This project is co-funded by
the European Union

Hate speech, gender, social networks and political parties

Map of the main political parties using hate speech against the "ideology of gender" through the social networks and internet in Europe.

Antigona Research Group – Autonomous University of Barcelona (UAB)
Center for Policy Studies (CPS) – Central European University (CEU)
Centro Interdisciplinare di Ricerca sulla Vittimologia e sulla Sicurezza (CIRViS) – University of Bologna (UNIBO)
Department of Law - University of Gothenburg (UGOT)
Department of Communication - University of Vienna (UNIVIE)

2021

This report is the result of the European project titled **"Hate speech, gender, social networks and political parties" (GENHA)** (Reference: 875388)
www.genha.eu

*This report was funded by the **European Union's Rights, Equality and Citizenship Programme** (2014-2020)*

The content of this report represents the views of the author only and is his/her sole responsibility. The European Commission does not accept any responsibility for use that may be made of the information it contains.

Consortium partners and researchers:

Antigona Research Group – Autonomous University of Barcelona (UAB)

Noelia Igareda González
Adrián Alberto Pascale

Center for Policy Studies (CPS) – Central European University (CEU)

Anna Fejős
Violetta Zentai
İlkin Cankurt – social media advisor

Centro Interdisciplinare di Ricerca sulla Vittimologia e sulla Sicurezza (CIRViS) – University of Bologna (UNIBO)

Raffaella Sette
Sandra Sicurella

Department of Communication - University of Vienna (UNIVIE)

Andreas Schulz
Claudia Wilhelm

Department of Law - University of Gothenburg (UGOT)

Moa Bladini
Astrid Helmstad
Eva-Maria Svensson

Index

Chapter 1

Users’ Responses to Anti-Gender, Homophobic and Sexist Facebook Posts in Hungary and Germany.

Center for Policy Studies (CPS) – Central European University (CEU)
Department of Communication - University of Vienna (UNIVIE)

2021

1 Introduction

Right-wing parties, politicians, right-wing networks and initiatives successfully spread their populist messages on social networks (Blassnig & Wirz, 2019). As shown in the content analysis (WP 3.1), right-wing actors work to shift the discourse towards an anti-libertarian, anti-gender, and anti-LGBTQI+ attitudes and beliefs. The results of the descriptive content analysis revealed the quantitative reach of the agitative, defamatory and rumor-mongering posts and the agendas they set. The qualitative analysis showed exemplary user reactions to such posts. For the case of Germany, we identified that anti-genderism and especially ‘early sexualization’ elicit stronger affective user reactions in the comments than for example homophobic posts. In a next step, it is necessary to take a closer look at the recipient's perspective and to understand which topics and types of hate speech lead to higher levels of user engagement (i.e., liking, commenting, sharing, or flagging of posts), which other attributes of the posts influence users’ responses and how users perceive this content. Following the content analysis, we conducted a comparative online survey in Germany and Hungary to analyze user engagement with anti-gender hate speech in social media.

Based on the content analysis of posts and user comments on the issues of sexism, homophobia and anti-genderism, we are interested (1) which content and individual characteristics influence users' perceptions and behavior when they are confronted with hate postings. We are also interested in the intentions behind social media interactions: (2) What are the personal motivations for sharing, commenting or flagging postings with gendered hate speech content.

We expect results to provide us with a more nuanced view on how gendered hate speech on social media is perceived and which gender-related topics are more engaging than others. Especially through the difference in support for right-wing populist policies and attitudes towards minorities we expect differences in the acceptance of hate speech but also with respect to relevance of the three main topics of gendered hate speech (Anti-Gender, homophobia, sexism) between the two countries (Takács & Szalma, 2020).

2. Study

To investigate which content and source characteristics influence the acceptance of hate speech, which topics are more likely to be flagged, commented on, shared or liked by social-media-users than others, we conducted a choice-based conjoint analysis. A choice-based conjoint design allows us to analyze how important specific content characteristics are to users when they judge hate speech content and respond to it (cf. Wilhelm, Joeckel & Ziegler, 2020).

We were particularly interested in the following aspects: Does it matter if a politician or an ordinary citizen spreads content? Is there a difference between a man and a woman spreading hate speech on social media? Does the type of hate speech affect users’ responses and if so, how do social media users react to agitation, defamation and rumors? And is it relevant whether the post already has many likes, comments and shares to respond to or not?

2.1 Participants

German (*n* = 515) and Hungarian (*n* = 740) participants were recruited by an online access panel provider. Due to the research interest, participants were asked in the beginning how often they use social media for any purpose. The majority of 74 percent (Germany) and 90 percent (Hungary) of the participants indicated to use social media at least daily. Both samples slightly deviated from the respective country populations: More female participants (GER: 60.6%; HUN: 54.2%) and more persons with a university education (GER 30.1%, HUN 35%) took part. The voting behavior, which we determined with an opinion poll, corresponded to the official opinion polls in both countries at the time of data collection in May and June 2021 (cf. statista, 2021; PolitPro, 2021).

2.2 Procedure of the study

The study participants were asked to respond to four pairs of Facebook posts. For each pair, they were asked which one they would ‘like’, comment, share, or flag as inappropriate, having the option to choose either one, both or none of the presented posts. If they chose one of the first three options, the participants were asked follow-up questions about their motivations to comment, share or flag the respective posting.[1] The postings were manipulated with respect to content and source characteristics. They varied by (1) gender of the source (female/male), (2) professional status of the source (politician/ordinary citizen), (3) topic of the post (anti-gender, sexism, homophobia), (4) type of hate speech (defamation, agitation, rumor), and the level of social media engagement indicated by web metrics (low vs. high number of shares, likes, comments). The specific content of the edited posts varied between the countries and represent the specific gender discourses in the countries that have been identified in the content analysis of WP 3.1.

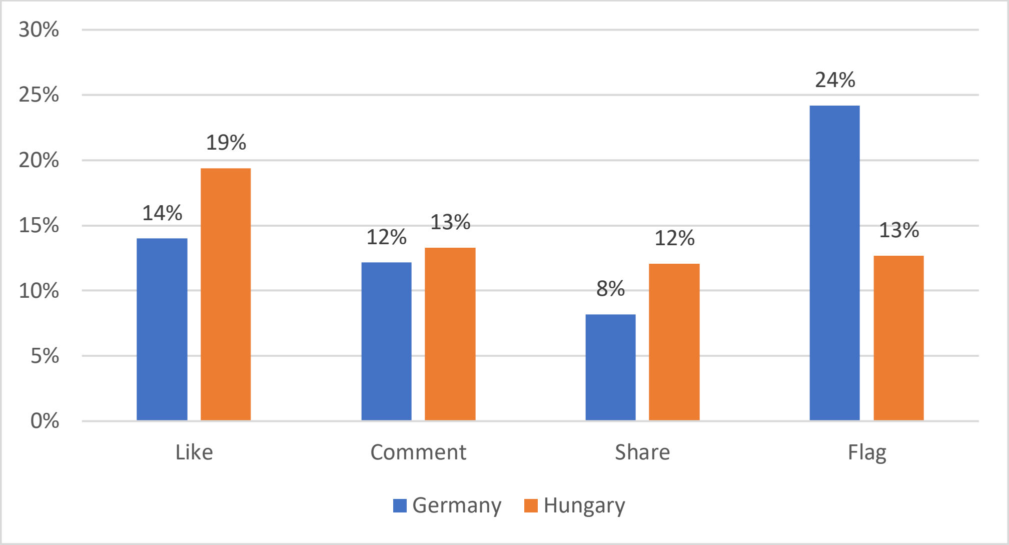
3. Results

3.1 Descriptive Analysis

In the German sample, participants indicated to like only 14 percent of the presented posts, compared to 19 percent in the Hungarian sample. In both countries, a comparable amount of comments has been made, so that the German participants commented 12 percent and the Hungarians 13 percent of all postings. We observed greater differences when it comes to sharing a post: Participants in the Hungarian sample shared 12 percent of all postings, whereas in the German sample it was only eight percent. The biggest difference was found in flagging behavior as the German participants reported almost twice as many posts (24%) as Hungarian participants (figure 1).

Figure 1 Average engagement per participant (in percentages)

[1] The motivation for 'Likes' was not queried as it is a very low-threshold activity compared to the other types of engagement.



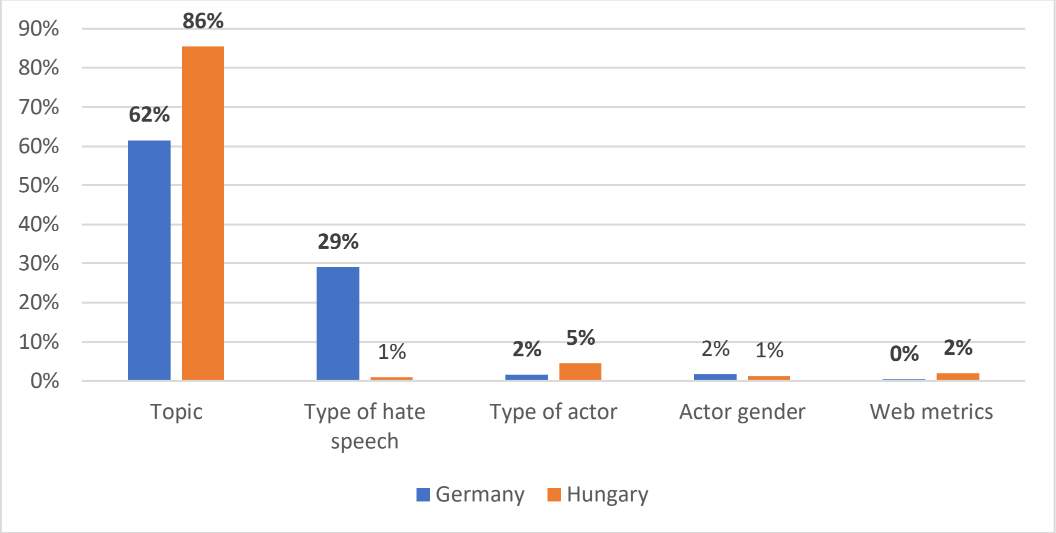
Note. Data includes all cases (GER: *n* = 515; HUN: *n* = 740).

With a focus on those participants who actively responded to the postings, there are even larger differences between the countries: 48 percent of Germans liked at least one post, compared to 60 percent in the Hungarian sample. In addition, German participants shared significantly less content (GER: 30%, HUN: 41%), but there were only slight differences in commenting behavior (GER: 40%; HUN: 43%). The majority of German participants flagged at least one of the posts they saw (73%). In contrast, among the Hungarian participants, only 44 percent indicated to flag a post as inappropriate.

3.2 Source and Content Characteristics that Influence User Engagement

Liking Posts. As shown in figure 2, in both countries the topic of the posting has the strongest influence on whether a post was liked (GER: 62%; HUN: 86%). German participants preferred to like anti-gender posts over sexist posts and showed less willingness to like homophobic posts. Hungarian participants preferred liking homophobic posts followed by anti-gender posts but were rather dismissive of liking sexist posts. For the German respondents, the type of hate speech also played a role in 29 percent of whether the post was liked, i.e., they were more likely to like agitating content than posts that include defamations or rumors. For Hungarians, the type of actor and the level of web metrics of the post also played a role. They preferred liking posts by politicians with a high number of likes, shares and comments over posts by ordinary citizens with a low number of likes, shares and comments.

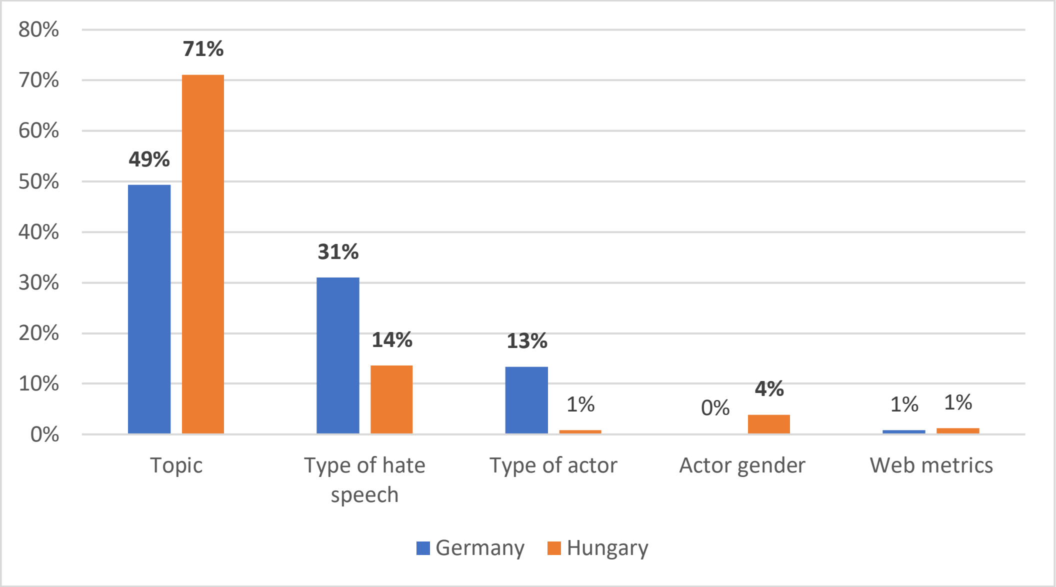
Figure 2 Determinants of liking gendered hate speech



Note. Percentages indicate the relative importance of content and source characteristics. Numbers in bold indicate significant effects, *p* < .05.

Commenting Posts. The topic also determined to a large extent whether a post was commented on, indicated by a relative importance of 71 percent for Hungarians and 49 percent for Germans (figure 3). In the German sample, anti-gender and homophobia topics were both preferred topics for commenting. In the Hungarian sample, homophobic content clearly increased commenting activity. Hungarians were more likely to comment on rumors, Germans were more likely to comment on defamatory content. For the German respondents, the type of communicator was more relevant than for the Hungarians: Posts of politicians were rather commented.

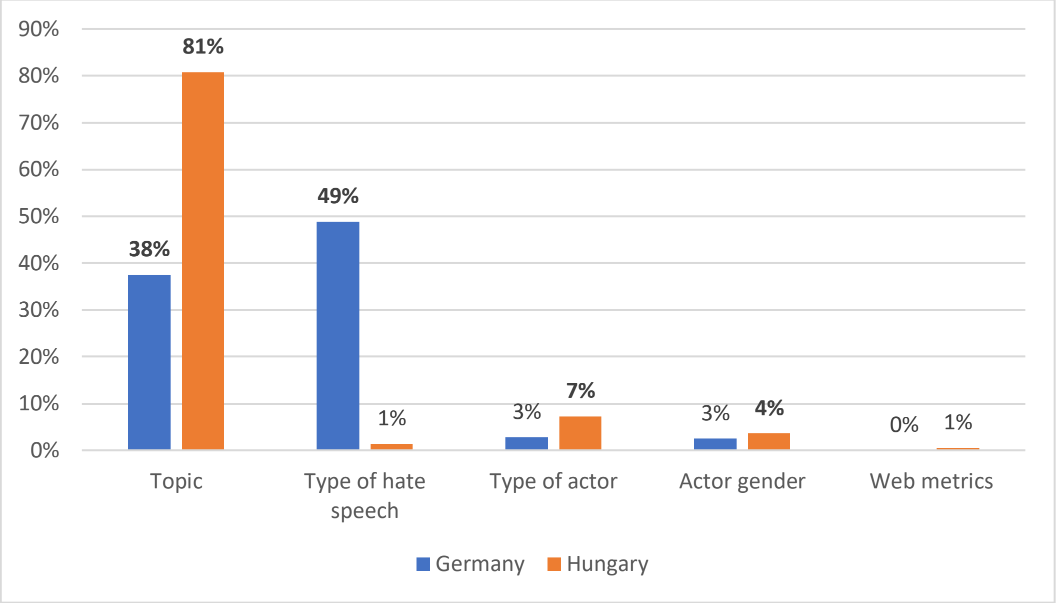
Figure 3 Determinants of commenting gendered hate speech



Note. Percentages indicate the relative importance of content and source characteristics. Numbers in bold indicate significant effects, $p < .05$.

Sharing Posts. For Germans, the type of hate speech was most important content factor for sharing a post (49%), i.e., they preferred to share defamatory content over rumors and agitative posts (figure 4). The topic was also relevant (38%) as anti-gender content increased intentions to share a post and homophobic content had a negative effect. In the Hungarian sample, 81 percent of whether a post was shared depended on the topic. Hence, Hungarian participants were more likely to share anti-gender and homophobic posts than sexist posts. The type of hate speech had no effect on Hungarian participants' intentions to share a post. In contrast, their sharing of gendered hate speech postings was influenced by the type of actor and the gender: Posts of male politicians were more likely to be shared than from ordinary citizens or female persons.

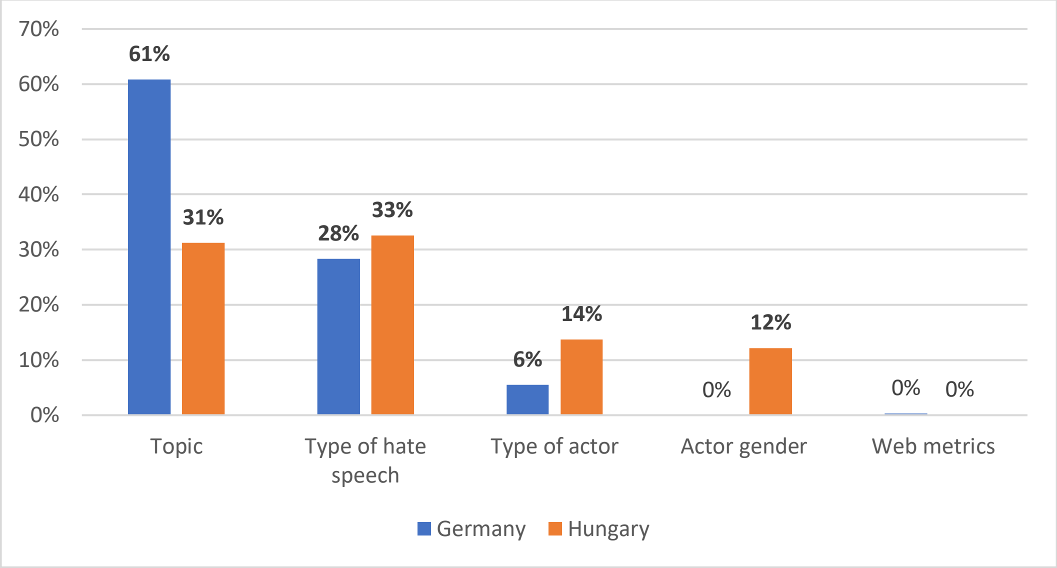
Figure 4 Determinants of sharing gendered hate speech



Note. Percentages indicate the relative importance of content and source characteristics. Numbers in bold indicate significant effects, *p* < .05.

Flagging Posts. For German participants, the topic of the posting (61%), the type of hate speech (28%) and the type of actor (6%) were significantly decisive when asked if they would flag a posting as inappropriate (figure 5). They were more likely to flag homophobic and anti-gender posts than sexist posts as well as posts by politicians than posts by ordinary citizens. In addition, defamatory content was more likely to be flagged than rumors and agitations. For Hungarian participants, the topic (31%) and the type of hate speech (33%) were almost equally important, showing similar preferences as the German participants (anti-gender and homophobic posts defamatory content were more likely to be flagged than sexist posts, rumors and agitations). The Hungarian participants showed a higher intention to flag posts by male politicians than posts by females and ordinary citizens (see figure 5).

Figure 5 Determinants of flagging gendered hate speech



Note. Percentages indicate the relative importance of content and source characteristics. Numbers in bold indicate significant effects, *p* < .05.

3.3 Differences in Participants’ Engagement with Hate Speech Posts by Social Media Usage, Political Orientation, Voting Intention, and Sociodemographics

Social media usage. The German data show that a higher level of social media use (daily or several times a day) increased the likeliness to flag posts by politicians. Also, German participants who indicated higher levels of social media use, were more likely to flag any of the presented posts than participants who reported lower levels of social media use. None of these effects emerged in the Hungarian sample.

Political Orientation. The more right-wing, the more German participants liked any of the posts as well as when the source of the posts was male. Right-wing orientation also generally increased intentions of German participants to share and comment a post, but decreased intentions to flag a post. The more right-wing, the more Hungarian participants liked homophobic posts, the more they generally liked, shared and commented posts and the less they flagged homophobic and defamatory posts.

Voting intention. Hungarian participants who indicated voting intentions for right-wing parties, were more likely to like, share and comment and less likely to flag any of the presented posts than participants who reported voting intentions for other parties. In the German sample, participants who indicated voting intentions for the AfD, were more likely to like and share homophobic posts and more likely to comment on anti-gender posts than participants who reported voting intentions for other parties.

Education. The Hungarian participants with university degree were less likely to like, share and comment but also less likely to flag any of the presented posts than participants on lower educational level. Lower educated Hungarian participants were more likely to like, share and comment posts with a high number of likes, shares and comments than higher educated Hungarian participants. In the German sample, education generally increased flagging intentions and had a negative effect on liking defamatory content.

Age. In the Hungarian sample, participants aged between 30 and 39 years were generally less likely to like any of the posts than other age groups. Also, younger participants (< 40 years) were less likely to share any of the posts than older participants. Hungarian participants aged 40 to 49 years were less likely to comment on any of the posts than other age groups. In the German sample, participants aged 30 to 39 years were less likely to share and comment any of the posts but more likely to flag homophobic posts than other age groups.

Gender. Regarding the gender differences, data from both countries shows that males show a stronger engagement with the posts than females. Analysis reveals that male participants in the Hungarian sample were more likely to like and share and but also more likely to flag any of the presented posts than female participants. German male participants were more likely to share and comment and less likely to flag any of the presented posts than female participants.

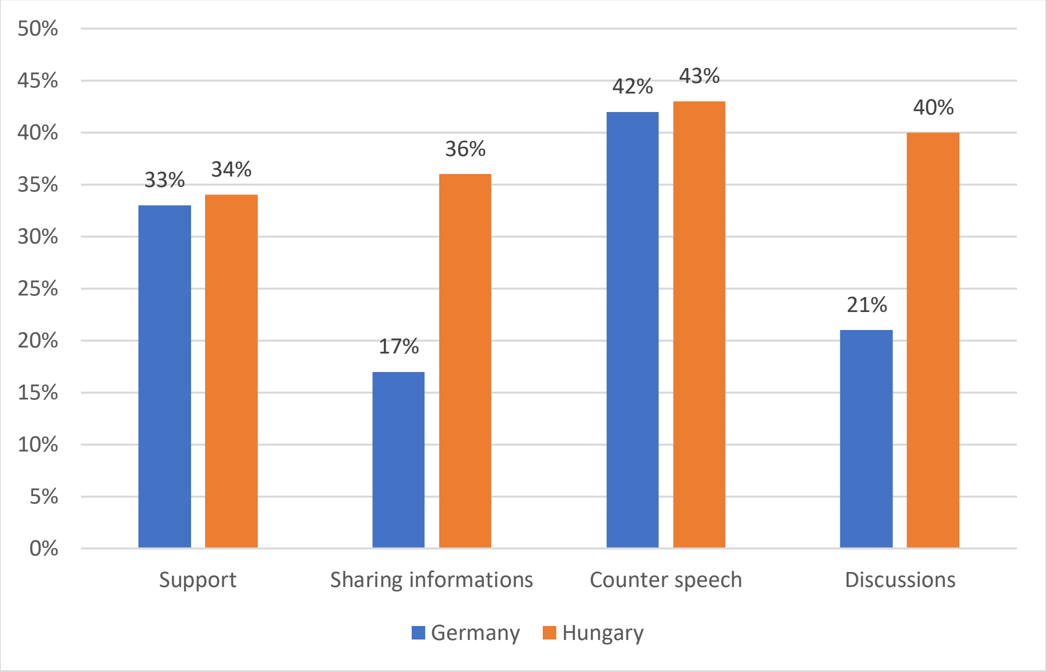
Religion. The more religious, the more Hungarian participants were likely to like, comment and share any of the presented posts. In the German sample, religiousness increased intentions to like posts composed by males and general intentions to comment a post. The more religious, the less Germany participants were likely to flag highly engaging posts (high number of likes, shares and comments). The opposite was the case in the Hungarian sample, religiousness slightly increased the likeliness of flagging highly engaging posts.

3.4 Motives for Engagement with the Postings

Commenting. Almost a third the German participants agreed or strongly agreed with the statement that they fully approve the content of the posting and indicated that it was important to support the opinion of the person who posted the content (figure 6). On the other hand, 47 percent tended or strongly disagree with the statement. Only 17 percent agreed with the statement that they had specific knowledge about the posting topic and would like to share it with other users. In contrast, 43 percent of the respondents rejected the statement. 42 percent have the opinion, that it is important to counter such postings.

Similar to the German sample, one third intended to write comment in support of the content and the posting person. In comparison to Germany, more than a double agree on the statement, that they have a specific knowledge about the topic in the posting and like it to share with other social-media-users. The agreement for the statement, that it is important to counter such postings is quite similar in both countries. Hungarian participants showed stronger motivations for discussing with other users about the posting issue (40%) than German participants (21%).

Figure 6 Motives for commenting a posting (in percentages)

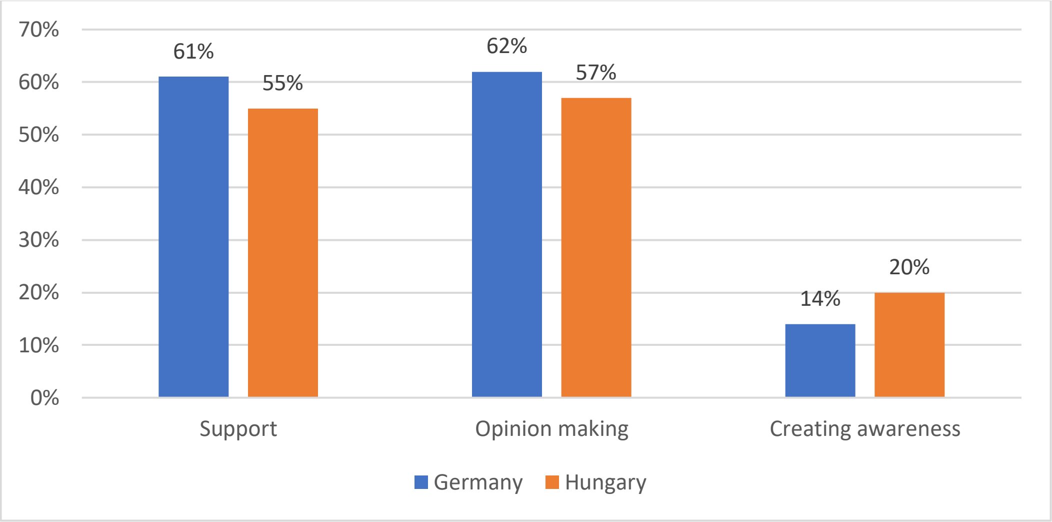


Note. Data includes all cases, who decided to comment a posting (GER: *n* = 208, 40,4%; HUN: *n* = 315; 42.6%).

Sharing. 61 percent in the German sample, who chose to share at least one of the presented postings tend to agree or strongly agree that the posting person is telling the truth, and they like to support it, while sharing the post (figure 7). 62 percent agreed on the statement that the posting reflects their own opinion on the topic and that it is important that other people read it too. Only 14 percent would share the posting for the reason that it does not reflect their opinion on the subject and they would like to make their community aware of it.

40 percent of the Hungarian sample indicated to share at least one posting. That are almost 10 percent more than we can observe for the German data. Fewer respondents than in the German sample agreed on the statements, that the posting person says the truth (55%) and the posting reflects the personal opinion (57%). For 20 percent of the Hungarians it is important to share the posting with their own community to create awareness, because it is not reflecting their personal opinion.

Figure 7 Motives for sharing a posting (in percentages)

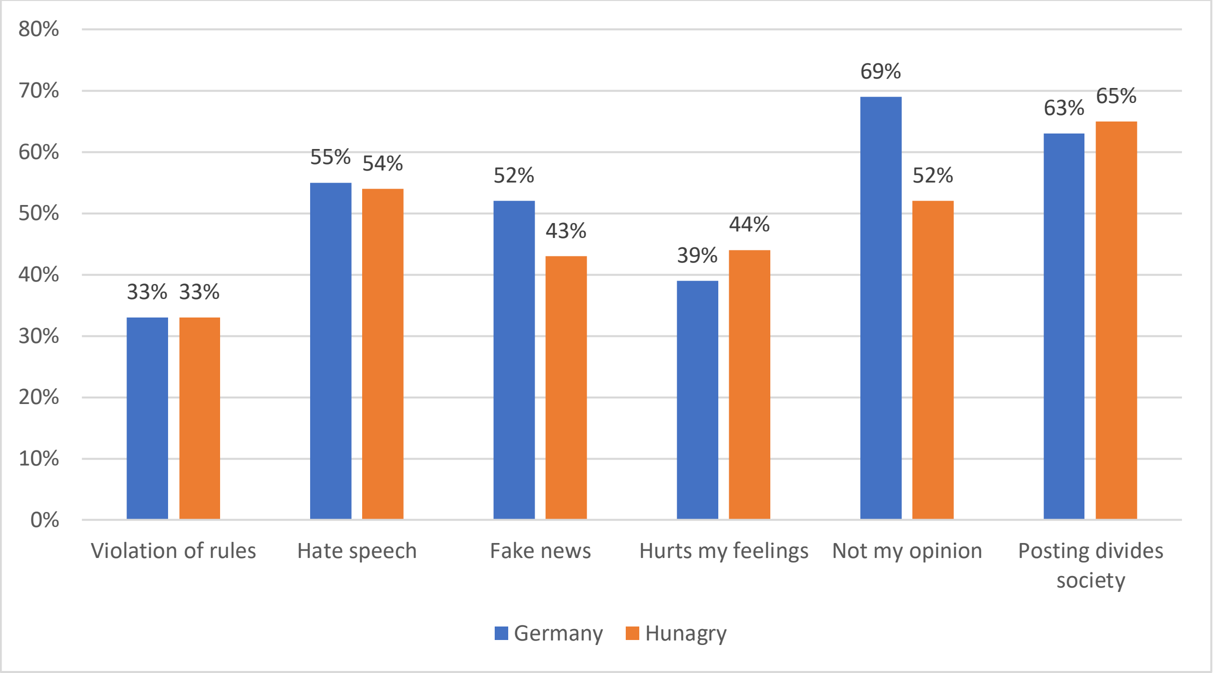


Note. Data includes all cases, who decided to share a posting (GER: *n* = 156, 30.3%; HUN: *n* = 302, 40%).

Flagging. One third of the participants, who indicated to flag one of the postings, shared the opinion, that the stimuli violate the community guidelines (figure 8). The majority of 55 percent agree that the flagged posting contains hate speech. 52 percent also agree that the posting contains inappropriate content such as fake news. The statement that the posting hurts the personal feelings is agreed of 39 percent. One third of the participants disagreed on that. More than two-thirds of the participants tend to agree or strongly agree that the posting does not reflect their opinion on the topic and 63 percent agree with the statement that people like that posting person are dividing the country with such posts.

The distribution in the Hungarian sample is similar in many respects to those of the German sample. But there are some crucial differences: Only 44 percent of the respondents flagged at least one of the postings - a difference of 28 percent. Only 43 percent tend to agree or fully agree on the statement, that the flagged posting contains wrong content - a difference of 9 percent to the German sample. There are also just 52 percent that agreed on the statement, that it is not reflecting their personal opinion on the subject; a difference of 17 percent in comparison to the German respondents.

Figure 8 Motives for flagging a posting (in percentages)



Note. Data includes all cases, who decided to flag a posting (GER: *n* = 374, 72.6%; HUN: *n* = 328; 44.3%).

4 Summary

The results illustrate clear differences in the perception of gendered hate speech between Germany and Hungary. First, it should be emphasized that the presented postings on anti-gender, homophobia and sexism are more strongly liked and shared by the Hungarian participants. Postings in the German study were reported almost twice as often as those in the Hungarian sample. This is also reflected in the number of participants who chose to flag a posting. The majority of German participants (73%) reported at least one posting, compared to 44 percent of the Hungarian respondents.

The topic of the posting was most relevant regarding liking and commenting on a posting. For the Hungarian participants the topic also played the most important role in sharing the posting, for the Germans the type of hate speech played a decisive matter. For the German participants was the topic the most important factor to flag a posting, in addition to the type of hate speech For the Hungarian participants, the type of hate speech was most important for flagging intentions, followed by the topic, the type and the gender of the posting actor.

Regarding the differences between social groups, we found that some of them are country-specific whereas others are consistent across countries: We observed similar effects of political orientation and voting intention in both samples such that the more right-wing participants political orientation and voting intentions are, the more they engage positively with the posts and refrain from flagging such content. In the Hungarian sample, there was a stronger engagement with homophobic posts that was fostered by a right-wing orientation and voting intention. We did not observe such an effect in the German sample. Moreover, the effects of social media use and religiousness are country-specific, also particular effects of gender and education.

In terms of reasons for commenting on a post, the Hungarians' self-assessment that they have specific knowledge about the topic and would like to share it, as well as that they would like to discuss the topic with other Facebook users, outweighs the German participants' self-assessment. In terms of reasons to share the post, the majority of respondents in both countries have been stated that the post reflects their opinion on the topic and that the posting person is communicating the truth. However, Hungarian participants were more likely to share a posting than German participants. People from both countries who decided to flag a post mainly shared the opinion that the posting person divides the country with such posts and recognize that the post contains hate speech. However, only one third of the participants agreed that the posting violates community rules.

References

Blassnig, S. & Wirz, D. S. (2019). Populist and Popular: An Experiment on the Drivers of User Reactions to Populist Posts on Facebook. *Social Media + Society*, *5*(4), Online first. Retrieved from: https://doi.org/10.1177/2056305119890062

PolitPro (2021). Wahltrends und aktuelle Sonntagsfragen für Ungarn [Election Trends and latest Sunday Polls for Hungary]. *politpro.eu*, June 20th, 2021. Retrieved from: https://politpro.eu/de/ungarn

statista (2021). Sonntagsfrage zur Bundestagswahl nach einzelnen Instituten 2021 [Sunday Poll for the 2021 Federal Election by Individual Institutes]. *statista.com*, June 23, 2021. Retrieved from: https://de.statista.com/statistik/daten/studie/30321/umfrage/sonntagsfrage-zur-bundestagswahl-nach-einzelnen-instituten/

Takács, J. & Szalma, I. (2020). Democracy deficit and homophobic divergence in 21st century Europe. *Gender, Place & Culture*, *27*(4), 459-478.

Wilhelm, C., Joeckel, S., & Ziegler, I. (2020). Reporting Hate Comments: Investigating the Effects of Deviance Characteristics, Neutralization Strategies, and Users' Moral Orientation. *Communication Research, 47*(6), 921-944.

Chapter 2

Analysis self-regulation of Facebook and Twitter

Antigona Research Group – Autonomous University of Barcelona (UAB)

Department of Law - University of Gothenburg (UGOT)

2021

1. Twitter

1.1 Introduction

Twitter was officially launched in July 2006. It is an online microblogging site where users can find and share information and messages with a maximum of 280 characters. Users can use the platform to write any statement, update, how they feel, status update, idea or point of view on the 'What's happening?' field in their profile. These updates are called *Tweets* and displayed on a user's profile. Users can also follow updates from other users and reply to another user using the symbol @ before the username or just send a private message. Other functions on this platform allow to Retweet (reshare) another user's comment and select 'favourite' Tweets (Skemp, 2020).

From 2006 to today, Twitter has considerably contributed to the importance of public debates in society. It started out as a very lightweight service for updating friends about one's thoughts or activities but it has become as huge company on which journalists, academics and politicians depend. It has also become a social and professional tool for public dialogue and relations. Twitter has been described as a 'nervous system for the planet' and a 'global newsroom' hosting a large amount of real-time data about social behaviour and public communication (Burgess and Baym 2020).

1.2 Impact

Since its creation, Twitter has had an immediately global effect. Initially in 2006, the community of users was very small and not fully aware of the public character and impact of this platform. Nowadays, after over fourteen years, Twitter is the fifth most visited site on the Internet (with over five billion visits). Five hundred million Tweets are sent every day (six thousand Tweets every second) and 83% of the world's leaders have a verified account on this platform.

1.3 Self-regulation on Twitter

One of the aims of this report is to identify and analyse the internal policies adopted by Twitter, how to deal with Tweets, comments and messages from far-right political parties inciting hate crimes and/or hate speech with an anti-gender content.

1.3.1Rules and policies on Twitter

The section Rules and policies includes two important sections for our review: Twitter Rules and policies and General guidelines and policies to be analysed below.

1.3.1.1 Twitter rules and policies

1.3.1.1.1 Twitter rules

Twitter Rules contain the behaviour which is allowed and not allowed on Twitter. It works as an ethical code where Twitter affirms its desire to create a safe and free space that allows public debate and communication. Twitter does not provide a definition of hate speech, hate crime, incited hate, sex violence, discrimination, threats or gender violence. It just describes which kind of behaviour is not allowed.

In the Safety section, the following appears:

- Hateful conduct: "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."

- Violence: "You may not threaten violence against an individual or a group of people. We also prohibit the glorification of violence."

- Abuse/harassment: *"*You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm."

In the Privacy section, the following appears:

- Non-consensual nudity: "You may not post or share intimate photos or videos of someone that were produced or distributed without their consent."

1.3.1.1.2 Hateful conduct policy

Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories”.

Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category”.

Twitter states that its mission “is to give everyone the power to create and share ideas and information, and to express their opinions and beliefs without barriers”. It recognises free expression as a human right, and consequently Twitter manifests that its role is to “serve the public conversation, which requires representation of a diverse range of perspectives”.

Twitter also admits that if people experience abuse on Twitter, it can jeopardise their ability to express themselves, and it also accepts that research has shown that some groups of people are disproportionately targeted with online abuse such as “women, people of colour, lesbian, gay, bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities”.

Twitter affirms that it is committed to “combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category”.

1.3.1.1.3 Non-consensual nudity policy

Since November 2019, Twitter has warned that “you may not post or share intimate photos or videos of someone that were produced or distributed without their consent”.

Here it affirms that sharing explicit sexual images or videos of someone online without their consent is a severe violation of their privacy and the Twitter Rules. This is sometimes referred to as revenge porn, and there is no mention here about women as a first target and victims of these kinds of practices: “This content poses serious safety and security risks for people affected and can lead to physical, emotional, and financial hardship.”

1.3.2 General guidelines and policies

1.3.2.1. Glorification of violence policy

For Twitter, glorifying violent acts could inspire others to take part in similar acts of violence. Twitter also states that “glorifying violent events where people were targeted on the basis of their protected characteristics (including race, ethnicity, national origin, sexual orientation,

gender, gender identity, religious affiliation, age, disability, or serious disease) could incite or lead to further violence motivated by hatred and intolerance”.
Twitter has therefore a policy against content that glorifies acts of violence in a way that may inspire others to replicate those violent acts and cause real offline harm, or events where members of a protected group were the primary targets or victims.

1.3.2.2. Abusive behaviour

A Twitter rule can be identified as a general warning that you may not engage in the targeted harassment of someone or incite other people to do so. Abusive behaviour is considered an attempt to harass, intimidate, or silence someone else’s voice.
One interesting example of what is considered abusive behaviour is **unwanted sexual advances**. Twitter prohibits “unwanted sexual advances and content that sexually objectifies an individual without their consent”. This includes behaviour such as “sending someone unsolicited and/or unwanted adult media, including images, videos, and GIFs; unwanted sexual discussion of someone’s body; any other content that otherwise sexualizes an individual without their consent”.

Once again, Twitter does not make any specific mention to the fact that women are the main and more common targets/victims of this kind of behaviour, and it simply talks about “individuals”.

1.3.2.3 Defending and respecting the rights of Twitter uses

This part focuses on three aspects: freedom of expression, privacy and transparency.
Twitter expresses its commitment to different legal frameworks such as the United States Bill of Rights and the European Convention on Human Rights. Furthermore, it obtains information from a number of additional sources including the members of their Trust and Safety Council.

1.3.2.4 Violent organisations policy

Twitter provides a definition of violent organisations as a collection of individuals with a shared purpose who systematically target civilians using violence.
There are some exceptions to this policy, mainly regarding groups with representatives who have been elected to public office through democratic elections: “This policy also doesn’t apply to state or governmental organizations”.

1.3.2.5 Violent threats policy

witter defines violent threats as “statements of an intent to kill or inflict serious physical harm on a specific person or group of people”. This definition may include statements like “I will”, “I’m going to”, or “I plan to”, as well as conditional statements like “If you do X, I will…”.

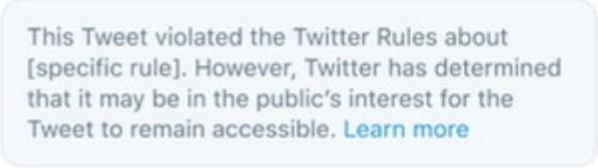
But statements that express a wish or hope that someone experiences physical harm, making vague or indirect threats, or threatening actions that are unlikely to cause serious or lasting injury are not actionable under this policy, unless it becomes abusive behaviour and hateful conduct.

1.3.2.6 About public-interest exceptions on Twitter

Twitter considers that some content could be considered of public interest if “it directly contributes to understanding or discussion of a matter of public concern”.

Twitter limits the exception to one critical type of public-interest content—Tweets from elected and government officials—given the significant public interest in knowing and being able to discuss their actions and statements.

In those cases, the platform places a disclaimer behind the Tweet context informing that the content violates some rules but nevertheless allows people to click through to see the Tweet. Placing this disclaimer also limits the possibility to engage with the Tweet through likes, Retweets, or sharing on Twitter, and makes sure the Tweet is not algorithmically recommended by Twitter.



This Tweet violated the Twitter Rules about [specific rule]. However, Twitter has determined that it may be in the public’s interest for the Tweet to remain accessible. Learn more

But not any Tweet is allowed. Twitter recognises that it is very difficult to balance rights between the public interest of a world-leader’s Tweet (with his/her freedom of expression included) and the violation of the Twitter internal rules. Twitter admits this is unprecedented and each decision constitutes a new precedent.

On top of that, maintaining those Tweets will help to have some sort of record of them and eventually contributes to the public interest: “maintaining a robust public record provides benefits to accountability”.

This dilemma becomes even more difficult being a private company, where the economic benefit increases as the number of users grow. In other words, if it starts deleting the Tweets that have a major impact on the general public, that will also affect its economic profits.

With this in mind, Twitter will remove Tweets where there is evidence that the content may be leading to actual or likely offline harm.

1.3.2.7 Coordinated harmful activity.

Twitter identify groups, movements, or campaigns that are engaged in coordinated activity resulting in occasional harm. It analysed these groups, movements, or campaigns against an analytical framework, with specific on-Twitter consequences if they determine that they are harmful. This can be Technical Coordination (one person uses several accounts at the same time and sends the same Tweet from all these accounts) or Social coordination (a group of people that disseminate a specific message within and outside Twitter).

Twitter divides this harmful activity into Physical, Psychological and Informational harm, and also divides the severity of the actions into low, moderate and high. The consequences of this type of harmful activity go from limiting visibility of the Tweets to suspending accounts whose primary use is to propagate and/or encourage engagement in the identified coordinated harmful activity.

1.3.2.8. Reporting false information in France

France is the only country in the European Union that has included a specific section in General guidelines and policies on Twitter. In November 2018, France passed the Against Information Manipulation Act[2]. This act focused on the election campaigns just before and during the elections and it aims to prevent the risk to unfairly influence the election results (as occurred during the 2015 American Presidential elections and during the Brexit campaign) and to protect democracy from the dissemination of fake news.

This act emphasises the obligation for transparency of digital platforms, which “must report any sponsored content by publishing the name of the author and the amount paid. Platforms exceeding a certain number of hits a day must have a legal representative in France and publish their algorithms”. The text establishes a duty of cooperation of the online platform operator[3] (Article 11), to force them to introduce measures to eliminate fake news and make these measures public.

[2] https://www.gouvernement.fr/en/against-information-manipulation

[3] The definition of an online social platform comes from article L11-7 of the French Consumer Code and establishes: “An online platform operator is any natural or legal person offering, in a professional capacity, paid or unpaid, an online communication service to the public based on: first, the classification or

This act brings more transparency, which means a more protective legal framework for situations not legally foreseen. For example, when these platforms censor certain contents, they are not obliged to explain the reasons behind this.

In this case, Twitter can remove the prohibited content and/or permanently suspend the account if the Tweet violates Twitter rules.

French law also requires a mechanism to report fake information that attempts to disturb the authenticity of the vote or the public order.

1.4. Safety and security on Twitter

1.4.1 Abuse

Twitter recommends users to unfollow and end any communication with an account which posts something that we do not like. If that behaviour continues, Twitter recommends we block the account and if we continue receiving unwanted, targeted and continuous replies on Twitter, and we feel it constitutes online abuse, Twitter urges us to consider reporting the behaviour.

Twitter also explains the procedure after reporting abusive behaviour. After we submit a report, Twitter will review the account in question and/or related Tweets and if it deems that this account or Tweet constitutes a violation of its policies, it will take action (ranging from a warning to permanent suspension of the account).

This procedure on how to report abusive behaviour on Twitter does not make any specific mention to sex or gender or refer to the complainer's gender or sex.

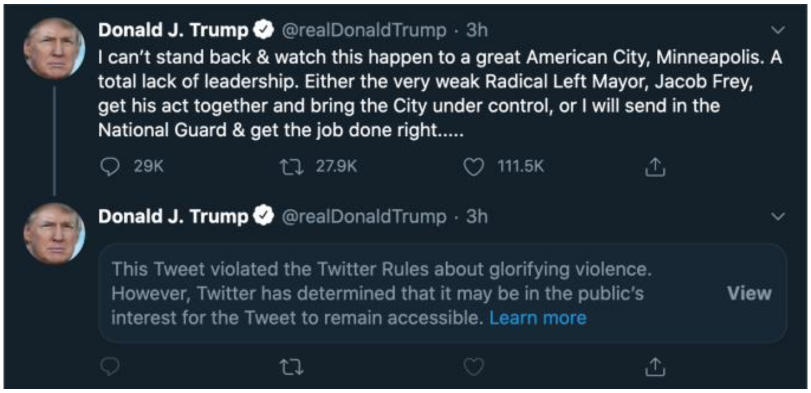
1.5 Twitter ads policies

In Twitter's Ads Content Policies, it is interesting to highlight the category Hateful content where Twitter "prohibits the promotion of hateful content globally". Examples of hateful content include hate speech or advocacy against a protected group, individual, or organisation based on, but not limited to, the following: "race, ethnicity, colour, national origin, sexual orientation, sex, gender identity, religious affiliation, age, disability, medical or genetic condition, status as a veteran, status as a refugee, status as an immigrant".

referencing through computer algorithms, of content, goods or services offered or put online by third parties.

1.6 World leaders on Twitter: principles & approach

Twitter puts in context every reported Tweet and tries to look for as many interpretations as possible. However, if a world leader's Tweet has a clear public interest value, (see the definition of public interest in section 1.4.2.6 of this review) and also breaches Twitter rules, the user can still access the content of this Tweet.



Donald J. Trump @realDonaldTrump · 3h

I can't stand back & watch this happen to a great American City, Minneapolis. A total lack of leadership. Either the very weak Radical Left Mayor, Jacob Frey, get his act together and bring the City under control, or I will send in the National Guard & get the job done right.....

Donald J. Trump @realDonaldTrump · 3h

This Tweet violated the Twitter Rules about glorifying violence. However, Twitter has determined that it may be in the public's interest for the Tweet to remain accessible. Learn more

Example of a Tweet of the former U.S. president Donald Trump after the street disturbances in Minneapolis caused by the death of George Floyd, with a disclaimer by Twitter. (Source:https://techcrunch.com/2020/05/29/twitter-screens-trumps-minneapolis-threat-tweet-for-glorifying-violence/).

Twitter states that no leader can be above its policies. Twitter underlines that certain issues cannot be exempted (even if the Tweet is in the public interest) such as threats of violence, sharing photos with intimate videos or child exploitation (among others).

Policies and approaches are under constant supervision, taking into account that the power of messages from world leaders could have an effect on the offline world.

Twitter concludes that it aims to comply with the rules in a critical and impartial manner. In this sense: "we aim to provide direct insight into our enforcement decision-making, to serve public conversation, and protect the public's right to hear from their leaders and to hold them to account".

1.7 Information regarding the 17th Transparency Twitter Report

Twitter publishes a transparency report every six months.

In this report[4], it was noted that the number of accounts actioned under the Hateful Conduct Policy by Twitter has decreased by 35%. The development of more restrictive rules on hate speech fulfilled their role of general prevention.

The number of accounts affected by the non-consensual nudity (NCN) policy decreased by 58%. Twitter announced it will continue to be more proactive in this area.

Twitter works with the Lumen database[5] to deal with the requests of content removal.

The report also points out that they have received 19% more reports based on local laws from trusted reporters and NGOs.

In the statement in the 14th report drawn up by Twitter[6], Twitter explained that "As the public discussion about regulation increases, we believe transparency is an essential part of ensuring you — the public — are able to see how these laws operate." And it also indicates that transparency is not an exclusive matter of private enterprises. Government and lawmakers have the duty of being transparent about their own actions, letting people know if the decision to eliminate a Tweet was a decision by Twitter or due to a Governmental request: this transparency is essential if we are to foster an informed debate and mitigate the risk of inappropriate use of state power.

2. Facebook

According to Skemp (2020), Facebook is "The world's largest online social-networking website where users can connect and share with friends". This social network allows us to create a profile that can be personalised and that carries out all its activity. Through Facebook, every user can share thoughts, opinions, links, notices, photos and videos. They can also use specific apps, online games, marketing programmes, monitor trade, etc. 36% of the people get news updates from Facebook. The possibilities are huge. But it is not only people who can have a profile. Profit and non-profit organisations can also have a Facebook page (Guerrero, 2014).

2.2 Impact

Co-founder and owner Mark Zuckerberg said that Facebook: "is a very early indicator that governance is changing—[and of how] powerful political organizations can form. These things can really affect people's liberties and freedom, which is kind of the point of government..." (Kirkpatrick, 2011).
Facebook became a place of hard debates on the political field, and politicians' Facebook profiles became a space for debate, contest and protest (Zurutuza et al, 2018).

[4] https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html

[5] The Lumen database (lumendatabase.org) is a partnership of Harvard's Berkman Klein Center for Internet and Society. This database provides all content reported and removed from online material, helping users to understand the law, study the prevalence of legal threats and let Internet users see the source of content removal.

[6] https://blog.twitter.com/en_us/topics/company/2019/key-data-and-insights-from-our-14th-twitter-transparency-report.html

2.3. Facebook as a social arena in Sweden and Spain

2.3.1 Sweden

he use of social media in general continues to increase, and in 2020 nearly 9 out of 10 Internet users were using social media in Sweden. Daily use has also increased steadily during the 2020 pandemic. Facebook, launched in Sweden in 2004, is the biggest social media platform, followed by Instagram and Snapchat. In 2019, 74% of the Internet users over 12 years old used Facebook occasionally and 51% on a daily basis (compared to Instagram 61/41% and Snapchat 39/24% respectively). In the 3rd quarter of 2020, 81% of Internet users over 16 years old used Facebook occasionally and 58% used it every day. Twitter was in sixth position, after LinkedIn and Flashback, with 24/7% in 2019.

The figures for Facebook have been relatively constant for several years (70-71-74-76-74% respectively from 2015-2019). And it seems like the activities on Facebook are stagnating or even decreasing in general over recent years, with the exception of sharing news and articles which is increasing. Young people (up to 25 years old) are the ones leading the disinterest.

Women use Facebook more than men: 79% of female Internet users compared to 69% male users. Men use Twitter and Flashback more than women: 29% compared to 19% and 37% compared to 27% respectively. Almost all students use at least one social media platform (98%), and the figure for Facebook among students is about 90%.

The use of Facebook is almost the same for urban and rural inhabitants, while Twitter is used less by people living outside cities.

Specifically, younger people are exposed to hate speech on the Internet, but there has been no increase of this exposure in recent years. Harassment through e-mail has, on the contrary, increased.

2.3.2 Spain

According to the 2019 report of Elogia & IAB[7], people who use Facebook in Spain do it mostly, for practicalities, and to a lesser extent, for gossip, fun and to connect with family and friends. Conversely, Twitter, in the same report, was claimed to be used for news and opinions.

In January 2020, Facebook became the most commonly used social-media platform in Spain with 2.449 million visits, above YouTube and WhatsApp (We Are Social ES 2020).

In January 2021, 52% of Facebook's users were women. Of these, almost a quarter were 25 to 44 years old (Statista, 2020).

[7] https://marketing4ecommerce.net/x-estudio-de-redes-sociales-2019-whatsapp-supera-a-facebook-como-la-red-mas-usada/ (text in Spanish).

Finally, from 2010 to 2019, Facebook declared income in Spain of an amount of 261.6 million Euros (Statista, 2020).[8]

2.4 Self-regulation on Facebook

2.4.1. Community Standards on Facebook.com

The Community Standards include hate speech; one of the most serious attacks is to compare a person to something, thus dehumanising the person. Hate speech and other violations and forms of online abuse must be accessed in relation to its context.

A concern Facebook had to deal with, which is relevant for the analysis of anti-gender hate speech, is the effect of the #metoo movement. As a result of the enormous amounts of testimonies of sexual harassment and sexual violence against women, there were large amounts of posts and comments of the type "men are scum" and "men are trash", which, in accordance with the Community Standards, were taken down. So, the speech that was aiming at drawing attention to the social movement of #metoo was caught up in the filter that blocks dehumanising attacks against gender (as hate speech). They argued that if it were allowed, there would be a risk of an increased gender-based hate speech targeting women.

Another interesting development in the context of anti-gender hate speech is how Facebook has dealt with violent rape jokes. In 2013, these were allowed due to policies that allowed for toxic speech which did not seem to provoke physical harm. Consequently, several companies withdrew their ads which led to Facebook removing the jokes and re-writing their policies (van Zuylen-Wood 2019).

A tendency, at least in Sweden, that might be of interest is that many extreme right-wing people and organisations have left Facebook due to its harsh policy and started accounts on the Russian based social media platform Vk.com, very similar to Facebook in its design and function.

In some cases, Facebook will allow content that breaches the community policy if it is newsworthy or of interest to the public. This is done after balancing between the interests and the potential harm and in line with international standards on human rights.

The freedom of expression is delimited by the following values:

- Authenticity
- Security
- Confidentiality
- Dignity

In Sweden, the feminist party Feministiskt initiativ had their most influential and active communities closed down after being reported to Facebook by at least 50 individuals (50 complaints being the amount needed to close down an account or page for further investigation of possible breaches). As a result, the communities lost their most important platforms for considerable time.

[8] https://svenskarnaochinternet.se/english/earlier-years-reports/

2.4.1.1. Violence and criminal behaviour

This section includes what is not allowed to be posted on Facebook regarding violence and criminal behaviour. Here Facebook affirms its desire to prevent offline harm that may be related to content on the social network.
In this section, we found several sub-sections with information about what is not allowed to be posted, including the sub-sections entitled Violence and Incitement and Dangerous Individuals and Organisations, analysed below.

2.4.1.1.1. Violence and incitement to violence

There is a list of types of contents which are not allowed to be posted:

Threats that can lead to death (and other forms of serious violence) targeting people or places, whereby threat is defined as any of the following:

- Statements of intent to commit high-severity violence, including content where a symbol represents the target and/or includes a visual of an armament or methods to represent violence.
- Incitements to serious acts of violence, where no target is specified, but a symbol represents the target and/or includes an image of weapons or methods that represents violence.
- Statements that advocate serious acts of violence
- Far-reaching or threatening statements about the performance of serious violence.
- Contents where services to kill someone are offered or requested.
- Confessions, declarations of intent, advocations or incitements to kidnap someone.
- Content representing kidnappings, when it is certain that they have not been undertaken by the victim or the family seeking for help, or not shared to inform, condemn the act or raise awareness.
- Threats of serious violence against real people in digitally produced images, including weapons, methods of violence or mutilations.
- Threats that lead to serious injury (mid-severity violence) towards private individuals, unnamed specified persons, less well-known public figures, vulnerable persons, or vulnerable groups,

The following are prohibited:

- Statements that contain an intention to commit violent acts.
- Statements that advocate violence.
- Incitements to mid-severity violence including content related to a specified target, but where a symbol represents the target.
- Far-reaching or threatening statements about committing violent acts.
- Content that targets other individuals, less well-known public persons, vulnerable persons or vulnerable groups, and all credible:
    - ▪ statements about an intention to commit violence

- incitements to violence
- statements that advocate violence
- far-reaching or threatening statements to carry out violent acts.

There is no definition of “vulnerable persons” or “vulnerable groups”.

Additionally, the list includes acts such as:

- Threats that can lead to physical harm, targeting private persons or less well-known public persons (intentions, incitements, advocates, or far-reaching or threatening statements about carry out violence).
- Content created to expose that a person is part of a group at risk, which is specified or possible to identify.
- Instructions to make or use weapons or high explosives if there is evidence of an aim to seriously harm or kill persons.
- Content related to violence targeting elections or certain places such as religious places, schools, or election administration etc.

It is noteworthy that content that exposes LGBTQI+ persons to risks by disclosing their sexual identity against their will or without permission is not allowed, but Facebook states that it needs more information and/or context to be able to enforce its community policy on these matters.

2.4.1.1.2. Dangerous individuals and organisations

There are several cases included in this section (terrorism, mass murder, human trafficking and organised violence or other criminal activities), but we will focus on **organised hate** for the purposes of the present report.
Hate organisation is defined in this section as: “Any association of three or more people that is organized under a name, sign, or symbol and that has an ideology, statements, or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, **gender**, **sex**, **sexual orientation**, serious disease or disability”.

2.4.1.1.3 Coordination and publication of criminal activities

Facebook divides and describes the actions in different sections; some are forbidden *per se* and others are prohibited but must be put in context before a decision can be made.

The first section, forbidden actions *per se*, includes, among other examples, images, acknowledgements or encouragements of actions that causes individuals physical harm, and participation in very dangerous viral challenges, or the spreading of contagious diseases. These examples might be of interest in the context of organised attacks on persons due to their sex, gender or feminist or gender ideology.

A second section addresses content that will be followed by a warning and concerns content that pictures risky viral challenges if the post is complemented with a text that condemns or raises the awareness of such challenges.

A third section addresses posts that need to be put in context for Facebook to be able to apply its Community Standards. The organised attacks must include severe violence. Although, from our perspective, also less severe attacks in an organised form might constitute a serious threat to gender equality and in the long run, to democracy (see, for example, Bladini 2017). Examples that are not included here are actions that at first sight seem to be less serious actions, but that in organised form affect the victim largely

2.4.1.2 Safety

This section lists what posts are not allowed on Facebook as regards safety. There are several chapters included in this section but we are going to focus on Bullying and Harassment, due to the aims of this report. Other sections deal with suicide and self-destructive behaviour, sexual exploitation and other forms of child abuse, sexual exploitation of adults, human exploitation, such as trafficking and breaches against confidentiality. Furthermore, the section Sexual exploitation of adults could be of relevance as part of sexist abuse online (image-based sexual abuse, for example) disseminated online as part of a strategy against a journalist, politician or other public person.

2.4.1.3 Harmful content

There is a section regarding **hate speech**.

Facebook defines **hate speech** as a “direct attack on people based on what we call **protected characteristics** — race, ethnicity, national origin, religious affiliation, **sexual orientation**, caste, **sex, gender, gender identity**, and serious disease or disability”.

“We do not allow hate speech on Facebook. We define hate speech as *violent or dehumanizing* speech, *statements of inferiority, calls for exclusion or segregation* based on protected characteristics or slurs. These characteristics include race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity and serious disability or disease. When the intent is clear, we may allow people to share someone else's hate speech content to raise awareness or discuss whether the speech is appropriate to use, to use slurs self-referentially in an effort to reclaim the term or for other similar reasons.”

It is interesting to highlight how Facebook denominates **protected characteristics** those categories involved in this report: sexual orientation, sex, gender and gender identity. Here, Facebook separates attacks into three levels of severity, worth emphasising in level 1:
- “Mocking the concept, events or victims of **hate crimes** even if no real
person is depicted in an image.”

- Transgender or non-binary people referred to as "it".

2.5 Advertising policies on Facebook

Facebook understands as prohibited content the following:

- Ads must not violate the Community Standards, analysed above.
- Discriminatory Practices: “Ads must not discriminate or encourage discrimination against people based on personal attributes such as race, ethnicity, colour, national origin, religion, age, **sex, sexual orientation, gender identity**, family status, disability, medical or genetic condition’.
- Personal Attributes: “Ads must not contain content that asserts or implies personal attributes. This includes direct or indirect assertions or implication about a person’s race, ethnic origin, religion, beliefs, age, **sexual orientation or practices, gender identity,** disability, medical condition (including physical or mental health), financial status, voting status, membership in a trade union, criminal record, or name.”

3. Discussion

Political hate speech, especially against gender theories, is not mentioned on either platform. The intersection of hate and anti-gender speech seems to be new on Facebook and Twitter. The biggest change on how these social platforms deal with freedom of expression of political leaders/parties was that, after the capitol riots in the U.S. in January 2021, the Twitter account of the former president of the United States of America, Donald Trump, was temporarily suspended for violating its rules.

Facebook does not provide a definition of hate crime, incited hate, sex violence, discrimination, threats and gender violence nor a definition of gender and far-right political parties. It simply describes what kind of behaviour is not allowed and what kind of content users are not allowed to post.

Twitter, for its part, does not provide examples when describing the incited hate speech.

The vagueness, generality and practical arbitrariness of many of the rules and examples included in the analysed internal rules of Twitter and Facebook have generated perplexity among experts and activists. They have reflected on the openly discretionary manner in which some of the currently most important platforms for the dissemination of opinions, ideas and views could, if necessary, censor or prevent access to certain content posted by its users.

European recommendations do not have any visibility (e.g., Council of Europe Resolution 2144 (2017) on Ending cyber-discrimination and online hate; European Commission Recommendation on measures to effectively tackle illegal content online C (2018) 1177).

The European Commission worked last year on a Bill called the Digital Service Act that was ironically published on Twitter[9] by the executive vice-president of the European Commission for a Europe fit for the Digital Age (Competition), Margrethe Vestager. She declared that this European Act “will create safe & trustworthy services while protecting freedom of expression. Give new dos & don’ts to gatekeepers of the digital part of our world - to ensure fair use of data, interoperability & no self-preferences”. This Act will include more transparency in all areas: informing every user why you recommend following one person or figure, or why the algorithms show us some videos or photos. Also, there should be a person in charge to verify any information that could be fake.

Still, the use of algorithms remains a problem for these platforms, and especially since they explain the visibility of some Tweets or posts. There is a lack of transparency on how they work, including the incidents of "Shadow Banning"[10] that Twitter echoed.

Facebook has announced the recruitment of more content moderators to prevent, in particular, the dissemination of fake news or other undesirable content. But it is still unknown who they are, and what their experience in that field is or what type of training they have[11].

The same happens on Twitter. They refer to the fact that “A cross-functional team including Trust and Safety, Legal, Public Policy and regional teams will determine if the Tweets are a matter of public interest.[12] They analysed the content of controversial Tweets, but it is unknown who the members of the team are, their background, their education on gender perspective and hate speech (for the purposes of the present report).

Moderators, analysts, experts and officers of these companies are the new *worldwide judges* in this field, but they follow a business criteria of justice. These hidden people have the power to ban some speeches, allow others, tell people what they should say and what they should not share on social networks, as a new form of agenda setting.[13]

But the big question remains whether the responsibility for prevention should remain in the hands of private companies or whether it is the states’ responsibility to guarantee a free, open and plural communication on social media with justice and without any unjustified restriction.

4. Conclusion

From a regulatory point of view, therefore, the imposition of general duties and responsibilities in terms of monitoring and control of content on social networks and other

[9] https://twitter.com/vestager/status/1338869405329936385

[10] https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html

[11] https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/

[12] https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html

[13] In 1972, Maxwell McCombs and Donald L. Shaw, in their article *“The Agenda-Setting Function of Mass Media”,* elaborated a theory called *Agenda Setting*, focusing on the mass media (in broadcasting era). The main idea of this theory claims that *what the public perceives as important is highly influenced by the media, who choose to cover certain stories over others.* So, in those cases, the media had the power to choose what issue would be important for the public to talk about, debate and so on, and what issue was not, therefore acting like a judge of sorts.

platforms could be counterproductive in some aspects and necessary in others. First, because it opens a dangerous door to private censorship and corporate control of the public sphere. However, on the other hand, it is clear that these enterprises have too much power in the public opinion and must be subject to some obligations and responsibilities. That involves the development of more education in how their users should participate in an open and plural debate.

This must be followed with the help of the public authorities. Public authorities must consider certain intervention mechanisms that guarantee a series of principles and values in a space as important as that of social media and similar platforms.

The case of France could be a good example. France, on its own, put on the table a serious legislation on fake news. Twitter had to accept it and changed its policies on fake news after and during the elections in France (see section 1.4.2.8 of this report).

Another point to consider is the flow of information on social media. Some authors talk about how to smash one's *echo chamber* and how to tackle hate (Bartlett 2018). They say that platforms can give us the opportunity to see and participate in debates (with moderation and respect) listening to others' points of views or ideas.

A wider and more structural analysis of speech is needed. Amongst other things, to prevent the *indirect censorship* that silences the voices of attacked groups. Also, a new frame of interpretation of freedom of expression that allows a diversity of voices without being subject to violence.

Because the diversity of voices is as crucial as a part of freedom of expression (Kenyon et al 2017).

Therefore, transparency and understanding on how algorithms work is urgently needed. People want a democratic mechanism to secure their accountability, so every state must create a powerful, technically skilled and well-resourced body capable of supervising their potential misuse.

To summarise, a solid European legal framework is needed. Also, future legislation may require a more proactive involvement of companies such as Facebook and Twitter. They cannot be expected to be the judges or the human rights guarantors on the Internet. But the EU's legal mechanisms on the Internet should include legally binding tools to include these social platforms as active actors.

5 Bibliography

Allen, R. (2017). *Hard Questions: Who Should Decide What is Hate Speech in an Online Global Community?* June 17, 2017. Retrieved from https://about.fb.com/news/2017/06/hard-questions-hate-speech/

Bartlett, J. (2018). *The people vs tech: how the internet is killing democracy (and how we save it*). Ebury Press.

Bladini, M. (2017) *Hat och hot på nätet. En kartläggning av den rättsliga regleringen i Norden från ett jämställdhetsperspektiv*. [Hate and threats online – a survey of the legal regulation in the Nordic region from a gender equality perspective] Gothenburg: NIKK. Nordisk information för kunskap om kön [Nordic Information on Gender], på uppdrag av Nordiska Ministerrådet.

Bladini, M. (2021) Silenced voices. Online Violence Targeting Women as a Threat to Democracy, in Svensson, Eva-Maria, Burman, Monica & Gunnarsson, Åsa (Eds.) *Special Issue: Exploiting Justice in a Transformative Swedish Society*, in *Nordic Journal on Law and Society*, Vol. 3 No. 02 (2020).

Burgess, J. & Baym, N. K. (2020). *Twitter: a biography*. New York University Press.

Davis, Antigone, 2018. Protecting People from Bullying and Harassment (https://about.fb.com/news/2018/10/protecting-people-from-bullying/)

Díaz Limón, J., (2019). *Derecho En Tiempos De Zuckerberg : Estudio Jurídico Sobre Las Condiciones, Políticas y Normas De Facebook*. 1st ed. Ciudad de Mexico: Tirant Lo Blanch, p.16.

Edström, Maria. 2016. Swedish Experiences of Mediated Sexualised Hate Speech in the Aftermath of Behring Breivik. *International Journal for Crime, Justice and Social Democracy* 5(2): 96-106.

Guerrero, D. (2014). *Facebook: guía práctica*. RA-MA Editorial. https://elibro.net/es/ereader/uab/106451?page=18)

Kenyon, Andrew, T., Svensson, Eva-Maria & Edström, Maria. 2017. Building and Sustaining Freedom of Expression. Considering Sweden. *Nordicom Review* 38(1): 31-45

Kirkpatrick, D. (2011). *The Facebook effect: the inside story of the company that is connecting the world* (1st Simon & Schuster trade pbk. ed). Simon & Schuster Paperbacks.

Skemp, K. M. (2020). *Twitter and Facebook*. Salem Press Encyclopedia.

van Zuylen-Wood, S. (2019) ”*Men are scum”: Inside Facebook's war on hate speech, in Vanity Fair*. Feb. 26 2019.

Zurutuza-Muñoz, C., & Lilleker, D. (2018). Writing graffiti on the Facebook wall:

Understanding the online discourse of citizens to politicians during the 2016. Spanish election. *Communication & Society*, 31(3), 27–42. (https://doi.org/10.15581/003.31.3.27-42)

6. Sitography

Community Standards Enforcement Report, Fourth Quarter 2020. Accessed on 11 February 2021. *About Facebook*. https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020/

Statista 2020. *Facebook: frecuencia de uso en España en 2020. (s. f.)*. (text in Spanish). Retrieved on 16 February 2021 from https://es.statista.com/estadisticas/1017708/frecuencia-de-uso-de-facebook-por-los-usuarios-de-redes-sociales-en-espana/

We Are Social ES 2020. Digital 2020 España. (s. f.). Retrieved on 16 February 2021 from https://wearesocial.com/es/digital-2020-espana