Elicitation of Requirements for an AI-Enhanced Comment Moderation Support System for Non-tech Media Companies

Marco Niemann(✉)

ERCIS, University of Muenster, Leonardo-Campus 3, 48149 Muenster, Germany
marco.niemann@ercis.uni-muenster.de

Abstract. Traditional (news) media companies are increasingly facing rising participation in their discussion sections and a simultaneous surge of abusive contributions. Legally required to prevent the dissemination of hate and threats, manual moderation is an increasingly daunting task for journalists and part-time community managers. Consequently, many comment sections are closed for economic reasons world-wide. While there is ongoing academic and practice research on machine learning (ML) systems to detect abusiveness or hate, the focus typically remains on this limited technical task. Integrations into systems for practical community management are still rare. Based on eleven semi-structured interviews with experts of four German newspapers of varying size (incl. an observation of their working patterns), complemented by insights from workshops on community management, we could identify five major functional requirements for creating such integrated systems. This range goes from the need for increased transparency and controllability to better support for team-based community management. In this paper, we outline each requirement's origin and implications for the development of integrated, artificial intelligence (AI)-enhanced comment moderation support system (CMSS).

Keywords: Abusive language · Hate speech · Comment moderation · Decision support system · Functional requirements

1 Introduction and Background

Journalism and journalists have a long-standing tradition of serving as gatekeepers. They filter the plethora of events and information occurring locally, nationally, and increasingly globally, deciding on the topics worthy of being reported upon [3]. However, the journalistic self-understanding goes beyond gatekeeping and extends to the facilitation of discourse and public debate [14,19,21]. Hence, participatory formats such as letters to the editor have a long-standing tradition [12,19], providing room for otherwise uncovered topics, criticism and feedback,

C. Stephanidis et al. (Eds.): HCII 2021, CCIS 1419, pp. 573–581, 2021.
https://doi.org/10.1007/978-3-030-78635-9_73

as well as appeals and calls to action [19]. As recent studies [12] confirm, even in 2020, such traditional participatory formats are still popular and contribute to both the image and economics of the publishing outlets.

However, letters to the editors did not fully fulfill their intended deliberative purpose, as requirements in terms of form, time to publication, and interactivity are still high. The internet and Web 2.0 flattened communication hierarchies and eliminated some of the hurdles pertinent to traditional formats such as letters to the editor [3,21]. Many newspapers added comment fora and debate sections to their websites to use this new participatory channel, hoping for active and insightful discussions between readers [6,17]. The high hopes linked to comment sections remained unfulfilled, as abusive and uncivil behavior soon became a massive problem as acknowledged by academics [5,18] and journalists [2,8]. As a result, beginning in the mid-2010s, a growing number of newspapers decided to shut down their comment sections, reaching a closure rate of up to 50% in Germany [16]. Despite the positive economic impact of participatory formats, comment moderation is highly resource-intensive—often to the point that financial and personal resources are deterred from the core news business: article writing [4,9,13]. This effect was found to have a more substantial effect on smaller outlets [13].

Throughout the past decade, an increasing number of academics and practitioners have set forth to experiment with ML and natural language processing (NLP) to use these upcoming technologies to tackle abusive comments [7,20]. While the research stream is reporting continuous improvements, there are still open challenges: Firstly, many publications focus on the technical refinement of used models [11,25]. While work in this direction is crucial, it is challenging to digest standalone for all interested parties lacking experts (typ. smaller outlets). Even approaches geared towards better interpretability are still more technical demonstrations than integrated solutions [22]. Secondly, a large part of the published works are carried out by major online media companies, working on solving problems faced by their internal teams [27]. Lastly, the integration of the research work into broadly applicable solutions is still ongoing. While several tools set forth to improve the commenting experience, many of them are tailored towards the creation of improved, more interactive, and insightful debates [15]. The moderation perspective, especially the AI-supported one, is mostly neglected ([24] being one of the few exceptions). This leads to the conclusion that news outlets in general and small ones in particular lack consideration in the ongoing research efforts.

To address this extant gap, this paper's research goal is the *elicitation of functional requirements of an AI-enhanced comment moderation support system for non-tech media companies.*

The remainder of this paper will unfold as follows: Sect. 2 presents the research method chosen for this paper. The identified requirements are presented and explained in Sect. 3 before the paper is concluded in Sect. 4.

2 Research Method

To obtain the requirements presented in this paper, semi-structured interviews were chosen as the primary source of information. These are among the most common and suitable methods to obtain functional requirements [1,10]. Furthermore, they provide valuable additional degrees of freedom to elicit potentially unexpected insights in the young domain of AI-enhanced CMSSs.

In total, eleven semi-structured interviews [23] were conducted with journalists and community managers of four newspapers. Among these were two local, one over-regional, and one national newspaper with circulations between ∼10,000 and >300,000 newspapers daily. The interviews were centered on understanding their current working patterns (mode of operation and central tasks to be supported) and probing how they could imagine an AI-supported CMSS to support their daily work. Each interview was recorded (*with the interviewees' explicit consent*) and subsequently transcribed for further assessment.

As secondary sources of information, two further commonly used requirements elicitation methods [29] were applied: Participation in journalistic workshops[1] served to get a better domain understanding. The goal was to uncover potential pain points by observing discussions without explicitly nudging the people towards algorithmic solutions. Furthermore, three of the participating newspapers allowed us to observe their staff working on daily community management. Here additional gaps and pain points of the existing software solutions could be identified—sometimes, the interviewees even remembered additional issues they could not mention in the interview before.

The collected data was subsequently structured, analyzed, and interpreted to understand the sometimes plainly stated but often hidden requirements towards a modern AI-supported CMSS. Identified requirements were initially validated through discussions with researchers working in the same field, and some of the initially interviewed community managers.

3 Requirements for AI-Enhanced CMSS

Based on the previously outlined information sources, we could identify five major requirements towards AI-enhanced CMSS (for non-tech media companies). Each of them subsumes multiple minor requirements, which have the same overarching goal and were clustered accordingly. This paper's focus will be on functional aspects of a CMSS, as many non-functional requirements for platform software (e.g., performance, usability, . . .) are domain-independent and already covered in other outlets. Each identified requirement will be described in more detail in the following subsections, enhanced by direct and indirect quotes. The interviews have been conducted in German, and the author translated quotes.

[1] Three workshops were visited before the COVID-19 pandemic. The first two had ∼20 participants—mostly journalists and newspaper personnel—and centered around hate speech prosecution and management. The third had ∼10 participants and centered on hate speech labeling and detection.

3.1 Team Moderation and Review Functionality

In nine out of eleven interviews, the massive team focus in practical community management is the most prominent insight. Currently, research on AI-supported community management mainly focuses on supporting individuals. However, there has been little indication so far that community management might require systems/CMSS to facilitate exchange and group work as well. Some interviewees stated that 70% of the comments get moderated individually, while for "the remaining 30%, I will ask my colleagues again". Other team leaders explicitly state that they "regularly encourage employees to share when there is a case that is not clear". As of now, many community managers have to copy comments to external tools (primarily *Slack*) to discuss them—typically lacking whatever meta-information their moderation system provides. What is missing is described as means to "Forwarding [comments to] different employees who can then take it over", as well as functionalities to enrich comments with internal information (e.g., about the user, or interpretations of the comment itself).

REQ$_1$: A CMSS should provide community managers with the ability to assign comments to other peers for review or assessment. This should be accompanied by a functionality to leave internal feedback for comments.

3.2 Interpretability and Transparency

More commonly reflected in extant literature was the demand for interpretability and transparency of AI and system decisions [26]. Six of our interviewees were concerned about having software "that [...] only tells me: delete the [comment/person]". From the interviews and observations, it could be derived that community managers are scanning and searching for specific keywords to make first decisions; as one interviewee put it: "We look of course for keywords that might be relevant". They would appreciate a solution that highlights words or passages that an algorithm deemed problematic to ensure a better and faster understanding of why a particular action should be taken with a comment. This increased transparency would also better accommodate the general moderation process, which the interviewees describe as "relatively fast scanning" with a targeted processing time per comment of sometimes 5 s and less. An alternative approach suggested is sorting of comments based on their criticality to ensure faster processing of out-of-control situations. Furthermore, interpretable cues and efforts towards transparency should be well thought-through, as one interviewee explicitly stated that poorly understandable flags and annotations get ignored.

REQ$_2$: A CMSS should always provide explanations for decisions or assessments created by the AI. This should be done through visual cues such as highlighting specific words or parts and providing additional information.

3.3 Control and Correction of Machine Decisions

Confronting community managers with ideas for the (semi-)automatic detection and deletion (or blocking) of comments typically led to reserved reactions. Many of our interviewees only interacted—if at all—with blacklists as automated solutions where they correctly observed that the context might get lost: “the danger: You have to see the whole context, how people write or use words of course.” Even after receiving additional information and understanding that there are more sophisticated and context-aware means to classify comments, one typical answer to the question of whether they would trust the machine was: “Yes, [it] is hard.” However, this does not indicate a rejection of machine support, as less intrusive measures were welcomed: “Where things are flagged as being of varying degrees of sensitivity. Something like that would be really good, I think”. The majority of the interviewees agree more or less plainly that automation is desirable but not without human control. One participant summarized it as: “So security mechanisms, but otherwise, yes, I think that would be a good thing.”

REQ$_3$: A CMSS should provide community managers with decisions in the form of suggestions that humans can override.

3.4 Decision Support Beyond AI

Throughout the interviews, one pattern kept reappearing: People explicitly asked for or outlined the desire for decision support mechanisms beyond an AI supporting them in comment assessment. Two of our interviewed newspapers are already using systems that provide them with information about the commenters: e.g., their registration time, number of comments, number of flags/interventions. They state that such information helps them interpret comments, make complex decisions, and justify them when comments themselves might be disputable. Another three interviewees of newspapers with nickname registrations listed the detection of duplicates as a pressing concern. The inability to permanently lock out people creates situations where comments have to be repeatedly assessed, taking away time from other tasks. Last but not least, an overview of moderation actions was described as a potential feature. Based on the impressions from some ticket-based systems, community managers see value in getting information on whether a comment—or even the overarching article and its thread—has already been dealt with. The described benefits include being able to track decisions even after users change comments and linked to REQ_1 to see whether colleagues already interacted with an ongoing discussion.

REQ$_4$: A CMSS should provide supportive information beyond AI decisions. This entails elements such as user statistics and an overview of moderation actions for comments and stories.

3.5 Openness of the System

The last aspect that kept reoccurring throughout the interviews can be summarized as the openness of the CMSS. Most of the newspapers we talked with operate multiple commenting opportunities (on-site, on Social Media, . . .) and/or are linked to other newspapers under one overarching corporation. Much of this has grown historically. Community managers often have to switch between systems which creates additional overhead, to the point of losing track of the accounts they have to check. Hence, four interviewees stated they would prefer to have one system to aggregate all the moderation work ("there are two systems now [...] would be cooler, of course, if it were just one system."). One outlet already uses a ticket system (*swat.io*) aggregating several sources, giving away a further requirement for this capability: CMSS cannot be content management systems at the same time but have to be independent.

Beyond this, three of the four outlets work with proprietary systems, and the two larger ones face issues regarding the extensibility and configuration of their systems. There were ongoing efforts to change to an open source system to gain additional degrees of freedom in one case. This does not only pertain to the overall CMSS but also the included AI aspect. While community managers are aware of existing proprietary solutions, they doubt their worth or that it even is a working AI system ("if it is real AI"). This indicates a so far uncovered gap: Academia can offer thoroughly evaluated AI models. There are available model registries [28]. However, no CMSS systems allow for configurable models. Thought further, even external model management and maintenance is up for consideration as most outlets lack internal knowledge to maintain an AI component.

REQ$_5$: A CMSS should provide the ability to be extended and enhanced by its users (without paying any software vendor). The AI part should be externally manageable to ensure being up-to-date and avoid black box models.

4 Discussion and Conclusion

This study aims to provide requirements for CMSS that can provide AI support to non-tech media companies. Despite the plethora of extant AI models, there is still little knowledge on how to provide software artifacts that are of practical use to community managers in non-tech media companies.

Based on eleven interviews with community managers and journalists of four German newspapers, five central requirements have been identified and outlined in this paper. An overview of the requirements and associated cluster elements is depicted in Fig. 1.

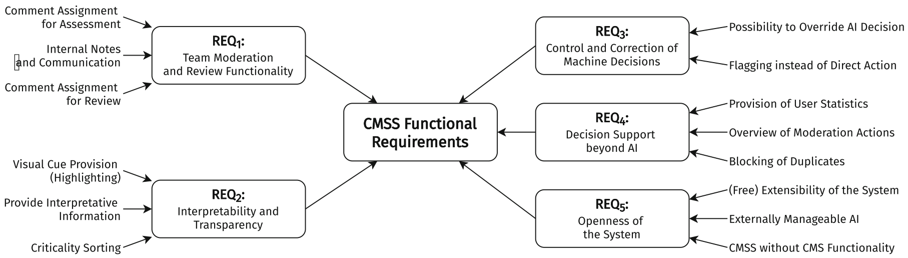
Comment Assignment for Assessment
Internal Notes and Communication
Comment Assignment for Review
REQ1: Team Moderation and Review Functionality
Visual Cue Provision (Highlighting)
Provide Interpretative Information
Criticality Sorting
REQ2: Interpretability and Transparency
CMSS Functional Requirements
REQ3: Control and Correction of Machine Decisions
Possibility to Override AI Decision
Flagging instead of Direct Action
REQ4: Decision Support beyond AI
Provision of User Statistics
Overview of Moderation Actions
Blocking of Duplicates
REQ5: Openness of the System
(Free) Extensibility of the System
Externally Manageable AI
CMSS without CMS Functionality

Fig. 1. CMSS Functional requirements and cluster overview

While there are systems fulfilling parts of these requirements, a CMSS adhering to all is unavailable to the best of our knowledge. Hence, our research agenda's next step is to develop a corresponding CMSS. With this software artifact, the requirements can be validated in a practice setting with our project partners (several German newspapers of various sizes). Based on their feedback, we expect to refine the identified requirements and complement them by additional ones emerging from the hands-on work with the artifact—which will again serve to refine the CMSS artifact. We hope to gain valuable insights into the acceptance of such systems in smaller newsrooms and which measures can be taken to increase the trust in both the system and the novel way of community management.

As a conceptual paper reporting about research in progress, some limitations remain: We are currently only considering the design of a CMSS, abstracting away all elements of such a system that would be commenter-facing. This is reflected in the exclusive consideration of the community manager's perspective. Communication and presentation to commenters and the generation of acceptance are issues to be addressed by future research.

Acknowledgments. 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 Jens Brunk and Vera Steinhoff for their assistance in carrying out and transcribing 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. Bilton, R.: Why some publishers are killing their comment sections (2014). https://digiday.com/media/comments-sections/
3. 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)
4. Braun, J., Gillespie, T.: Hosting the public discourse, hosting the public: when online news and social media converge. Journal. Pract. **5**(4), 383–398 (2011)
5. Coe, K., Kenski, K., Rains, S.A.: Online and uncivil? Patterns and determinants of incivility in newspaper website comments. J. Commun. **64**(4), 658–679 (2014)
6. Einwiller, S.A., Kim, S.: How online content providers moderate user-generated content to prevent harmful online communication: an analysis of policies and their implementation. Policy Internet **12**(2), 184–206 (2020)
7. Fortuna, P., Nunes, S.: A survey on automatic detection of hate speech in text. ACM Comput. Surv. **51**(4), 1–30 (2018)
8. Gardiner, B., Mansfield, M., Anderson, I., Holder, J., Louter, D., Ulmanu, M.: The dark side of Guardian comments (2016). https://www.theguardian.com/technology/2016/apr/12/the-dark-side-of-guardian-comments
9. Goodman, E.: Online comment moderation: emerging best practices - a guide to promoting robust and civil online conversation. Technical report, World Association of Newspapers (WAN-IFRA), Darmstadt, Germany (2013)

10. Hadar, I., Soffer, P., Kenzi, K.: The role of domain knowledge in requirements elicitation via interviews: an exploratory study. Requir. Eng. **19**(2), 143–159 (2014)
11. HASOC: Call for Participation (2019). https://hasocfire.github.io/hasoc/2020/call_for_participation.html
12. Hayek, L., Mayrl, M., Russmann, U.: The citizen as contributor-letters to the editor in the Austrian Tabloid Paper Kronen Zeitung (2008–2017). Journal. Stud. **21**(8), 1127–1145 (2020)
13. Hermida, A., Thurman, N.: A clash of cultures: the integration of user-generated content within professional journalistic frameworks at British newspaper websites. Journal. Pract. **2**(3), 343–356 (2008)
14. Juarez Miro, C.: The comment gap: affective publics and gatekeeping in the New York Times' comment sections. Journalism, 1–17 (2020)
15. Kim, J.: Moderating the uncontrollable. Intersect Stanford J. Sci. Technol. Soc. **10**(3), 1–9 (2017)
16. Köffer, S., Riehle, D.M., Höhenberger, S., Becker, J.: Discussing the value of automatic hate speech detection in online debates. In: Tagungsband Multikonferenz Wirtschaftsinformatik 2018. MKWI 2018, Lüneburg, Germany (2018)
17. Kolhatkar, V., Taboada, M.: Constructive language in news comments. In: Waseem, Z., Chung, W.H.K., Hovy, D., Tetreault, J. (eds.) Proceedings of the First Workshop on Abusive Language Online, pp. 11–17. ALW1, ACL, Vancouver (2017)
18. Muddiman, A., Stroud, N.J.: News values, cognitive biases, and partisan incivility in comment sections. J. Commun. **67**(4), 586–609 (2017)
19. Nielsen, R.K.: Participation through letters to the editor: circulation, considerations, and genres in the letters institution. Journalism **11**(1), 21–35 (2010)
20. Niemann, M., Welsing, J., Riehle, D.M., Brunk, J., Assenmacher, D., Becker, J.: Abusive comments in online media and how to fight them. In: van Duijn, M., Preuss, M., Spaiser, V., Takes, F., Verberne, S. (eds.) MISDOOM 2020. LNCS, vol. 12259, pp. 122–137. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-61841-4_9
21. Papacharissi, Z.: Democracy online: civility, politeness, and the democratic potential of online political discussion groups. New Media Soc. **6**(2), 259–283 (2004)
22. Risch, J., Ruff, R., Krestel, R.: Offensive Language Detection Explained. In: Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying, LREC 2020, pp. 137–143, ELRA, Marseille (2020)
23. Robson, C., McCartan, K.: Real World Research: A Resource for Users of Social Research Methods in Applied Settings, 4th edn. Wiley, Hoboken (2016)
24. Schabus, D., Skowron, M.: Academic-industrial perspective on the development and deployment of a moderation system for a newspaper website. In: Proceedings of Eleventh International Conference on Language Resources and Evaluation, LREC 2018, pp. 1602–1605. ACL, Miyazaki (2019)
25. Vidgen, B., Yasseri, T.: Detecting weak and strong islamophobic hate speech on social media. J. Inf. Technol. Polit. **17**(1), 66–78 (2020)
26. Wich, M., Bauer, J., Groh, G.: Impact of politically biased data on hate speech classification. In: Proceedings of the Fourth Workshop on Online Abuse and Harms, pp. 54–64. ALW4, ACL, Stroudsburg (2020)
27. Wulczyn, E., Thain, N., Dixon, L.: Ex machina. In: Proceedings of the 26th International Conference on World Wide Web, WWW 2017, pp. 1391–1399. ACM Press, Perth (2017)

28. Zaharia, M., et al.: Accelerating the machine learning lifecycle with MLflow. Bull. IEEE. Comput. Soc. Tech. Comm. Data Eng. **41**(4), 39–45 (2018)
29. 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