

Funded
by the European Union
and the Council of Europe



Implemented
by the Council of Europe



Report on research on how to identify incidents most likely to generate waves of hate speech

Report prepared by ActiveWatch
with additional data collected by Agenzia di Promozione Integrata per i Cittadini in Europa (APICE) and Neue Deutsche Medienmacher e.V. (NdM)

March 2021



Table of contents

I.	General introduction	4
II.	Aim of the research	5
III.	Methodology.....	6
IV.	Findings of the research.....	7
1.	Romania.....	7
	Diagram 1 – Localisation of the event.....	7
	Diagram 2 – Social network posts with hateful comments	9
	Diagram 3 – Rate of hateful comments on social network posts.....	9
	Diagram 4 – Posts with CAN messages based on incidence of perceived threats	9
	Diagram 5 - Social network posts with hateful comments	10
	Diagram 6 – Rate of hateful comments on social network posts.....	11
2.	Germany.....	12
	Diagram 7 – Hateful user generated posts on social networks	13
	Diagram 8 – Rate of hateful comments on social network posts.....	14
	Diagram 9 – Posts with CAN messages	15
	Diagram 10 – Hateful user generated posts on social networks based on relation with country where the event took place	15
	Diagram 11 – Social network posts with hateful comments based on perceived double characteristics by perpetrator.....	16
	Diagram 12 – Posts with CAN Messages	16
	Diagram 13 – Hateful user generated posts on social networks	17
	Diagram 14 – Rate of hateful comments on social network posts.....	17
	Diagram 15 – Hateful user generated posts on social networks	18
	Diagram 16 – Social network posts with hateful comments	18
3.	Italy.....	19
	Diagram 17 – Social network posts with hateful comments	20
	Diagram 18 – Rate of hateful comments on social network posts.....	21
	Diagram 19 – Posts with CAN messages	22
	Diagram 20 – Posts with CAN messages based on perceived threats	23
	Diagram 21 – Rate of hateful comments on social network posts based on perceived characteristics	23
	Diagram 22 – Rate of CAN messages in posts where CAN is present based on perceived characteristics	23
	Diagram 23 – Hateful user generated posts on social networks	24
	Diagram 24 – Social network posts with hateful comments	25
V.	Overall conclusions of the research.....	26
	Appendix 1: Further information on the members of the consortium for the “WE CAN for human rights speech” project.....	28
	Appendix 2: Research on hate speech going viral -data collection methodology-	29

I. General introduction

The Council of Europe's discrimination monitoring body, the [European Commission Against Racism and Intolerance \(ECRI\)](#), identified during its last monitoring cycle that racist, xenophobic and populist discourse is increasing across all member states of the Council of Europe, which includes EU member states.

In line with [ECRI's General Policy Recommendation No 15 on Combating Hate Speech](#), more than legislation is needed to push back against hate speech, a coherent and comprehensive approach includes the use of counter speech, awareness raising and educational efforts.

The Council of Europe's [No Hate Speech Movement \(NHSM\) youth campaign \(2013-2018\)](#) was instrumental in raising awareness, mobilising civil society and developing educational tools for preventing and countering hate speech. Following the evaluation of the campaign, national level NHSM committees and activists expressed the need for sustainability, further capacity building and enhanced networking to consolidate their work and to improve their actions against hate speech using effective counter and alternative narratives that speak for human rights on and offline.

The "[WE CAN for human rights speech](#)" project aims to meet these needs, the project is funded by the European Union's Rights, Equality and Citizenship Programme (2014-2020) and the Council of Europe. It is being implemented from February 2020 to January 2022 by a consortium of partners: Active Watch, APICE, CEJI, and Neue Deutsche Medienmacher e.V. (NdM), led by the Council of Europe's No Hate Speech and Co-operation Unit (more information about the partners can be found in Appendix I).

The project aims to:

- help organisations and young activists fighting hate speech to become more efficient by providing them user friendly tools (developed via evidence-based research and simplified and updated previous materials generated by the No Hate Speech Movement) and with trainings
- consolidate the cooperation among these organisations and activists at the European level and help develop new partnerships with social media companies, other networks of NGOs and national authorities.

Between February and November 2020, the consortium conducted research on how to identify situations most likely to generate waves of hate speech, and on timing and ways of engaging in counter and alternative narratives to have maximum impact.

The aim of this research is to help activists and NGOs be prepared for action and be more effective in their activities against hate speech. The research will also feed into the development of tools and trainings to analyse hate speech, develop human rights-based narratives and communication strategies which should be available in Spring 2021.

The research methodologies were prepared before the COVID-19 pandemic and were subsequently adapted to take into account the new situation and potential new waves of hate speech. The period of research was also extended from an initial 6-months to 10-months in order to follow the potential new trends linked to COVID-19.

II. Aim of the research

Hate speech has become a ubiquitous phenomenon online, raising more and more concerns for regulators and social media platforms as well as for the Non-Governmental Organisations (NGOs) and activists who are trying to combat it. Strategies aimed at curbing this phenomenon can be very broadly split into three categories: the ones relying on taking action against the speaker (administrative, civil or criminal proceedings or limiting the speakers' actions on internet platforms due to their breaching of the terms and conditions), the ones aimed at censoring expressions (of which the most common form is the deletion of content on social media platforms) and the ones relying on communication to either offer new perspectives on the situation / victims of the hateful messages (also known as alternate narrative strategies) or to dismantle the arguments used in the hateful narrative (also known as counter narratives).

While strategies relying on punishing speakers or on censorship offer quantifiable, short term results, they tend to be less efficient in the long run. Some of the reasons for this are: content that has been taken down tends to resurface again quickly on other platforms and on other websites, and the fact that many speakers find ways to circumvent the mechanism aimed at restricting hate speech or even that these types of action do sometimes contribute to the radicalization of audiences, turning haters into martyrs and hateful narratives in truths targeted by "global elites".

Of the three types of strategies, we consider that the one relying on more communication is the most effective one when aiming for long-term results, as it is the one with the highest potential to actually shift the perspective of the audiences and, as a result, neutralize the effects of hateful narratives. Unfortunately, there are two challenges to deploying this strategy. First of all, it is very dependent on the skills of the communicator to build and disseminate powerful and compelling messages to the intended audiences. In other words, it is much harder to do, than to simply report a content or file a complaint against an extremist. Secondly, matching the type of strategy (counter narratives or alternative narratives) with timing and with the beliefs of the audiences can also be hugely challenging.

Therefore, an "early warning" system that would give activists a clue as to when a wave of hate speech is about to strike would increase their chances of deploying the right narrative at the right time, to the right audiences. And while there is no way to foretell precisely when influencers decide to spontaneously express their own hateful beliefs, it is easily noticeable that waves of hate speech online tend to be correlated with events happening in the real world.

The research undertaken in our project has aimed at trying to identify the specific characteristics that an event would need to have in order for it to generate hate speech online. We have collected data from 3 countries (Romania, Germany and Italy) and we tried to identify both national and transnational trends.

III. Methodology

The methodology of the research is based primarily on grounded theory, which involves the collection then the analysis of data. In order to reach our conclusions, we first collected data over the spring and summer of 2020, focusing on social network posts related to events involving social groups which are usually targeted by hate speech online: Roma ethnic minorities, sexual and gender minorities, Jews, Muslims and in the case of Romania, Hungarian ethnic minorities.

Each partner in the research collected the data using their own methods:

- In Romania, ActiveWatch used the scrapping service ZeList Monitor. Social media content from Romanian Facebook pages and Twitter and Instagram accounts was retrieved based on keywords (accepted and pejorative denominations for the social groups targeted by hate speech)
- In Germany and Italy, the content was retrieved manually by NDM and, respectively, APICE staff. Some of the content was flagged by NDM's and APICE's network of activists, and some was identified using keyword searches.

The content which referred to events involving the social groups of interest to our research was kept, and all the other content was discarded. Then, each of the events was described based on the following characteristics:

- Description (what happened)
- Scope of the event (local / regional / national / European / other)
- Perceived threat generated by the event
- Materialized threat generated by the event
- Number and characteristics of the direct victims of the event
- Characteristics of the people pictured by the hateful narratives as the (potential) aggressors involved in the event
- The context in which the event took place

Also, the public discourse on social media related to each of the identified events was coded based on the following characteristics

- The approximate number of the social media posts covering the event (the extent of the coverage)
- The most probable audience of the messages (only in those cases in which this was possible to infer)
- The extent to which the social media posts that were identified contained hate speech
- The rate at which the comments to the social media posts contained hate speech
- The types of influencers involved in generating negative emotions towards the groups targeted by hate speech
- The extent to which counter and alternative narratives were deployed

The limitations of our research have to do primarily with how the data was collected and "coded" by the researchers. The use of a scrapping tool to retrieve online content in Romania allowed the identification of a larger number of events, but which had a smaller number of hate speech messages associated with them. Also, the Romanian way of collecting data allowed for the monitoring of only "professional accounts", not those of private individuals, which can also explain the generally hate-free nature of the posts themselves (not of the comments). In Germany and Italy, however, the data collection methodology allowed also for the exploration of accounts considered as "usual suspects" when it comes to the online dissemination of hatred. The result is a lower number of events but a higher incidence of hate speech.

As far as coding the data is concerned, the main limitations were generated by the hybrid qualitative-quantitative nature of the research which led to the partners having to harmonize the way they interpreted the identified content. To achieve this harmonization, the internal

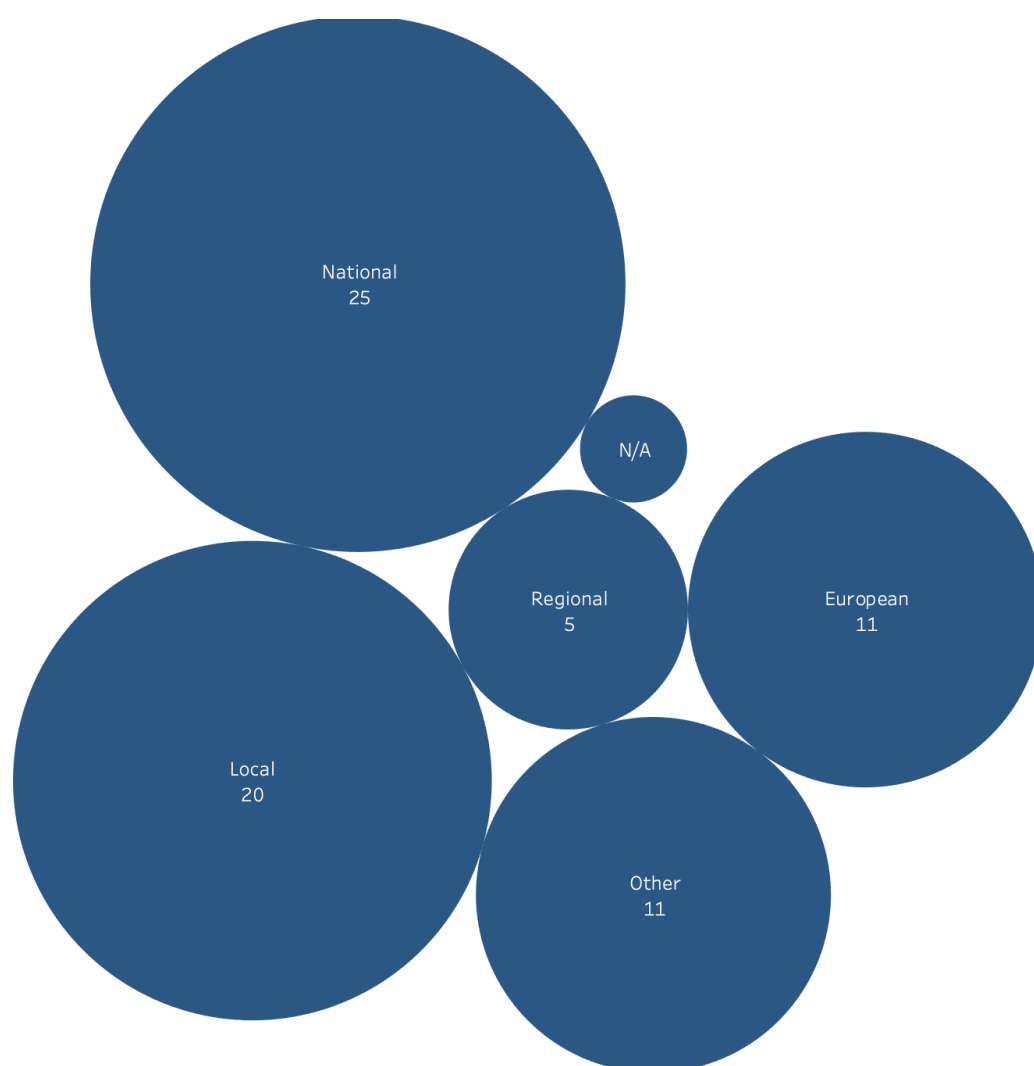
research methodology included a “researcher’s manual” with explanations on how the coding grill should be filled out. Also, during the consortium meetings dedicated to the research component, this issue was one of the topics of discussion and several coding examples were examined together. However, we believe that a complete harmonisation of data interpretation was not achieved, mainly because of the differences in language, cultural context and number of staff members and volunteers working on the research.

IV. Findings of the research

1. Romania

73 events were identified and analysed for the chapter of the research dedicated to the situation in Romania. These events generated hundreds of posts, the vast majority of them being Facebook posts and thousands of comments.

Diagram 1 – Localisation of the event



Hate is very seldom present in the actual posts on social networks, being more prevalent in the comments to the posts. This finding should, however, be taken with a grain of salt as the data collection methodology for Romania (using a software tool to retrieve content from Facebook pages and not people’s profiles) means that the researchers had access mostly to content produced by communication professionals, people who are either more concerned

with respecting deontological norms or are highly skilled at packaging hate in a way that protects them from the legal consequences of breaching national law on hate speech.

Interestingly, events taking place and covered by outlets or social media accounts with an audience in the historical region of Transylvania tended to generate more hate speech than the ones in other regions in Romania.

Posts describing events involving or concerning the LGBT community were more likely to get shared by a larger number of people than posts describing events relating to any other minority group (29% of the events relating to LGBT issues were each shared by more than 100 accounts - some of these were even shared by thousands of accounts, while only 11% of the Roma related posts and none of the posts related to the Hungarian or the Jewish minorities gathered the same amount of attention). This can be partially explained by the fact that both LGBT supporters and conservative groups opposing LGBT rights are well organised in Romania and, as such, tend to re-share content rapidly and in a coordinated manner once it is published. What was also interesting to note is that most of the people re-sharing content do so by just sharing the link to the articles and not writing any description of it. Other two consequences of the existence of these two very well-organised communities (the progressive one and the conservative one) are that (1) posts on LGBT issues were more likely to gather high rates of hate speech in the comments section than the ones targeting the Roma and (2) posts on LGBT issues were also more likely to contain counter or alternative narratives either by themselves or in the comments section.

The events that were associated with more hate speech towards the LGBT community had to do with actions that can be interpreted by conservatives as threats to the traditional way of life (gay marriages, gay adoptions, public acceptance of sexual minorities). When it came to the events relating to Roma ethnic minorities, the ones that generated most hate had to do with depictions of theft or acts of physical violence. Both of these findings are consistent with the existing stereotypes and prejudices in Romanian society with regards to these two groups vulnerable to hate speech. In the case of the hateful posts against the Hungarian minority, the main concerns expressed by the commentators were related to the possible discrimination of Romanian ethnic minorities in the areas where Hungarians represent a local majority. Also, many hateful posts referred to the supposed wish of Hungarians to take Transylvania away from Romania, a common stereotype regarding the Hungarian minority in Romania and Hungary itself as a country.

Another very interesting finding has to do with the localisation of the events that generated the most hate speech. In the case of the LGBT community, events taking place outside Romania (either in the EU or elsewhere) were more likely to generate hate speech than events taking place in Romania. This could be explained by the fact that conservative groups perceive the LGBT rights movement as being in an incipient stage right now in Romania and "imported" by pressure groups from outside the country. This is consistent with the narrative that there were no sexual minorities in the traditional Romanian society and that sexual minorities are invented by radical left-wing groups (called "neo-marxists" by Romanian conservative groups).

In general, but especially when members of the Roma community were involved, events taking place at the local level were more likely to generate comments containing hate speech than the ones taking place nationally or at European or global level. This is confirmed both by the rate of social media posts containing hate speech comments, and by the proportion of the hate speech comments to these posts.

What this suggests is that the proximity of the events to the readers strongly influence the likelihood that they would be emotionally triggered and would engage in hate speech. This is also supported by the fact that, although national events tended to generate slightly more posts containing hate speech than local events (see graph below), the commentators did not react to them as often as they did to the local ones.

Diagram 2 – Social network posts with hateful comments

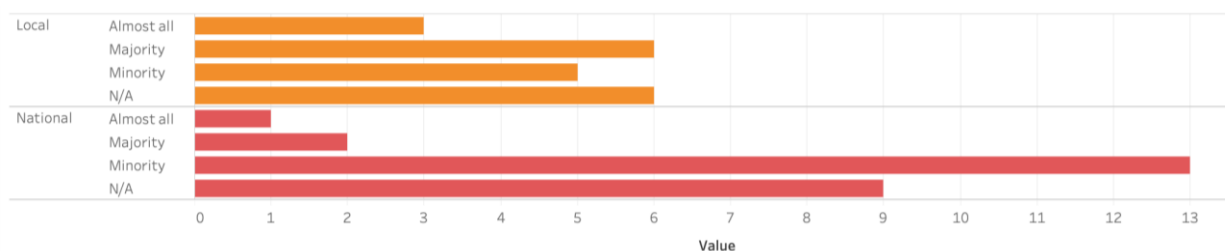
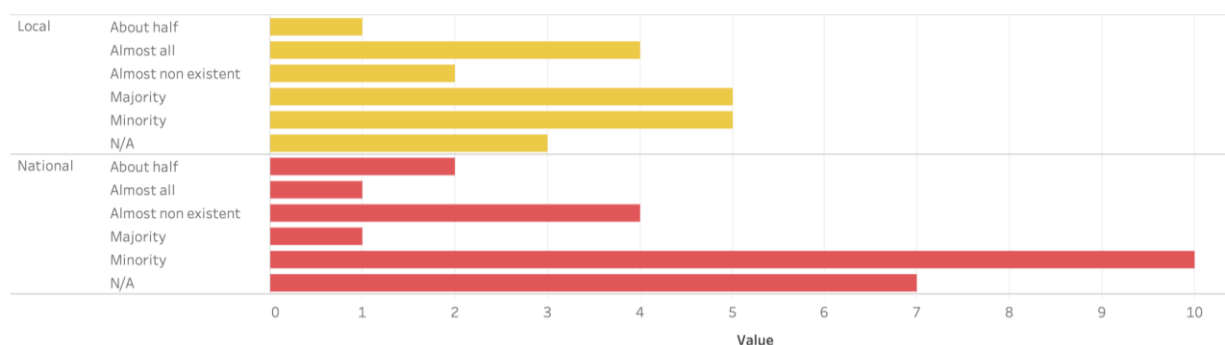


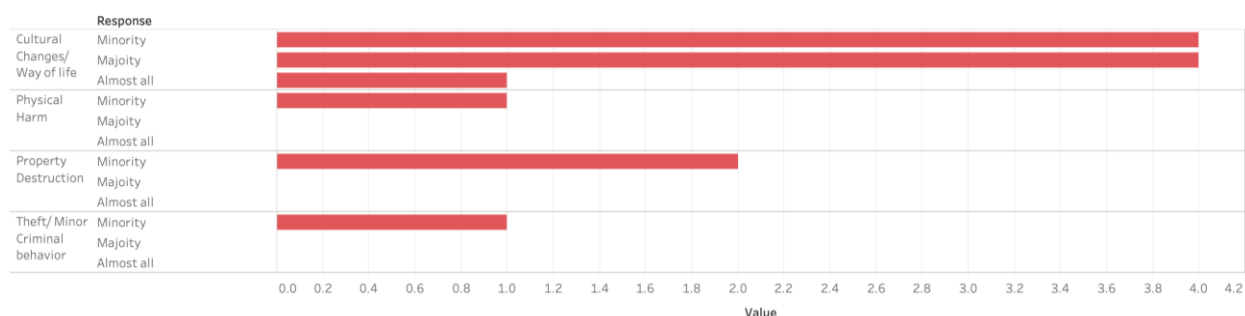
Diagram 3 – Rate of hateful comments on social network posts



The type of perceived potential threat that was associated with the highest levels of hate speech was that of theft, or minor criminal behaviours. Nevertheless, based on the data collected, 41% of the events associated with theft or minor criminal behaviours generated hateful posts in a majority of cases compared to 30% for events associated with cultural changes or way of life. However, all of the events associated with this type of threat involved members of the Roma community, so it is possible that it is not this threat itself that stimulates hateful messages, but rather that hate was expressed at the Roma community.

Also, from the point of view of the perceived threats associated with the events, the lowest levels of hate speech were correlated with perceived threats of cultural change or changes to the way of life. This was especially true when it came to the rate of hate speech in posts discussing these events. Not only that, but the highest levels of counter or alternative narratives were observed here. It is important to note that all of the events in this category had at their centre members of the LGBT community or of the Hungarian minority, and so, as mentioned above, it is possible that the real reason for the strong support to be the fact these communities and their supporters are more active online.

Diagram 4 – Posts with CAN messages based on incidence of perceived threats



Surprisingly, higher levels of hate speech were observed when the perceived threat that could potentially emanate from the event was singular, rather than when the events had the potential to generate two or more types of perceived threats.

Diagram 5 - Social network posts with hateful comments

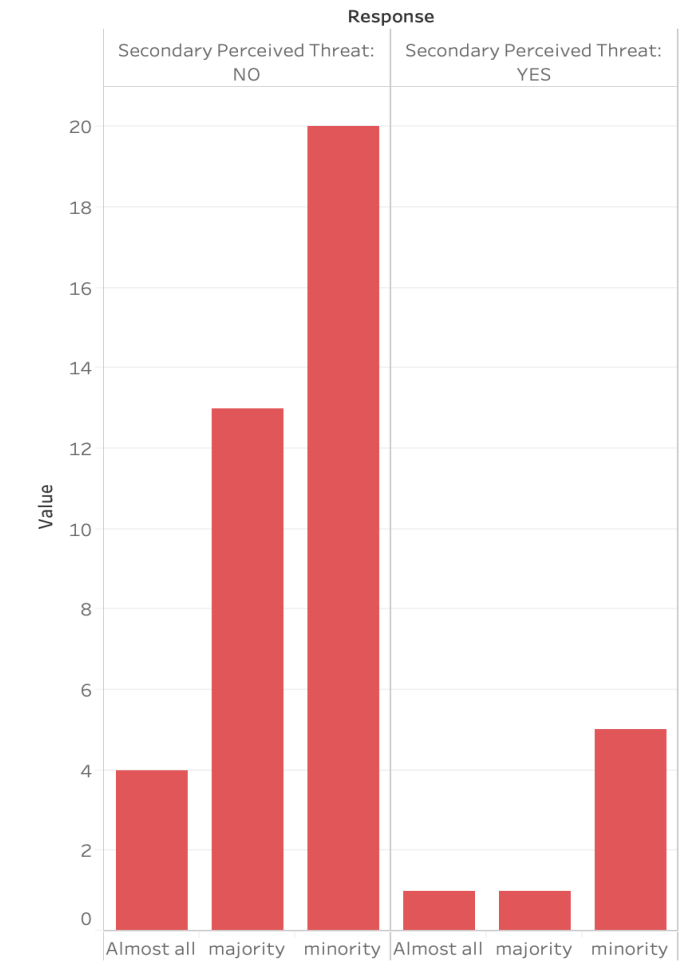
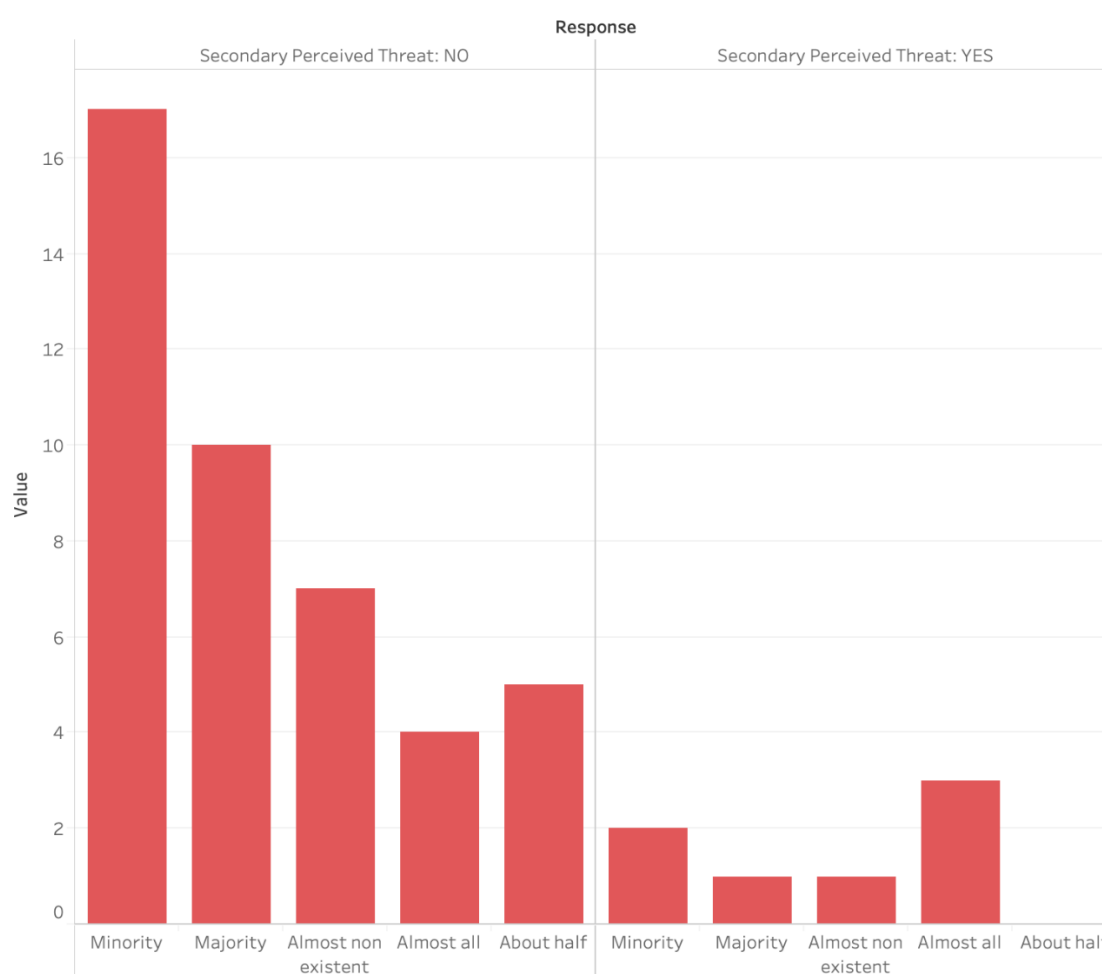


Diagram 6 – Rate of hateful comments on social network posts



Another surprise of the research was that the levels of hate speech were not influenced by whether an event involved concrete antisocial acts or not. This confirms once again that hate speech tends to be motivated not by facts but by subjective perceptions.

Interestingly, events taking place and covered by accounts with an audience in the historical region of Transylvania tended to generate more hate speech than the ones in other regions in Romania.

Recommendations

Romanian activists should pay special attention to events taking place at a local and national level, since they are more likely to engage people in hate speech online. This is particularly true for events taking place in cities and towns in Transylvania. However, the activists who are focused on supporting the LGBT community should channel their efforts to monitoring the online discourse around international events.

While still very few, counter and alternative narratives tend to be used more often by activists supporting the LGBT and the Hungarian minorities.

Given that hate speech tends to be more prevalent in the user generated comments to posts on social networks, the managers of these accounts should allocate more resources for content moderation activities.

2. Germany

The German language part of the research was done on a corpus of 15 events occurring mostly in spring 2020. These events generated multiple posts on social networks (mainly on Facebook and Twitter) and hundreds of comments. Most of the events were susceptible to primarily generate fear regarding cultural change or changes to the way of life in Germany (11 out of 15 events). Concrete antisocial acts or behaviours were observed in only one event out of the 15 we have analysed. This confirms once again that hate speech tends to be motivated not by facts, but by subjective perceptions, which also coincides with the changing political landscape as right-wing parties utilise factors such as fear and anger to gather support. Out of the 15 events, 9 had the potential to generate hate speech against people based on their ethnicity and 5 based on their religion.

One third of the events concerned refugees, a disproportionately large number that could point to the vulnerability of the public perception of asylum seekers in German society, meaning that the opinion of the general public as regards refugees and asylum seekers can, seemingly, easily be swayed by populist rhetoric. Two thirds of the events involved minorities, either historical or more recent. Almost two thirds of the events happened at a national level (9 out of 15) and almost one third at a European level (4 out of 15).

Hate speech was more frequently found in comments on social media posts than in the posts themselves. While the discrepancy between the quantity of hate generated by the people writing the posts and the one generated by their audiences is smaller than in the case of Romania (this difference being explained by the data collection method being used), it is still a strong indicator that people tend to pay more attention to what they publish on their pages or profiles than what they write in comments online. This also supports our general recommendation that page managers should invest more time and energy in moderating the comments on the pages they manage or on their profiles.

The events taking place at a national level tended to generate more hate speech, which makes sense insofar that it might affect users in Germany directly whereas there still is a widespread disconnect to the European level. This is, for example, visible in the low voter turnout in European elections, but might also be related to the aftermath of the financial crisis and the subsequent media coverage on Germany's responsibility to support its European partners, which first allowed Eurosceptic parties and groups to gain traction. This was observed both in regards to hate expressed in the posts themselves (hate speech was expressed in the majority of the posts associated with 4 out of 5 events taking place at a national level) and hate speech in the comments to these posts (almost all comments or a majority of comments to posts regarding 7 out of 9 national events represented hate speech). Yet, counter and alternative narratives tended to be used less frequently in relation to national events than to European ones.

Surprisingly, the perceived threat of cultural change or changes to the way of life was a better predictor to the occurrence of hate speech than perceived threats of a more concrete nature. This might be explained by the continuous spreading and reiterating of fears, but also conspiracy theories, by right-wing politicians and influencers, warning of an increasing "Islamisation" and the threat of new cultures replacing "German traditions".

Hate speech was found in either all or the majority of the posts of 5 out of 11 events having the potential to generate fear of cultural changes or the way of life. Only 1 out of the 4 events associated either with fear of physical harm or fear of theft led to the majority of the posts associated with it to contain hate speech. Not only that, but for these types of events, the rate of hateful comments to the posts describing them tended to be very high.

Diagram 7 – Hateful user generated posts on social networks

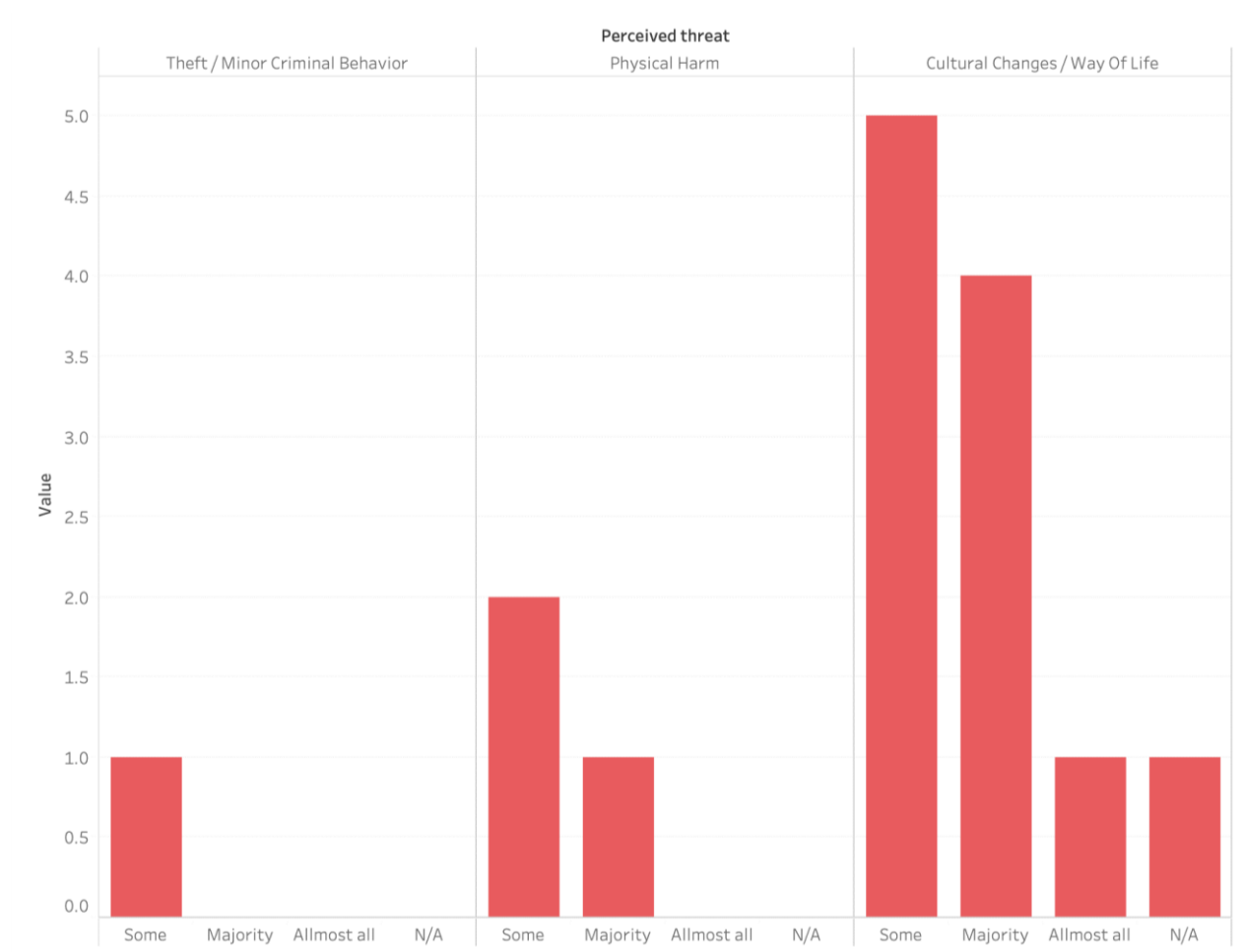
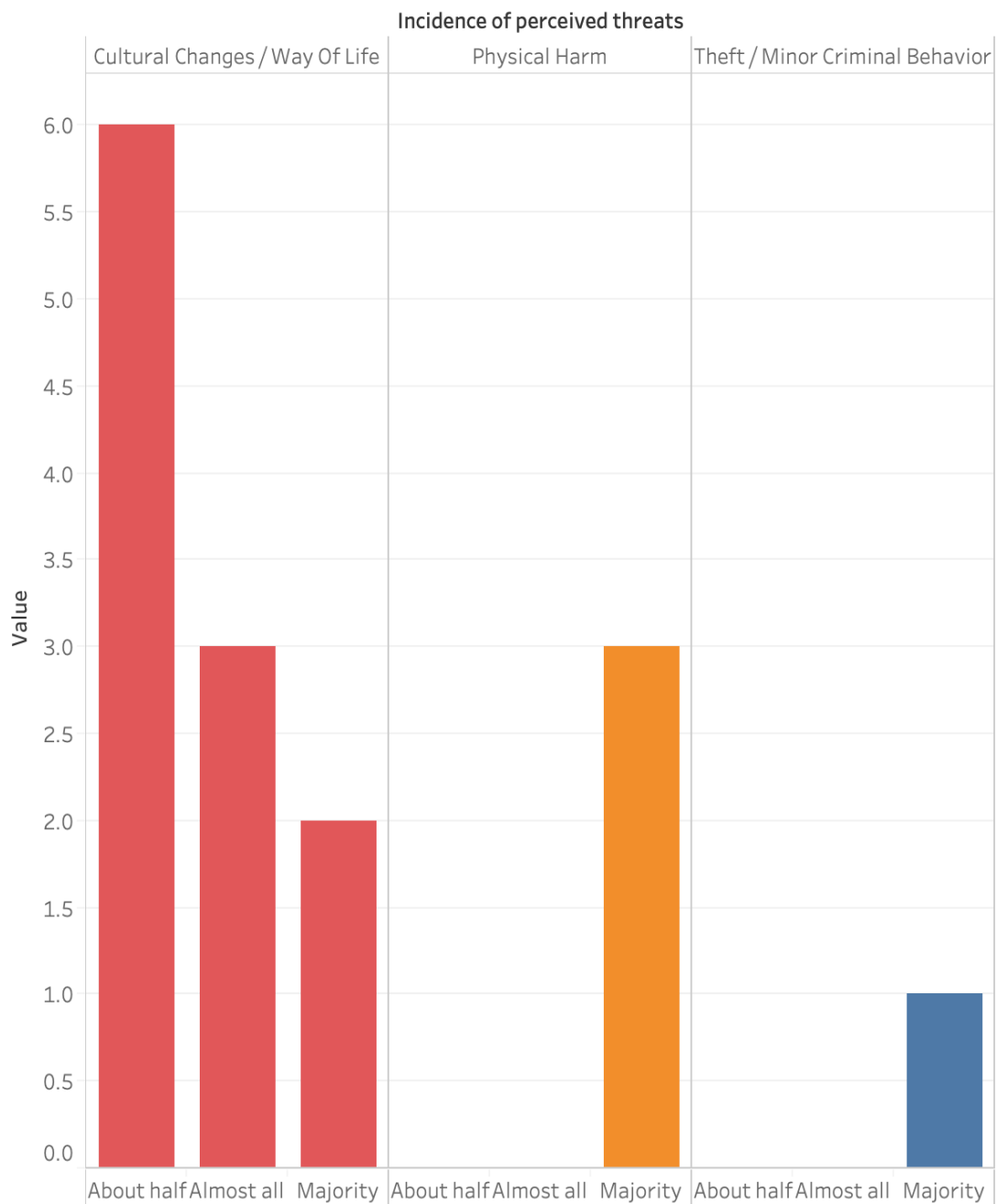
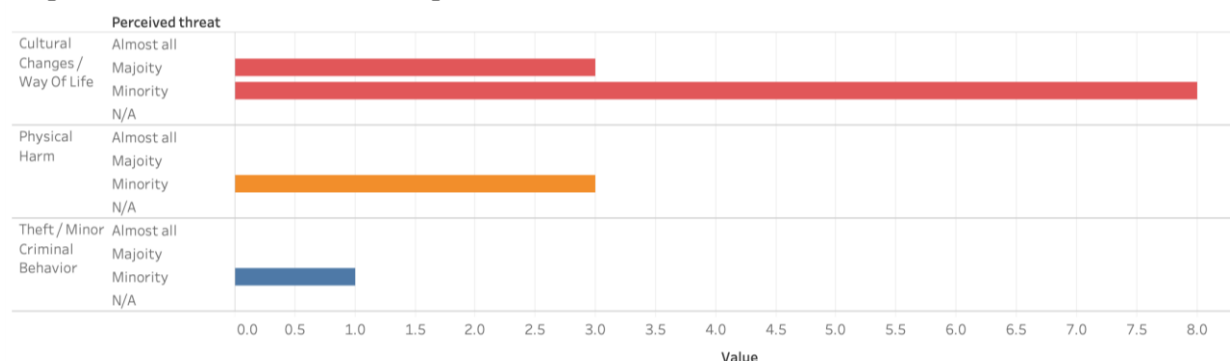


Diagram 8 – Rate of hateful comments on social network posts



However, it is also worth mentioning that counter and alternative narrative strategies were used more often for online content depicting events that could make people feel their way of life or culture would be challenged.

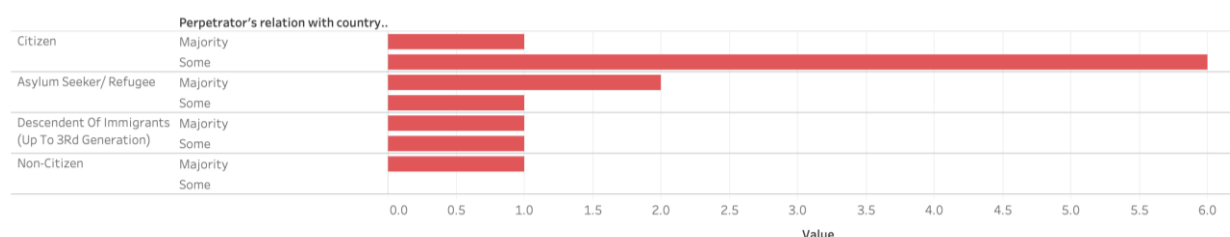
Diagram 9 – Posts with CAN messages



Events that could consolidate multiple types of fear were slightly more prone to see a more widespread occurrence of hate speech in the comments section of the posts relating to them when compared to events with the potential to consolidate only one type of fear. Almost all, or a majority, of the social network posts related to 7 out of 9 events associated with at least two types of threats contained hate speech comments, compared to 4 out of 6 events associated with one type of threat.

Events involving refugees and asylum seekers generated higher rates of posts containing hate speech and generating hateful comments than events involving citizens, non-citizens and descendants of immigrants.

Diagram 10 – Hateful user generated posts on social networks based on relation with country where the event took place



An explanation for this might be the increased media coverage on the influx of refugees in the EU, and more specifically in Germany, as well as the depiction of refugees and asylum seekers in the media as reports, more often than not, focus on negative news whereas news on other minority groups, while also triggering hate speech, are less visible to the general public. Consequently, hate speech targeting refugees also seemed to mobilise more people to up speak against hate.

If the people involved in the events were described based on two or more characteristics (such as being both an ethnic and religious minority at the same time), higher occurrences of hate speech were observed. At the same time, more counter and alternative narratives were also observed to be used in the defence of people targeted by hate speech, even though such efforts were sparse.

Diagram 11 – Social network posts with hateful comments based on perceived double characteristics by perpetrator

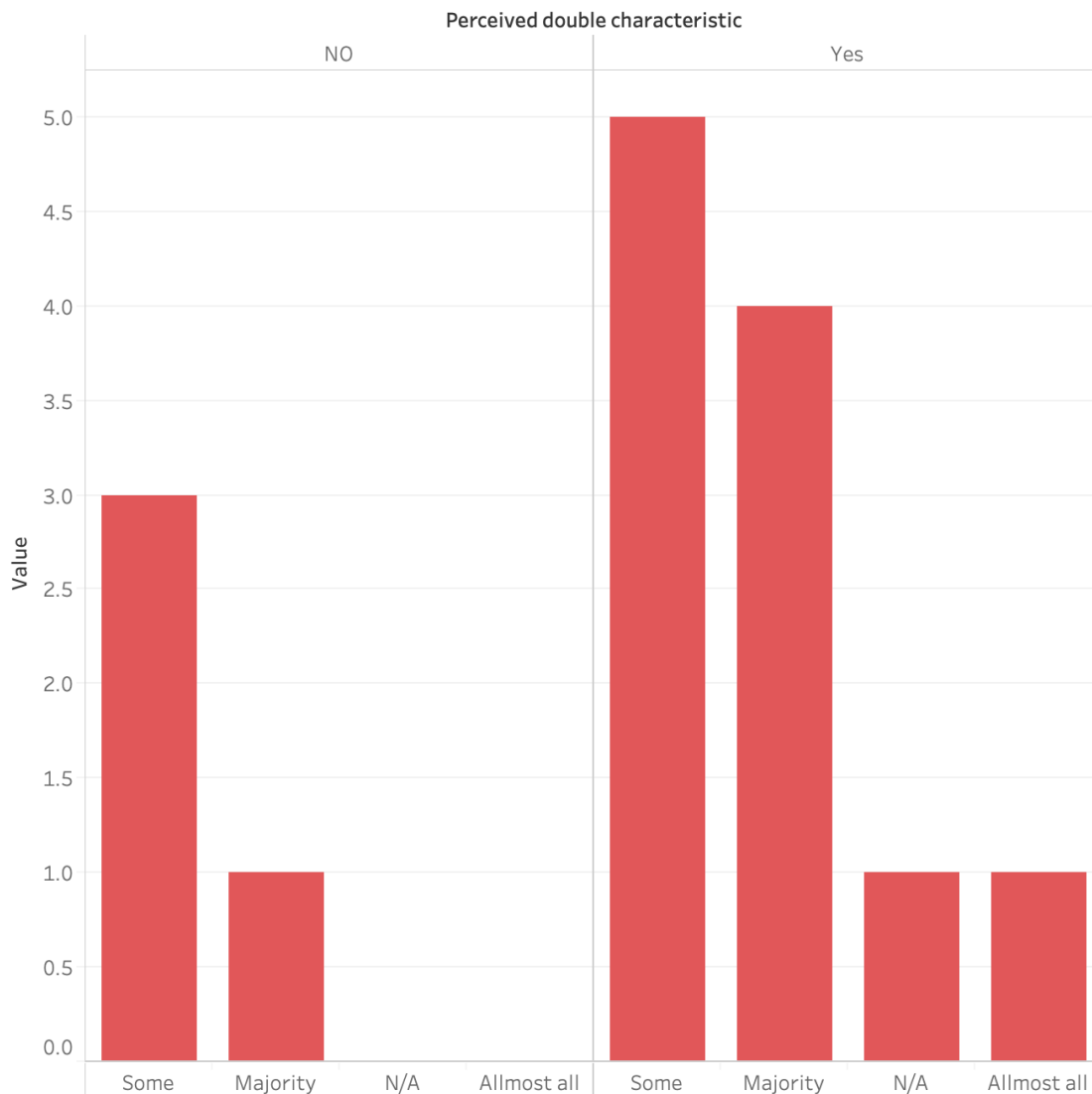
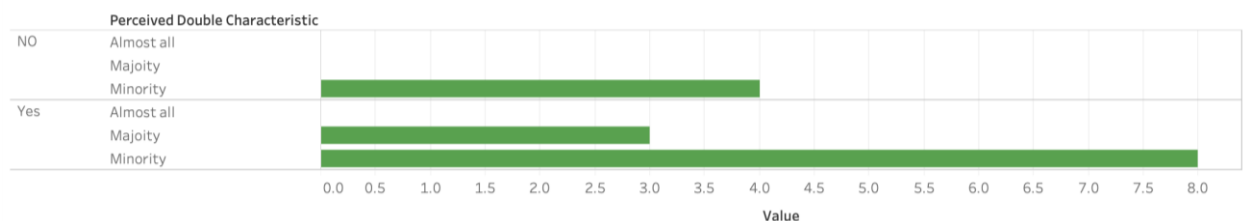


Diagram 12 – Posts with CAN Messages



Another strong predictor for the occurrence of hate speech has been the minority status of the people involved in the event. Events involving people belonging to recent minorities have generated higher proportions of hateful posts; however, this might also be due to the increased reporting on recent minorities. Also, posts discussing these events were marginally more likely to attract comments containing hate speech.

Diagram 13 – Hateful user generated posts on social networks

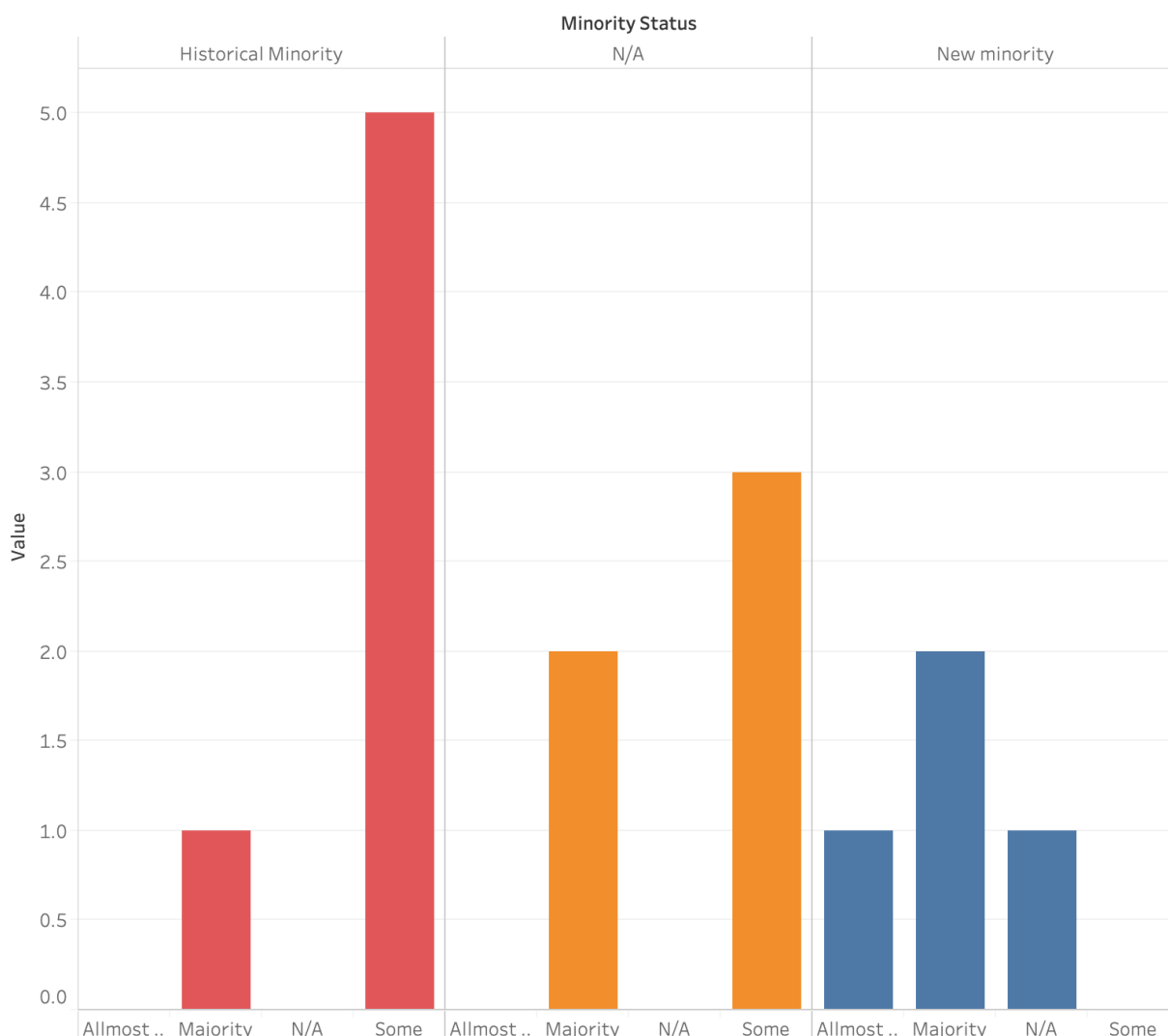
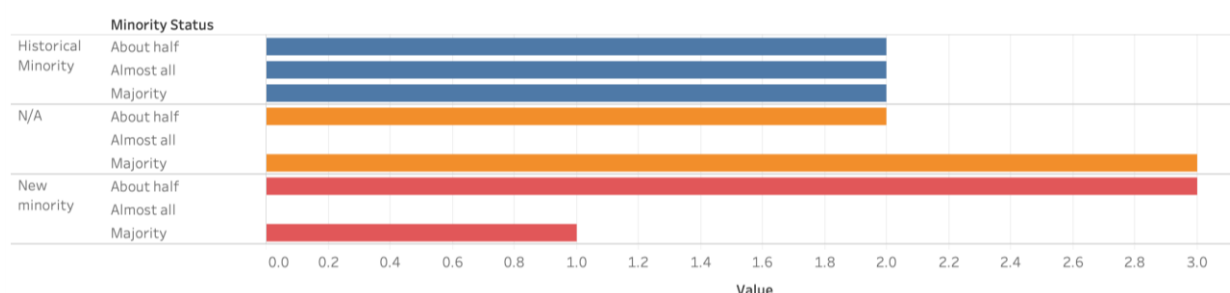


Diagram 14 – Rate of hateful comments on social network posts



It is also important to mention that reports on Roma minorities also triggered widespread hate speech as anti-Roma discrimination is not only widespread but almost internalised by German society. Consequently, efforts aiming at defending this minority or countering anti-Roma hate speech are scarce.

However, where present, hateful comments represented a smaller proportion from the total number of comments when compared to posts discussing events involving people belonging to historical minorities.

As far as the consequences of the involvement of public figures in generating hate speech is concerned, two conclusions can be drawn. As expected, events discussed by extremist

influencers have led to higher numbers of hate speech posts compared to events discussed by other types of public figures. The surprise comes from the events where media or entertainment public figures have engaged in hate speech: posts reporting or discussing these events were more likely to attract hateful comments. This is in line with a current societal debate on political correctness in Germany as efforts of marginalized groups, for example as regards renaming public places and streets, are diminished and criticised - a development visible (and equally criticised by more conservative actors) across Europe in light of the Black Lives Matter movement.

Diagram 15 – Hateful user generated posts on social networks

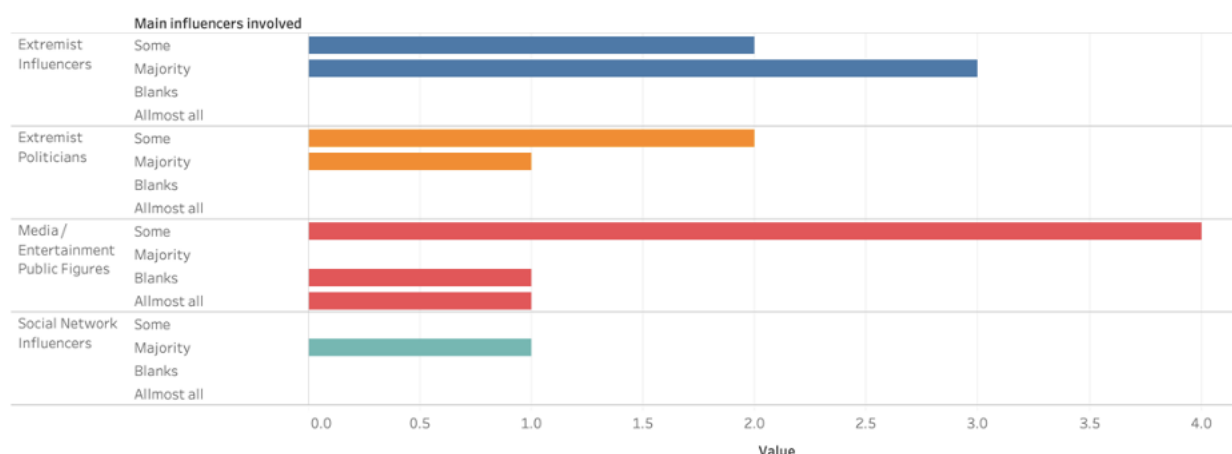
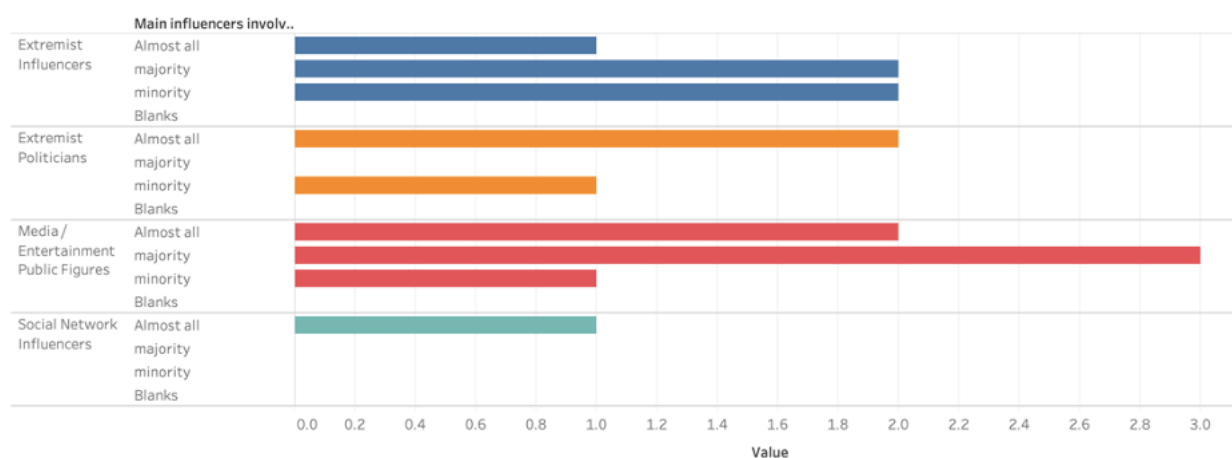


Diagram 16 – Social network posts with hateful comments



Conclusions

German-speaking activists involved in fighting hate speech should expect higher volumes of hate speech from events taking place at a national level, involving refugees, asylum seekers or people depicted as belonging to more than one group vulnerable to hate speech. Particular attention should be paid to events that can consolidate the fear of threats to the current way of life in Germany or of cultural changes. This development or status has been a rather long time in the making as movements and parties exploited fears of the general public in light of the increasing influx of refugees in Europe since 2015 as well as the New Year's Eve sexual assaults in Germany (2015-2016). The spread of racist stereotypes and speech targeting refugees has further been aided by the media adopting the speech of populist politicians, repeating their points of view without classifying it. Furthermore, media offices neither have the financial or personnel resources to moderate their social media postings, allowing often for the unopposed spread of hate speech.

It is also worth highlighting that, although counter and alternative messages are very rarely seen, they tend to occur where most needed, being rather well correlated to the situations in

which hate speech was most likely to manifest. Yet, the outlook is optimistic as more and more media offices and page managers increase their efforts on moderating and community management. Similarly, activists are increasingly organised in order to respond to hate speech; for example, the activist group #ichbinhier organises counter speech efforts on a daily basis on Facebook, getting actively involved in debates that already triggered a lot of hate speech, to allow for more balanced perspectives in the commenting sections.

3. Italy

For the Italian part of the research, 21 events were included in the corpus, collected between March and September 2020. The events were discussed in hundreds of posts on social networks (mainly on Facebook and Twitter), which generated thousands of comments.

In line with the findings on the situation in Romania and Germany, the most common type of threat that these events could generate was fear about changes to the current way of life in Italy or fear of cultural changes (16 out of 21 events). Other types of threats evoked by the events were fear of property destruction (2 events), fear of theft / other minor criminal behaviour (2 events) and fear of physical harm (1 event).

Concrete antisocial acts carried out by the people vulnerable to hate speech involved in the events were observed in only 3 out of the 21 events. Reactions in the form of hate speech were more frequently found for these events than for those which did not end up in concrete antisocial actions.

The most common types of perceived characteristics of the people involved in the events were race (8 events) and ethnicity (5 events), followed by gender (4 events), sexual orientation (2 events) and age (1 event). Almost all events (18 out of 21) involved Italian citizens. 1 event involved asylum seekers, 1 involved an immigrant and 1 involved a non-citizen.

Almost all of the events happened in Italy, either at a national level (14 events) or at a local level (4 events). One of the events included in the research happened in the USA (the murder of George Floyd by police officers which sparked the Black Lives Matter protests).

In strong contradiction to the findings in Romania and Germany, hate speech was frequently found in the posts discussing the events, as in the comments to these posts. Not only that, but there was also a very strong, almost 1 to 1 correlation between the rate of hateful posts associated with an event and the rate of posts containing hateful comments associated with that same event.

Diagram 17 – Social network posts with hateful comments

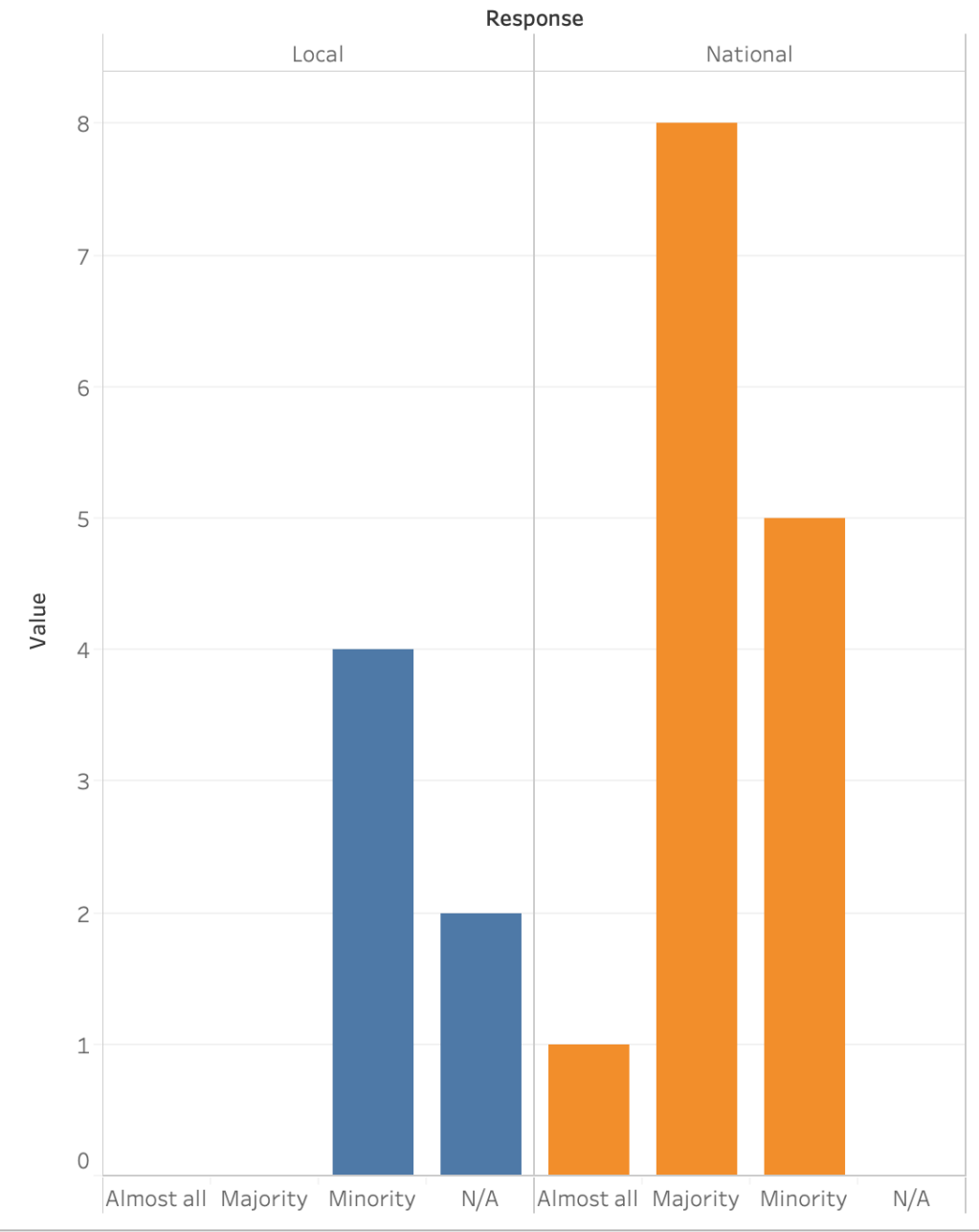
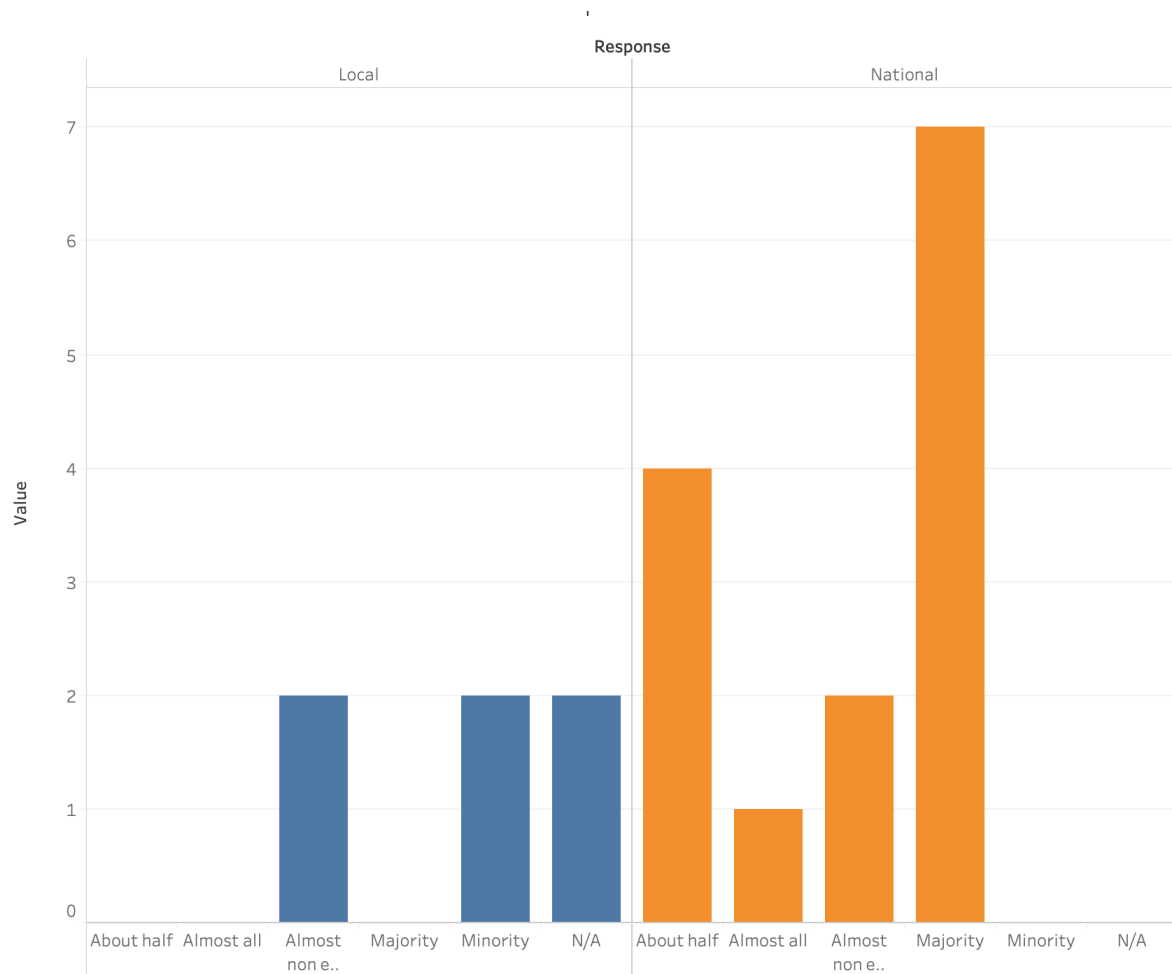


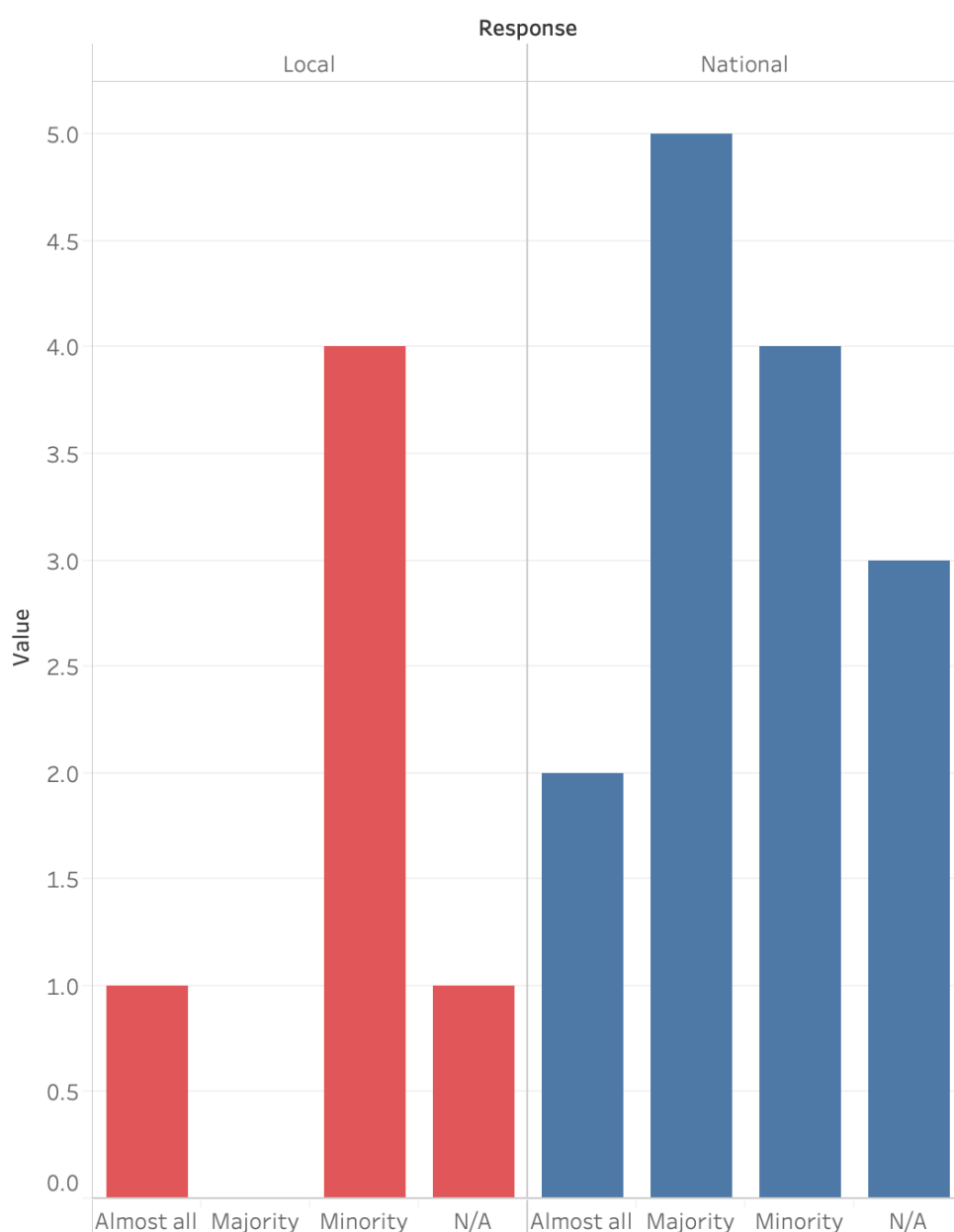
Diagram 18 – Rate of hateful comments on social network posts



Significantly more hate speech was generated by events taking place at the national level than at the local one. This was observed both in what regards hate expressed in the posts themselves (hate speech was expressed in the majority or in all of the posts associated with 8 out of 14 events taking place at a national level, compared to none in the case the events taking place locally) and hate speech in the comments to these posts (almost all comments or a majority of comments to posts regarding 9 out of 14 national events represented hate speech, again compared to none in the case of events taking place locally).

However, national level events also attracted more counter and alternative messages aimed at stopping them or at offering alternatives to them.

Diagram 19 – Posts with CAN messages

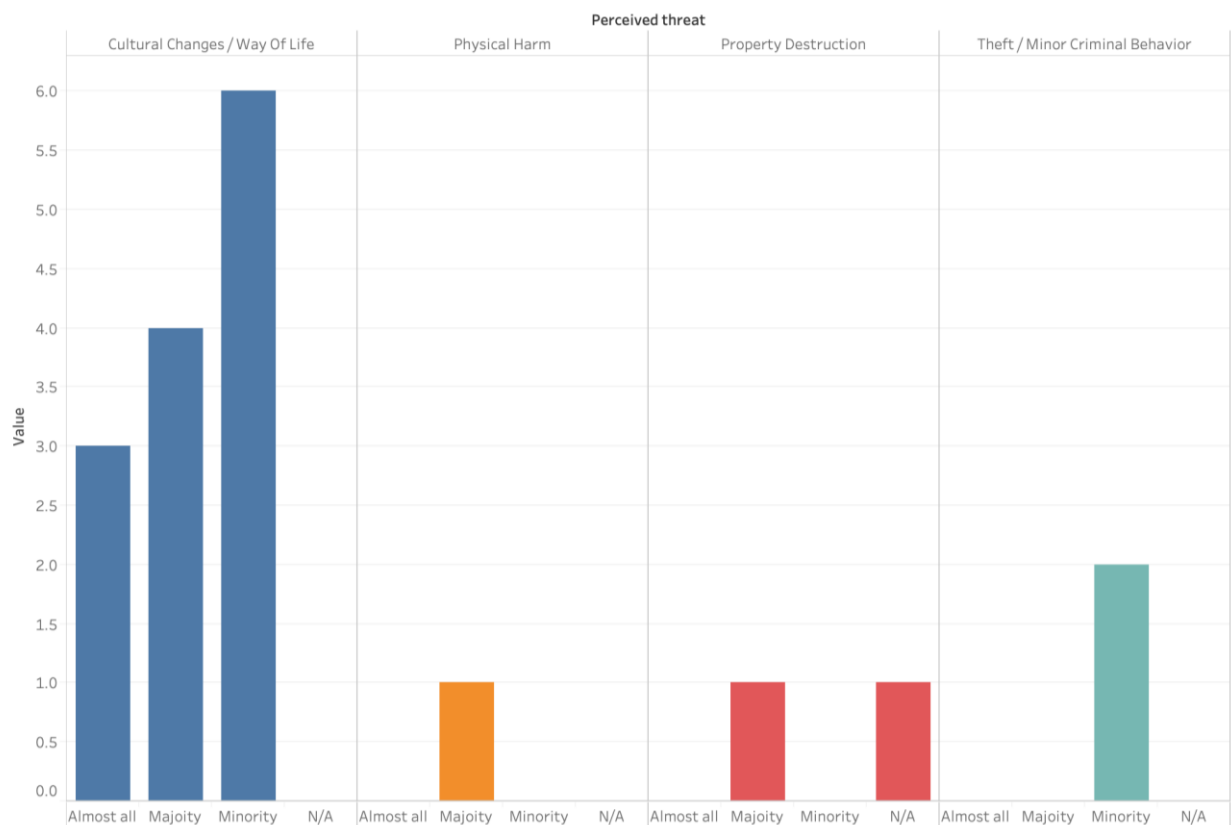


This may be explained by the strong polarisation that is now present in the public discourse in Italy, where some relevant actors seem to voluntarily intercept certain events to strengthen their position, either for negative purposes or to spread positive messages. People are often looking at these influential peoples' (journalists, opinion leaders, media people, etc.) comments on articles and/or news items rather than at the event itself.

Events instilling a fear of theft or of other minor criminal behaviours were more likely to generate higher rates of posts containing hate speech, and these posts were marginally more likely to attract hate speech comments, or that there would be more hate speech comments than non-hate speech ones.

Events likely to generate a fear of cultural changes or of changes to the way of life in Italy were significantly more likely to attract counter and alternative narratives.

Diagram 20 – Posts with CAN messages based on perceived threats



Higher rates of hateful posts were observed around events involving people prone to be victims of hate speech based on their race and gender. In 3 out of 4 events involving women, the majority of the posts had attracted at least one misogynistic comment. However, the rates of hate speech comments to regular comments were higher in the events involving people susceptible to be targeted based on their race. Events involving persons that could be targeted by hate speech based on their ethnicity attracted marginally more consistent counter and alternative messages.

Diagram 21 – Rate of hateful comments on social network posts based on perceived characteristics

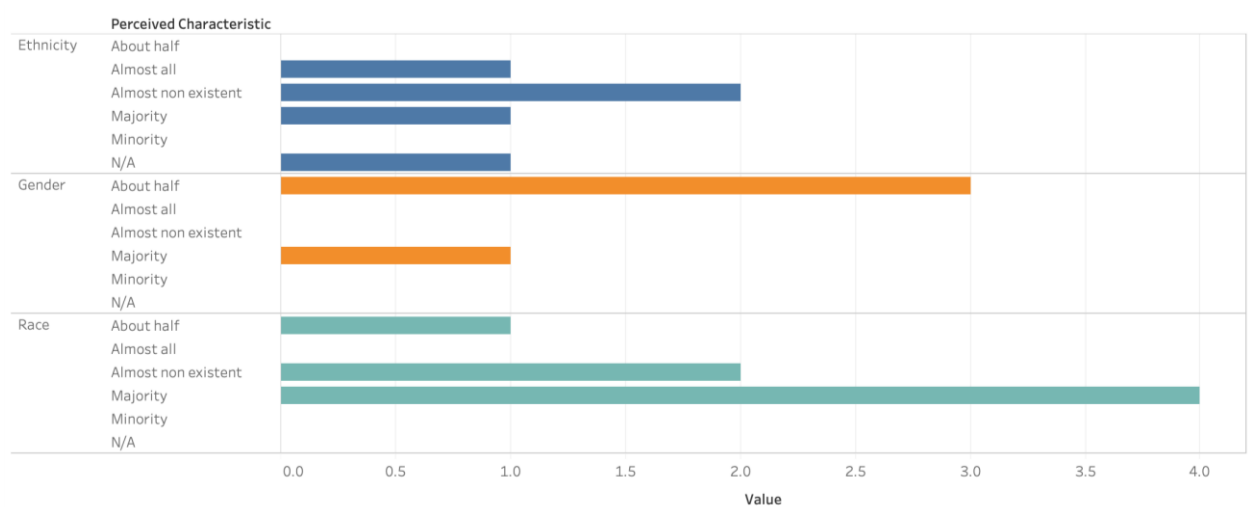
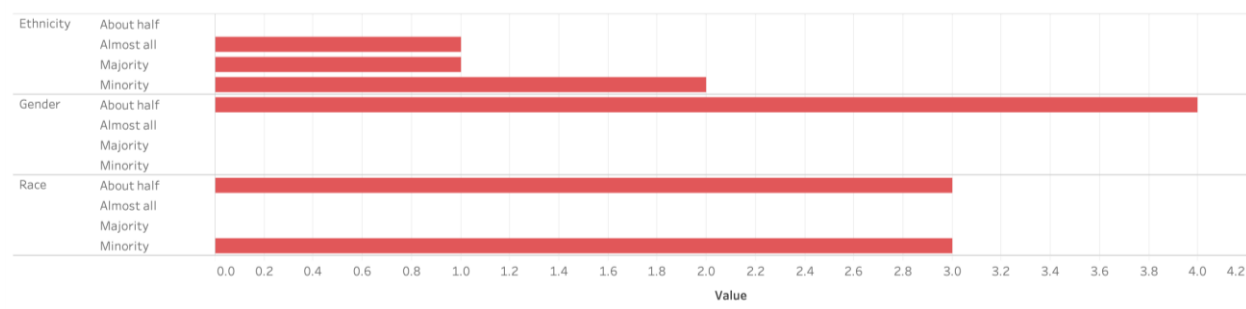
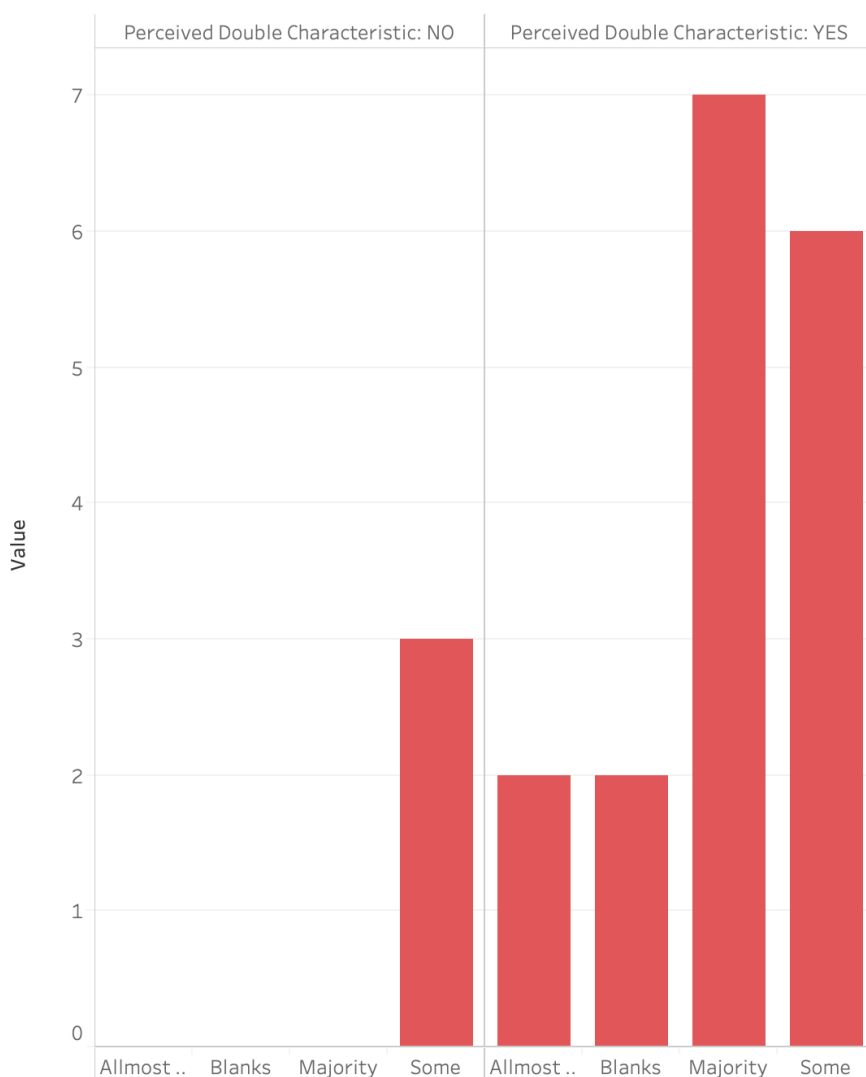


Diagram 22 – Rate of CAN messages in posts where CAN is present based on perceived characteristics



One of the two major predictors of the occurrence of hate speech was whether the people at the centre of the events could be targeted by hate speech based on more than one characteristic or not. All the indicators associated with the occurrence of hate speech in relation to events involving persons identified by two or more characteristics showed worse results when compared to events involving people with just one trait vulnerable to hate speech.

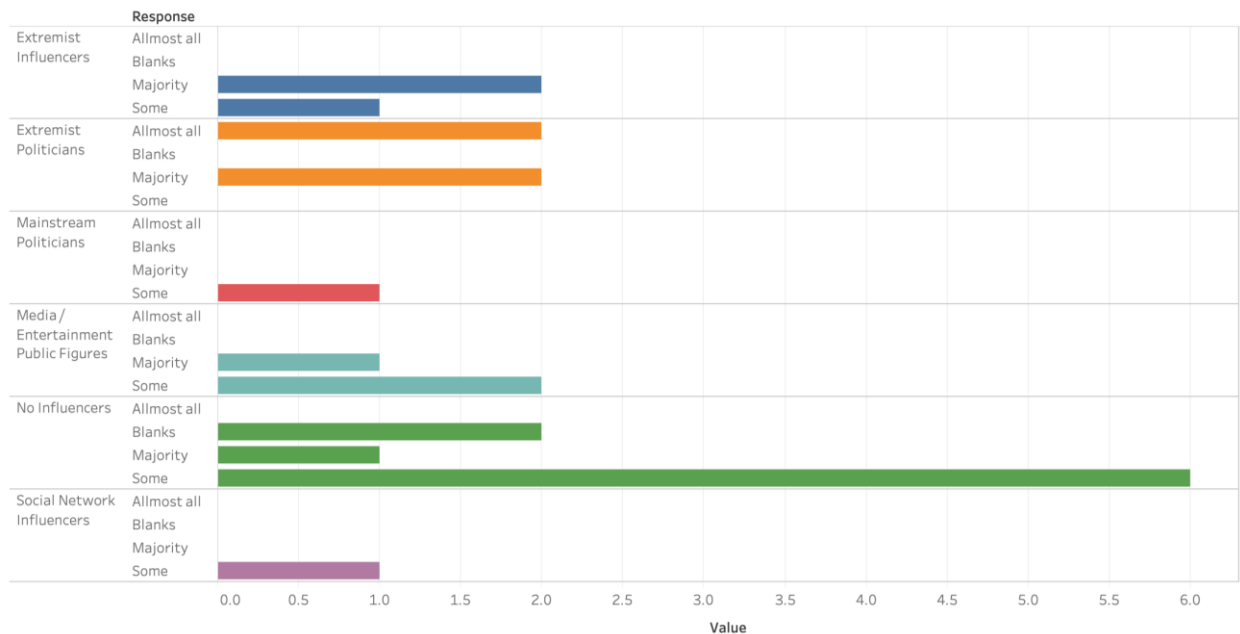
Diagram 23 – Hateful user generated posts on social networks



The other major indicator was the existence of more types of public figures involved in generating hate speech. In fact, the smallest levels of hate speech were noticed when no public figures were involved in generating hate speech. The presence of just one type of

public figure generating hate speech worsened the situation slightly, with worse results observed, predictably, when extremist influencers or politicians expressed their views regarding the events.

Diagram 24 – Social network posts with hateful comments



Conclusions

Italian-speaking activists involved in fighting hate speech should pay particular attention to events that either attract more types of public figures to express hateful comments or that have at their centre persons vulnerable to hate speech based on more than one characteristic. Other types of events that pose more risks of generating hate speech are those happening at a national level, those which generate fears of theft or minor criminal behaviour as well as those that involve concrete antisocial actions.

Counter and alternative narratives are less frequent in case of events that could generate fears of concrete, direct threats and become more frequent to combat threats of cultural changes or changes to the way of life.

V. Overall conclusions of the research

The analysis carried out on the events taken place in Germany, Italy and Romania resulted in few common findings. This holds true even when comparing any of the two countries side by side. This meant that the research was not able to identify any “model” that would predict, at a European level, which events would result in hate speech, nor was it able to do this by grouping the three countries on grounds of cultural similarities, religious majority or geo-political localization.

This result illustrates the importance of national context in terms of identifying hate waves, proving that there are no “one size fits all” solutions that can be employed by activists. While on the one hand, this can further complicate the work of activists, on the other hand, by comprehensively engaging with their national contexts, activists will be better equipped to respond to hate storms.

Therefore, a key finding of this research has to do with the primacy of culture when analysing hate-speech related phenomena. Activists must then remember to always adapt the materials and recommendations they may come across to their local and/or national contexts. This will ensure both better analyses and responses to online hate waves.

One finding observed in all countries was that the most common types of threats that could be perceived by people reading about the events was that of cultural changes or changes to the current way of life in the three countries. This shows once again that hate speech does not require to be fuelled by concrete violent or antisocial acts. Thus, in the cases analysed it appears that hate speech is often motivated by subjective perceptions and/or beliefs. This finding also further demonstrates the importance activists should pay to local/national context and the general social landscape within these. By analysing the general perception of the public regarding certain social groups, activists can be one step ahead of hate storms.

However, with the exception of Germany, the events with potential to instil fear of cultural changes or of changes to the existing way of life, generated lower levels of hate speech and even attracted the most consistent efforts aimed at showing support for victims of hate speech or at combating the hateful narratives (a trend that was noticed for all three countries).

Another conclusion that can be drawn from the research is that events happening at a national or local level tend to attract more hate speech than international events. This highlights the fact that people might react more strongly to events that they perceive as closer to their surroundings. Conversely, events taking place at the European or international level seem to spark less hateful reactions. This might be due to a perception of detachment from events which do not directly influence one’s life. The only exceptions to this trend are events involving the LGBTI+ community featured by Romanian social media accounts (which generated more hate speech when taking place outside Romania). Nevertheless, even in this context, this finding is strongly linked to the perception of the Romanian public regarding the LGBTI+ community.

Lessons learnt and recommendations for future research

Therefore, we encourage activists looking to identify early warning signs regarding online hate speech to apply the research methodology (links below), to their national context.

Nevertheless, from our experience we also recommend that activists apply this methodology to a higher number of events and cases. While we were able to draw conclusions from our research, which in Italy and Germany specifically was based on a small number of events, we believe that including more events in the analysis will provide more accurate results. Moreover, even if the results turn out to be similar, applying the methodology to a greater number of cases will surely increase the legitimacy of the results and the actions that will follow the research.

It is important, however, to note that it is not an easy feat to conduct this type of research. Even when applying the methodology to a limited number of cases, the process can be tedious and time consuming. Thus, we recommend that activists create a data base which they can constantly update with events (and the ensuing hate responses where these appear) that spark online hatred. For activists who already monitor online hate-speech, this would imply coding these events when they come across them. This might allow for more data to be collected on a constant basis, hence providing more nuanced results.

Just as importantly, we do not recommend that this type of research be viewed as a one-person job. CSOs should train multiple members of staff or volunteers on how to collect this type of data. By doing this, more people could contribute to the data base meaning that more data would be collected. Collaboration is thus highly crucial and so CSOs should make use of their activist networks.

Another lesson that resulted from conducting this research has to do with the manner in which the differences in language, cultural context and number of people working on the research is going to influence the harmonization of the data interpretation. Hence, we recommend that activists pay special attention to the national/local culture and language when coding these events and fill in the information in the coding grid in a nuanced manner.

Lastly, in terms of similar research projects that could be conducted in the future, we recommend that activists increase their research efforts both at the national and European level. In this sense, at the national level what could be very interesting in terms of research results would be if CSOs from different regions of their countries could collaborate in terms of documenting more local events that spark hatred. This would provide useful insights both in terms of increasing data gathering but also in terms of local context, which in this scenario would imply that the data would be coded in a very nuanced and contextual manner.

At the European level, comparative national research could also provide interesting observations. This type of research should be conducted in the context of similar events taking place in multiple European states. Examples of such events are European Elections, national elections (when these occur during a similar period), European or regional summits, meetings, or conferences on different topics (e.g. climate change, human rights, COVID-19, finance and economy just to name a few). Such comparative research could help activists better understand why certain events spark more hatred in some countries while the same events do not produce similar reactions in others. Moreover, such research could also give hints about an indexation of national triggers, which could afterwards better inform activists' campaigns and responses. Other aspects that could be documented and compared are, for example, media involvement and manner of reporting as comparing media biases might also help explain the manner in which the general public perceives a certain event. The same can be said about political rhetoric, CSOs' campaigns and other public communicators' discourses.

As our research illustrated, hate waves are very complicated and complex phenomena that vary greatly from country to country. However, researching hateful responses to European and/or regional events at the national level might also lead to some common findings since in this scenario there would be a common denominator. Regardless, this type of research would surely benefit activists and further spark pan-European collaboration on the topic of hate-speech.

[Methodology for research into events most likely to generate waves of hate speech](#)

[Link to the coding grill](#)

[Link to the corpus collection table](#)

Appendix 1: Further information on the members of the consortium for the “WE CAN for human rights speech” project

The **No Hate Speech and Co-operation Unit** of the Council of Europe supports member states and NGOs to address hate speech, hate crime and discrimination through a range of co-operation projects.

ActiveWatch is member of the No Hate Speech Movement (NHSM) national committee in Romania and has gained extensive experience with monitoring hate speech. Their annual national report on hate speech provides qualitative content analysis of the hate narratives and its potential impact on the state of human rights and democracy in the country.

Agenzia di Promozione Integrata per i Cittadini in Europa (APICE) is the national coordinating organisation of the NHSM national committee in Italy. Since the publication of ‘We CAN!’, they have led three successful European trainings based on the manual and several national trainings, among others funded via Erasmus+. Together with the organisations of the national committee, Amnesty International Italy and others, they have rolled out online counter narrative campaigns responding to hate incidents.

Neue Deutsche Medienmacher e.V. (NdM), is the coordinating organisation of the NHSM national committee in Germany. NdM developed expertise in design of long-term counter narrative strategies. With a successful media strategy in place the organisation also developed expertise in responding to hate comments on their channels.

CEJI - A Jewish Contribution to an Inclusive Europe stands with people of all backgrounds to promote a Europe of diversity and respect. With over 25 years of experience in the field of anti-discrimination education and advocacy, CEJI offers specific trainings on unconscious bias, religious diversity, overcoming antisemitism and overcoming islamophobia. Through the Facing Facts project series, that aims to improve responses to hate crime and hate speech in Europe, the staff developed a unique expertise in facilitating multi-stakeholder co-operation across a diverse range of communities.

Appendix 2: Research on hate speech going viral -data collection methodology-

1. Research aims

- identify the criteria that should be taken into consideration when mapping the types of situations that tend to generate more hate speech online
- quantify the predominance (occurrence factor) that each identified criterion (or complex of criteria) has on the probability that a certain situation would lead to hate speech going viral
- try to identify early warning signals with regards to events or situations that tend to generate more hate speech online

2. Acquiring the corpus for the research

National press will be scanned daily in order to identify **current** events that could be included in our research. In addition to this, a google news search will be performed daily, restricted to news from the partner's country. Keywords indicating groups vulnerable to hate speech will be used (e.g. Muslim, gay, refugee etc.). News published over month ago from the date of the search will be discarded.

We shall also collect data on the occurrence of the events on social networks. If partners have access to a scrapping service, they are encouraged to use it. If not, the search functionalities of the dominant social networks in each of our countries can be used.

3. Coding process

Each of the events included in the corpus of the research will be coded in the coding grill annexed to this methodology. Each event will be reserved a separate line in the coding grill. The coding grill is composed of the following items:

1. Description of the event (max 200 characters) – a brief description of the event, in English, limited to 200 characters. Information should be given as to what happened, where it happened, who was involved

2. Date of occurrence – either a single date or a timeframe in which the events took place

3. Localization of the event – a dropdown menu. Partners must select where the event took place (local level, regional level, national level, European level, other). An example of a local level event would be a rape committed in a county in Romania; a European level event would be the refugee crisis of 2016.

4. Perceived threat (dominant) – a dropdown menu consisting of: physical harm, property destruction, theft / minor criminal behaviour, cultural changes / way of life. This section will be filled with the dominant threat that is expressed in the media articles / reader's opinions regarding the event.

5. Perceived threat (secondary) – same as the above, but the second most common threat should be listed.

6. Materialized threat – a dropdown menu consisting of: verbal violence, psychological violence, physical violence, murder, terrorist attack, rape, theft, other antisocial behaviour. This section will be filled out only if the event had actual implications in real life on a general population / people that can be directly harmed by it.

7. Number of victims – a dropdown menu consisting of an approximation of the number of victims. Partners should use the information offered in the media to assess the number of victims directly affected by the event.

8 Victim type (1) – a dropdown menu consisting of: women, children, men, multiple. This is to assess a first level of emotional engagement based on sensibilities relating to traditional family roles.

9. Victims type (2) – a dropdown menu consisting of military / law enforcement, politician, public official, highly esteemed professions / influencer / regular citizen. Same as above, but correlated to sensibilities related to social roles of professions. By influencer, we mean any public figure that has the capacity to emotionally or rationally influence people (actors, pop icons, vloggers etc.). Highly esteemed professions will be assessed based on the national context of the partners, but they tend to include teachers, doctors, professors, nurses etc.

10. Perceived characteristic of the perpetrator – a dropdown menu consisting of: ethnicity, religion, gender, sexual orientation, economic status, race. The main trait that describes the

(alleged) perpetrator's perceived "otherness" compared to the general population / victims. The perpetrator must not necessarily have this trait, but the general public must think that the perpetrator does (e.g. a person might not be of a certain ethnicity, but the readers assume that she or he is).

11. Perpetrator's relation with country of attack – a dropdown menu consisting of: citizen, asylum seeker, immigrant, descendent of immigrants (up to 3rd generation), non-citizen. This is used to assess the (alleged) perpetrator's perceived "otherness" in relation to the state in which the event took place.

12. Minority status of the perpetrator – a dropdown menu consisting of N/A, majority, historical minority, new minority. A more in-depth look at the (alleged) perpetrator's perceived otherness. If the (alleged) perpetrator is an asylum seeker or a non-citizen of the country in which the event took place, N/A shall be filled in.

13. National context - a brief description of no more than 500 characters of the national context in which the event took place should be offered. It should cover similar events that have happened close to the date of the one being analysed, or that have contributed to the social climate in which the event took place, general stereotypes regarding the characteristics of the (alleged) perpetrator, etc.

14. Global context – same as above, but at a global scale.

15. Date of first identified articles

16. Number of press occurrences – a dropdown menu consisting of 1-20, 20-100, more than 100. An estimate of the number of press stories generated by the event will be offered.

17. Duration of media interest – a dropdown menu consisting of less than a week, 1-2 weeks, 2-4 weeks, more than a month. An estimate of the period of time in which the media has focused on the event and has produced content related to it.

18. Channel of communication (dominant) – a dropdown menu consisting of TV, written press, social networks. Partners will choose the most important medium through which the event was covered. Importance is expressed in terms of the number of audiences reached and the perceived impact on the audience.

19. Channel of communication (secondary) – similar to the above

20. Influencers involved in generating negative emotion (dominant) – a dropdown menu consisting of: mainstream politicians, extremist politicians, extremist influencers, media / entertainment public figures, social network influencers. If negative messages were spread, partners will select the most important category of people that have done so. Importance is expressed in terms of the number of audiences reached and the perceived impact on the audience.

21. Influencers involved in generating negative emotion (secondary) – similar to the above.

22. Dominant type of articles – dropdown menu consisting of informative, opinion, satire. Partners will evaluate what type of articles dominated the media coverage around the event.

23. Dominant tone of articles – dropdown menu consisting of very negative, negative, neutral, positive, very positive. Partners will evaluate what was the tone that the majority of articles had towards the group that is vulnerable to hate speech, not towards the perpetrator(s) involved in the event.

24. Articles with hateful comments on newspaper forums – dropdown menu consisting of: minority, majority, almost all. This is to assess the predominance of articles to which at least one user has generated comment that can be considered hate speech. Partners are to select minority if less than half of the articles have had hate speech comments and almost all if over 90 % of the articles have had hate speech comments. If comments have been deleted, partners can try to evaluate if the deleted comment contained hate speech based on the other comments surrounding it.

25. Rate of hateful comments on affected articles – a dropdown menu consisting of: almost non-existent, minority, about half, majority, almost all. Partners must rely on their common sense knowledge and assess the median prevalence of user generated hate speech comments on the articles that contain at least one such comment.

26. Social network posts with hateful comments – similar to 22. Articles with hateful comments on newspaper forums, but applied to the relevant social networks in each of the partner's countries

27. Rate of hateful comments on social network posts – similar to 23. Rate of hateful comments on affected articles, but applied to the relevant social networks in each of the partner's countries

28. Hateful user generate posts on social networks – a dropdown menu consisting of: minority, majority, almost all. Partners will assess the prevalence of user generated posts on social networks than contain discriminatory remarks / hate speech and that are associated to the event that is being coded.

29. Articles and posts with CAN messages – dropdown menu consisting of: minority, majority, almost all. Partners will assess how many articles contain counter and alternative narratives (CAN) in their comments section.

30. Rate of CAN messages in articles and posts where CAN is present – a dropdown menu consisting of: almost non-existent, minority, about half, majority, almost all. Based on the articles identified at point 29, partners will assess the rate at which hate comments have been countered or alternative messages have been deployed.