 introduces the research method utilized in this paper. Afterwards, the results are presented in Sect. 3. Finally, Sect. 4 discusses the findings and concludes the paper.

2 Research Method

In order to gain insights into the daily work of community manager and their requirements towards an NLP-model store semi-structured interviews were conducted. Semi-structured interviews are a widely accepted form of building a basis for requirement elicitation [1,27]. Furthermore, this form of interview allows for deeper insights into the necessary domain knowledge of community managers [11].

Community managers from three different newspapers participated in the interviews: One national newspaper (among the three largest newspapers in Germany), one regional newspaper (with daily circulation $> 250,000$), and one local newspaper (with daily circulation $< 10,000$). During the interview the interviewees were asked questions focusing on their daily work in comment moderation, their current practices, and possible automation possibilities. Later, they were shown a comment moderation system with a functioning automated moderation support and were then asked which requirements they would have towards such a system. Following up, questions were aimed at their current moderation system and how they could envision an integration of ML-models for automated comment moderation. Afterwards, questions regarding their possible capabilities in creating their own models as well as a possible envisioned procurement process for such models were asked.

The interviews were conducted in a virtual setting via Zoom[3] and, after approval by the interviewee, recorded with both video and audio. After the finished interview each was transcribed and analyzed.

To gather additional information, we participated in three domain-specific workshops. Here, insights regarding practitioner's views and issues in their current work were gained. These insights were subsequently structured to allow for further analysis.

Combining the two sources of information, the data was analyzed to obtain the requirements. They were then evaluated by discussions with community managers.

3 Results

Utilizing the input of the interviews and the workshops, we elicit five requirements for NLP-model stores for abusive language detection. The requirements are explained in more detail in the following.

[3] https://zoom.us/.

3.1 Ease of Understanding

As mentioned above, community managers are rarely experts in the domain of NLP. While the community managers expressed a rather positive attitude towards ML in general they also mentioned insecurities about inner workings of ML. Furthermore, they mentioned that also the dark side of ML was not obvious to them either. However, even with these concerns, all community managers expressed the willingness to include ML in their moderation process. Thus, eliminating possible insecurities about ML and informing the community managers about the workings of the presented ML-models could pave the way for broader acceptance.

REQ_1: An NLP-model store for abusive language detection should be conceptualized in such a way, that community managers are guided towards an understanding of the necessary ML specific technical terms and workings and are thus able to make informed decisions.

3.2 Explainable Models

While transparency, regarding the moderation decision, towards the users varied between the different community managers, all community managers stated that transparency in moderation decisions is of crucial importance within their teams. Traceability between different community managers are important in order to make similar moderation decisions and present a consistent image towards their users. When confronted with model explainers[4] the community managers appreciated the quick overview provided [15,19]. However, as soon as too much text was highlighted all agreed to lose focus on the important passages. Furthermore, the community managers pointed out different aspects which they focus on during their moderation process which should be reflected in possible word highlight. Thus, highlighting the different ML outputs on the NLP-model store examples could provide the community managers with valuable feedback on model performance and fit to their specific moderation habits.

REQ_2: An NLP-model store for abusive language detection should include the option to add model-agnostic explainers in order to make the ML output understandable and transparent for community managers.

3.3 Multiple Vendors

All community managers were utilizing different moderation systems, many are tailor-made or customized for the specific newspaper. Some already had experiences with different kinds of ML input in their moderation decision. However, none expressed the desire to switch to proprietary system, sourced from one company, which would support the comment moderation with ML. Rather than

[4] The presented moderation system was equipped with explainers, highlighting different words or passages which were crucial to the models moderation decision.

a new solution, the community managers preferred the integration of ML into their system, either as a full integration or an interface in order to "keep control" over their own data and infrastructure. However, they expressed the desire for external support as long as they were not dependent on a single supplier.

REQ_3: An NLP-model store for abusive language detection should include the option to buy ML-models from multiple vendors in an easy to use format, preventing lock-in effects and allowing for the integration into existing moderation platforms.

3.4 Comparable Metrics

The community managers stated difficulties in evaluating ML-models. This becomes even more difficult, as researchers utilize many different metrics in evaluating different ML-models. As non data-scientists, the community managers prefer metrics which are easy to understand and also appreciate consistency when it comes to the metrics utilized when comparing different models. This is also echoed by existing literature (cf., [26]).

REQ_4: An NLP-model store for abusive language detection should offer comparable and understandable metrics for all offered models to enable community manager to find the best model suited for their specific business needs.

3.5 Customized Models

The community managers detailed significant differences between their respective moderation processes. These were also reflected in the community guidelines, the moderation process, and the harshness/lenience in moderation decision making. Thus, the community managers pointed out that a ML-model should reflect both the newspaper's community guidelines and the community managers judgment. Consequently, not every model will fit every newspaper. Therefore, the community managers expressed the desire commission tailor-made models, based on their own previous decisions, which could better gauge the desired discussion culture on their respective comment sections.

REQ_5: An NLP-model store for abusive language detection should offer the option to commission tailor-made ML-models for community managers, based on data gathered from their own newspapers.

4 Discussion and Conclusion

The goal of the research presented in this paper is the elicitation of requirements for an NLP-model store for abusive language detection to support both community managers in their daily work and provide data scientist with the necessary input to improve existing models or develop entirely new ones. By connecting these distinct stakeholder groups new collaboration possibilities should arise which could benefit both sides.

Towards this goal, three interviews with community managers were conducted and three workshops on content moderation were visited. Based on the acquired knowledge, five requirement have been elicited and presented in this paper. These requirements should guide researchers and practitioners in creating an NLP-model store for abusive language detection.

However, the presented research in progress does not come without limitations. First, we only interviewed community managers. This excludes the views from data scientists from our results and, therefore, does not include important aspects and requirements they have towards such an IT-artifact. Second, only three interviews with community managers have been conducted up-to-date. Consequently, the insights acquired from these interviews should be validated with different community managers from other newspapers. Third, the presented research is purely conceptual and should be demonstrated and evaluated by implementing a functioning model store.

Thus, our next steps in our research agenda include conducting more interviews; both with additional community managers to validate the current findings, as well as with data scientist to include their viewpoints. Next, utilizing design science research [12], the NLP-model store needs to be developed based on the generated requirements. Utilizing the developed store the requirements and further insights can be evaluated [23]. Lastly, the developed store will be integrated into the currently existing moderation system[5] to offer practitioners the ability to procure models from data scientists.

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).

Furthermore, I would like to thank Marco Niemann and Holger Koelmann for their assistance in conducting the interviews used in this paper.

References

1. Agarwal, R., Tanniru, M.R.: Knowledge acquisition using structured interviewing: an empirical investigation. J. Manag. Inf. Syst. **7**(1), 123–140 (1990)
2. Agarwal, S., Sureka, A.: Using KNN and SVM based one-class classifier for detecting online radicalization on Twitter. In: Natarajan, R., Barua, G., Patra, M.R. (eds.) ICDCIT 2015. LNCS, vol. 8956, pp. 431–442. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-14977-6_47
3. Assenmacher, D., Niemann, M., Müller, K., Seiler, M., Riehle, D.M., Trautmann, H.: RP-mod & RP-crowd: moderator- and crowd-annotated German news comment datasets. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021)
4. 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)
5. Bartlett, J., Reffin, J., Rumball, N., Williamson, S.: Anti-social media. Demos **2014**, 1–51 (2014)

[5] moderat.nrw.

6. Bloch-Wehba, H.: Automation in moderation. Cornell Int. Law J. **53**(1), 41–96 (2020)
7. Boberg, S., Schatto-Eckrodt, T., Frischlich, L., Quandt, T.: The moral gatekeeper? Moderation and deletion of user-generated content in a leading news forum. Media Commun. **6**(4), 58–69 (2018)
8. 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)
9. Gitari, N.D., Zuping, Z., Damien, H., Long, J.: A lexicon-based approach for hate speech detection. Int. J. Multimedia Ubiquitous Eng. **10**(4), 215–230 (2015)
10. Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data Soc. **7**(1), 1–15 (2020)
11. Hadar, I., Soffer, P., Kenzi, K.: The role of domain knowledge in requirements elicitation via interviews: an exploratory study. Requirements Eng. **19**(2), 143–159 (2012). https://doi.org/10.1007/s00766-012-0163-2
12. Hevner, A., Chatterjee, S.: Design science research in information systems. In: Hevner, A., Chatterjee, S. (eds.) Design Research in Information Systems, pp. 9–22. Springer, Boston (2010). https://doi.org/10.1007/978-1-4419-5653-8_2
13. der Justiz, B.: Gesetz zur verbesserung der rechtsdurchsetzung in sozialen netzwerken (netzwerkdurchsetzungsgesetz - netzdg) (2017). https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html
14. Loosen, W., et al.: Making sense of user comments: identifying journalists' requirements for a comment analysis framework. Stud. Commun. Media **6**(4), 333–364 (2017)
15. 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)
16. 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. HT 2017, pp. 85–94. ACM, Prague (2017)
17. Niemann, M., Müller, K., Kelm, C., Assenmacher, D., Becker, J.: The German comment landscape. In: Bright, J., Giachanou, A., Spaiser, V., Spezzano, F., George, A., Pavliuc, A. (eds.) MISDOOM 2021. LNCS, vol. 12887, pp. 112–127. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87031-7_8
18. 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. WWW 2016, pp. 145–153. ACM Press, Montreal (2016)
19. 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)
20. Ruckenstein, M., Turunen, L.L.M.: Re-humanizing the platform: content moderators and the logic of care. New Media Soc. **22**(6), 1026–1042 (2020)
21. Salminen, J., Veronesi, F., Almerekhi, H., Jung, S.G., Jansen, B.J.: Online hate interpretation varies by country, but more by individual: a statistical analysis using crowdsourced ratings. In: Proceedings of 5th International Conference Social Networks and Analysis Management and Security. SNAMS 2018, pp. 88–94. IEEE, Valencia (2018)

22. Ting, I.H., Chi, H.M., Wu, J.S., Wang, S.L.: An approach for hate groups detection in Facebook. In: Uden, L., Wang, L., Hong, TP., Yang, HC., Ting, IH. (eds.) The 3rd International Workshop on Intelligent Data Analysis and Management, pp. 101–106. Springer, Dordrecht (2013). https://doi.org/10.1007/978-94-007-7293-9_11
23. Venable, J., Pries-Heje, J., Baskerville, R.: Feds: a framework for evaluation in design science research. Eur. J. Inf. Syst. **25**(1), 77–89 (2016)
24. 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)
25. Wulczyn, E., Thain, N., Dixon, L.: Ex machina: personal attacks seen at scale. In: Proceedings of the 26th International Conference on World Wide Web, pp. 1391–1399 (2017)
26. Xiu, M., Jiang, Z.M.J., Adams, B.: An exploratory study of machine learning model stores. IEEE Softw. **38**(1), 114–122 (2020)
27. Zowghi, D., Coulin, C.: Requirements elicitation: a survey of techniques, approaches, and tools. In: Aurum, A., Wohlin, C. (eds.) Engineering and Managing Software Requirements, pp. 19–46. Springer, Heidelberg (2005). https://doi.org/10.1007/3-540-28244-0_2