HATEFUL COMMENTING ONLINE.

CONTROL STRATEGIES FOR NEWSROOMS

Leif Kramp & Stephan Weichert

LANDESANSTALT FÜR MEDIEN NRW
Der Meinungsfreiheit verpflichtet.

The comprehensive results were published in German by the VISTAS Verlag in December 2018: :

Kramp, Leif und Weichert, Stephan:
Hasskommentare im Netz. Steuerungsstrategien für Redaktionen. Leipzig: 2018
(Schriftenreihe Medienforschung der Landesanstalt für Medien NRW, Bd. 79),
ISBN 978-3-89158-647-1, 26,- €

HATEFUL COMMENTING ONLINE.

CONTROL STRATEGIES FOR NEWSROOMS

Leif Kramp & Stephan Weichert

With the assistance of
Viviane Harkort and Lara Malberger

LANDESANSTALT FÜR MEDIEN NRW
Der Meinungsfreiheit verpflichtet.

CONTENTS

FOREWORD

People share their personal feelings and thoughts, they grieve and celebrate together, and are likely to react spontaneously. However, they also pass on unchecked or even incorrect information –intentionally or against their own better judgement. Digital audience relations (specifically the commenting attitudes of media users) have radically changed the culture of public debate. Particularly for young target groups there is a potential for diverse and improved opportunities to participate in social discourse, as well as take part in forming opinions on public policy. At the same time, this culture of debate also poses great risk to democratic discourse rationality.

For newsrooms, their editorial staffs, and journalism as a whole, while dialogue with the audience (including viewing, moderating, re-viewing, and activating comments) does repre-sent a creative exercise, it is mainly just frust-rating. Although researching and publishing proprietary information are considered among the central tasks of serious reporting, editorial staffs still perceive the moderation of user discourse in the comment sections of their news sites as an extra unpleasant task due to the lack of resources.

On behalf of the Media Authority of North Rhine-Westphalia and with the financial support of Google Germany, we studied the discussion behaviour of users of the leading online news brands and their specific moderation strategies. The analysis comprised the editorial websites and one social media presence each of Deutschlandfunk Kultur, RP ONLINE, RTL, and tagesschau.de. We were mainly interested in the issue of how journalistic media can use targeted strategies and editorial control mechanisms (including moderation, community management, audience engagement, and deletion policies) to be able to monitor user discourse constructively and regulate escalating disputes. The study was intended to show editorial staffs starting points in (potentially) hate-driven discussions to allow them to integrate them easily into their everyday work in the editorial room.

The authors thank: Viviane Harkort, Lara Malberger, Daniel Moßbrucker, Lisa-Marie Eckardt (project collaboration); Dr. Nicola Balkenhol and Torben Waleczek (Deutschlandfunk Kultur); Daniel Fiene, Julia Nix, Mathias Schumacher, and Henning Bulka (RP ONLINE); Christian Beissel (RTL Inter-active); Christina Elmer, Torsten Beeck, Eva Horn, Philip Löwe, and Werner Theurich (SPIEGEL ONLINE); Dr. Kai Gniffke, Christiane Krogmann, Rike Woelk, and Sven Königsmann (tagesschau.de); Sabine Frank and Anika Lampe (Google Germany); Patricia Georgiou and Lucas Dixon (Perspective API); Dr. Meike Isenberg, Marie-Franca Hesse and Mechthild Appelhoff (Media Authority of North Rhine-Westphalia).

Bremen/Hamburg im Juni 2018,
Leif Kramp & Stephan Weichert

MODERATION STRATEGIES – RECOMMENDATIONS FOR NEWSROOMS

The ever increasing demands posed by audience engagement, community development, and platform management require a great deal of energy, impressive tact, and nerves of steel. Often, editorial staffs are simply lacking in time and resources, while sometimes they do not have specific techniques and tools at their disposal to handle situations where they aim to make user discourse more objective by means of arguments, rather than watching an emotionally charged debate escalate. Codes of conduct in place to leave as little room as possible for hate speech, harassment, extremism, slander, and marginalization are one helpful element based on the continuous exchange of experiences both within the newsroom and beyond. Promoting user comments constructively while widely preventing conflict through moderation is primarily a matter of employing a comprehensive strategy for the moderation process. Should social media editors intervene at an early stage or allow conversations to run their course? Are moderators striving to delete hate speech or to have discussions with haters, and are these moderators acting as individuals or on behalf of the media brand? How can community managers protect themselves against verbal aggression? Should trolls be banned, should their comments be deleted, or should they even be prosecuted?
On the basis of our empirical evidence, we are presenting a **10-point plan** to combat hate speech with the objective of repairing the often problematic debate culture in the comment sections of news sources.

1 Decisive moderation

To promote the contribution and argumentation of serious commenters, employ an unambiguous and resolute moderation style. You don't have to remain silent while haters, troublemakers, and trolls make your life difficult with their primitive nonsense. Defuse and banish linguistic intimidation from your comment section by applying dauntless yet factual moderation which indicates who is in charge on your site (including any social media offshoots).

2 Direct approach

Speak up frequently rather than just concentrating on deleting or blocking problematic comments. Try to address malicious and abusive users directly and call them to order. Often the mere fact that haters realise that they are being observed and monitored is enough to civilise their discourse. Make it clear to your moderators that you will only accept comments on your website and your social media accounts which adhere to the editorial staff's netiquette policies and exhibit courteous and fair interaction..

3 Encourage counterspeech

It is essential to reward counterspeech creators who speak up against hate speech in your comment section. Seek out the loyal and active commenters among your users and join forces with them or demonstrate your solidarity with them. Entrust them to co-moderate discussions which have a risk of escalating but could be mediated at an early stage by means of communicative self-regulation. Pay particular note to the presence of activist groups such as #ichbinhier[1] as they provide constructive intervention to discourse and actively crack down on hate and harassment. The targeted empowerment of counterspeech and constructive communication can increase the self-healing powers of the Internet and strengthen the immune system of debate culture.

[1] The Facebook Group #ichbinhier *("I'm here")* is a non-partisan action group promoting better debate culture and combatting harassment on social media.

4 Campaigns to combat hate speech

Devise journalistic programmes, formats, and events which address the root causes of hate. Ignorance, a lack of understanding, and disappointment often lead to emotionally charged abuse. Measures such as Tagesschau's "Sag's mir ins Gesicht" [2] campaign are valuable experiments against hate speech which demonstrate that open discussion with frustrated critics can definitely pay off. Only flexible editorial work on equal footing can invalidate allegations of being "state media" or "fake news" .

5 Minority-led dominance

Realise that a loud minority dominates the digital space for discourse with their hateful posts while the majority of users remain silent. Although the realisation of this imbalance does not dismantle online hate and harassment, just this awareness can help you in perceiving the deterioration of debate culture as much more benign. Blocking posts (promptly) can serve to specifically isolate haters and troublemakers.

6 Constructive journalism

Feel-good and fluffy journalism cannot stop the coarsening of commenting culture in the hard news business, but the regular publication of solution-oriented comment posts can – especially posts which relate to the actual social environment of users. Academic studies has shown that the more constructive the reporting is, the more engaged users tend to be. This also encourages more harmony in the debate culture overall since commenters get less worked up over problems and crisis-ridden trends. Instead, they relay more of their positive experiences and perspectives.

7 Man-maschine filter

Although algorithms and artificial intelligence cannot replace the human assessment of comments, they can make it easier by means of prefilters. The automated channelling of user feedback using technological systems based on speech or syntax recognition can sort out the bad comments and even delete them as necessary. Whoever has to read, moderate, and analyse several thousand comments each day will see this as a welcome relief and be able to dedicate themselves to positive user posts.

[2] As part of the "Sag's mir ins Gesicht" *("Say it to my face")* campaign,Tagesschau journalists encouraged commenters to confront them in a video chat which was aired live on Facebook.

8

Zone free of sarcasm and cynicism

One of the greatest challenges for comment moderation is the fact that it is difficult for both computers and many users to recognise sarcasm. Therefore, healthy user discourse dictates that sarcastic moderation elements should only be applied with consideration and that no users should be ridiculed – not even the unkind ones. This applies even when it sometimes requires a great deal of self-discipline not to take the wrong tone when dealing with pernicious commenters. Your inner cynic is not a good advisor, either, because it often tempts you to simply stifle discussions rather than encouraging objective posts.

9

Provide resources

Holding onto the dialogue with users and continuing to promote it to utilise the full creative potential of the commenting possibilities for your own media brand is mainly a matter of making the needed capacity and infrastructure available. The double staffing of social media editors for each platform as well as the targeted cooperation of comment moderators and authors can enable a focus on the content-related aspects of debate. Rather than succumbing to emotional provocation, moderators should use their considerable resources to react to the motives of commenters..

10

Earn respect

Especially in times of harsh communication, it is not only a question of communicating with one another on equal footing. Conveying a sense of wrongdoing to users also involves consistently showing repeat offenders their boundaries. Continuously active trolls and haters who contaminate entire online news sites with their hateful posts should be banished from the comment sections and prosecuted where appropriate. Campaigns such as "Verfolgen statt nur Löschen" [3] from the Media Authority of North Rhine-Westphalia are important steps for collaborations with the goal of effectively combating hate crime and facilitating safe and fair discourse online.

[3] This first-of-its-kind initiative was founded in 2017 with the objective of setting a clear signal against lawlessness and ruthlessness on the Internet and consequently promoting freedom and democracy online. Its name encourages "Persecution rather than just deletion". In addition to the Media Authority of North Rhine-Westphalia, the initiative's collaborators and founding members include the cybercrime central department and contact point for North Rhine-Westphalia (ZAC NRW) at the Cologne public prosecutor's office, as well as the media companies Rheinische Post, Mediengruppe RTL Deutschland, and WDR. These participants are also in interaction with the platforms Google and Facebook.

RESEARCH DESIGN

With the online editorial staffs of Deutschlandfunk Kultur, RP ONLINE, RTL, and Tagesschau, four archetype cooperation partners were acquired for the study from various types of quality media with editorial offices located in North Rhine-Westfalen and Berlin (DLF: Cologne/Berlin, RP ONLINE: Düsseldorf, RTL Aktuell: Cologne), as well as in Hamburg (tagesschau.de). Furthermore, SPIEGEL ONLINE (Hamburg) is another knowledgeable journalistic partner available for our panel of experts. SPIEGEL ONLINE boasts an extremely high volume of regular comments and many years of experience in processing comments, for example in pertinent forums. During the initial phase of the study, a total of twelve expert panels including a pre-test were held with select social media editors and the managers of these newsrooms. During the second phase, select online discourse processes from the second half of 2017 were analysed. During the third phase of the study, we conducted an experiment with one of the online editor staffs as a sample in spring 2018.

The sample of the comment analysis represents the core element of the study. It comprises a total of 24 different online discourse processes relating to 16 relevant articles (cf. Table 1, p. 12). In doing so, the relevant platform and moderation strategies of the cooperation partners were taken up. For example, preliminary talks with the responsible staffers already indicated that no considerable staff resources were allocated to the social media moderation of YouTube, Twitter, or Instagram. Rather, the strategic considerations were primarily concentrated on their Facebook pages. In terms of the current development and practice of moderation strategies, as well as relating to the newsroom management of user discourse, other platforms like Twitter, YouTube, and Instagram are of no importance or only play a negligible role. Therefore, our analysis covered the user discourse and commenting on the Facebook offshoots of the aforementioned media offerings, as well as in the comment sections of the newsrooms' own websites at RP ONLINE and tagesschau.de.

DISCOURSE TYPOLOGY

The user discourse on the fan pages of each medium relating to the relevant articles can be broken down into five categories (cf. Fig. 1):

Figure 1: Typology of the analysed user discourse

1 Hateful or conflict-ridden discourse relating to articles about sensitive sociopolitical subjects (strongly characterized by many destructive comments)

2 Discourse which does not tend to escalate but is characterised by strong negativity and individual comments

3 Discourse with a self-regulation effect by users (destructive comments are neutralized by constructive, solution-oriented comments)

4 Discourse with primarily constructive, solution-oriented and/or affirmative comments (involving great/little moderation effort by the editorial staff)

5 Discourse relating to realistic everyday issues and a high volume of primarily neutral comments

Figure 1: Characterization of the analysed user discourse

Discourse characterization	Medium	Beitragstitel	Veröffentlichung
Hateful or conflict-ridden discourse relating to articles about sensitive sociopolitical subjects (strongly characterized by many destructive comments)	RP Online	„Studie bescheinigt Muslimen Erfolge auf dem Arbeitsmarkt“ (Topic: Muslims on the job market)	24.8.2017
	RTL Aktuell	„Flüchtlingsunterkünfte oft sehr mangelhaft“ (Topic: Many refugee shelters inadequate)	7.12.2017
	RTL Aktuell	„GEZ-Schock: Neuer ARD-Chef fordert höheren Rundfunkbeitrag“ (Topic: Possible rise in TV licence fees)	30.12.2017
	Tagesschau.de	„Milliarden-Deal mit Israel” (Topic: Billion-euro deal made with Israel)	23.10.2017
	Tagesschau.de	„Kai Gniffkes Kommentar zur AfD” (Topic: Kai Gniffkes kommentar on the right-wing party AfD)	12.11.2017
Discourse which does not tend to escalate but is characterized by strong negativity and individual comments Discourse with a self-regulation effect by users (destructive comments are neutralized by constructive, solution-oriented comments)	RTL Aktuell	„Sexuelle Belästigung: 16 Frauen erheben sich gegen Donald Trump“ (Topic: Women alleging harassment by Donald Trump)“	12.12.2017
	Deutschlandradio	„Thea Dorn zur Sexismus-Debatte: 'Ein neuer Totalitarismus'“ (Topic: Op-ed relating to sexism debate)	10.11.2017
	Deutschlandradio	„Sollen wir die AfD wie jede andere Partei behandeln? Liane Bednarz vs. Michel Friedman” (Topic: Debate on the right-wing AfD party)	31.10.2017
	Deutschlandradio	„Die Staatsfunk-Kampagne wird weitergehen. Ein Kommentar von Brigitte Baetz“ (Topic: Op-ed on state-funded media outlets)	20.10.2017
	RP Online	„Flucht aus Syrien: Wiedersehen nach 1162 Tagen“ (Topic: Reunion of Syrian refugee family)	17.7.2017
	RTL Aktuell	„GroKo wäre eine Koalition der Verlierer – aber wo sind die Alternativen“ (Topic: Criticism of Germany's grand coalition)	29.12.2017
	Tagesschau.de	„Alle Jahre wieder: Gerüchte über Weihnachtsmärkte“ (Topic: Rumours about Christmas markets)	17.11.2017
Discourse with primarily constructive, solution-oriented and/or affirmed/encouraging comments (with great/little comments moderation effort by the editorial staff)	Deutschlandradio	„Was ist Ihre Lieblingsband aus der DDR?“ (Topic: favourite bands from the former GDR)	16.10.2017
	RP Online	„DEG Winterwelt in Düsseldorf: Eisbahn auf der Kö wird ab nächste Woche aufgebaut“ (Topic: Winter ice skating rink in Düsseldorf)	2.11.2017
Discourse relating to realistic everyday issues and a high volume of primarily neutral comments	RP Online	„Unfall: Frau verursacht Totalschaden wegen Spinne im Auto“ (Topic: Woman totals car due to spider)	5.4.2017
	Tagesschau.de	„LKW der Zukunft?“ (Topic: Lorries of the future)	17.11.2017

ONLINE DISCOURSE ANALYSIS: KEY RESULTS (EXTRACT)

1 There is **hardly any moderation by the editorial staff** in terms of active discussion posts, which results in the newsroom only exercising a minor influence on the development of the public discourse relating to its news sources.

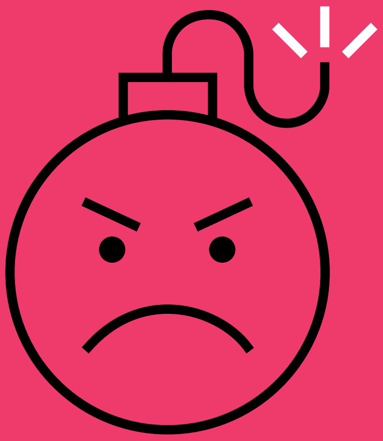
2 **Active discussion participation** by the respective newsroom has a direct impact on the positioning of the respective main comment and its discussion thread in the current algorithmic sorting process. The respective editorial comments are automatically ranked higher. In this way, an editorial staff is able to help certain discussion threads garner more attention by means of **targeted commenting**.

3 The **accusation of propaganda and fake news** was observed in nearly all of the analysed discourse. **Regardless of the topic** of the article, users accuse the journalists of deliberate manipulation and subjective reporting.

4 **A maximum of one-third of the comments has a thematic relevance** to the respective article. In contrast, many comments pick up on formulations which newsrooms use in their articles, modify them, and/or put them in new contexts. The majority of these comments comprise off-topic opinions, often including slander and harassment.

5 There are only a **few influential yet consistently negative commenting users,** who mostly post similar comments invoking certain viewpoints and/or actions. Commenting like this appears to indicate typical 'troll' behaviour. These repeat offenders are peculiar characters whose motives range from a desire for self-affirmation to missionary fervour.

6 Nearly all comments are posted on the **first day following the publication of an article**. After that, users only post comments occasionally and sporadically. Furthermore, the later comments appear in the discourse, the shorter they are and the less relevant they are to the article.



EDITORIAL MODERATION STRATEGIES TO COMBAT HATE SPEECH

In social (and dissocial) discourse on the Internet, skilled comment moderation can filter out the best of the arguments, assessments, positions, and viewpoints of users – and moderate out or delete the worst. This is usually the expectation of newsrooms which hope for added value from their professional community management. However, in the specific practical experience of journalists, there are often not enough editorial resources or sufficiently trained editorial staff available to be able to fulfil these expectations.

Frequently, social media editors and their (student) temps are continuously exposed to drastic stressors in connection with explicit words and images (especially hate speech, racism, anti-Semitism, calls for violence), some of which is targeted towards them personally. For the most part, editorial staffers must decide within just a few seconds whether they should delete or publish a user's post. What they experience at an accelerated pace could potential have grave psychological consequences. There are very few therapeutic services in place for newsrooms to process these traumatic experiences or share them with other affected parties. The opinion that the development of automatized filter systems could represent *the* long awaited solutions to control this kind of content and thus civilise discourse using deletion practices is contentious.

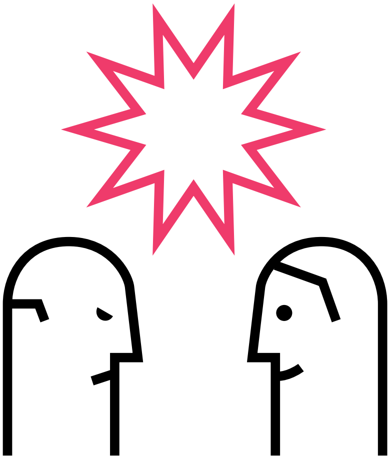
We will introduce two moderation strategies below with a total of ten moderation elements which could enable newsrooms to not only cope with the increasing proliferation of hate speech in user discourse, but also effectively control it. In cooperation with RP ONLINE, we tested strategies derived from discourse analysis in practice. They can be deployed successfully against hate speech.

The quantitative-qualitative discourse analysis which is the foundation for this white paper prepared on behalf of the Media Authority of North Rhine-Westphalia pertains both to the comment sections on the websites of the four analysed high-quality media, as well as the comment sections of select social media platforms managed by editorial staffs. As a result of the empirical study, two distinct debate cultures have crystallized with regard to the two options for publishing comments:

> a) with regard to *comments on their own websites*, this relates more directly to the moderation of discursive content provided by the corresponding news site. It could also be feedback about the newsroom as an organisation or about individual authors or editors, or even the media brand as a whole.
>
> b) with regard to *comments on the analysed social media channels*, the discussions are dominated by users, which means that the newsroom plays a subordinate moderation role because the outsourced debates relate to real life more generally and users refer to one another more often.

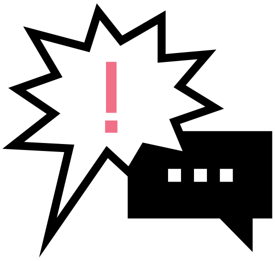
The fact that the deliberative quality of comments can also vary (it was lower on social media channels) and that there is a direct correlation between the barriers to participation and the civility of user debate is primarily due to the nature of the respective commenting infrastructure and its technology: On the social media channels of the medium, users mainly post from their own accounts and often following their personal networks (and thus receive more feedback from their "friends" to these articles due to the algorithmised logic). In contrast, on the newsrooms' websites, users are frequently subjected to complex registration processes or a rigorous review of their profiles by the staff. On social media channels, newsrooms also make less use of their "home field advantage”. As a general rule, staffs also make less of an effort to ensure that their netiquette is observed. During the twelve expert panels, it was confirmed time and again that this was above all a question of resources. Surprisingly, many newsrooms give up simply in light of the volume of comments, claiming it is no longer manageable.

In the experimental live operation at RP ONLINE, it became evident that in particular the principles of *empowerment* and *disempowerment* of commenters, entire discourses, or individual discourse elements should be applied as practicable moderation strategies. Psychologists have found that a system of rewards and punishments can be quite successful in dealing with haters. Field research has shown that psychological incentives



can be effective for social and against anti-social behaviour in comment sections by following the motto of *"Don't feed the trolls!"* and not giving haters a platform for their public abuse. Overall, with regard to comment moderation, we can specify at least three archetypical approaches to limit hate speech and inspire constructive discourse. In general, these can be broken down into *protective* (defending users), *disqualifying* (excluding users) and *supportive* approaches. It has been demonstrated that the range of provocative approaches – for example inordinately coarse communication provoked by the moderation – intensifies the trouble caused by haters and trolls.

Accordingly, when testing the moderation strategies for the civilisation of online discourse, we primarily concentrated on the principles of protecting, supporting, and excluding users. However, during the practical test, provocative elements were also tested occasionally in the form of irony. As a whole, depending on the intensity and range of actions, ten different moderation elements can be entered on a graph according to the four-field diagram of strong/weak "disempowerment" vs. strong/weak "empowerment" and a (relatively) strong vs. (relatively) weak "degree of editorial effort" (Fig. 2).



REGULATING STRATEGY: DISEMPOWERMENT

Punishment

Consequences for hate speech under criminal and civil law

This approach is a component of an exclusionary strategy in which comment moderators are required to warn users or have them prosecuted if warranted by the opinions they voice. Punishment features a relatively moderate degree of editorial effort on the y-axis but the highest level of disempowerment on the x-axis. Defamation, libel, and sedition are not covered by freedom of speech any more than calls to crimes or violent acts are trivial offences. Rather, they are active offences which can be prosecuted and punished with imprisonment. This applies to editing staffs' social media channels, as well as to posts on their own websites. Potential culprits tracked down via their IP addresses may be subject to criminal penalties such as imprisonment or fines pursuant to Sections 185, 186, and 187 of the German Penal Code (StGB).

Counter Speech

Conclusive arguments against hate speech

In the moderation process, this is one of the most complicated strategies since it requires active communication and in particular close attention and dynamic action by the comment moderators. However, it is also the most important attempt to transform dissocial discourse into constructive discourse without blocking individual users or deleting their posts. Accordingly, counterspeech –also including the moderators countering with their own rude comments, although we do not recommend that – is characterised by a comparatively high level of editorial effort and powerful disempowerment. Making conclusive arguments against hate speech – both from the user and newsroom side – is also shaped by the attitude of not putting up with everything posted by haters. A smart alternative for low-resource newsrooms is rewarding counterspeech among their users by posting encouraging comments (cf. "embracing") rather than staffers continuously posting counterspeech themselves. This bolsters this loyal group of users and has a positive impact on the self-regulating debate culture.

Deconstructing

Deconstructing hateful commenters and hate speech

Deconstructing hateful commenters can be disheartening and is a strategy which even well-staffed editorial teams are barely able to accomplish because they entail the thorough management of haters and trolls. On one hand, hardly any media companies can invest the time and money required to employ this strategy. On the other hand, it would seem reasonable to refute the expression of opinions from this loud minority down to the last detail so that they do not influence the silent minority, e.g. with populist misstatements. The expense, meaning the high level of editorial effort required for this approach, only seems justified if the deconstruction principle is effective against hateful posters and demonstrates a sustainable impact for the entire debate culture for the respective medium. Disempowerment is somewhat less pronounced in this regard since moderators have to "sink to" the level of haters and read into their lines of argument.

Blocking/Deleting

Muting hateful posters, blocking/deleting hateful posts

This strategy is currently one of the most common among the German media companies as it entails relatively little time and effort, while also being one of the few sustainable approaches. Whether this policy is carried out internally by editorial staffs working in shifts or it is outsourced to a service provider, deleting hate speech has prevailed, especially for comments. However, any success here is fleeting: many hate posters and troublemakers keep returning with newly created social media profiles with different names and carry on where they left off. The following approach has proven to be an efficient alternative: instead of deleting hate speech, its authors are muted (by means of additional editorial functions) and their comments are hidden from other users while remaining visible to them. The level of editorial effort is also comparatively low for this strategy, while the disempowerment is quite high.

Ignorance (withdrawal of attention)

Ignoring hate speech rather than reacting to it

Completely ignoring negative comments can also be a moderation element, even if it is rather passive. It represents a minimum of editorial effort and at the same time limited disempowerment. The issue for newsrooms is if a user discussion which could include criminal offences unfolds in a comment section managed by an editorial staff, ultimately the newsroom could potential be made (jointly) responsible for it. Instead of ignoring hate speech entirely, it can be more advantageous for moderators to deliberately turn a blind eye to individual troublemakers and not react to their hateful posts to deprive their arguments of any attention. This withdrawal of attention is a courageous, very basic editorial decision because it can cause the momentum of discourse in the newsroom's scope of responsibility to intensify dramatically, especially when dealing with social media.

EMPOWERING STRATEGY: EMPOWERMENT

Ironization (humour)

Ironizing hate speech and facing it with humour

Although ironizing hate speech is not a desirable moderation strategy, this method is often applied nonetheless. In many newsrooms, moderators combat negative comments from users with irony, and in the meanwhile even with sarcasm and cynicism – sometimes out of frustration, and sometimes out of helplessness. With an ironic undertone and ironic language, using GIFs, emojis, and likes, they attempt to flip dissocial discourse into the positive, which can potentially be counterproductive. In written exchanges, irony is often met with misunderstanding, especially when the debaters are not personally acquainted with one another. Although the comparatively low level of required editorial effort for this kind of moderation element is tempting, the professional standards for journalism also apply for written comment moderation in general. Irony should only be deployed rarely by moderators and not used to disparage or mock troublemakers and haters because in case of doubt it could fall back negatively on the newsroom, particularly as the empowerment effect is limited here.

Understanding

Determining the background of hate speech, de-emotionalising the debate

A further strategic element which requires a relatively high degree of editorial effort and thus editorial resources is showing understanding to hateful posters, bullies, and troublemakers. Within the spectrum of empowerment, when seen as a transfer of arguments, this approach is not as time-consuming as an intensive dialogue. However, even targeted questioning relating to the background of hateful comments, the authors' motives, and the assessment of the underlying (moral) standards with a carefully moderated discussion of values require tenacity, as well as strong nerves and trained personnel. In this context, it can be worthwhile to persistently de-emotionalise poisoned discussions and provide clarity with regard to hate speech by insisting on the exchange of arguments and personal justifications. This is reflected not least in the attitude of the respective media brand that is being represented to the outside world. This kind of attitude can be positioned and defined by the newsroom if it employs a strategy like this.

Dialogization

Meditating between opposite standpoints, encouraging dialogue between users

Dialogization is the ultimate in strategic elements. For example, dialogization encompasses mediating between opposite standpoints, which could occasionally include hate speech. Unlike "understanding", moderators do not only attempt to correct user behaviour here. Instead, they also aim to intervene to encourage dialogue which does not tail off into emotional skirmishes. The hope is that negative commenters can also be won over for constructive dialogue if they move to a level of objective arguments. This requires moderators to open themselves up to extreme standpoints, even if they are based on incorrect facts or insults. Because pondering standpoints like this can be extremely fraught and an extremely high degree of editorial effort is required, double staffing social media moderation shifts is generally recommended. You run the risk of frustration from all sides, but with the prospect that even haters are ultimately human and their honour will cause them to join a civil discussion at some point. When this happens, this strategic element has been successful and exhausted its high level of empowerment potential accordingly.

Solidarization

Showing solidarity with the affected parties and opponents of hate speech

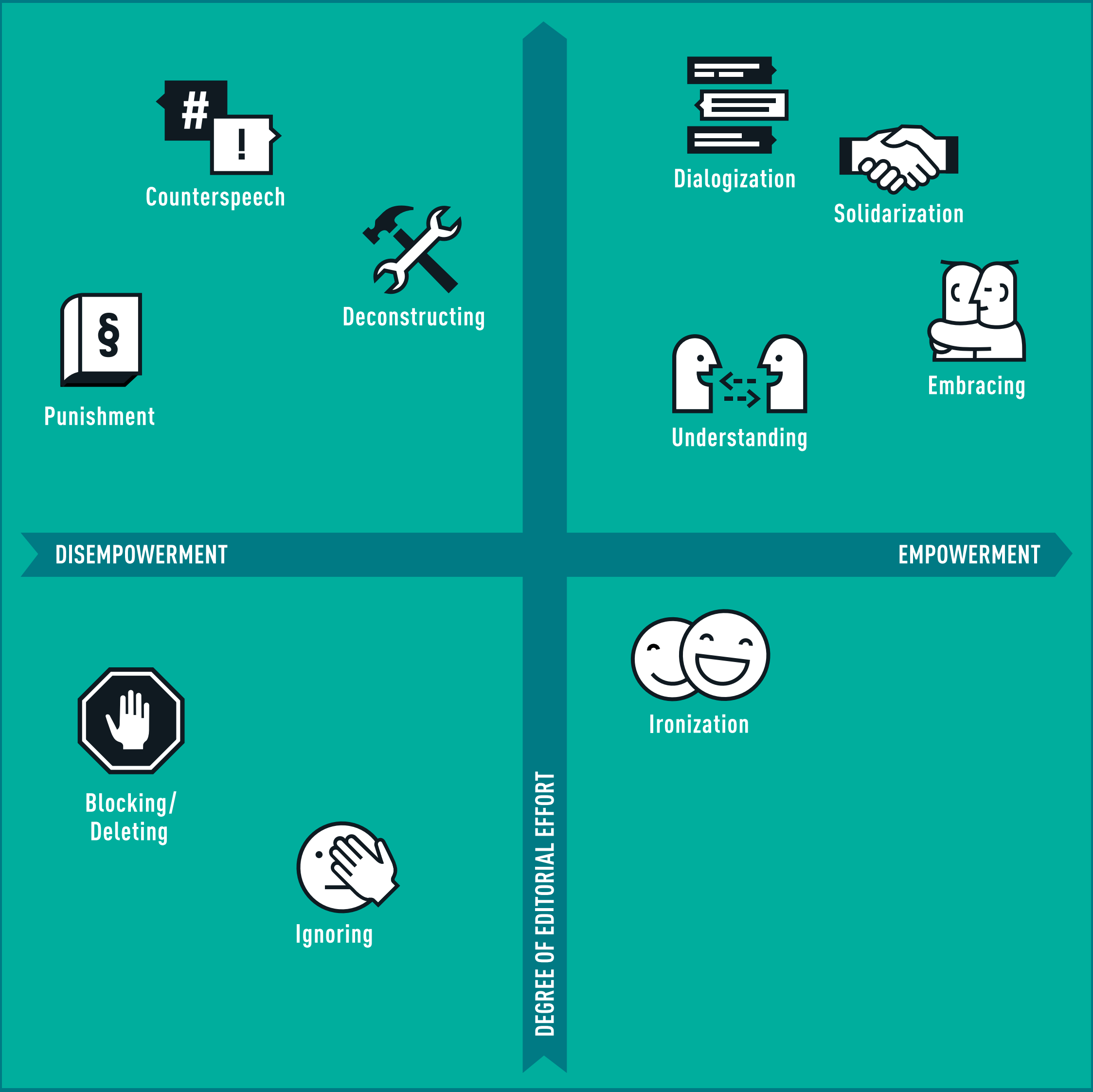
At times, this moderation element transcends the borders of journalistic professionalism. However, it cannot be ruled out because of its strategic impact. In journalism, showing any solidarity with an issue (even a good one) is considered improper, but it can also entail interesting effects on user discourse. Firstly, this method proves constructive because comment sections can suddenly develop into hate-speech-free zones by no longer tolerating haters and troublemakers because of the strong solidarity shown by the editorial staff to their opponents (who post counterspeech). Secondly, the group of hate speech opponents could also potentially become an extremely loyal user community who not only weigh into discussions with more authority but also identify more strongly with the media brand and dedicate additional users to it. In other words, solidarization concepts such as the #metoo movement could grow into waves of solidarity, which would benefit the medium and its journalistic attitude. In accordance with the strategy matrix (Fig. 2), this method is associated with a high level of empowerment, which can be raised by putting in a correspondingly high level of editorial effort.

Embracing (motivating)

Purposefully bolstering affected parties and counterspeech in discourse

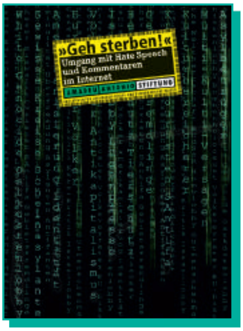
The embracing method refers to users posting counterspeech and the objects of hate speech and means purposefully supporting them in their argumentation and ultimately in their communication against and vis-à-vis hateful posters. However, unlike overt counterspeech, this strategy foregoes explicit languages, warnings to hateful posters, propaganda, and the like. Instead, the side opposing hate speech is verbally reaffirmed and shown moral support through the presence of moderators. This also requires a great degree of editorial effort to allow this empowerment to take hold. As part of this approach, haters and troublemakers are not ignored, blocked, or punished. The discourse is improved by the fact that they ideally disqualify themselves with their dubious arguments, which makes their targets feel better protected. It would seem sensible to combine this strategy with regulating moderation elements such as counterspeech or additional empowerment through solidarization.

Figure 2: Editorial Moderation Strategies to Combat Hate Speech – Matrix of regulating and motivating moderation elements



ANNEX

LEITFADEN FÜR JOURNALIST*INNEN IM UMGANG MIT HATE SPEECH IM NETZ



»Geh sterben!«
Umgang mit Hate Speech und Kommentaren im Internet
AMADEU ANTONIO STIFTUNG

WORLD EDITORS FORUM
WAN IFRA
Do Comments Matter?
Global Online Commenting Study 2016

How To Create
Your Community
Your name:
Project name:
Date:
A Workbook from The Coral Project

GUIDELINES AND AUTOMATIZATION TOOLS FOR NEWSROOMS WHEN DEALING WITH HATE SPEECH

Guidelines

NO HATE SPEECH MOVEMENT:
Leitfaden für Journalistinnen und Journalisten im Umgang mit Hate Speech im Netz (2017)

The guidelines of the "No Hate Speech Movement", coordinated by the non-profit New German Media Professionals (*Neue deutsche Medienmacher*), mainly comprise specific recommendations for action for newsrooms and individual journalists when dealing with online hate speech. It is currently only available in German and has the objective of providing advice and recommending options for dealing with hate speech on social networks, particularly Facebook – both for hate speech which personally targets the newsrooms and articles and for hate speech which does not have a direct reference to the journalists as individuals or the organisation.

→ Download as a PDF

AMADEU ANTONIO STIFTUNG:
„Geh sterben!“ Umgang mit Hate Speech und Kommentaren im Internet (2016)

This brochure, which was published by the Amadeu Antonio Foundation and is currently only available in German, discusses the topic of hate speech as a phenomenon in a society that is increasingly shaped by online communication. The publication aims to provide an overview of the debate on the topic of hate speech and formulate possible solutions. In this framework, it provides definitions for the categorisation, lists features of hate communication to make it easier to detect, and offers insight relating to how affected parties perceive hate speech thanks to the experiences reported by journalists and victims of viral controversies.

→ Download as a PDF

WAN-IFRA:
Do Comments Matter? Global Online Commenting Study (2016)

In its study titled "Do Comments Matter?" the World Association of Newspapers and News Publishers addresses the added value of comments on online news in an editorial context and primarily analyses how journalists worldwide act towards user discourse. The objective of the empirical study is pointing out assistance for the editorial moderation of comments and showing examples of media organisations succeeding in promoting constructive discussions with their target groups. For these purposes, 78 media companies from 46 countries worldwide were surveyed relating to their impressions and practical handling of user comments by means of personal interviews and an online questionnaire.

→ Download as a PDF

THE CORAL PROJECT:
Community Guides for Journalism. Instructions and ideas for better engagement, written by experts (2017)

The Coral Project is a joint initiative of the Mozilla Foundation, the New York Times, and the Washington Post with the intention of improving the quality of online discussions. Within the scope of the project, open source tools have been developed and provided to editorial staffs free of charge for use in managing their communities. The "Guide" section of the Coral Project website contains comprehensive guidelines for newsrooms relating to online interaction with their communities. Its mission is to assist newsrooms when building and managing their communities. It includes approximately 70 different management strategies, cases studies from media professionals worldwide, and additional sources describing how to handle online readership successfully, ranging from target group strategies to the use of analytic tools.

→ Download as a PDF

Automatization tools

CONVERSARIO:

"More time for positive dialogue"

Because it is nearly impossible to manage the comment volume and momentum of online discourse manually, Conversario promises "proactive protection against hateful and spam comments". Conversario is the first German start-up to focus on AI-based comment moderation on social media. The technology firm ferret go GmbH, which is located in Bernau bei Berlin, Germany, is behind Conversario. ferret go primarily deals with natural language processing, machine learning, and automated services. With Conversario, the firm has developed a tool for automated community management which intends to have a positive impact on user dialogue. According to its website, ferret go has collaborated with leading German publishers and media companies, including Focus Online, FAZ.net, n-tv, Berliner Zeitung, and rbb. The implementation of Conversario is also being considered at tagesschau.de to screen for harmless comments on Facebook and on their own website.

→ https://conversar.io/en/index.html

TALK:

"Have better conversations"

The "Talk" software was financed by the non-profit John S. and James L. Knight Foundation and originally developed in cooperation with The Washington Post and The New York Times under the leadership of the Coral Project, which now manages and markets the software. The tool aims to bring newsrooms and their communities closer together, consequently making online discourse more constructive, and in general improve the debate climate in terms of editorial audience engagement. "Many of our most loyal readers are commenters. The combination of Talk and ModBot (moderation software developed by The Post and based on artificial intelligence) will allow us to get to know them better, more easily interact with them, and quickly find and highlight thoughtful and insightful comments for all readers to see," said Emilio Garcia-Ruiz, managing editor at The Post, when the tool was introduced. "This is a first-of-its-kind commenting system that takes a comprehensive approach to comments, giving us the technical capability to connect with commenters in a deeper, more meaningful way at scale."

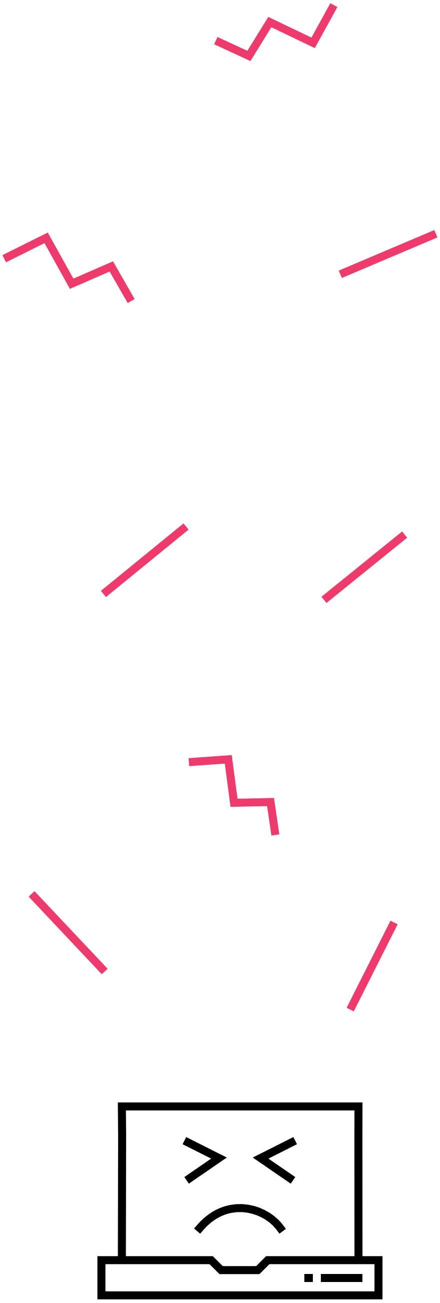
→ https://www.coralproject.net/talk

PERSPECTIVE API:

„What if technology could help improve conversations online?“

No software provider has yet been able to develop an intelligent system to use algorithms to significantly improve the hygiene of online discourse and help prevent discrimination. The Google-created company Jigsaw (formerly Google Ideas from 2010 to 2015) is a technology incubator based in New York City. It has developed an ambitious API (application programming interface) to rate text comments on a toxicity scale spanning from 0 to 100 to make pre-moderation and the assessment of comments easier for newsrooms. The first cooperating partners included reputable media companies such as The New York Times, The Economist, The Guardian, and the Wikimedia Foundation, which runs the online encyclopaedia Wikipedia, which initially made 115,000 discussion posts available for Perspective's first machine learning process. The tools should be able to recognise whether users would rate a comment as inappropriate and furthermore under which adverse conditions those users would leave a discussion platform. Although Perspective API sorts and classifies automatically, newsmakers make individual decisions (as controlled by human moderators) regarding whether the comments which were categorised as toxic should be deleted, persecuted, published, or used for other purposes.

→ https://www.perspectiveapi.com



THE AUTHORS

Photography: Beate C. Koehler

Dr. phil Leif Kramp is a senior researcher and research coordinator at the Centre for Media, Communication and Information Research of the University of Bremen (ZeMKI). Kramp was among the founding directors of the Verein für Medien- und Journalismuskritik e.V. *(Association for Media and Journalistic Criticism)*, is a member of the board of directors of VOCER Innovation Medialab (a funding programme for aspiring journalists), as well as an author and co-editor of numerous textbooks and studies relating to the transformation of journalism. He was a member of the nominating committee for the 2018 Grimme Online Awards, the jury of the NETZWENDE Awards for sustainable innovations in journalism, and the jury of Initiative Nachrichtenaufklärung e.V.

Contact: kramp@uni-bremen.de

Photography: Jörg Möller

Prof. Dr. phil. Stephan Weichert leads the Digital Journalism master's programme, the Urban Storytelling Lab, and the Digital Journalism Fellowship Programme at the Hamburg Media School (HMS). Since 2008 he has been a professor of journalism in Hamburg. Weichert is the founder of the think tank VOCER.org and founding director of the VOCER Innovation Medialab, a scholarship programme for aspiring journalists. As a scientist and journalist, he has dealt with the consequences of digitalization on media, journalism, and society for the last 20 years. Weichert was presented with the 2014 Medienethik-Award for his excellent journalistic efforts on the topic of digital society.

Contact: stephanweichertpost@gmail.com

Leif Kramp and Stephan Weichert have recently published an extensive study collection on the media use of young target groups. Their book "Der Millenial Code. Junge Mediennutzer verstehen – and handeln" was published in German by the VISTAS Verlag in 2017.

GLOSSARY

Deutschlandfunk Kultur
Deutschlandfunk Kultur is one of the three radio programmes of Deutschlandradio broadcast throughout Germany. It provides content for listeners interested in culture.

Mediengruppe RTL Deutschland
The RTL Deutschland media group is the leading German provider of video content and is headquartered in Cologne. Its majority shareholder is the Luxembourg-based RTL Group, which is one of the world's market leaders in the programming, content, and digital businesses.

RP ONLINE
RP ONLINE is an far-reaching German news site supplementing the print version of the Rheinische Post newspaper with up-to-date reports. The Rheinische Post is a daily regional newspaper located in Düsseldorf.

SPIEGEL ONLINE
SPIEGEL ONLINE is an Internet service of the news magazine Der Spiegel with its own independent newsroom. It is one of the most widespread German-language news websites.

tagesschau.de
tagesschau.de is the central online news site of the ARD. The ARD was founded in 1950 as a network of the public service broadcasters in Germany. It is financed by means of television licensing. The network comprises nine regional broadcasters which provide and air joint television programmes (e.g. Das Erste), as well as their own regional television and radio programming.

IMPRINT

Publisher:
Media Authority of North Rhine-Westphalia
Zollhof 2
40221 Düsseldorf / Germany
T +49 211 77007-0
F +49 211 727170
info@medienanstalt-nrw.de
www.medienanstalt-nrw.de

26.06.2018

Project management:
Sabrina Nennstiel (Head of Communications)

Editing:
Dr. Meike Isenberg (research),
Marie-Franca Hesse (communication)

Translation:
Nicole Groß Translations, Neuss

Design:
ressourcenmangel an der Panke GmbH, Berlin

Printing:
Börje Halm, Wuppertal

With support from Google Germany GmbH