Harmless Cursing or Hateful Comments: International perspectives on hate speech

Authors

Ina Weber, Graduate Research Assistant, Erasmus University Rotterdam

Aquina Laban, Student Intern, Erasmus University Rotterdam

Gina M. Masullo, Associate Director, Center for Media Engagement

João Gonçalves, Lecturer, Erasmus University Rotterdam

Marisa Torres da Silva, Assistant Professor, NOVA University Lisbon

Joep Hofhuis, Assistant Professor, Erasmus University Rotterdam

Summary

With this study, we want to clarify what is perceived as hate speech in different countries in order to find out how the moderation of harmful online content can be improved. Our findings give some guidance to platforms for how to make their moderation guidelines as effective as possible and resolve some of the confusion of their users about why certain comments are removed while others are allowed.

For this purpose, we tested ten different social media posts in terms of their perceived hatefulness and profaneness. We found that there is no universal understanding of hate speech as perceptions clearly differ between the US and Europe. But also, within Europe, perspectives on what is hate speech do not necessarily overlap.

The Problem

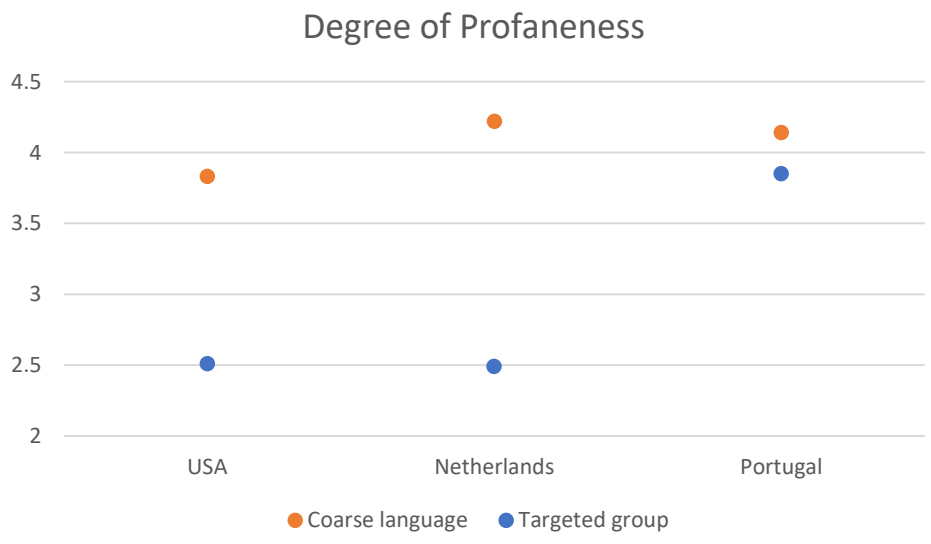
Social media platforms and news outlets want online comment sections to be productive spaces for discussions, so they use content moderation to remove hateful speech. However, users are often confused and angry when their content is removed. As much as social media platforms try to limit the spread of hate speech, protecting users from discrimination and offense while not limiting them in their freedom of expression is a challenge that is all the more complex because of the lack of an universal definition of what hate speech is. European legislation understands hate speech as expressions which denigrate, harass, promote negative stereotypes or generally incite

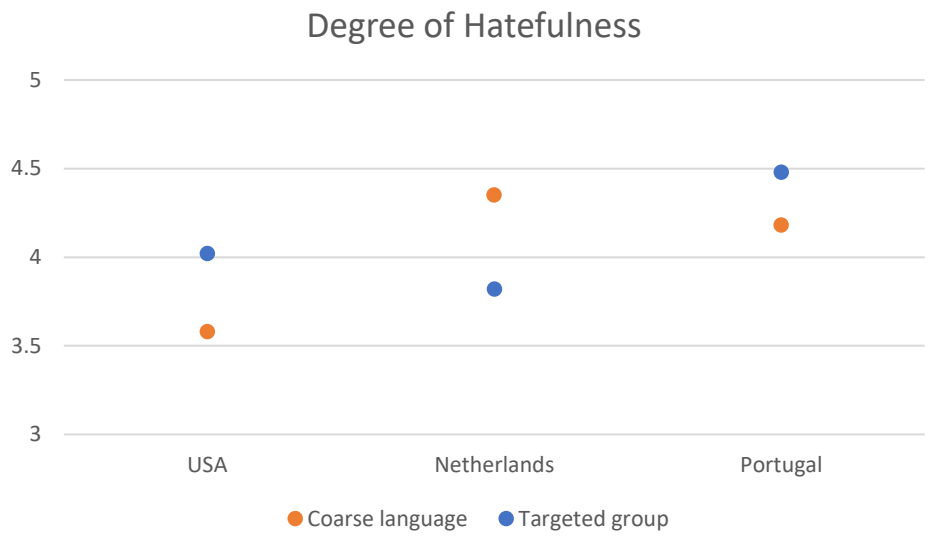
hatred of a group of people defined by personal characteristics such as, among others, their race, ethnicity, gender, sexuality, religious beliefs, national origin or age. However, this approach is not adopted worldwide, including in the United States.

Erasmus University Rotterdam teamed up with researchers from the Center for Media Engagement at the University of Austin at Texas and NOVA University Lisbon to figure out how people from these three countries understand hateful speech. We conducted an experiment where we showed internet users from the three countries a set of social media posts, and they rated the hatefulness and degree of profanity of each post. Facebook funded the project, although we worked on it independently.

Key findings

- Americans had different perceptions of profanity and hate speech than Europeans. Americans saw coarse language or swear words as substantially less hateful than derogatory comments about immigrants or comments that incited violence
- For participants in both European countries, this distinction was less clear. They perceived posts with profanity as both profane and hateful.
- In Portugal the line between profanity and hateful speech was particularly blurry.
- In all three countries, people had no trouble identifying profanity and they perceived attributes of profanity consistently.
- Overall, Americans had a clearer perception of what they considered hate speech than Europeans. This is notable because the US does not have a legal definition of hate speech, while Europeans countries do.[i]





Implications

Having a better understanding of their users' perspective on hate speech and profanity is useful for social media platforms and news outlets to optimize their moderation guidelines and practices. By being able to clearly distinguish between hate speech and profanity, posts which are perceived as harmful by most users can be removed more effectively while not limiting users in their freedom of expression. Knowing how these perceptions differ across countries is particularly relevant for international platforms with users from all around the world.

Our findings offer global guidance for social media platforms and news outlets regarding how to effectively create moderation guidelines to limit confusion among users about why certain posts and comments are removed while others are allowed.

- Content moderation guidelines should be tailored to the culture of specific countries. For example, platforms in Portugal and the Netherlands should highlight their definitions of hate speech more prominently because that distinction is not clear for users in those countries
- Users should be informed about the definitions of profanity or hateful speech when they agree to use the platform, so they are clear about what is permitted.
- Users should be told what was wrong with the content if it is removed, so they will learn how profanity or hateful speech is defined on that platform.

The experiment

A total of 304 participants[ii] from the three countries were randomly assigned to view either five posts that contained derogatory barbs against immigrants or incitements of violence or five

posts that contained swear words, name-calling, or words in all capital letters to indicate shouting that were not targeted at a protected group defined by specific characteristics[iii] All posts were created to appear as if they were posted to a social media platform like Facebook and translated into Dutch or Portuguese for those participants. To make the posts as believable as possible, we referred to minorities which are particularly vulnerable and likely to be stigmatised with a stereotypically negative view. The word "Mexicans" was replaced with "refugees" in the Dutch experiment and "Brazilian migrants" in the Portuguese experiment.[iv]

Examples of the comments in the United States

Comment targeted against a specific vulnerable group	Comments with swear words, insults, and words in all capital letters
"Mexicans come from an uncivilized, backward society. They are filthy criminals, molesting innocent American women and menacing entire neighborhoods. For the sake of our safety, they should all be beaten up and rot in jail forever. We need to protect ourselves."	*"I can't believe how our stupid politicians do nothing to improve the situation in our country. Our welfare system is a fucking joke, our society is divided, integration is a huge fail… so many issues but they're not making the SLIGHTEST F#CKING EFFORT to find solutions. These damn idiotic office sitters are giving zero fucks about us!! All they do is lame talking but this requires some ACTION, Jesus Christ is that so difficult??!?!?!"*

Participants rated each of the comments they viewed on how profane or how hateful they perceived it on a 1 to 5 scale with a higher number meaning the comment is seen as more hateful or more profane.[v] Analysis of the data revealed the following:

What we found

- In all countries, posts containing profanity were seen as similarly profane. However, only participants from the US made a clear distinction between expressions considered profane and expressions considered hateful. This suggests Americans think of hate speech and profanity as two distinctly separate concepts.[vi]
- In both European countries, perceptions of profanity and hatefulness overlapped. For Dutch participants, comments containing insults, all caps, and swear words were perceived as profane but also as distinctly more hateful than comments that attacked specific vulnerable groups.[vii]

- Portuguese participants considered comments that targeted specific vulnerable groups and those that contained profanity equally hateful and profane.[viii] Overall, the Portuguese seemed to make the least distinctions between hate speech and profanity. They perceived profane language as more hateful than American participants and expressions targeting specific vulnerable groups as more hateful than Dutch participants.[ix]
- Internet users in different European countries appear to have a rather diffused understanding of hate speech. The lack of a distinction between profanity and hate speech in Portugal and the partial overlap of the two concepts in the Netherlands show us that we cannot generalize opinions on hate speech across European countries. Thus, what makes an expression profane or hateful is dependent on national and cultural contexts.

Methodology

The experiment was conducted on March 5 and 6, 2020, using Qualtrics. Our sample was recruited through Dynata, an online survey company that provided participant samples matching the demographics of the population of each country in regard to age, gender and distribution of population over regions. A total of 149 people from the U.S., 93 from the Netherlands, and 62 from Portugal participated.[x] Participants accessed the survey experiment through an online link and completed it on their own computers.

Participant Demographics

	US	Netherlands	Portugal
	n = 149	*n* = 93	*n* = 62
Gender			
Male	47.0 %	36.6 %	40.3 %
Female	52.3	62.4	59.7
Other	0.7	1.1	

Age			
18-29	11.5	17.4	21.3
30-49	49.3	39.1	54.1
50-64	34.5	35.9	23.0
65+	4.7	7.6	1.6
Education			
High School or less	32.2	46.3	40.4
Bachelor's	40.3	39.8	29.0
Master's	20.8	7.5	27.4
Doctorate	6.7	6.5	3.2

[i] ECRI. (2016). ECRI General Policy Recommendation 15 on combating hate speech. Strasbourg: Council of Europe.

[ii] Erasmus University in the Netherlands granted Ethics Review Board approval for the project on February 3, 2020. Institutional Review Board approval for the project was granted by The University of Texas at Austin on February 26, 2020. We accounted for ethical concerns and informed participants about the nature of the comments they were about to see in the introduction of the experiment.

[iii] We based these categories on definitions of hateful speech from ECRI. (2016). ECRI General Policy Recommendation 15 on combating hate speech. Strasbourg: Council of Europe; Article 19. (2015). *Hate speech explained. A toolkit:* London; Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech.* Paris: UNESCO. Our profanity condition was based on the conceptualization of incivility from Chen, G.M. (2017) Online incivility and public debate: Nasty talk. Palgrave Macmillan and Muddiman, A. (2017). Personal and Public Levels of Political Incivility. *International Journal of Communication*, *11*, 3182–3202.

[iv] For Portugal, we chose Brazilian migrants because Brazilians have a long history of immigrating to Portugal, and they constituted the largest group of immigrants in Portugal in 2019 (Silva, M.T. (2019). *Fact sheet*

Portugal. Country report on media and migration. New Neighbours. Retrieved from: https://newneighbours.eu/research/). For the Netherlands, we chose refugees because this minority group has affected public opinion strongly even though numbers of incoming refugees have declined since a peak in 2015 (d'Haenens, L., Joris, W., & Heinderyckx F. (Eds.). (2019). *Images of immigrants and refugees in Western Europe. Media representations, public opinion, and refugees' experiences.* Leuven: University Press.). The U.S. has had long-time tension over immigration from Mexico that heightened during the presidency of Donald J. Trump.

[v] For each comment, we asked the participants the following questions: "What do you think is the author's attitude towards immigration?", "To what extent would you say does this post contain profanity?", "To what extent would you say does this post contain hate speech?" and "How realistic is it that someone would post a text like this one on a social media platform?"

[vi] An independent samples *t* test for our U.S. sample shows that participants perceived comments that attacked specific vulnerable groups or incited violence as significantly more hateful [$t(147) = -2.33$, $p = .021$, $M = 4.02$] than comments containing swears, name-calling, or words in all capital letters ($M = 3.58$).

[vii] Independent samples *t* tests for the Dutch sample indicate that the two types of comments were seen as significantly different, regarding both how profane [$t(86.47) = 7.92$, $p < .001$] and how hateful [$t(76) = 2.55$, $p = .013$] they were. Comments with insults, swear words, and words in all capital letters were seen as more profane ($M = 4.22$) and hateful ($M = 4.35$) than comments that attacked specific vulnerable groups or incited violence for profaneness ($M = 2.49$) and for hatefulness ($M = 3.82$), respectively.

[viii] Independent samples *t* tests for the Portuguese sample show no significant differences between the two types of comments in terms of profaneness [$t(60) = 1.15$, $p = .256$] or hatefulness [$t(60) = -1.51$, $p = .136$].

[ix] Results of an ANOVA for profaneness of comments that attacked a specific vulnerable group or incited violence [$F(2, 158) = 14.33$, $p < .001$, partial $\eta^2 = .15$] showed that participants in the three countries had different perceptions of how profane this speech was. Tukey post hoc tests showed that participants in Portugal rated these comments as significantly more profane ($M = 3.85$) than participants in the Netherlands ($M = 2.49$, $p < .001$) or than participants in the U.S. ($M = 2.51$, $p < .001$). Means for the Netherlands and the U.S. were not significantly different ($p = .995$). Significant differences were also found for ratings of hatefulness of comments that attacked a specific vulnerable group or incited violence, [$F(2, 158) = 3.51$, $p = .032$, partial $\eta^2 = .04$]. Tukey post hoc tests revealed that participants in Portugal rated these comments as significantly more hateful ($M = 4.48$) than those in the Netherlands ($M = 3.82$, $p = .026$) but not when compared to U.S. participants ($M = 4.02$, $p = .122$). Ratings from participants in the Netherlands were not significantly different from those in the U.S. ($p = .599$).

[x] Initially, there were 334 participants in total, but data from 30 participants had to be removed because they completed the survey too quickly to be reliable.