MODERATION IN THE "LESS HATE, MORE SPEECH" PROJECT: HOW IT WORKS

- SUMMARY -

Raluca Toma, Marina Popescu
This report explains how comment moderation in the “Less Hate, More Speech” project works: what experiments we tried and what drove and informed the design of our comment system, what our main principles and goals are, what key concepts inform our moderation procedures and how, exactly, moderators decide what can and cannot be published.

The implementation of moderation was a test of whether institutionalized rules can be effective if implemented and explained over a long period of time. The research also involved various moderation experiments, as well as engaging with users directly to find out more about what and how they think.

Core website GSP.ro entered moderation as part of the “Less Hate, More Speech” project in April 2015, followed by the second core website, Tolo.ro, in September. “Less Hate” moderation of Blogsport.ro and Paginademedia.ro began in February 2016. The supervised moderation phase, overseen by the research team, ended in June 2016.

We started from the premise that comments and reader-engagement are valuable for publishers, journalists and to readers themselves, and that it is desirable from both a practical and a normative point of view to encourage dialogue, while taking responsibility for the content that is hosted and the environment that is fostered in the comment section.

MODERATION GOALS:

- Allow commenters to share opinions and information, debate passionately, be ironic or critical, have a laugh, if they do so in a civil manner. Aim to show users that we do not seek to suppress passion or fun or any specific opinions.
- Encourage more people to post more comments, by making the comment section a safer space, more free of aggression, hate, trolling, spamming and other anti-social behaviours.
- Encourage higher quality user contributions, by opening up the space to the more civil and thoughtful commenters to express themselves.
- Make commenters reconsider before impulsively posting an aggressive, hostile or vulgar message.
- Help users learn what is not acceptable on the moderated websites and how to express themselves without falling astray of the commenting rules, by giving them clues as to where they erred and in what way.
- Encourage the acceptance and internalization of the commenting rules among the users, so that they may effectively become community-accepted norms and that users can eventually police themselves and each other.

WHAT WE MODERATE, IN BRIEF:

- Hate speech
- Incitement to violence
- Incitement to discrimination or exclusion
- “Inferiority” (comments painting certain groups as inferior, deviant, sick, etc)
- Slurs and insults that exhibit intolerance
- Dehumanization
- Highly vulgar language
- Threats and harassment
- Highly specific accusations that are not backed up by evidence available from the commenter or the public record
WHAT WE DO NOT MODERATE:

- Criticism of the publication, authors, the moderation
- Harsh criticisms and even insults that are not libellous, intolerant or dehumanizing
- Off-topic comments
- Views that run contrary to those of the publication or the authors but are voiced in a civil manner

In short, the implemented moderation does not promote a certain ideology or agenda but, rather, simply seeks to instil civility and respect for the equal rights and dignity of all.

The procedures developed collaboratively with the journalists from Gazeta Sporturilor were informed by the literature on intolerance, previous work on the content analysis of uncivil and intolerant discourse (including online comments), relevant legislation and court decisions and socio-psychological and behavioural insights.

KEY „LESS HATE, MORE SPEECH” MODERATION PRINCIPLES:

- Don’t take the fun out of the game, as long as things stay reasonably civil. Criticism, playfulness, irony are not discouraged, as long as comments don’t cross the line.
- Moderate the comment, not the commenter. Comments are to be judged on a case-by-case basis and one’s opinion of the user should be suspended.
- Moderate your feelings. Moderators strive to limit the impact of their own predispositions (thoughts and feelings) about a person or issue.
- Strive not to change the meaning of the comment. While it is part of the logic of moderation that a message may sometimes become confusing or harder to understand after asterisks are applied, moderators should be careful not to change a negative comment into a positive one, or the other way around.
- Moderate speech, not thought. Examine what is written, not what could have been written or what the commenter may think. When facing “borderline” comments, where one suspects a commenter is racist, xenophobic or sexist because of what they say or imply, if their speech itself does not break the rules, one cannot moderate their statements.
- Dislike ≠ intolerance. Simply giving voice to dislike or even resentment of a person or group should not draw moderation, as long as the message is not demeaning or discriminatory.
- Context matters. Know the lingo of the comment section, know the news context and mobilize that knowledge to judge the appropriateness of a message.

Our moderation methods may not be the only or the best solution for all platforms, and checking all comments may not be feasible for all publishers.

Still, many websites now find it necessary to moderate at least some comments, and here our experience offers some useful insights. While every website’s environment and needs are slightly different and moderation policies may vary, we believe moderation has greater chances of success if it is: respectful of the commenters and transparent; systematic and driven by detailed procedures; and in touch with the characteristics of the online audience it is addressed to.

DETAILS ABOUT THIS RESEARCH REPORT:

Report developed as part of the project “Less Hate, More Speech: An Experimental and Comparative Study in the Media and Political Elites’ Ability to Nurture Civil, Tolerant, Pro-Democratic Citizens,” financed through the EEA Financial Mechanism 2009-2014, EEA Research Programme. For more details, access lesshate.openpolitics.ro.