WE CAN for human rights speech

NO HATE

Report on the effects of timing of launch, formats, and messages of human rights narratives to maximise target audience reach

Report prepared by the Agenzia di Promozione Integrata per i Cittadini in Europa (APICE) and Neue Deutsche Medienmacher e.V. (NdM)

March 2021

ActiveWatch

Apice
Agenzia di Promozione Integrata
per i Cittadini in Europa

A JEWISH
CONTRIBUTION
TO AN
INCLUSIVE
EUROPE

ndm | neue deutsche
medienmacher*innen

Table of contents

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.   Introduction to the research

"Hate online is organised – by a small minority. Involved accounts [in Germany] are active in the right-wing extremist scene, amongst others on the pages of the Identitarian movement."
(Conclusion of the research conducted by #ichbinhier and NdM in 2019)

Reports by investigative journalists and monitoring efforts of activist movements and associations all over Europe confirm what has been common knowledge amongst activists and organisations working in the digital sphere since years: hate speech and digital violence are, more often than not, coordinated and strategically organised, not only in one country, but often also across borders and time zones (Schwarz, 2020). However, counter speech efforts, though widely spread, are less coordinated, and consequently, neither reach a large audience nor fulfil their purpose and potential to rapidly, as well as efficiently, respond to hate speech online, triggering the question: Why? And how can we improve those efforts?

Recent studies suggest that organised counter speech in commenting sections have a stabilizing effect on online discourse, thus being an effective tool against hate speech online (Garland et al. 2020; Ziegele et al. 2019). Those findings not only suggest that more coordinated efforts of civil society actors and activists are needed to effectively respond to hate speech online, but also imply that counter speech campaigns in general should be more organised and structured, than just developing and publishing content. In order to reach a wider audience and have an actual impact on public online discussions, in the short-term, and on attitudes and behaviours, in the long-term, campaigns promoting counter or alternative narratives should be more structured.

However, one of the challenges when introducing counter and alternative narratives in online debates is that activists as well as organisations have to act fast: in order to defuse digital attacks and hate speech, one must usually respond right away to prevent hate and hateful narratives from further spreading. This makes efforts of more coordinated and structured campaigns an even bigger challenge, leading to questions about the timing of counter narrative campaigns and formats of content and messages to benefit from algorithms of social media platforms as well as the interest of users in the topic.

During 2019, the No Hate Speech Movement in Italy introduced the so-called counter action days as a new model of coordinating activities online, using a structure similar to the one of the action days[1] of the Council of Europe's No Hate Speech Movement and introducing actions related to the production, diffusion and application of counter-narrative online. By monitoring the results of the counter-action days, the Italian No Hate Speech Movement observed that similar counter narrative contents had a different impact online, seemingly depending on the time when they were published, in relation to the occurrence of the hateful incident. During the build-up and then "explosion" of the hate storm, counter narratives were not really effective, however, it started to penetrate the public opinion and produce reactions once the hate storm was calming down and the circulation of fake news about the incident were progressively decreasing. Based on these observations, the window of time during which counter narratives are more successful could be identified as between 24 and 48 hours after the initial incident.

Meanwhile, the No Hate Speech Movement in Germany, being coordinated by NdM, worked closely with online activists and experts in order to collect data on effective counter speech strategies. In addition, NdM has launched several long-term, as well as short-term, counter narrative campaigns during recent years, with and without cooperation partners, which identified that the success (and reach) of campaigns not only depends on messaging, but also on format. This hypothesis was further substantiated by the experiences of a separate NdM project, namely Handbook Germany, which drew similar conclusions. It further triggered an increasing focus on developing campaigns in line with recommendations by social media companies as to what works best on their platforms.

These experiences of both project and research partners in recent years led to the following research question:

Which factors influence the success and reach of counter or alternative narrative campaigns in general?

The previous experience further allows us to focus the research aim on the specific factors of timing, messaging and format:

How do the variables of timing, messaging and format affect the reach and impact of counter or alternative narrative campaigns?

However, it is not only necessary to discuss the effects of the different variables individually, but more importantly to analyse the relevance of timing in combination with factors such as format and messaging when developing counter or alternative narrative campaigns. The combination of these three factors (timing, content and messaging) could completely change the impact and reach of counter and alternative narratives.

The research, therefore, aims to identify the influence of timing, format and messaging on the reach and uptake of the counter or alternative narratives. It aims at supporting activists and NGOs with recommendations regarding these three factors in order to benefit the reach and overall impact of their respective counter or alternative narrative campaigns. The researchers' assumption is that, despite the content of campaigns, it is rather technical considerations that have to be kept in mind and which might, in the end, determine the performance of a campaign.

The following research report is split in two parts, mainly prior work and data collection as well as analysis:

1. The first part of the report provides background information that led to the research at hand and prior research on the effects of timing and formats on the performance of counter or alternative narrative campaigns. The report further looked into research on political campaigning and marketing to identify relevant knowledge.

2. The second part focuses on the methodology and data collection method, but also on the analysis of the collected data on timing, format, and messaging.

Lastly, the report ends with final remarks and recommendations for future research, but also provides tips for NGOs and activists planning to launch counter and alternative narrative campaigns.

III. Part 1: Reasons for research and prior scholarly work

Before diving into the data collection and analysis regarding the performance of counter and alternative narrative campaigns, the report provides research on counter narratives, but also presents some insights into the background leading to the research at hand. The presented report further drew lessons from political campaigning and digital marketing strategies.

1. Prior research

Previous research on counter speech and narrative campaigns not only benefited the data collection, but also provided the framework for this research; yet, it is important to emphasize that research on hate speech in general, but more specifically on counter and alternative narrative campaigns and their impact is still scarce. Therefore, the research provided by the project WE CAN for human rights speech is a valuable and unique addition to existing scholarly work, especially as it considers the practical application and perspective of civil society organisations and activists from different EU countries, allowing not only for a comparative overview, but more importantly, supporting the practical work and setting new impulses for future research. Yet, before presenting research results, a general overview on previous research, specifically on format in campaigning and in counter speech efforts, is provided.

In their 2015 report on counter speech, Jamie Bartlett and Alex Krasodomski-Jones analyse content challenging extremism online across Europe, arriving at useful recommendations for activists and civil society organisations when developing counter speech campaigns. Their analysis focused rather on the style of messaging/ tone of messaging than format in general; it showed that the most popular counter speech content were questions, followed by commentary - though in France attacks against haters were similarly popular. A report by the Institute for Strategic Dialogue further corroborates that the tone of message has a significant impact on the success and reach of counter speech and online interventions. The research team identified that antagonistic messages prove unsuccessful, recommending activists to utilise casual or sentimental messaging. They further suggest the use of personal stories (Frenett and Dow (2015), in 2019 the International Centre for Policy Advocacy (ICPA) implemented a so-called narrative change lab, consulting civil society organisations when setting up narrative change campaigns on topics such as immigration and migration in Germany. The ICPA later evaluated the implemented campaigns, offering the summary that not only do personal stories matter, but also the authenticity of the messenger as well as the messenger's portrayal is highly important.

The research of Bartlett and Krasodomski-Jones also suggests a funny or satirical tone. The researchers further examined which topics fare especially well in regard to counter speech, arriving at the conclusion that topics such as immigration, race and religion are receiving most interactions, especially when counter speech is satirical. However, it is important to mention that news reports on immigration, race and religion also generate more hateful reactions and content, thus awareness on those issues is generally higher, as a study by Freie Universität Berlin indicates. Additionally, research partner NdM worked closely with activists of the civil-rights movement #ichbinhier in order to analyse coordinated online attacks, but more importantly to identify effective counter speech strategies, identifying humour as a viable strategy, but nonetheless outlining warnings when relying on satire and cynicism as a response to hate speech. The data and experiences collected as part of this cooperative research fed into the research endeavour as outlined in the introduction of the report but was supplemented by further insights on formats and messaging of counter narrative campaigns.

In regard to formats, Bartlett and Krasodomski-Jones' report indicates further that photos and videos are "the most effective type of content to post to reach a broader audience", urging

counter speech pages to adopt more visual content. Their research further suggests that counter speech organisations and initiatives have to produce and publish content more regularly to benefit from the algorithms introduced by the social media companies, which also urges for research on the timing of counter or alternative narrative campaigns; in the qualitative interviews with representatives of the respective companies, this claim has been repeated and substantiated. Their report further arrives at the conclusion that "if counter-speech page administrators and users were more active, and changed their content slightly, it could dramatically increase the reach of their messages" (Bartlett & Krasodomski-Jones 2015).

Counter and alternative narrative campaigns are a vital tool in combating hate speech and digital violence online, but also offline, as the reach of viral campaigns often surpasses the boundaries of the digital space. Furthermore, recent studies have shown that coordinated counter speech efforts might have a significant impact on hate speech online, being able to positively affect online discussions and societal discourse on social media (Garland et al. 2020; Ziegele et al. 2019).

However, counter speech campaigns and narratives have to tackle several obstacles, among them the timing of the response to hate speech as well as the algorithms of social media platforms. Additionally, resources such as time, but also money, are often scarce, thus, content has to be developed under pressure in order to stay relevant – often leading to the disregard of recommendations for format and messaging, and thus putting the counter speech efforts at a disadvantage.

A general overview of marketing and political communications research further suggests that visual content benefits from the social media networks' algorithms, but also that the identification of objectives and the target audience are of the utmost importance, implying that the success and reach of counter or alternative narrative campaigns ultimately depends on the knowledge one has about the audience as well as the functioning of social media.

The research at hand focuses on the importance and effectiveness of format and messaging in counter speech and alternative narrative campaigns, aiming at being the basis of future efforts of activists and civil society organisations by also providing samples and recommendations in order to benefit from the inner and inherent workings of the social media platforms.

2. Reasons for research at hand

The aim for this research is not only grounded in the lack of scholarly work in general as well as the failure to include practical experiences, but also due to the efforts of the involved research partners, namely APICE.

APICE, the Italian organisation set up the national support group of the No Hate Speech Movement in Italy in 2015, has been implementing the movement in Italy during the last five years, after being recognised by the Council of Europe (CoE) as the official coordinator. The first three years (2015 - 2017) of the movement were dedicated to preventing and combating hate speech online, through monitoring and reporting online hate speech, implementing action days as well as human rights education activities, using the Bookmarks manual provided by the CoE.

Yet, from 2018 on, the Italian Movement started to increasingly focus on counter online hate speech, introducing counter and alternative narratives by utilising the CoE's We Can manual. During those activities, especially in 2019, APICE noticed that the performance of counter and alternative narrative campaigns seemed to be connected to the timing of launching the

respective campaigns, meaning that the reach and impact of narrative campaigns is not only dependent on the content itself, but rather on the time frame within which it is published.

Based on this understanding, APICE developed a model that would allow for the collection of useful data, aiming at identifying specific data to determine the influence of timing on the performance, and ultimately success, of counter narrative activities in an activist context. The developed model, if successful, would then benefit other campaigns and initiatives, in Italy and abroad. In order to determine whether the model and the collected data is viable, the organisation monitored the performance of digital campaigns during the so-called Counter-Action Days, observing that the success of same counter narrative contents had a different impact online, depending on the time during which they were published, in relation to the occurrence of the hateful incident: during the build-up and then "explosion" of the hate storm, counter narratives were not really effective. However, it started to penetrate the public opinion and to produce reactions once the hate storm was calming down and the circulation of fake news about the incident were progressively decreasing. Based on these observations, the window of time during which counter narratives are more successful could be identified between 24 and 48 hours after the initial incident.

Consequently, APICE decided to collect data on the performance of counter-narrative content online, on the Facebook and Instagram accounts of No Hate Speech Movement Italy, during three different time-slots (within 12 hours after the incident; in between 12 to 24 hours after the incident; and in between 24 to 48 hours after the incident) to determine the influence of timing on the success and reach of counter narrative campaigns, which ultimately led to the research at hand.

The objective of the research conducted by APICE is to analyse the best timing for the launch of human rights based narrative campaigns on a social media platform after incidents of hate speech. Therefore, the research will seek to

- observe and collect data on the effects of the application of human rights-based counter narrative online within a different time slots going from 12 to 48 hours after a hate speech incident
- analyse the time frame during which counter narratives are most effective
- identify the better time slot for the launch of online counter narratives

The research on timing will further be supplemented by and connected to the data collected on the effects of formats and messaging, which will be conducted by NdM. Ultimately, both research efforts will be combined to also identify how the different variables might be interdependent.

IV.   Part 2: Methodology and data collection

As the research effort focuses on three variables, namely timing, format and messaging, in three different countries (Germany, Italy and Romania), different data collection methods were introduced for the respective variables. Subsequently, the methodology and data collection in regard to the different variables are presented.

1. Methodology and data collection on timing

For the methodology, we were supported by researchers at the Universities of Padova and Verona, who proposed scientific indicators and a structured model to overcome the problem of not proposing the exact same content in the 3 different time slots.

The general idea was to produce for each time slot two contents (a card and a short video), with the same form but different messages. In this way, it would be possible to collect a variety of data that would give a good measurement of the differences between the time slots, even with contents that are technically different.

We set a minimum target for the data collection, to be collected directly from the insight's tool of the platforms, especially Facebook[1], and we estimated to have:

- 6 Counter narrative products (card or video) responding to 6 hate speech incidents;
- 6 data-collection packs, 1 per each case, with data and samples of the social media posts in the form they were published;

Below you can find the practical examples of the content produced for the Facebook page and that were made in the time slots indicated in order to collect data for the research.

Content 1 (12 hours slot - Card)[2]
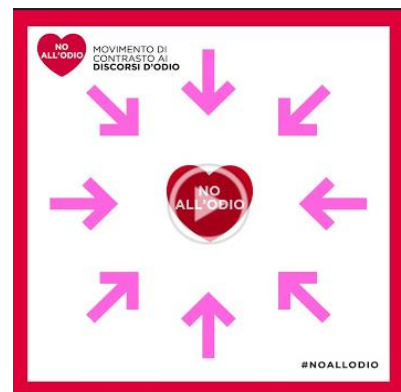
Content 1 (12 hours slot - Video Frame)

---

[1] We later decided to include only Facebook data in the final research report because it gave us more detailed insights compared to Instagram.

[2] The messages of all the cards and short videos (which are quite similar to follow the research model), are playing with the words used by an offensive title on an Italian newspaper, as well as with the name of the newspaper itself, which is famous for its offensive contents. The title in question was referring to a supposed "ethnic replacement" of Italians by migrants, following the deaths caused by COVID19.

Content 2 (24 hours slot - Card)

Content 2 (24 hours slot - Video Frame)

Content 3 (48 hours slot - Card)

Content 3 (48 hours slot - Video Frame)

In drafting the model, researchers advised also on the composition of the teams in charge to produce the content. In accordance to their advice, this factor is important for the research because the counter messages[3] could have been influenced by the personal bias of the activists if the same teams would have worked in the same time slots, causing a possible falsification of the data results.

Such a dynamic is very important in order to avoid the occurrence of a variable that would affect the impact of the counter narratives. Luckily, the working methods in the activist group is based on a voluntary participation in the action teams, which are open when hate incidents happen or around action days. In this way, the different activists are never assigned to certain topics or slots, thus eliminating their possible personal bias from the equation.

In order to ensure the number of cases needed for a reliable analysis, it was decided to collect data of 3 types of counter narrative actions
1. Counter narrative produced by our activist group to respond to a hate incident (counter narrative actions);
2. Counter narrative produced by our activist group in preparation for a counter action day;

---

[3] The message is the element that could be affected because for the general aesthetics we follow the usual lines and motifs that are now typical of the visual identity associated with the campaign, thus not influencing the success of the diverse contents.

3. Support actions, such as sharing products made by other organisations/users, to counter a hate incident.

While the data collection phase was originally foreseen for 3 months, due to the lockdown starting in March, following the Covid19 outbreak, this phase was extended to cover 6 months until 21 September.

At the end of the data collection process we gathered 20 data packages, including
- 9 cases of counter narrative actions to hate incidents/storms
- 3 cases of support actions
- 8 cases of counter action days

The difference among the 3 types of actions is the following:
- Counter narrative actions and support actions respond to hate incidents in the moment, counter action days offer counter narrative products to be used in case of hate incidents related to a specific topic, that means that counter action days are not related with timing because they are not responding to a specific hate incident. The reason why we decided to include them in the analysis is because of the possibility to double check specific hypotheses (e.g. is this result related more with timing or with Facebook algorithm?);
- Counter narrative actions and counter action days are planned, prepared and implemented by our activists, support actions are initiated by other people/organisations who responded to hate incidents and we decided to support them by sharing their contents. This is relevant for the research because it could tell us if the No Hate Speech Movement Italy strategy influences the performance of the counter narrative, being more or less successful than other counter narratives on the same topic in the same time slot or if they perform all according to a pattern (e.g. timing).

Each data pack includes:
- The counter narrative product (card, video, infographic…)
- The screenshot of the message accompanying the product
- The screenshot of the post as it was published
- The collection of data about reach, interactions and reactions

All the visuals have been stored in drive folders, the data collection and the messages have been gathered in the research data set, in excel format.

The research data set is quite broad and gives space for analysis and investigations beyond the scope of the timing related interest, such as the influence of the time of the post on the success of the counternarratives (e.g. is it more efficient to post a counter narrative in the morning or in the evening?), or the format combined with other variables (e.g. when posting in the morning, a card is more efficient, when posting in the evening, a video is more efficient), but for the scope of this research, we analysed those variables regarding the type of reactions generated or their level of penetration into the public discourse.

<u>Implementation</u>

Researchers from different Universities supported the production of the model of our research. These academics are internationally recognised as knowledgeable on the topic since they are fellow partners of the Rete Nazionale per il Contrasto ai Discorsi e ai Fenomeni d'Odio (National Network for contrasting hate speech and hate phenomena); they listened to our needs and, taking into account our working method, helped us to design the 6 content models presented above.

Nonetheless, basing our activities on the grassroot work with young volunteer activists, when trying to implement the model, we noticed that it was not feasible to produce the necessary content at the prescribed times (12-24-48 hours since the occurrence of the hate storm), because especially for the first slot, 12 hours proved to be insufficient to group the necessary people and to gather quality ideas in the pandemic scenario. Especially during the first months of the COVID19 pandemic in Italy, the lockdown heavily affected the possibility to organise work and to engage young people and activists, sometimes not only for practical reasons connected with logistics and communication, but because of the radical change in the personal agendas of many people, without even mentioning the burden in terms of physical and mental health that many people experienced. These factors often impacted the drafting process, allowing the teams to post only after roughly 24 hours.

In producing counter narrative contents, the team upheld certain quality standards, coming both from our campaign experience, which prescribe special attention to the tone and the language, to make sure they respect human rights values, but also from social media standards, which require some minimal criteria regarding aesthetics, especially when targeting young people and professionals already involved in the fields.

Counter narrative contents appearing on No Hate Speech Movement Italy accounts are always produced following the principles coming from the 7th Chapter of the We CAN manual, "Doing it Yourself Step by Step", according to the method and the practice that was established during the last years. We have already produced, in fact, our own Italian template for assessing the oppressive narrative, following the first step of the guide and making it more accessible for Italian audiences. This comes from a consolidated approach in our activist strategy that is heavily depending on the principles and values promoted by the No Hate Speech Movement Campaign during the years and that put human rights at their very core.

This adherence to such principles and values set one of the main indicators through which we evaluate the contents we produce; for this reason when deciding to perform the research in a concrete scenario, using our usual social media accounts, we were confronted with the issue of making sure not to lower our standards for posting and publishing contents that would be coherent with our usual activities, not only in order to satisfy our general audience, but also not to risk the falsification of our data collection, since seeing something artificial or too much different from our usual contents would have altered the response surrounding the actions, thus not allowing us to collect the needed data.

During the research, the interference of Facebook algorithm on the process of posting was also noted, which automatically favours contents that get the most likes and interactions and that, in our case, were also those which appeared first on the page. In the beginning, we were not sure whether there was a real pattern, since it had happened that contents posted later produced more interactions and reactions, but as it is possible to see from the data below, regarding content posted during action days, there is clearly a gap between the first content of the day and those following. Faced with this data, we were forced to reconsider the relationship between contents appearing in the 12 and 24 hour slot, since the second one would be certainly penalised in terms of reach, not because of the timing in relation to the hate storm, but for the timing related to the previous post, from which it could have been hidden.

In the table below it is possible to see that in 7 out of 8 cases the first post performed always better in generating reactions[4], regardless of the format used (card, article, video, etc.), the topic of the counter action day, or the exact time of the day in which the post was published.

---

[4] The number of reactions is a better indicator related with reach, because it gives us the number of users who performed a proactive action, expressing publicly an emotion or a comment, and this means that the counter narrative provoked a response;

Table 1. Engagement of the 8 counter action days - reach, interaction and reactions

TOPIC	FORMAT OF COUNTER NARRATIVE	DAY	TIME (CET)	REACH	INTERACTIONS	REACTIONS
Racist hate speech	Concept (original)	21.03.2020	11:30	555	122	**83**
	Card (original)		12:30	1491	117	82
	Youtube Video (shared)		15:00	356	19	13
	Image (shared)		16:00	190	24	12
Victims of slavery	Image (shared)	25.03.2020	10:00	866	76	**42**
	Article (shared)		14:00	94	4	4
	Article (shared)		16:30	108	5	5
Anti-gypsy hate speech	Video (original)	08.04.2020	11:00	2392	193	**102**
	Video (shared)		14:30	406	34	21
	Link to website (Shared)		16:00	319	25	20
	Link to website (Shared)		17:00	320	13	17
	Image (shared)		17:58	305	20	26
Press freedom day	Card (original)	03.05.2020	10:30	4267	171	**134**
	Link to article (shared)		13:00	537	72	46
	Card (original)		15:30	860	48	32
	Card (original)		19:00	767	36	27
Homo-Trans-Bi-phobia	Card (original)	17.05.2020	09:59	18263	1438	**931**
	Card (original)		12:00	5971	949	383
	Infocard (Original		14:30	2933	206	111
	Video (shared)		17:00	529	31	23

the reach only tell us how many users were passively reached, without any evidence on change or emotion experienced. The reach only tells us about the popularity obtained.

Refugee day	Card (original)	20.06. 2020	09:30	3122	183	**124**
	Infocard (Original		11:00	1617	70	44
	Card (original)		13:00	1229	76	49
	Card (original)		16:30	917	42	33
	Video (shared)		18:30	235	17	15
Victims of hate crimes	Card (original)	22.07. 2020	09:00	2282	115	77
	Card (original)		10:00	2216	187	90
	Card (original)		12:00	1768	92	62
	Card (original)		14:30	1505	93	41
	Card (original)		16:30	1666	157	95
	Webinar (Original)		18:34	3054	311	95
Peace day	Card (original)	21.09. 2020	09:00	1632	82	**622**
	Card (original)		10:30	903	53	35
	Card (original)		13:30	1003	588	40
	Card (original)		16:30	998	41	24
	Card (original)		19:02	517	20	16

These considerations were mostly confirmed when we launched counter narrative content between 12 and 24 hours as standalone posts, with much better results, showing us that no real measurements and recording was possible in such scenario, deducing that three posts regarding the same topic, with the same format and similar content appearing close to each other, would certainly be competing in terms of reactions, but inevitably favouring the first one while penalising the third, which would appear as a repetition and that would be unlikely to meet audience's enthusiasm.

On the other hand, by observing the data coming from counter narrative actions appearing as standalone posts in different time slots, it is possible to see that, with the exception of a post consisting of a reshared content (in row 4), all the posts made in the timeslot between 12 and 24 hours slot performed significantly better than the others, in terms of reach, interactions and reactions.

Table 2 Counter narrative actions appearing as standalone post - reach, interactions, reactions

CASE N.	TOPIC	TYPE OF ACTION	FORMAT OF COUNTER NARRATIVE	DAY	TIME (CET)	TIMING - Hours after the Hate Incident	REACH	INTERACTIONS	REACTIONS
1	**Hate speech against young people during COVID-19**	Counter narrative responding to hate wave	Card (original)	10.03.2020	16:28	12-24	3345	235	**116**
2	**Revenge porn**	Counter narrative responding to incident	Video (original)	05.04.2020	12:33	12-24	2304	240	**111**
3	**Homophobic hate speech**	Counter narrative action responding to hate wave	Cards (original)	20.04.2020	13:27	24-48	1749	120	75
4	**Islamophobia (welcome back Silvia Romano)**	Counter narrative action responding to hate wave	Video (original)	11.05.2020	17:00	12-24	2392	319	**207**
5	**Homo-Trans-Bi phobia**	Counter narrative action responding to hate incident	Video (shared)	16.07.2020	11:00	12-24	727	56	26
6	**Racist hate crime**	Counter narrative action responding to hate incident	Cards (original, shared)	09.09.2020	18:28	after 72	439	41	24

According to our ongoing analysis and evaluation, the idea of posting standalone counternarratives in different time slots (within 12 hours, between 12 and 24 hours, and between 24 and 48 hours) proved to be the only viable strategy to be able to obtain reliable data on timing, since the proposed research model had to be adapted to the current scenario.

Even by adapting our methodology, we would have needed a higher number of hate incidents to happen in order to better assess our evaluation and conclusion.

Such conditions were practically impossible to achieve, because despite the everlasting presence of hate in public discourse and the explosion of certain specific cases, those were always obfuscated by the attention towards COVID19 related topics and arguments, especially during the time in which our research was happening.

This is the reason why we changed our strategy and we went much beyond the original idea of collecting only 6 data packs, as we started to gather data also on action days and support actions to analyse the way in which they performed, in order to have more terms of comparison with the counter narrative contents.

In the end we included 20 data packs in our research: 9 counter narrative actions, 3 support actions and 8 action days. Out of the 20 cases, the 9 counter narrative actions are the ones strictly related to our research on timing, the other cases will be used to check variables, different from the timing, to prove or deny possible findings related with other dimensions.

2. Methodology and data collection on formats

The data collection is based on a combination of quantitative as well as qualitative methods to identify the effectiveness of different formats of counter speech campaigns. Furthermore, data was collected in three different countries, namely Italy, Germany and Romania, during two fixed dates, among them the International Day against Homophobia, Biphobia, Intersexism and Transphobia on 17 May 2020 and the International Day for the Victims of Hate Crime on 22 July 2020. In Germany, further data has been collected on International Women's Day, 8 March 2020, as well as the International Human Rights Day on 10 December 2020. During the so-called action days, the project partners published formats and messages on their respective social media accounts; the following analysis of performance, including metrics such as engagement and reach, provided not only insights into the usage of formats in general, but also data on the effectiveness of different formats in the respective countries.

A repeated and continuous social media monitoring and recording is required in order to sufficiently collect relevant data to further identify whether and how formats can affect the performance of counter speech and alternative campaigns. Therefore, a comprehensive coding log was provided for all involved partners to collect data, allowing for the measurement of the following indicators, which will be defined and further discussed in the following section: quantitative metrics, such as awareness, engagement and/ or interaction, and qualitative metrics, such as comment analysis and feedback, differentiating between positive and negative responses. The qualitative analysis is further supplemented by logging the number of exchanged arguments/comments, the number of involved accounts and the use of hashtags.

Next to the above-mentioned metrics, the data coding log additionally asked to specify the location of the interaction as well as the used format and narrative/message. The log also recorded the tone of the narrative/message.

Even though the term format has been used before in this report, it is necessary to provide a definition to ensure a common understanding. When talking about format, it means the way content is presented and arranged; however, in more detail and more differentiated than just the organisation of words on a page or picture. Different formats, at least in this research, encompass different styles and kinds of graphic content, such as quotes, empowering statements and photos. Furthermore, also written content and videos can be counted as versions of formats, though videos can further categorised into animated videos, statement videos or explanatory videos.

An example of the format used for IDAHOTB 2020

In the course of the data collection for this research, no videos were developed. However, when formulating recommendations at the end of the report, videos will be considered as a way to present and publish content, drawing from the experiences of NdM, who launched an extensive video campaign when introducing the No Hate Speech Movement in Germany in 2016, but also from the feedback of the representatives of the social media companies.

a. Quantitative data collection

The quantitative data collection focuses on common metrics used to measure the performance of campaigns, drawing from research on and experience from marketing campaigns and the analysis of previous counter speech campaigns. As formats were tested across platforms, the research aims at considering the platform-specific terminology; furthermore, the Counter-Narrative Monitoring and Evaluation Handbook of the Institute for Strategic Dialogue recommends putting the metrics into the two broad categories awareness and engagement. Therefore, definitions of the most commonly used metrics are provided prior to presenting the results of the data collection as part of this research.

Awareness

Consequently, awareness is defined as "the total number of people who view your campaign content" (ISD, 2016). The Institute for Strategic Dialogue, amongst others, further recommends analysing the awareness metrics in order to identify and understand the audience reached with the shared content, to either adjust or improve your campaign. Awareness metrics include reach and impressions, though both need to be differentiated, despite tracking closely, depending on the platform you are using. None of the mentioned awareness metrics actually reveal whether users interacted or engaged with the shared content.

On Facebook, reach is officially defined as the "number of people who saw your ads [or published content] at least once", further specifying that reach "gives you a measurement of how many people were exposed to your message" (Facebook Business Help, accessed 2020). The platform differentiates between organic and paid reach (in form of advertisements), but also considers viral reach, which means that users might have seen content due to interactions of friends with the content (Hootsuite, 2018). Meanwhile, Facebook defines impressions as "the number of times that your adverts [or published content] were on-screen" and seen by your target audience (Facebook Business Help, accessed 2020). Therefore, the number of impressions is more relevant when spending money on advertisement campaigns to identify, first of all, whether the target audience has been reached and, secondly, how well it has been reached. Instagram treats reach and impressions similar to Facebook.

Twitter does not (yet) trace reach, thus, one has to rely on the number of impressions while being careful not to get confused. Twitter defines an impression "as the moment a Twitter user sees your tweet" (translated, Hootsuite, 2018). Consequently, it is rather recommended

to reply to your very own tweets than other tweets as this allows for an increase in impressions.

Nonetheless: the number of impressions on Twitter do not depict the reach of content, thus comparisons in regard to reach between Facebook and Twitter should not be conducted or, if done, conducted with caution.

Engagement

The Institute for Strategic Dialogue further defines engagement "as the volume and types of interaction between audience members, campaigners or campaign material", which includes likes and shares, comments, but also email responses (ISD, 2016). Despite having a quantitative component, the engagement metric can also include qualitative aspects such as the evaluation of positive and negative responses. However, this is covered in the qualitative data collection, thus, the engagement metric focuses at this point on quantitative data such as the number of clicks, comments and interactions, demonstrating the performance and the audience's reaction to counter speech or alternative narratives.

Facebook defines engagement as the "number of reactions, comments, shares and clicks on your post" (Facebook Business Help, accessed 2020), demonstrating how much users interacted with your content. Instagram has similar engagement parameters, namely the number of likes, comments, shares and saves.

On Twitter, engagement is defined as the "total number of times a user interacted with a Tweet. Clicks anywhere on the Tweet, including Retweets, replies, shares, likes, links, cards, hashtags, embedded media, username, profile photo, or Tweet expansion" (Twitter Help, accessed 2020). The engagement rate is further measured as the number of engagements divided by impressions.

The engagement rate is of high importance on all platforms as it demonstrates whether content resonates with the audience, in general; a qualitative analysis of comments and responses further allows to identify if the audience likes or dislikes the content, including also feedback on format and message of content.

	Facebook	Instagram	Twitter
Awareness			
1. Reach	"number of people who saw your ads [or published content] at least once"	Similar to Facebook	no measurement of reach
2. Impressions	"the number of times that your adverts [or published content] were on-screen"	Similar to Facebook	"the moment a Twitter user sees your tweet"
Engagement	"number of reactions, comments, shares & clicks on your post"	"number of likes, comments, shares and saves"	"total number of times a user interacted with a Tweet; Clicks anywhere on the Tweet, including Retweets, replies, shares, likes, links, cards, hashtags, embedded media, username, profile photo, or Tweet expansion"

b. Qualitative data collection

Previous research on the impact of counter and alternative narrative campaigns suggests that quantitative data is not enough to measure its impact, suggesting qualitative analysis to identify whether narratives and/ or campaigns actually affect the users' attitudes and behaviours. Consequently, the research at hand aimed at combining sentiment analysis, focusing on the tone and style of comments and feedback left by social media users, with expert interviews with the representatives of social media companies, such as Twitter, Facebook, Instagram and Google/ YouTube to gather more insights on the inner workings of the platforms.

Sentiment analysis

Sentiment analysis generally refers to the use of language processing and text analysis to systematically identify and quantify the "polarity (i.e., positive or negative) of a message" (Gonçales et al. 2010), thus determining whether users or clients approve or disapprove of a product or message. This form of analysis is mostly conducted with machine learning and computer processing large amounts of data, applied mostly in analytical domains ranging from customer care over financial investors aiming at identifying market opinions to politics to measure how potential voters respond to candidates and messages (Tumasjan et al. 2010).

Research utilising sentiment analysis generally relies on automated tools to extract the respective sentiment of messages; however, as neither the resources nor large sets of data were given, a manual analysis has been conducted as part of this project. Despite aiming at considering dimensions of analysis as used in previous scientific studies, amongst them positive emotions, negative emotions, sadness, anxiety, anger, tentativeness and certainty, the research at hand mainly distinguished between positive and negative emotions for reasons of simplicity.

In addition, the analysis of the content shared on Facebook allowed for further categorisation as users can select different so-called reactions (thumbs up, love, laughing, surprised, sad, angry and since recently: care) to respond to content. This allows for the monitoring of the sentiment of the users, though data interpretation has to be cautious as users might misuse the so-called reactions.

Expert interviews with representatives of social media companies

During the research meeting in the beginning of April 2020, representatives of the largest and most popular social media companies provided valuable insights, benefitting the research on format and messages. Subsequently, the recommendations of Twitter, Facebook, Instagram and Google/ YouTube are listed, adding also the different national contexts of audience profiles, as far as possible:

Twitter

The social network Twitter is hashtag-based, thus, in order to engage in counter or alternative narrative campaigns, one has to introduce or adopt existing hashtags.

In Germany, being a comparatively small market for Twitter, counter speech campaigns are successful on Twitter when activists react fast and 'capture' hashtags introduced by right-wing extremists and so-called haters; a manual analysis of a sample campaign demonstrated that counter speech activists seizing a right-wing hashtag responded within three hours.

However, user numbers in Germany for Twitter are comparatively low, so the 'capturing' of hateful hashtags might not work in larger markets; an additional concern to keep in mind is that hateful content is reproduced. Therefore, it might be more useful to rather introduce a positive hashtag than keep reproducing negative ones. Yet, this only works in smaller markets such as Italy and Germany, but most likely not in Spain and the United Kingdom. Furthermore, Twitter's recommendations on format and messaging, next to the introduction or adoption of popular hashtags, are the use of emoticons as they work better than GIFs or graphs on Twitter.

Facebook

Facebook is the largest social network in Europe, having the largest user numbers, though the audience's demographic changed significantly over time as younger people increasingly chose picture and video-based platforms such as Instagram, Snapchat and Tik Tok.

Recommendations provided by Facebook's representatives in regard to format as well as content development in general focused especially on the definition of the target audience, based on demographic and geographical factors, interests and user engagements, as different audiences require different format and appeals.

Prior research further suggests that videos and picture-based content in general works best on Facebook; however, the content has to have a high quality and credible, or even surprising, speakers to reach a larger audience. In addition, pictures and graphs should not have too much written content as they might otherwise be put at a disadvantage by the algorithm while videos should include subtitles and strong images to capture users, as they are mostly checking Facebook when commuting to work or university.

Instagram

Instagram is especially popular amongst younger audiences, especially as it continuously adopts new features. Additionally, social networks such as TikTok and Snapchat are also popular amongst young people, thus it is relevant for future research to also consider those networks to provide effective recommendations for civil society organisations as well as pedagogics aiming at reaching a young audience.

Instagram's recommendations focus mostly on "rather working with the algorithm, than beating it", implying that the regular publishing of content benefits reach, interaction and impact. Furthermore, previous research and analyses of successful campaigns on Instagram suggest that, on the one hand, prominent and newly adopted features, hashtags and formats should be picked up to increase reach outside of the very own bubble of supporters, and on the other hand that, in terms of format, content should be branded or have a similar style to create a sense of recognition.

Additionally, Instagram's representatives confirmed that infographics on debunking conspiracy myths regarding the Covid-19 pandemic gained increasing popularity, demonstrating that, for one, it is beneficial to respond to current challenges on the platform, and second, that the pandemic significantly affected the research or, more precisely, the conditions surrounding the research.

V.   Part 3: Evaluation

The evaluation section of the research paper covers the analysis and evaluation of the data collected, allowing for the pooling of findings. The outline of the subsequent section is as follows: firstly, data collected by APICE will be presented and findings on timing will be

displayed. The conclusion reached by APICE is then supplemented by the data collected in Germany on formats.

Italy

As concerns **timing**, despite the challenges, limitations and the adaptation previously described, the data packages collected have shown evidence of a certain pattern when it comes to the impact of counter narrative action applied in different time slots, especially in terms of interactions and reactions.

Out of the 20 data packages analysed, 9 were strictly related to counter narrative actions.

Depending on the hate happenings, we could identify 3 different types of counter narrative actions performed:

1. Counter narrative responding to hate incident;
2. Counter narrative responding to hate wave;
3. Counter narrative responding to hate storm;

We consider the hate incident as a case of rising hate speech online, but not a wave; the hate wave has a wide resonance and rises a higher level of fear and anger, the hate storm includes the increasing violent speech, threats and polarized emotions.

Following the data observation, we noticed that the 3 data packages related with the hate storm on inclusive languages (case N. 7 and 8) registered a very different performance, in terms of data, timing, duration in time, as it is visible in the table below. For these reasons, we decided to analyse it as a case study, and this is included as an appendix to this report.

We also decided not to include the data comparison on timing for the counter narrative actions posted on 6 to 8 May 2020 (case N. 9), during which we applied the original model with the repetitions of 2 formats per each timeslot, because of all the previous arguments, but we included the case when we checked and compared other variables.

In the table below, including the 9 cases as described, it is possible to compare the reach, reactions and interactions for each counter narrative action posts considered valid (Case N. 1 to 6), showing that the time slot between 12 and 24 hours after the occurrence of the hate incident is the one with the highest engagement rate and number of reactions (notice that the only post with lower engagement within 12 and 24 hours' time slot is the video shared on 16 July 2020 and this might be explained by the fact that it is the only non-original content used for a counter narrative action).

A higher number of reactions and interactions is usually linked with the idea that a content has ignited a more significant response, leaving a heavier footprint in the public discourse, while a high number of the research only shows us the popularity of the counter narrative. Such an impression is further confirmed when it is possible to notice that reactions and interactions are high in numbers also in the resharing of the post (whenever it was possible to follow the path taken by the sharing).

Table 3. Counter narrative actions - engagement per time slot

CASE N.	TOPIC	TYPE OF ACTION	FORMAT OF COUNTER NARRATIVE	DAY	TIME (CET)	TIMING Hours after the hate incident	REACH	INTERACTIONS	REACTIONS	ENGAGEMENT (Average 5,05%)
1	Hate speech against young people during COVID-19	Counter narrative responding to hate wave	Card (original)	10.03.2020	16:28	12 - 24	3345	235	116	3,47
2	Revenge porn	Counter narrative responding to incident	Video (original)	05.04.2020	12:33	12-24	2304	240	111	4,82
3	Homophobic hate speech	Counter narrative action responding to hate wave	Cards (original)	20.04.2020	13:27	24-48	1749	120	75	4,29
4	Islamophobia (welcome back Silvia Romano)	Counter narrative action responding to hate wave	Video (original)	11.05.2020	17:00	12-24	2392	319	207	8,65
5	Homo-Trans-Bi phobia	Counter narrative action responding to hate incident	Video (shared)	16.07.2020	11:00	12-24	727	56	26	3,58
6	Racist hate crime	Counter narrative action responding to hate incident	Cards (original or shared)	09.09.2020	18:28	after 72	439	41	24	5,47
7	Hate storm on inclusive language	counter narrative action responding to hate storm	Card (original)	04.08.2020	13:57	48	87891	22917	5533	6,30

#	Category	Description	Media	Date	Time					
	responding to hate storm	counter narrative responding to hate storm	Screenshots (original)	07.08.2020	13:57	72	2298	448	**140**	**6,09**
8	**responding to hate storm**	Counter narrative action responding to hate storm	Cards (original)	12.08.2020	10:06	*5 Days	4173	695	**238**	**5,70**
9	**Press freedom of expression**	Counter narrative action responding to hate incident	Card (original)	06.05.2020	19:15	12	**1209**	87	**63**	**5,21**
	Press freedom of expression	Counter narrative action responding to hate incident	Video (original)	06.05.2020	19:16	12	1324	82	60	4,53
	Press freedom of expression	Counter narrative action responding to hate incident	Video (original)	07.05.2020	08:57	24	809	23	16	**1,98**
	Press freedom of expression	Counter narrative action responding to hate incident	Card (original)	07.05.2020	08:57	24	706	27	23	3,26
	Press freedom of expression	Counter narrative action responding to hate incident	Video (original)	08.05.2020	10:02	48	661	29	19	2,87
	Press freedom of expression	Counter narrative action responding to hate incident	Card (original)	08.05.2020	10:02	48	**764**	51	**36**	**4,71**

Another interesting finding is related to the **type of reactions** registered within the different time slots, showing the diversity of emotions the users experienced after 12 hours, between 12 and 24 hours and between 24 and 48 hours from a hate incident and during the hate storm. On Facebook, the types of possible reactions are, in order, LIKE, LOVE, HUG, AHAH, WOW, SIGH, GRR, COMMENT and SHARE. The order in which they appear when interacting with the social media reflects the frequency of use of each one.

During 5 out of 9 incidents, the only types of reactions registered were LIKE, LOVE, COMMENT and SHARE, and the comments were quite positive, including emoticons or a few words. During 4 incidents we also registered AHAH and SIGH reactions, and only in 2 cases, the ones related with the hate storm, we registered WOW and GRR reactions.

This distribution of reactions shows the intensification and polarization of emotions when topics are getting more sensitive and become extreme, passing from a hate incident to a hate storm.

Table 4. Distribution of types of reactions per each case and timeslot

CASE N.	1	2	3	4	9	9	9	9	9	4	5	7		7	8	6
Timing	12-24	12-24	24-48	0-12	0-12	12-24	12-24	24-48	24-48	12-24	12-24	24-48		<48	<48	<48
Total	116	111	75	63	60	16	23	19	36	207	26	5533		140	238	24
Like	77	61	55	46	43	10	19	14	28	121	21	680		32	84	16
Love	17	20	12	3	7		2	2	2	44	2	82		3	21	5
Comments	5	6								20		2926		89	111	0
Sharings	17	24	8	11	9	6	2	3	6	21	3	266		10	21	3
Sigh				3	1							71		2		
Ahah										1		1416		1	1	
WoW												54		1		
Grr												39		2		

Elaborating on these previous observations, it is possible to explore the data of the different reactions in relation to the timing.

If we observe the LIKE and LOVE reactions, the data shows that in between 12 to 24 hours' time slot we can register a higher number of LIKE and LOVE reactions compared with the other time slots, even if we need to remember that the application of the original model during the days between the 6 to 8 May 2020 slightly altered the data, decreasing the performance of the contents published on the 7 May. For this reason, they are not displayed.

Figure 1 Comparison of LIKE reactions on original and shared counter narrative posts

We CAN for Human Rights Speech - REC PRoject - APICE Research
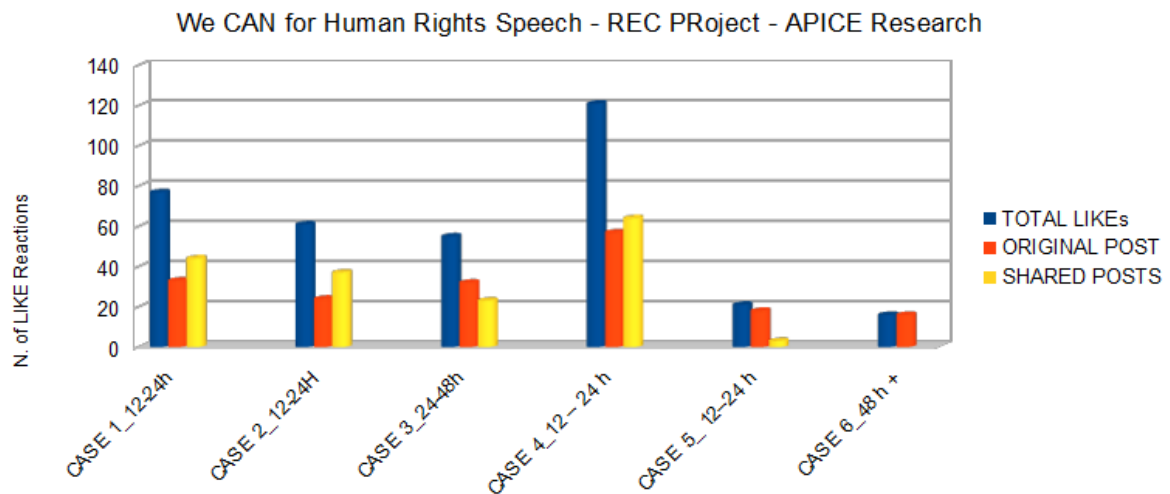
It is also visible that LOVE reactions are less likely than LIKE reactions, which is normal because the LOVE reaction expresses a more intense emotion.

Figure.2 Comparison of LOVE reactions on original and shared counter narrative posts

We CAN for Human Rights Speech - REC Project - APICE Research
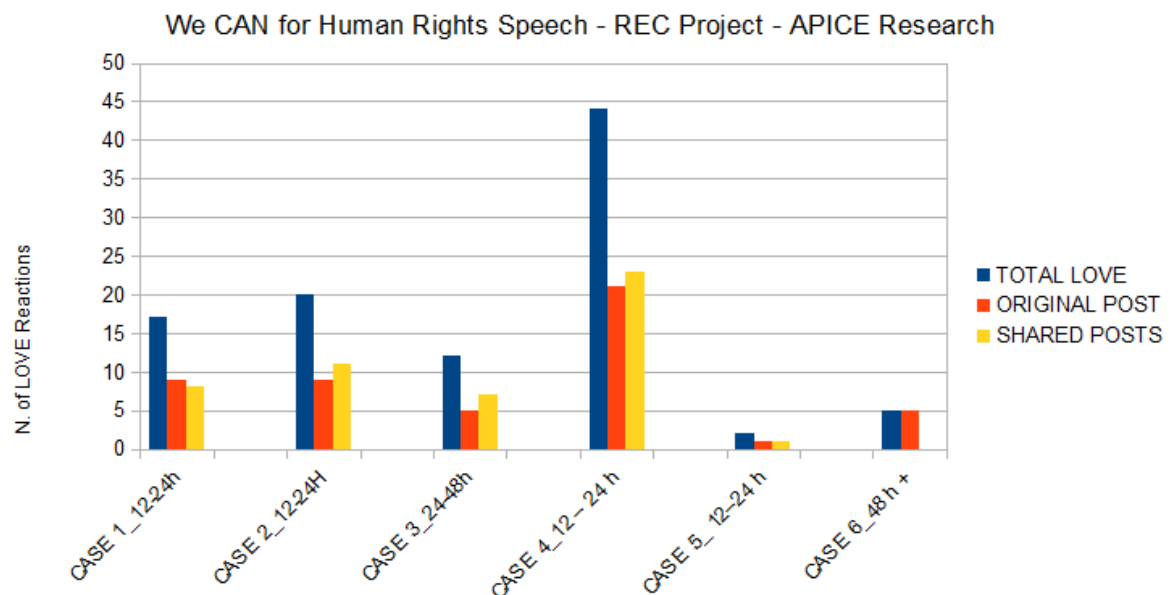
The sum of LIKE and LOVE reactions cover always over 80% of the total reactions except for 1 case only: the hate storm we experienced from the 4 to 12 of August 2020, during which we had the highest number of sharing and comments, most of them negative and aggressive, and the appearing of more strong and polarized emotions, expressed with the reactions GRR, SIGH and WOW.

It is important to notice that, except for the Hate Storm case, the positive reactions were resonating online out of the activist circle and usual public of the No Hate Speech Movement Italy page, as the positive reactions on shared Posts are higher than the ones on the original post on our Facebook Page. In other words, a counter narrative action posted between 12 and 24 hours on our Facebook page produces positive reactions, the majority of them are expressed on shared posts and this can be interpreted as an indicator of success of the

counter narrative in that precise time slot: a wider collection of data focusing on this aspect is required in order to confirm this finding.

Germany

The evaluation of the data collected on the Facebook and Twitter accounts of project partner Neue deutsche Medienmacher:innen e.V. (NdM) in Germany during four different action days, namely the 8 March, 17 May, 22 July and 10 December, allows for various recommendations for future counter or alternative narrative campaigns. The evaluation of the German data focuses on the impact of formats on the performance of counter narratives on three so-called action days.

The quantitative data collected during the first three action days imply that the used formats do not have a significant effect on reach and user interaction, indicating that impact among NdM audience is limited as only 1 000 to 2 000 users have been reached generally and no interaction in form of comments or messages has taken place, rendering a qualitative assessment in the form of a sentiment analysis of the data impossible and unnecessary.

☐	Für uns ist jeder Tag #Weltfrauentag, aber…	1154 People reached	39 Post engagements	8 March 2020 at 17:16	Public
☐	Die wunderbare Nalan Sipar erzählt in die…	2273 People reached	250 Post engagements	8 March 2020 at 15:21	Public
☐	Für uns ist jeder Tag #Weltfrauentag, aber…	1353 People reached	73 Post engagements	8 March 2020 at 12:55	Public
☐	Für uns ist jeder Tag #Weltfrauentag, aber…	1108 People reached	40 Post engagements	8 March 2020 at 09:55	Publ…

Figure 3: Quantitative data for the formats shared during 8th of March

A slight increase in reach and engagement can be, however, detected when sharing a video (see in the graph above), which either implies that short videos generally attract a larger audience or that the reliance on a credible and well-known speaker or messenger affects reach and engagement. Most likely, both factors are relevant and reinforce each other.

The impression that short videos (or even audio files) fare better with the audience of NdM in Germany is strengthened during the last action day on 10th of December; even though it was not planned to include the data in the research, the numbers collected demonstrate that audio or video files of credible speakers are appreciated by the NdM audience, implying that simple text postings as well as graphs and pictures are less popular among our audience. Generally, reach and engagement improved significantly across all platforms, namely Facebook, Instagram and Twitter, (as can be seen in the graph) as individual postings reached more than 7 000 users. In addition, the formats shared on 10 December further allowed for a sentiment analysis as users commented and shared the provided content; feedback has been generally positive as users commended the reliance on authentic messengers and brief messaging, once again reinforcing the notion that the selection of speakers and messengers is key when developing a counter or alternative narrative campaign.

Figure 4: Quantitative data for the formats shared during 10th of December

However, the success of the narrative campaign on the international human rights day can also be explained by the surrounding measures that have been taken to support the performance of the campaign, namely the activation of cooperation and project partners, providing them beforehand with the material and encouraging them during the action day to re-share the NdM content. Thereby, the material provided by NdM experienced an even bigger circulation as several civil society organisations and political actors such as Federal ministries shared the material.

The evaluation of the data in Germany indicates several things: first and foremost, that formats matter, as hypothesized in the beginning of the research. However, the selection and use of formats is highly dependent on the audience you want to attract with your narrative campaign, but also the messengers you choose.

Furthermore, narrative and alternative narrative campaigns on social media benefit significantly from measures taken outside of social media, meaning: In order to increase the reach and impact of narrative campaigns, it is recommended to involve partners that support your cause.

Romania

In Romania, data was collected only during one action day, namely the 17th of May. Data collection was further limited to Facebook as the main social media platform of the Romanian partner ActiveWatch. Nonetheless, conclusions can be drawn as we can compare the performance of the content shared to other postings on ActiveWatch's Facebook page, though we can only compare the number of likes and comments across the various postings as more insights are missing.

Figure 5: Content shared during 17 May

In comparison, the quantitative metrics, including reach and engage, imply that the selected format, in this case a picture with a text message (*Hate will never win #lovewins)*, performed only slightly better than the average of content shared on ActiveWatch's Facebook page, especially when we limit the comparative aspect to graphic content only. This conclusion is based on the number of likes and comments.

	General info	Specify			
Format Group	Picture	w text			
Narrative/ Message	*Hate will never win #Love wins (Ura nu va castiga niciodata #Dragostea castiga)*				
Date of testing	17.05.20				
Country of testing	Romania				
Used hashtags Hashtags:	*#IDAHOTB2020 #DragosteaCastiga #FaraUra*				
Facebook - [What did we share?]					
Post link	https://www.facebook.com/ActiveWatch/photos/a.264336760262328/3385234971505809/?type=3&theater				
Reach	2.003				
* paid	N/A				
* organic	2.003				
Engagement (from Ad manager)	112				
Post Engagement (reactions + comments)	31				
Top Audience					
Number of Shares	2				
Sentiment Analysis					
Comments					
Positive reactions (likes, ha ha, love)	26				
Negative reactions (Oh, cry, anger)	2				
Comments					
Number of comments	3				
Number of positive comments	1				
Number of negative comments	1				
Number of inbetween comments	1				
frame 1 in the comments	pro-lgbt ngo				
Example theme 1 comment	Good to know. Now I've unliked your page				
frame 2 of the comments					
Example theme 2 comment					

Figure 6: Quantitative data on the post, collected in the coding grid

However, the comparison between the suggested format as part of this research and previous postings by ActiveWatch demonstrates, once more, the importance of the target audience and the perception of the NGO by its audience, as ActiveWatch publishes open letters to Romanian politicians, and detailed analysis of hate storms and political developments on their Facebook account, so their audience is used to long text-based posts, and they fare much better, reaching a higher number of likes and comments. This implies that, while formats are important, it is not the decisive factor determining whether a counter or alternative narrative campaign is successful.

VI. Part 4: Limitations and Challenges

Despite the efforts to reduce bias of the data collection methods, it is necessary to point out several limitations of the conducted research, first and foremost the differences in audiences, national contexts as well as general data collection biases.

Even though the project partners aimed at establishing similar conditions by selecting international action days in order to test timing, formats and messaging, the collected data

might be flawed as all involved organisations have a different target audience and operate in different national contexts, thus affecting engagement and interaction rates. Consequently, the obtained data and recommendations might not be transferable and applicable to all national contexts, suggesting that NGOs and activists aiming at launching counter or alternative narrative campaigns might conduct an in-depth analysis of their target audience in order to improve their reach and impact.

Furthermore, measuring the actual impact of counter speech and alternative narratives on attitudes and behavioural changes requires longitudinal studies, measuring the effect of campaigns on a specific set of audience over years. As the research on formats at hand has been rather conducted over a set of days than years, focusing on campaigns run during so-called action days, the results are biased in so far that the quantitative data weighs more as qualitative data is challenging to obtain. Additionally, the qualitative data gathered as part of the sentiment analysis might be misleading as trolls and haters might have organised to dominate the comment sections; therefore, it might be useful to test formats and narratives in different settings.

In addition, the current circumstances related to the Covid-19 pandemic and the subsequent spread of misinformation provided challenges for the implementation of the project in general as well as the different building blocks of the research; regarding the research on format and messaging, a general observation has been that content focused on debunking conspiracy myths and fake news, presented in infographics, received more attention (and also benefited more from the adjustments of the respective social network companies as they aimed at increasing the reach of fact-checking and debunking content) than content on hate speech, in general. These observations contradict previous research and analyses, which suggested that especially humorous pictures and GIFs, but also videos are the most popular formats online. The increasing popularity of infographics, especially amongst younger audiences, has also been mentioned by the representatives of the social media companies, most notably Instagram.

Considering these factors, the research setting and methodology and taking into account specifically the findings aforementioned, but also remembering the definition of counter-narratives themselves and their dependence on the hate narratives they seek to counter, it clearly appears that all the work of activists and organisations is heavily influenced by the nature of hate and the forms it assumes in the public discourse.

This is one of the most peculiar aspects of counter narratives in comparison to alternative narratives, especially according to the definition on which the activist practice adopted is based and that comes from the We CAN manual (Chapter 5 p. 81).

Alternative narratives do not focus so much on the reaction to single events, but rather work to build an alternative mind-set to sustain the change they want to bring into society. Another important goal is to occupy public space by letting an alternative (coherent and structured) position be heard in public discourse.

Counter and alternative narratives co-exist: activists intending to produce changes in society will eventually need to go from reactions to emergencies, to the development of wider alternatives sustaining their actions; at the same time, movements engaged into a long-term path cannot avoid reacting to specific challenges, events or outbursts.

Counter narratives are in fact intelligible and make sense when are directly related to specific hate incidents and hate storms and are inherently bound to these manifestations to be able to intercept specific elements and to work on them to produce new messages that can serve as a reply. For this reason in a global scenario where the COVID19 pandemic absorbed most of the audience's attention and energy, hate found different ways to act, going beyond usual patterns in order to keep promoting certain messages, using also content different in style, as in the case with the already cited infographics.

In the first part of the pandemic and consequent lockdown though, most of the news and of the relevant events were aiming to share hope and unity in the population, calling for collective responsibility in the face of a tragedy that was affecting the whole world.
This caused an apparent decrease of hate, which was still present but more subtly, as we are trying to argue later.

From an activist perspective this was the perfect moment to launch more alternative messages, trying to gather the population around human rights based values and sentiment, while engaging in counter narrative actions targeting minor hate occurrences (always present), would have risked to highlight such hate, amplifying negative instances only with the idea of measuring the performance of such posts.
For the whole activist group it was clear from the beginning that while performing the research it wouldn't be possible to compromise the fundamental meaning of the activist mission and to risk going against the common practice and interest, even for the sake of scientific accuracy.

Such reflections did not come as a blind adherence to a prescribed way of activism but rather to escape a more performative vision of it, which usually doesn't consider the necessary time to reflect and evaluate its own actions. Even more so during an unprecedented time, which challenged our public and private lives, affecting every aspect of our behaviour and urging us to find new ways to communicate, to behave and to transmit messages that would allow to promote change and human rights principles.

Considering these elements our common decision, as a partnership, to keep conducting the research has become the occasion to gather data that apart from helping us in the demonstration of our thesis, could present us with a picture of the ways in which such situation influenced the work with hate speech, as it has then been further proved by many studies and research that were conducted and are still ongoing.

The most significant impact of the pandemic is surely connected with the absolute number of people reached by the messages, which were still present and moving towards the fortification of some of the most powerful hate narratives (e.g against migrants and Roma communities or against women), but were received in a more passive way, not generating much debate and/or media attention, unless they were related to the COVID topic, like the counternarrative action carried out on the 10 March responding to a hate wave against young people, accused of behaving irresponsibly causing the rise of COVID infection.

Such tendency is proven by the fact that some of the most notable hate incidents we observed in Italy, in terms of both visibility and endurance in the public discourse, happened as soon as the heavy lockdown has been lifted and there has been a general sense of relief in the population, with the assumption that things could go back to normal and the COVID19 had been defeated.

Apart from what it is possible to observe in relation to counter narrative actions to hate incidents, we noticed such tendencies also on the contents produced for our social media pages, especially in relation to Counter Action Days, which registered a rise in the numbers of reach and reactions in the general audience of the Facebook page and a general increase in the number of likes and followers.
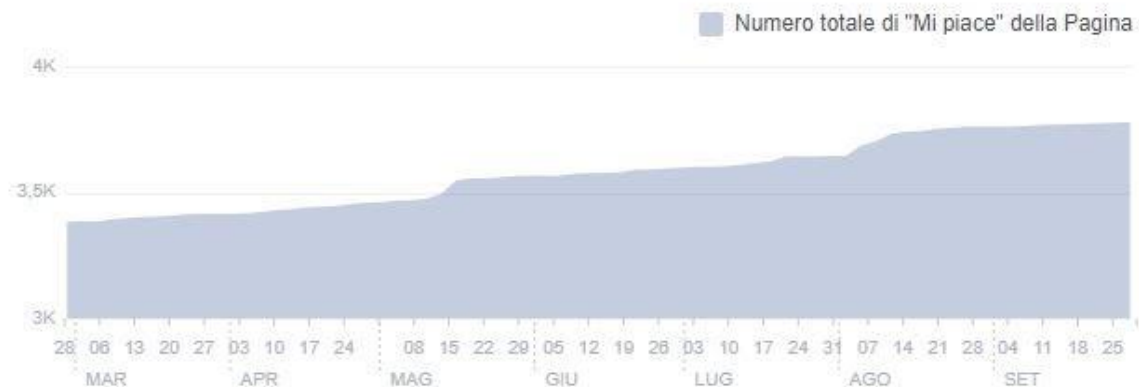
Followers by 1 March 2020: 3285 (pandemic outbreak)
Followers by 16 May 3549 (end of 1st lockdown period) +164
Followers by 30 September 2020: 3766 (End of the flattening curve) +217

If we compare such results with the same time in 2019 or 2018, we could notice that summer months are usually less popular for online activism, with few exceptions connected with certain international days or anniversaries.

It is usually difficult to catch the audience's attention and engagement, since in Italy the months between June and September mark the holiday season. Such a shift is a further proof in the change of habit caused by COVID19 and it is the reason why we would ideally repeat our research, so that it would be possible to assess if we are in front of a so-called "new normality" or to understand if what we witnessed has been a temporary effect that would soon be absorbed by the usual and historical patterns we're more accustomed to.[5]

Lastly, the insights and recommendations provided by the representatives of the social media companies primarily focused on Germany, though similar developments can be observed across Europe. However, user profiles and audiences differ slightly on the different platforms while other platforms, such as TikTok, become increasingly popular, especially amongst young people, but have not yet been considered in the research.

Nonetheless, the obtained data provides some general insight when drafting a counter speech or alternative narrative campaign, offering general guidelines that might benefit reach and impact.

VII.   Conclusion: lessons learned and recommendations

Despite the fact that the circumstances and developments due to the Covid-19 pandemic significantly influenced data collection and the research results in general, the current

---

[5] An interesting study conducted during the pandemic highlighted the more organised manifestations of hate connected with COVID19 and was focused around the use of social bots to influence narratives, especially in the US and the Philippines. The study interests a different geographical area and it analyses Twitter, which is usually the most popular social media for research purposes as shown by previous projects and researches conducted on the topic hate speech, also in Italy . See Contro l'odio: https://controlodio.it/ and the maps produced by Vox Diritti: http://www.voxdiritti.it/la-nuova-mappa-dellintolleranza-5/.
Nonetheless this study highlights an important factor in the adapting nature of hate narratives that we could witness directly when we became primary targets of a hate storm on our Facebook page. As showed and argued in the research in fact:
Without discounting that the broader public may organically engage in hateful talk online [71, 72], the evidence that bots exert significant influence in driving digital toxicity nonetheless makes it important for researchers to examine their potential impacts in the context of the pandemic [74, 79]. For instance, from a BEND perspective, bots can promote networked hate by building hate groups and backing hate-promoting opinion leaders. In addition, they could shape narratives through various manoeuvres such as distort or dismay to increase the volume of messages spreading hate."
(https://link.springer.com/article/10.1007/s42001-020-00087-4#Sec6)
More information on the August hate storm can be found in Appendix II.

situation also emphasised the importance of acts of solidarity, online as well as offline – which is actually in line with research conducted by NdM and #ichbinhier in Germany in 2019. As part of the data analysis of digital online attacks in 2019, victims of coordinated digital violence have been interviewed, all of them concluding that acts of solidarity have been the most powerful and effective help against continuous hate messages. Even though the spread of hate speech and conspiracy myths increased in recent months, it has also been observed that solidarity campaigns gain more traction, suggesting that future counter or alternative narrative campaigns should emphasise the empowerment of and solidarity with people targeted by hate speech more strongly.

However, empowerment and solidarity campaigns should consider the following aspects: It is of the utmost importance to include credible partners and speakers, though their reach and popularity might not benefit a campaign significantly. In their Review of programs to counter narratives of violent extremism, Rachel Briggs and Sebastian Feve (2013) of the Institute for Strategic Dialogue conclude that authentic, credible messengers need to have an authority with the target audience. The International Centres for Policy Advocacy analyses sample campaigns as part of the project [Reframe the Debate](#) further showed that it is not enough to list credible speakers, but that their voice and perspective needs to be seen/heard. Furthermore, the speakers should rather be seen 'in action' in their daily life rather than from the back or artificially positioned.

The research on formats further indicates that even though formats are a relevant factor to consider when drafting narrative campaigns, they are not the decisive one or at least not decisive in their own regard: Instead of solely focusing on developing creative and innovative formats to attract an audience, the research suggests to, first and foremost, identify the target audience and its characteristics: Where is the intended audience active, what kind of content and which speakers are popular? If there are more target groups, especially when they are active on different platforms, different formats and messaging might be needed and beneficial. Furthermore, social media campaigns in general benefit greatly from measures such as network activation, increasing the chance to reach a wider audience and having an impact.

Analysing the results coming from the research on timing instead, it is possible to conclude that an online counter narrative responding to a hate incident is more effective if applied in a time slot of 12 to 24 hours after the incident. Responding before the 12 hours can be ineffective, if the counter narrative is hidden by hateful emotions and reactions, fake news, polarization and confusion or, in the worst case, if not accurately designed and launched, it can feed the trolls and fuel an hate wave leading to a hate storm. Responding after 24 to 48 hours instead can result in a reduced impact, because the incident is already digested and the users already formed their opinion about it. In this case the proper intervention to be done would be an alternative narrative campaign, aiming to reverse the hateful opinion consolidated in the general public. All these conclusions do not apply in case of a hate storm, because after 12 hours from the incident the storm is still ongoing.

In the case of a hate storm (see Appendix II for more information) instead, you might need more time in order to get out of the storm and to find the way to identify the right moment to counter it, which should be defined according to the characteristics of the hate storm: more case study investigation is needed in order to identify the best timing for a counter narrative intervention toward a hate storm. In any case, we recommend not to post a new counter narrative during the hate storm as it could trigger the hate wave again.

Another conclusion is that it is better to avoid the posting of multiple counter narrative products in a short period of time: the first one will always have a better performance because of Facebook algorithms and because users will not be as drawn as they would be by new contents on a different topic. The best strategy is to prepare a standalone counter narrative

and to post it after 12 to 24 hours from the occurrence of the hate incident because it's usually the time in which the audience is still engaged but emotionally more stable and open to reflect.

Despite this research benefitting practitioners and activists, more research on case studies of hate incidents are urgently needed: digital phenomena such as hate speech and misinformation, though widely spread, are still not fully understood and investigated, especially their virality in social networks. Similarly, research on the efforts to respond to harmful content and violence online is almost non-existent, though academia slowly, but surely focuses on it, which is a commendable development.

Nevertheless, more research is needed as we can see that democracies around the world are under threat and the violence and hate incited on social media transgresses to the analogue world.

Appendix I: Further information on the members of the consortium for the "WE CAN for human rights speech" project

APICE is a national youth NGO composed of young people under 35 years of age, aimed at promoting a common European culture by raising awareness among citizens and by disseminating and promoting the fundamental values of the European Union and of the Council of Europe. APICE contributes to the improvement of active youth participation, social development, dialogue and cooperation, by supporting the participation of citizens, NGOs, public and private authorities, to initiatives and programs promoted by the EU and the CoE, particularly in the field of youth, human rights education, social inclusion, access to the rights of young people from disadvantaged backgrounds, engagement of youngsters at risk of criminalization, inclusion of migrants, refugees and asylum seekers, youth participation in decision making processes, youth mobility, activism combating hate speech online and offline, media literacy, internet governance, urban regeneration and graphic arts. APICE also supports public institutions for the implementation, coordination and management of Europe Offices, Eurodesk, Europe Direct and other similar offices, providing experts, project managers and a Pool of trainers and facilitators. .APICE is the national coordinating organisation of the No Hate Speech Movement Italy. Since the publication of 'We CAN!', the organization have led three successful international 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, APICE rolled out online counter narrative campaigns responding to hate incidents. APICE is a funding member of the No Hate Speech Network, contributes to the organization of Youth IGF in Italy and has been represented at EURODIG event during several editions.

Neue deutsche Medienmacher e.V. (NdM) is a Berlin-based, but nation-wide operating NGO in Germany. Primarily focusing on increasing diversity in media and editorial offices, NdM mostly implements projects targeting journalists and media officials. Since 2016, the NGO coordinates the No Hate Speech Movement in Germany, working closely with ministries, civil society organisations and activists. As part of the No Hate Speech Movement, the NGO interviewed experts on hate speech as well as online activists to identify common strategies of people spreading hate and racism online, but also collect effective counter speech strategies. In an additional measure, the team of the No Hate Speech Movement Germany analysed, in cooperation with civil society movement #ichbinhier, coordinated online attacks and talked to the victims of so-called hate storms to identify which measures actually benefit the people targeted by hate speech and discrimination online. Their findings were published in an all-encompassing webtool, the so-called helpdesk, aiming at facilitating counter speech in commenting sections on social media platforms.

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.

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.

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.

Appendix II - Description of August Hate Storm and tips for counter action

In describing the research model on timing and the type of contents produced to collect data, it has been mentioned how these represented only part of our usual online presence, meaning that there were still posts and content from the social media accounts of No Hate Speech Movement Italy dedicated to general activities, mainly information and dissemination of good practices and initiatives, together with support and solidarity actions related to hate speech.

In accordance to such general online presence, two posts were shared on the Facebook page of the campaign to express support to the work of a famous professor and sociolinguist, Vera Gheno, who has written several books, collaborated with many newspapers, research hubs and institutions such as L'Accademia della Crusca. The decision to support of Professor Gheno was linked to her contribution to the discussions about inclusive language in Italian, for a long-time she has urged Italian linguists to research and put more effort into finding ways to express the feminine and the non-binary in Italian language, which uses an "universal male" that includes the notion of collectivity, symbolically erasing every other experience other than the masculine one.

Her theories are not going to be presented here in too much detail because it is beyond the scope of this research[6], but to synthesize some of the main points of her possible solution towards a more inclusive use of Italian, it is important to understand that she proposed to add to some of the elements already used in activists groups and initiatives (mainly trans-feminist ones). These strategies include the use of the letter -u as a neutral (adopted also within the activist group of No Hate Speech Movement Italy), the use of the @ or the *. Among these, Professor Gheno introduced the use of a phonetic symbol, the schwa, ə, already common in some dialects to end some words, either masculine or feminine.

Prof. Gheno has always stated that this, similarly to the other innovations, adopted in other countries like Sweden or Spain, is just a proposal for action, something that should serve the purpose to initiate discussions and debates around the inclusivity of language, which is an important representation of who we are and who we want to be as people. Such a proposal has been mocked and attacked from the beginning and seen as an assault towards the "purity" and the "beauty" of the Italian language. Considering that such attacks were often personal and against Professor Vera Gheno herself, the group of NHSM Italy decided to reshare one of the interviews in which she explained her proposal and general approach towards this issue, in combination with a statement of support to show solidarity with her and her position. On the same day, 4 August 2020, a card produced for the action day of the 17th of May, against Homo-bi-transphobia (IDAHOBIT) was shared, since it was part of a post in which the campaign had already expressed the decision to use a more inclusive language in its activist practice and to use the letter -u for some collective forms and greetings whenever it would not be possible to use gender neutral forms.

To further support Professor Gheno and her work, the activists made a twin card, using the ə in place of the -u, while also reinforcing the idea that everything concerning language is always a process that implies dialogue, listening and discussion and that while there are no fixed decisions on the matter, there should always be the attempt to go for the opening of new spaces that would allow as many people as possible to be able to express themselves and their reality in freedom and safety.

---

[6] For a more detailed knowledge of Vera Gheno's work, we recommend reading her book and articles, which are unfortunately mostly in Italian. Here though there is an article on her general views and approaches in English: https://medium.com/redshirts/power-to-the-words-6b21f7e72398

In the usual style of the campaign page, the cards were posted together with a comment expressing in more detail the activist group's take on the matter and clarifying the position of dialogue and openness.

Being used to an audience made essentially of other activists, experts, professionals or people interested and sensitive to the topic of hate speech, the social media pages of NHSM Italy have never faced severe backlash for any posts, apart from some rare cases. Given the attention that inclusive language and the introduction of the schwa was seeing in the public discourse though, these particular cards caught the attention of a couple of people who populate Facebook pages and groups dedicated to "shitposting" and where there is the rejection of everything considered as "politically correct", generally seen as a negative thing that want to regulate people's freedom of expression making them speaking in a "ridiculous way".

The cards were soon reshared, getting thousands of reactions and interactions, attracting many offensive and derogatory comments, questioning mostly the sound of the proposed expression and with some genuine criticism towards the idea in general, very negative but not necessarily hateful and/or offensive.

After a few hours though, while the activists were responding to some of the strongest comments by using data, debunking and humour (a common strategy applied in comments sections and that the NHSM Italy experience shares with other Italian initiatives, such as the Task Force Hate Speech led by Amnesty International Italy[7]), it was possible to notice a change in the tone of the comments. Together with general mocking and slurs, popular

---

[7] https://www.amnesty.it/entra-in-azione/task-force-attivismo/

memes and GIFs started to appear, some of them coming from alt-right fora and using precise homo-transphobic language and images to diminish the level of the discussion and trying to delegitimise the inclusion of non-binary people, especially trans individuals.

This tendency was even more visible after one day, when trans-exclusionary radical feminists (TERFS), together with another couple of accounts that it was possible to connect with such ideology, started to post more and more comments, endorsing each other's views and attacking the proposal but in a subtle way, using fake arguments, such as that supporting inclusive language would erase women and the idea of feminine to favour the "perverted" trans nature and other very radical views, very violent but sustained by very long comments that would strongly seek to appear scientifically accurate and logical, guided by "common sense". Shortly after, even pro-fascist accounts joined the discussion, with their expressions and symbolism.

In registering such reactions, the campaign's coordination called an emergency meeting with activists, to decide what to do with the post, since it was clearly attracting dangerous views and providing a platform to share them, with the risk to amplify them when responding directly. Keeping in mind this risk, it was decided to leave the post but to stop commenting to the attackers, while preparing a counter-narrative post to expose the danger and the violence of such views and also looking for the support of Prof. Gheno herself through Federico Faloppa, Co-ordinator of the Rete Nazionale per il contrasto ai discorsi e i fenomeni d'odio (National Network to counter hate speech and hate phenomena), who has previously co-operated with the socio-linguist and who wanted to support the campaign after the harsh attack it experienced.

Following the decision on the strategy, a collection of some of the worst comments was posted on 7 August, 3 days after the cards were first published, with a commentary saying that considering the reactions that the original post had collected and the fact it had become a platform for homophobic, transphobic and fascist propaganda, as a campaign there should be the acknowledgment that something went wrong and that if a post made to show support and initiate a constructive debate became a host for hate speech, as an activist group there was a duty to produce a clearer content, which would not be misunderstood or be open to manipulation and that would give justice to the important debates surrounding the use of inclusive language and the people it would benefit.

No Hate Speech Movement Italia
Published by Debora Barletta [?] · 7 August 2020 · 🌐

📣 SPAZI DI CONTRONARRAZIONE

Da campagna attivista fondata sul rispetto e la diffusione dei diritti umani, abbiamo sempre messo al centro delle nostre attività di contronarrazione l'ascolto e il dialogo: per questo motivo il nostro operato è aperto a critiche e confronti che riteniamo possano portare ad una crescita mutuale delle parti coinvolte.

Non ci aspettavamo però che i commenti al nostro post potessero assumere toni tanto violenti, travalicandone e fraintendendone lo sc...

See more

Performance for your post

2,332 People Reached

140 Reactions, comments & shares ⓘ

	On post	On shares
👍 Like **32**	21	11
❤️ Love **3**	0	3
😄 Haha **1**	1	0
😮 Wow **1**	1	0
😢 Sad **2**	2	0
😠 Angry **2**	2	0
Comments **89**	36 On Post	53 On Shares
Shares **10**	8 On Post	2 On Shares

308 Post Clicks

This post attracted incredibly less reactions and engagements, but still 36 comments, mostly from the same extremist accounts who initiated a stronger propaganda under the original post and to whom other activists and supporters of the campaign decided to respond directly.

Nonetheless and as promised, the activist group worked in collaboration with Professors Faloppa and Gheno to produce a new post with a more comprehensive explanation about what inclusive language is, what it stands for, and why it could never be associated with an oppression or an attack on anything, but actually as an attempt to expand our perception in order to improve our common experience as members of a diverse and complex society.

On 12 August, 8 days after the cards were posted, the counter-narrative content was shared, with a discreet amount of success, especially thanks to other activists, groups and initiatives involved with countering hate speech and that supported the campaign. Nonetheless, as it is possible to see in the insights, there has been one "hide post" feedback and 111 comments, 41 of which it was not possible to monitor or analyse since they appeared on private pages/groups and/or profiles, limiting our possibility to further assess the level of penetration of our counter narrative outside our usual area of influence.

Not having the opportunity to really analyse the sentiment connected to those comments and to what could not be traced, it is possible to confirm that the peak of attention towards the topic was certainly diminished, even though hate speech and attacks to Prof. Gheno and inclusive language still continue, there is no account of another hate storm or of other strong hate waves on the topic in the subsequent weeks and months.

Following this occurrence there was then the need to reflect on the whole episode, to understand the reasons behind the hate but also to assess the impact of the activist's work in countering such hate manifestations.

All the strategy came from the application of the activists' practice principles but it had to face a violent attack that exposed activists directly, engaging them in long and stressful discussions with people who appeared to be organised and to pursue a precise goal.

The commitment towards the mission and the vision of the campaign had an unifying effect and helped the group to feel more connected and even motivated to keep up with their activist work, but for those in charge to coordinate the group and the campaign it represented a very challenging moment, since it put the whole movement under scrutiny and if it had not been handled well it could have compromised months of work.

Conclusions

Hate storms have proved to be connected with topics trending in public discourse but their ignition can come from different sources, even activist spaces, which can be suddenly invaded by extremists and violent instances working in organised attacks. These occurrences challenge the work of activists and organisations countering hate speech, especially when they are trying to perform their work in accordance with human rights values,

as they demand a huge amount of preparation, confidence, trust and endurance, but also the ability to remain calm and to strategize in order to respond in an appropriate time frame (such as the 24 hours suggested by the research above), and introducing a valid counter narrative.

Tips for Action

According to the experience of the No Hate Speech Movement Italy it is possible to assume that even in the most difficult situations, there should be a set of principles to follow. You can find them listed below, with the hope that a more comprehensive guide on how to deal with such situations could be produced soon.

1. Keep track of the sharing of your post and research the possible affiliations of accounts and pages connected with the hate storm. Do they belong to some more organised initiatives? Do they seem to be "casual" but using certain languages and symbols belonging to extremists/organised initiatives?
2. Keep an ongoing communication with the activist groups/the people in charge of moderating the content on your account. Make sure they always feel safe and in control of the situation.
3. Find allies and friends who can support you in responding to comments as well as experts that can give you the necessary insight to produce valid counter narrative contents.
4. Be ready to reconsider your content, be critical of your own work and ask yourself if the hate was involuntarily helped by some inaccuracies and/or manipulations. Was it the right content to deliver your message? Sometimes defending unconditionally our work can do more damage than having a critical and unbiased (as much as it is possible to have one) reflection on what you did. Showing honesty and the ability to be open to discussion can go a long way in creating opportunities for meaningful dialogue and even a significant change in attitudes.

Appendix III: Literature & further reading

Bartlett, J. & A. Krasodomski-Jones (2015): *Counter-Speech: Examining content that challenges extremism online.* London: Demos. Available via: https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Briggs, R. & S. Feve (2013): *Reviews of Programs to Counter Narratives of Violent Extremism - What works and what are their implications for government?* London: Institute for Strategic Dialogue. Available via: https://www.dmeforpeace.org/peacexchange/wp-content/uploads/2018/10/Review-of-Programs-to-Counter-Narratives-of-Violent-Extremism.pdf

Council of Europe (2017): *WE CAN! Taking Action against Hate Speech through Counter and Alternative Narratives.* Available via: https://rm.coe.int/wecan-eng-final-23052017-web/168071ba08

Facebook Business Help Centre, Definition of Post Engagement: https://www.facebook.com/business/help/251837856080253, last accessed on 29.01.2021.

Facebook Business Help Centre, Definition of Reach: https://www.facebook.com/business/help/710746785663278, last accessed on 29.01.2021.

Freie Universität Berlin, Department of Political and Social Sciences/ Institute for Media and Communication Studies: *No Hate – Overcoming crises in public communication about refugees, migration, foreigners*. Presentation of current status of the research project during the meeting of Germany's National Campaign Committee in November 2018.

Frenett, R. & M. Dow (2016): *One to One Online Interventions: A Pilot CVE Methodology.* London: Institute for Strategic Dialogue. Available via: https://www.isdglobal.org/wp-content/uploads/2016/04/One2One_Web_v9.pdf

Garland, J.; K. Ghazi-Zahedi, J. Young, L. Hébert-Dufresne & M. Galesic (2020): *Countering hate on social media: Large scale classification of hate and counter speech.* Available via: https://arxiv.org/pdf/2006.01974.pdf

Goncalves, P.; M. Araujo; F. Ribeiro; F. Benevenuto & M. Goncalves (2010): *A Benchmark Comparison of State-of-the-Practice Sentiment Analysis Methods.* ACM Transactions on Embedded Computing Systems 9 (4), available via: https://www.researchgate.net/publication/286302059_A_Benchmark_Comparison_of_State-of-the-Practice_Sentiment_Analysis_Methods

International Centre for Policy Advocacy, Reframing Migration Narratives Toolkit: http://www.narrativechange.org/, last accessed on 29.01.21.

Kreißel, P. & S. Laubenstein (2019): *Digitaler Hass gegen die Demokratie.* #ichbinhier, available via: https://www.ichbinhier.eu/digitaler-hass-gegen-die-demokratie

Lawal, M. (2018): *Reichweite versus Impressions: Worin besteht der Unterschied (und was sollten Sie tracken)?*, Blog Hootsuite, available via: https://blog.hootsuite.com/de/reichweite-versus-impressions-was-sollten-sie-tracken/

Reynolds, L. & H. Tuck (2016): *The Counter-Narrative Monitoring & Evaluation Handbook*. London: Institute for Strategic Dialogue. Available via: https://www.isdglobal.org/wp-content/uploads/2017/06/CN-Monitoring-and-Evaluation-Handbook.pdf

Schwarz, K. (2020): Hasskrieger – Der neue globale Rechtsextremismus. Freiburg: Verlag Herder.

Uyheng, J. & K. M. Carley (2020): *Bots and online hate during the COVID-19 pandemic: case studies in the United States and the Philippines.* Journal of Computational Social Science 3, p. 445-468, available via: https://link.springer.com/article/10.1007/s42001-020-00087-4#Sec6

Tumasjan, A.; T. O. Sprenger; P. G. Sandner & I. M. Welpe (2010): *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* Association for the Advancement of Artificial Intelligence, available via: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.309.532&rep=rep1&type=pdf

Twitter Help Center, About your Activity Dashboard: https://help.twitter.com/en/managing-your-account/using-the-tweet-activity-dashboard, last accessed on 29.01.2021.

Ziegele, M.; P. Jost; D. Frieß & T. Naab (2019): *Aufräumen im Trollhaus: Zum Einfluss von Community-Managern und Aktionsgruppen in Kommentarspalten.* Düsseldorf Institute for Internet and Democracy, available via: https://diid.hhu.de/wp-content/uploads/2019/04/DIID-Precis_Ziegele_V3.pdf