Empowering NGOs in countering online hate messages

Yi-Ling Chung [a,b], Serra Sinem Tekiroğlu [a], Sara Tonelli [a,*], Marco Guerini [a]

[a] *Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy*
[b] *University of Trento, Italy*

ARTICLE INFO

ABSTRACT

Studies on online hate speech have mostly focused on the automated detection of harmful messages. Little attention has been devoted so far to the development of effective strategies to fight hate speech, in particular through the creation of counter-messages. While existing manual scrutiny and intervention strategies are time-consuming and not scalable, advances in natural language processing have the potential to provide a systematic approach to hatred management. In this paper, we introduce a novel ICT platform that NGO operators can use to monitor and analyse social media data, along with a counter-narrative suggestion tool. Our platform aims at increasing the efficiency and effectiveness of operators' activities against islamophobia. We test the platform with more than one hundred NGO operators in three countries through qualitative and quantitative evaluation. Results show that NGOs favour the platform solution with the suggestion tool, and that the time required to produce counter-narratives significantly decreases.

1. Introduction

The rapid growth of social media platforms makes communication and interaction easier and faster. Still these platforms can also become a fertile ground for hatred content [1]. While a huge number of users passively receive and digest news and messages from social media, some can actively generate inflaming comments and post them publicly on any subject while remaining anonymous [2]. The side effect is that hate groups may use social media platforms as a propaganda tool to spread hate speech, appeal to a wider audience, and marginalize specific communities [3,4]. This scenario has motivated authorities of many countries in the search for effective ways of fighting online hatred. For example, UNESCO published a report on countering online hatred [5], and the Council of Europe coordinated the No Hate Speech Movement[1] and Get The Trolls Out project[2] to address hate speech.

Countering online hate involves two main challenges: monitoring and intervention. The first task concerns mainly collecting information on who posts which hate messages, trying to understand topics, trends and communities from a large amount of social media data. The second includes several possible actions to concretely counter online hate speech. Both activities are challenging, since it is extremely difficult during monitoring to identify potential online threats in real-time and to further stop them from spreading. Standard approaches mainly rely on users' reporting activity that can lead to the deletion of hate content or suspension of haters' accounts. Still, manual intervention cannot govern the magnitude of the problem, and empowering censorship on social media may limit the right to freedom of speech and diverse opinions. An alternative and novel strategy is to directly intervene in the discussion with counter-narratives. Counter-narratives (CNs) are non-negative responses to hate speech, targeting and counteracting extreme statements through fact-bound arguments or alternative perspectives [6,7]. This approach is preferable for several reasons: the right to freedom of speech is preserved, stereotypes and misleading information are rectified with reliable evidence, the exchange of diverse viewpoints is encouraged (also among possible bystanders) [6,7].

Not all responses to a hate message can be seen as 'good' counter-narratives. We show an example below, where a hate-message (HM) is reported with two possible responses. The first, R1, is a non-negative and appropriate response, which we consider in this work as a good counter-narrative. The second response, instead, is not appropriate because it contains abusive language, and it therefore does not represent the kind of counter-speech that should be used to address online hatred:

HM: *The world would be a better place without Muslims. They are only killing and raping our children.*

R1: *The world would actually be a very dark place without Muslims who contribute a lot to our society. What about our Muslim doctors, scientists, philanthropists, job-creators?*

R2: *You are very stupid and dumb to speak negatively about Muslims. Hell is where you belong!*

Currently, several NGOs are engaging in hate countering activity on social media platforms. They typically rely on volunteers and operators who, after a specific training, learn to (i) detect harmful conversations, (ii) identify relevant hate accounts and (iii) either report the hate content or produce effective counter-narratives, in most cases using fake profiles for safety reasons. This manual intervention strategy is clearly not feasible on a large scale, since it would become too costly and time-consuming for NGOs to handle. We therefore propose in this work a novel online platform that enhances NGO operators' online activities (similar to the concept of 'digital augmentation' presented in [8]), by automatizing many tasks in monitoring and counter-narratives writing. This is done through a combination of natural language processing, social network analysis and advanced data visualization tools.

While several research works have addressed the automatic detection of online hate speech using natural language processing [9–12], limited studies have explored automation of hatred countering [13,14] due to the scarcity of hate-speech/counter-narrative pairs. The goal of employing a suggestion tool lies in providing automated responses that reduce the time and efforts of NGO operators in responding to offencive posts, thus yielding more systematic and sustainable impact against online hate speech. Furthermore, a partial automation of the 'seek and reply' task of NGO operators would alleviate the emotional burden of having to interact with hateful content for hours, a situation that has proven to affect with lasting psychological and emotional harm content moderators exposed to extreme content [15]. The issue of moderators' well-being has been tackled in the past via automatic image modification, see for example [16] that used image blurring, or [17] that tested the effects of simple interventions like greyscaling. To our knowledge, no previous work has addressed the problem by automating the creation of hate countering messages through an interactive and real-time interface.

The novel ICT platform presented in this work has been designed to systematically support NGOs in monitoring, analysing and countering online hatred in English, French, and Italian, focusing on Islamophobic messages. The platform includes two main features: the first allows users to promptly monitor hate speech on Twitter targeting Muslims and Islam. Then, users can rely on the second feature, a novel counter-narrative suggestion tool, to automatically compose suitable responses to hate content. The platform was tested by 112 operators across 3 different countries through an extensive evaluation that includes open-ended questions, a user experience questionnaire, and time measurement. The experimentation showed that the platform can effectively empower operators in the crucial task of hate countering.

NOTE: This paper contains examples of language which may be offencive to some readers. They do not represent the views of the authors.

2. Background and related work

In this section we briefly review the main works related to hate speech in terms of monitoring and countering.

2.1. Online hate speech monitoring

Online hate speech monitoring is a crucial task that is difficult to carry out manually due to the severe magnitude and scale of the problem. We identify two types of monitoring: real-time and retrospective. The former aims to provide continuous monitoring of social media, and it is more difficult to achieve because of the ever-changing nature of social media, communities and shared contents. Among the best known projects dealing with real-time monitoring there is the Umati Project [18], concerning the run-up to Kenya's general elections in March 2013. It involved several operators that manually checked selected blogs, forums, online newspapers, Facebook, and Twitter daily, in English and seven other languages, to track and report hate speech. Also the COSMOS platform [19] offers real-time data collection from

Twitter and data analysis both at tweet level, such as user sentiment, gender, and geospatial location analysis, and at corpus level, including keyword frequency analysis and hashtag usage. Despite the relevance of the project, however, human error was an issue and scaling up the monitoring operations was reportedly expensive. Other similar projects have been carried out around the world, including Myanmar, Nigeria, Sri Lanka, Ethiopia, Czech Republic and US. In some cases, toolkits and guidelines developed within the Dangerous Speech project [20] have been used, with the goal to classify and analyse hate speech and to diminish its harmful impact. However, no specific platform to support the monitoring process has been developed within this initiative.

More recent hate monitoring platforms feature classification and visualization of hate speech applying machine learning techniques. For instance, Contro l'odio [21] focuses on detecting anti-immigrants discourse in Italy and providing the temporal and spatial reports regarding hate messages on Twitter. Similarly, the system introduced in [22] supports the educators in high schools by identifying cyberbullying phenomena on Instagram and depicting the network of interest. MANDOLA [23] is a platform that emphasizes the capability of performing large-scale classification and illustration of hate activities. Although the aforementioned tools conduct analysis on hate incidents, as opposed to the platform introduced in this work, their functionalities are served for specific languages: English or Italian.

While for real-time monitoring we could not identify any ICT platform used by institutions and NGO activists to automatically track hate speech in real scenarios, more works exist related to retrospective monitoring, probably because storing and cleaning data before analysing them make the output more accurate and visualizations can be better controlled. Monitoring platforms have been developed so that users explore hate speech datasets after they have been collected and analysed with natural language processing techniques [1,22,24–27]. To our knowledge, no platform exists that provides both real-time and retrospective monitoring like the one presented in this paper. Furthermore, no one was also designed to respond to incidents.

2.2. Counter-narratives against hate speech

Existing strategies for combating hate speech on social media are mostly limited to content deletion, shadow banning and account suspension. A few recent works have proposed an alternative approach, i.e. responding to hate speech using counter-narratives that users can manually create following some guidelines and best practices [5,28,29]. Several works examined the effects of counter-narratives on social media, such as Facebook [7], Twitter [6,30], and Youtube [31,32], and proved that they are indeed effective in fighting online hate [7,30,31, 33,34]. Also, counter-narrative datasets have been created for research purposes [13,35,36].

While using counter-narratives is an advised approach, composing a counter-narrative is time-consuming and demanding of wide knowledge and training. Although this training is part of the activities of NGOs who want to fight online hate, this is generally not sustainable on a large scale. A traditional scenario for an expert operator in NGOs to compose a counter-narrative includes reading hate messages, searching for and filtering related information online, and composing a counter-narrative. While a few recent studies have attempted to automatize the creation of counter messages using natural language generation techniques [13,14], to the best of our knowledge, no work has tried to integrate this process in the workflow of NGO operators and evaluate it in a real scenario. In this work, we therefore aim at filling this gap by shortening the time and effort needed for locating hate content and composing proper counter-narratives.

Table 1
Islamophobic messages monitored between October 2018 and May 2019 in English, French and Italian.

	Tweets	Replies	Retweets	Total
En	108,062	94,874	1,004,550	1,207,486
Fr	151,738	268,548	1,206,347	1,626,633
Ita	35,569	27,562	213,713	286,844

3. A platform against islamophobic discourse

We present in this section the online platform we developed and its two main functionalities: islamophobic speech *monitoring* and *countering*. NGO operators can access the platform using a password-protected account. The monitoring functionalities focus on Twitter, a choice that is mainly motivated by restrictions by other social media platforms. Indeed, neither Facebook nor Instagram, two major players dealing with online hate speech, make their APIs available to researchers for data collection and analysis. Another reason why we focus on Twitter is that it is the social media platform that removes the least content after users report hateful messages, i.e. only 43.5% of reported messages are deleted compared to 71.7% by YouTube, Facebook, and Instagram on average.[3] Consequently, the issue of automatic hate speech detection on Twitter becomes particularly urgent and relevant.

The countering functionalities, instead, can either take Twitter data as input through a link between the two functionalities, or process any message pasted or written by NGO operators. The input message calls the underlying counter-narrative engine, which suggests possible responses to the hateful post.

3.1. Monitoring functionalities

The three NGOs that were involved in this study devote part of their activities to monitor islamophobia on social media platforms. In particular, they are interested in tracking ongoing discourse against Muslims, understanding which communities are more active in spreading hateful messages, and monitoring the events that trigger peaks in islamophobic messages. While designing the monitoring functionalities of the platform, we took these requirements into account, trying to display analytics that operators would not be able to obtain manually because they cross-reference and summarize information extracted from large amounts of data. On the other hand, operators need also a fine-grained view of specific days, terms or users, therefore requiring a close–distant reading paradigm [37].

In order to account for all these aspects, the monitoring functionalities cover both real-time Twitter activity and past messages, collected between October 2018 and May 2019. The analyses are available for English, French and Italian data, which were filtered by using the 'language' tag provided by the Twitter API. As a preliminary step, a list of around 20 keywords and hashtags related to islamophobic discourse was collected from social scientists and NGO operators for each of the three languages. These include for example #banislam, #stopIslam, #NoMoschee and #IslamHorsDEurope. The keywords and hashtags have been used to query the Twitter APIs (https://developer.twitter.com/en/docs/api-reference-index) and retrieve content containing at least one of the hashtags/keywords of interest over a period of time starting from October 2018. A summary of the monitored data is reported in Table 1, showing the number of tweets containing at least one islamophobic term or hashtag, divided into original messages, retweets or replies for each language.

If a user wants to monitor islamophobic discourse *in real time*, the Twitter API is called on the fly searching for one of the key terms or

hashtags described before, and several analyses are displayed. These include the 5 islamophobic keywords and hashtags that have been most recently used, the network of hashtags and keywords that are currently being used in association with the one of interest, as well as the most recent tweets posted online that contain an islamophobic term. A screenshot of this view is displayed in Fig. 1.

The monitoring activity can also be *retrospective*, i.e. concern islamophobic messages posted online in the past. The platform includes therefore the possibility to select a period of time of 10 days and a predefined hashtag/keyword of interest, and to display the online activity around this term, such as the network of users that most posted or retweeted messages containing this term, the hashtags that most co-occurred with the one of interest (Fig. 2), and the most popular messages containing it. It is also possible to display the frequency with which the term has been posted online in the last months, with peaks flagged with a red dot when the frequency is above the average plus one standard deviation (Fig. 3), so that operators can use this information to connect peaks with past events occurring on specific dates.

From a technical point of view, the platform relies on a relational database and a Tomcat application server. The interface is based on existing javascript libraries such as C3.js (https://c3js.org), D3.js (https://d3js.org) and Sigma.js (http://sigmajs.org).

3.2. Counter-narrative suggestion approach

The main hate-countering feature consists in assisting operators in writing counter-narratives (CN) based on automatically generated suggestions. There is a specific tab of the interface that can be either activated from each hate tweet (see the blue bubble button present for each tweet in Fig. 1) or via menu. While in the former case the content of the hate tweet and CN suggestions are directly displayed, in the latter case the user must copy-paste the hate content of interest. Fig. 4 illustrates how the platform offers a list of suggestions for a given hate content. As displayed in the figure, if a user enters a hate message in the top box and clicks on the dialogue icon on the right, the platform will output a list of counter-narrative suggestions below the hate content. For each suggestion, a user can directly modify/delete the suggestion if necessary or copy it to any social media platform for intervention. At the end of the page of the platform, a user can add a new counter-narrative from scratch if needed.

For implementing the counter-narrative suggestion tool, we have adopted a data-driven approach and we have constructed a database of hate message/counter-narratives written by NGO operators. The collected dataset consists of ≈6K hate message/counter-narrative pairs for English, ≈5K for French, and ≈3K for Italian, see [35] for further details on this dataset.

As a suggestion methodology, we have implemented a tf–idf information retrieval model [38] that, given a hate message as input, selects the most suitable responses (i.e. counter-narratives) from a database. tf–idf is a statistical measure that shows the importance of a term for a given document in a collection of documents and it is calculated as the product of the term frequency (tf) and the inverse document frequency (idf). We built our tf–idf model by calculating the tf–idf word-document matrix using the collected hate message/counter-narrative pairs as the document collection. For retrieving the counter-narrative suggestions, we first generated the tf–idf vector of the hate message given in input by an operator. Given the set of all hate message/counter-narrative pair vectors, the ones with the highest cosine similarity w.r.t. the input hate message vector are selected as the output pairs. Then, the counter-narratives of the top 4 pairs are presented in the platform together with an empty response text box in order to let operators write their own message if none of the suggestions presented is to their liking. Such new messages will be acquired as novel examples and included in our database.

We use a standard information retrieval approach since we aim to provide carefully created, grammatically and semantically correct responses to the operators. Although neural response generation [39,40]

Fig. 1. Screenshot of the real-time monitoring view: after selecting the #FUCKISLAM hashtag, the network of co-occurring hashtags is displayed, together with the most recent tweets containing it. A small bubble button can be clicked to directly obtain CNs to that particular tweet.
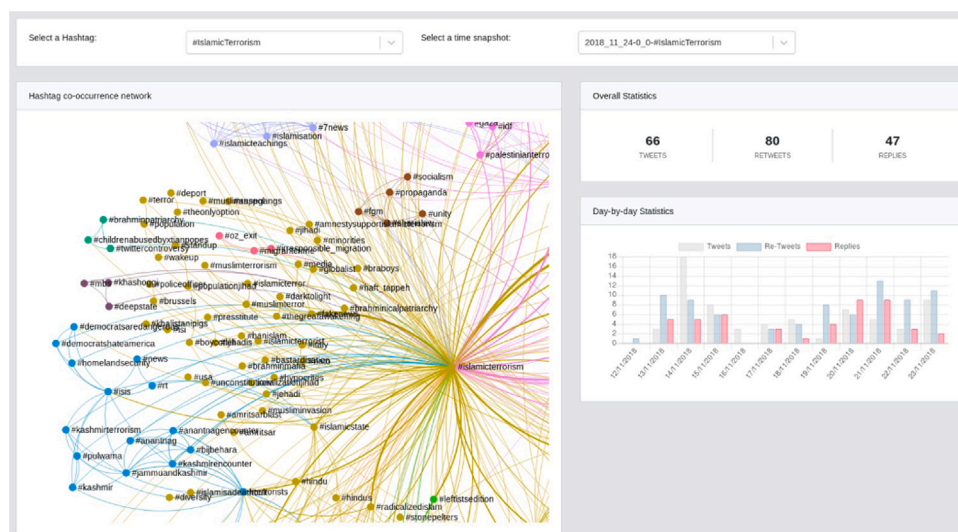

Fig. 2. Screenshot of the monitoring view for past tweets: after selecting the #IslamicTerrorism hashtag and the time slot, the network of co-occurring hashtags is displayed, together with overall statistics on the use of the hashtag in the period of interest.
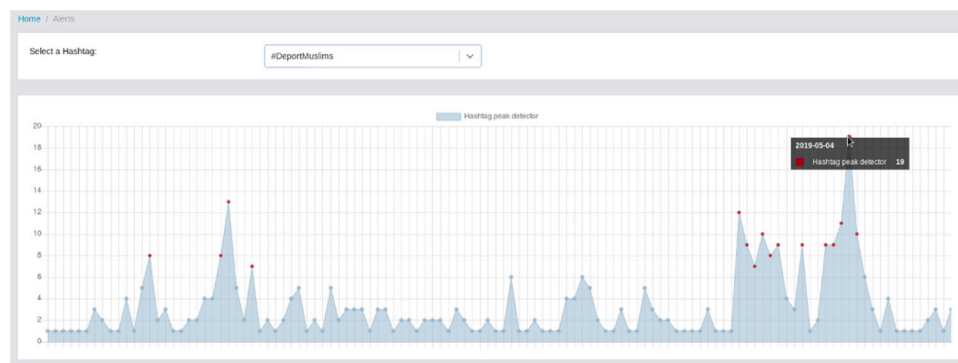

Fig. 3. Screenshot of the peak detector view: after selecting the #DeportMuslims hashtag, the number of occurrences for the given hashtag on each monitoring day is reported. Peaks are highlighted in red, with the associated date.

and argument generation [41,42] techniques have also been proposed recently, they still have limitations that might lead to substandard generations. For instance, the tendency to produce responses that are dull, generic and not engaging [43–45], or to overproduce negations [42] would be impractical to effectively counter online hatred. Furthermore,

since deep neural models are inherently data-hungry, they need much more data than what we collected to scale up properly. From an argumentation point of view, [42] discusses that automated generation systems have not been able to fully capture the organization
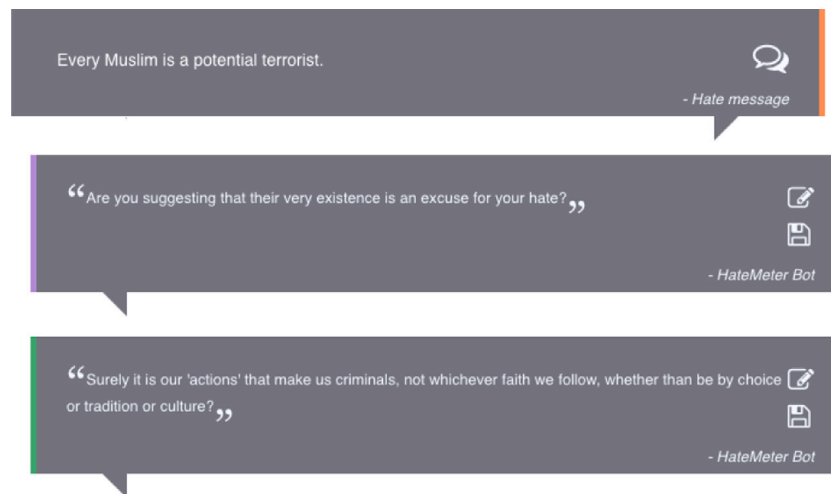
Fig. 4. On top the hate content "Every Muslim is a potential terrorist". Below a list of possible counter-narratives automatically suggested for that hate content.

and complexity of human arguments yet. Considering the aforementioned issues, we adopt an information retrieval strategy, which guarantees informative, engaging and well-organized responses for our counter-narrative suggestion tool.

4. Experiment setup

The goal of our experiment is to evaluate the effect of using our platform – and in particular the suggestion tool – in daily activities of NGO operators to fight hatred online. In our experiments we decided to focus on the CN aspect of the platform instead of monitoring. Since some of the monitoring functions are simply not doable by hand, it is not possible to obtain a clear control condition over them. On the contrary, the counter-narrative activity has always been carried out manually by operators and we can clearly measure the effect of its partial automation. In particular, we test if the CN suggestion approach is effective to help operators construct CNs in terms of quantity, usability (and quality).

Design. To test this effect we set up an evaluation framework with a between-subjects design and two conditions: using the platform with or without the counter-narrative suggestion tool (CN^+ and CN^- conditions respectively).

Subjects. In total 112 operators from three NGOs in different countries (UK, France, Italy) took part in the experiment. Specifically, **19** from UK, **42** from France, and **51** from Italy. These subjects are all volunteers that regularly contribute to NGO activities to fight islamophobia online and undergo specific training to comply with NGO code of conduct and fairness when it comes to online behaviour. Subjects were randomly assigned to one of the conditions, so that each operator was only exposed to a single interface version. In particular, 62 subjects were assigned to the CN^+ condition and 50 to the CN^-.

Procedure. For the experiment procedure, a deployment day was set up for the CN^- version and another one for the CN^+ version of the platform. Operators were asked to gather in NGO's premises on the day of the evaluation, and everyone was equipped with a computer connected to the platform. The first part of the deployment day was meant to introduce the platform and its functionalities. After the introduction, a phase of some usability exercises/tasks was carried out to prime each operator in the same way and to ensure that they tried all the characteristics of the platform. Finally a questionnaire was administered to the subjects. For each language/NGO we followed the same procedure and a total of 6 deployment days (2 for each NGO) were thus organized.

Instructions. The usability tasks were meant to let operators inspect all the functionalities of the platform. For example one task was "*Starting from the 'Hashtag trends' window, please analyse the key hashtags and keywords related to the concept of 'Islamisation', checking the data obtained from the network analysis and the graphs that the platform accomplishes. Later, please identify the hate speakers who mostly utilize the hashtags and keywords that you have identified*". For the CN^+ condition, in two of these tasks we asked operators also to identify the most relevant tweets and use the suggestion tool to answer them, as they would have done in their normal CN activity.

Measurements. For each condition we collected a mix of subjective (questionnaire) and objective (based on activity logs) measurements. Subjective responses were both quantitative and qualitative.

5. Questionnaires

To evaluate the operators' experience with the platform, two questionnaires were used: the User Experience Questionnaire (UEQ) [46] and an additional questionnaire paired with some qualitative questions specifically developed for the hate countering task of our interface.

5.1. UEQ

UEQ is a questionnaire designed to measure subjective user experience of interactive products. It covers 6 scales that evaluate attractiveness, hard usability aspects (perspicuity, efficiency, and dependability), and soft user experience aspects (stimulation and novelty). Each scale consists of several items with the form of semantic differential that is presented in 2 terms with opposite meanings. For each item, operators were asked to give a score between −3 and 3 on a 7-point Likert scale. The 6 scales used in UEQ are: ATTRACTIVENESS, general impression and aesthetics of the platform; PERSPICUITY, the difficulty to familiarize the platform; EFFICIENCY, the effort required to solve tasks; DEPENDABILITY, the level of control and trust on the platform; STIMULATION, the interest to use the platform; NOVELTY, the novelty about the platform.

5.2. Additional questionnaire and open-ended questions

In addition to the UEQ tool, we used rating scale questions and open-ended questions to examine the qualitative performance of the platform on specific aspects related to our task/scenario, specifically:

- MONITOR: How satisfied are you with the reliability of the platform in monitoring anti-Muslim hatred online?
- ANALYSIS: How satisfied are you with the reliability of the platform in analysing anti-Muslim hatred online?

- SUPPORT: How satisfied are you with the reliability of the platform in its ability to support NGOs in preventing and fighting online Islamophobia?
- EASE OF USE: How satisfied are you with the platform's ease of use?
- LOOK: How satisfied are you with the look of the platform?

Moreover, in addition to above 5 questions, we asked one extra question, in the CN^+ condition, to evaluate the counter-narrative suggestion tool:

- SUGGESTION: How satisfied are you with the counter-narrative suggestion tool?

For each question, operators were asked to give a score between 0 and 10 on a 10-point Likert scale and the reasons for assigning this score.

6. Data analysis

In this section, we first describe the outcome of the usability analysis for the platform according to the two questionnaires mentioned above. Then, we describe the analysis of the activity logs for the counter-narrative generation part.

6.1. UEQ analysis

As stated, the analysis was performed using the UEQ analysis tool.[4] We first performed inconsistency analysis and discarded inconsistent responses following the standard procedure indicated by the questionnaire. After inconsistency analysis, 9 responses were removed (2 responses from the CN^+ condition and 7 responses from CN^-).

In Table 2, we report the statistics of UEQ scales comparing the two conditions. The range of possible values is comprised between −3 (extremely negative) and 3 (extremely positive). According to UEQ tool, values between −0.8 and 0.8 correspond to a neutral evaluation, while values above 0.8 represent a positive and values below −0.8 express a negative evaluation. As can be observed in the table, all scales show a positive evaluation from users in both conditions, but the setting with the counter-narrative tool got a more positive evaluation. Among the scales, the one which received the worst judgements is Dependability, i.e. whether the user feels in control of the interaction. Possible explanations related to this aspect are reported in the open-ended questionnaire. The highest scores in both conditions concern the Attractiveness, i.e. overall impression of the platform, and the Stimulation, i.e. whether it is exciting or stimulating to use the system. In general, all values show that operators have a positive overall impression of the platform. This is rather expected, since previously monitoring analyses were performed manually by operators and many of the operations available with the platform can simply not be done with the standard social media search functions (e.g. hashtag trends going back to several months, or identification of sub-communities spreading islamophobic messages).

With regards to counter-narratives, operators reported higher evaluation for the condition where the CN suggestion tool is present, with Novelty and Stimulation scales showing most improvements. Users may feel stimulated to adopt the CN tool especially when no suitable counter-narrative idea is available to operators or time is a concern. On the other hand, one limitation of the automated suggestion mechanism is that the collected responses may fall within the scope of suggested responses (i.e. be similar to the suggested counter-narratives), since operators were primed by given counter-narrative references. Hence, they are more likely to think and write towards the reference provided.

Table 2
UEQ scales.

Scale	CN−		CN+		CN−/+	
	Mean	SD	Mean	SD	Mean	SD
Attractiveness	1.62	0.77	1.74	0.70	2.11	0.28
Perspicuity	1.21	1.14	1.33	0.75	1.36	0.49
Efficiency	1.31	0.85	1.47	0.73	1.69	0.68
Dependability	0.98	0.71	1.17	0.52	1.11	0.47
Stimulation	1.58	0.84	1.81	0.74	1.78	0.32
Novelty	1.27	1.17	1.75	0.67	1.83	0.42

The UEQ scales can be further grouped into three categories: Attractiveness, Pragmatic quality (including Perspicuity, Efficiency, Dependability) and Hedonic quality (including Stimulation and Originality), where pragmatic quality assesses goal-directed quality aspects and hedonic quality measures quality aspects that are not related to the platform goal of monitoring and countering hate speech. In Table 3, we provide the mean of the three pragmatic and hedonic quality aspects. As can be observed in the Table, all scales receive scores higher than 1.1, and the condition with CN is better evaluated than the one without CN for all the categories.

We further provide a comparison between our platform and the existing benchmark for the UEQ tool to better measure the relative performance of our platform. The benchmark dataset involves 18,483 participants judging 401 products, including web pages, social media, and business software. The benchmark categorizes a product into five groups: (1) excellent, which falls within the first 10% results, (2) good, where 10% of the results in the benchmark are better than the result for the evaluated product, while 75% of the results are worse, (3) above average, where 25% of the results in the benchmark are better than the result for the evaluated product, while 50% of the results are worse, (4) below average, where 50% of the results in the benchmark are better than the result for the evaluated product while 25% of the results are worse, and (5) bad, which is among the last 25% results.

The benchmark chart is displayed in Fig. 5. The analysis shows that all scales for our platform without suggestion tool (CN^- cond in the chart) achieve above average performance except for Perspicuity and Dependability. The latter scale in particular shows that the platform without counter-narrative functionalities supports operators in their activity only to a limited extent. This is however mitigated in the CN^+ condition, achieving an evaluation above the average for all scales, and even an excellent evaluation in the Attractiveness, Stimulation and Novelty scales. This suggests that users consider the counter-narrative tool to be innovative and creative, and that they prefer the platform including the suggestion engine than the one without it.

To further validate this finding, we run a post hoc experiment with 10 operators by replicating the experiment also in a within-subject design (i.e. after few weeks, some operators that were assigned to the CN^- condition were administered also with the CN^+ condition). With this experiment, that we indicate with $CN^{-/+}$, we tested if decisions under separate evaluation diverge from decision under joint evaluation. In this way, we controlled for the possible problem of insensitivity [47,48] that tends to affect specifically between subject design and we weighed the potential difficulty that operators have in accurately perceiving interface differences in a between design (operators in CN^- condition did not have to write counter-narratives during testing), by forcing individuals to experience scope differences in a within design. The results of this second evaluation are also reported in Tables 2 and 3, and also in Fig. 5. Results show, as expected, that for most scales the differences with respect to CN^- condition are more evident: when possible masking effects of novelty and insensitivity are controlled, the advantages of integrating the CN suggestion tool appears even more clearly.
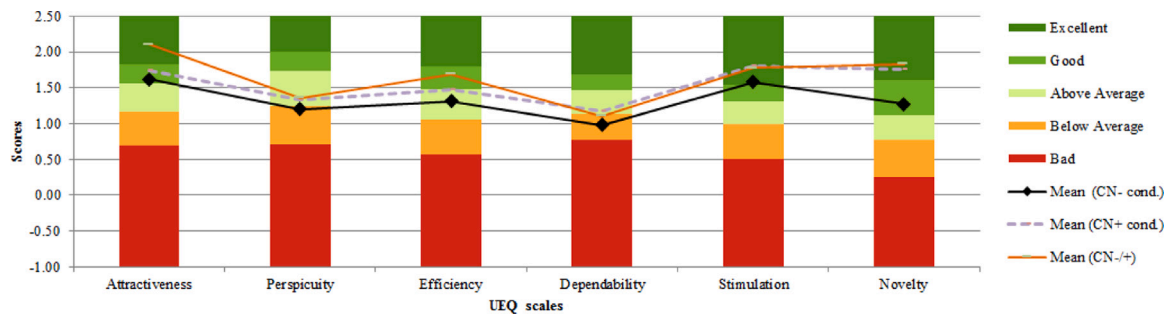
Fig. 5. Statistics on CN^- condition, CN^+ condition and $CN^{-/+}$, compared with the benchmark for UEQ scales.

Table 3
Pragmatic and hedonic quality.

Scale	CN^-	CN^+	$CN^{-/+}$
	Mean	Mean	Mean
Attractiveness	1.62	1.74	2.11
Pragmatic quality	1.17	1.32	1.39
Hedonic quality	1.43	1.78	1.81

6.2. Additional questionnaire and open questions analysis

We performed an additional qualitative evaluation to assess how the platform can be used to monitor online islamophobia and support NGOs to fight anti-Muslim speech, using rating scale questions and an open-ended questionnaire. The results for rating scale questions, discussed in the previous section, are reported in Table 4. Overall, the platform received positive reviews on all dimensions considered in the evaluation, with at least a median value of 7 and a mean value of 6.8. Regarding the suggestion tool condition, users reported higher ratings overall in contrast to a scenario where no suggestion tool is available, except for a slightly lower value concerning ease of use. Also in this case the results from $CN^{-/+}$ configuration show that the differences with respect to CN^- tend to increase.

Explanations of the reasons behind this evaluation and additional comments have been provided by operators in the replies to the open-ended questionnaire, which we summarize below.

6.2.1. Monitor

For all the three languages, users generally concluded that the platform is useful and reliable in monitoring anti-Muslim tweets. As expected, no remarkable differences in the comments can be found between CN^- and CN^+ conditions. The most appreciated features are the possibility to spot spikes in the hashtag frequency and to show users divided into sub-communities. Some operators are not completely satisfied with the choice to focus only on Twitter, and with the hashtag-based monitoring process, arguing that islamophobic messages that do not contain one of the predefined hashtags are not included in our view.

6.2.2. Analysis

Users commented that the platform overall is innovative and useful for analysing anti-Muslim hatred. In particular, diagrams are deemed useful in providing very clear analyses and are easy to interact with. Some operators would like to have more user analysis features, in order to profile specific haters and monitor them. However, this task poses ethical questions when it is done at scale, and this is also forbidden by the European General Data Protection Regulation 2016/679.[5] Like in the *Monitor* evaluation, there is no difference in the comments between CN^- and CN^+ condition because both the analysis and the monitoring tasks are not among the goals of the counter-narrative engine.

6.2.3. Support

Comments related to the ability of the platform to support NGOs in preventing and fighting Islamophobia online are overall positive, although operators tend to specify that this is just an impression and that they would need more time to test the effectiveness of the platform in the long run. Most operators appreciate the novelty of the approach compared to the manual one, since aggregated data provide an added knowledge on Islamophobic trends that the monitoring of single users or hashtags does not cover. In some comments, they suggest introducing the possibility to export such data for preparing reports and supporting long-term analyses. In the CN^+ condition, some comments concern the counter-narrative tool, which is deemed helpful since it is often difficult to form new responses and critical judgements from scratch. Furthermore, with the suggestion tool operators can decrease time and effort spent on writing the responses.

6.2.4. Ease of use

As for the ease of use, participants commented that the experience in using the platform is very intuitive and user-friendly. They appreciated in particular the fact that complex analyses where made understandable in an easy way. Few suggested to add a *help* function where more details are provided about how the displayed analysis was exactly extracted. In the CN^+ condition, the comments were similar to the ones in the CN^- condition, apart from one that mentioned explicitly that the counter-narrative tool was intuitive to use.

6.2.5. Look

While some users reported that the aesthetics of the platform is not as important as its usage and functionality, the overall evaluation was positive. The use of colours was appreciated by the operators, and some suggested that they could be used event more to guide the interpretation of the results. In the CN^+ condition, no specific comment concerned the counter-narrative interface alone.

6.2.6. Suggestion

The question about the counter-narrative suggestion tool was present only in the CN^+ condition. The evaluation by operators was very positive, with some comments stating that this is the most useful feature in the platform and that it could be a game changer in fighting islamophobia online, especially for operators who are not experienced and have to undergo training on counter-narrative strategies. Some comments state that the provided answers should be rather used as a template or a suggestion, and not be posted verbatim. Nevertheless, operators believe that this would be less time consuming than writing counter-narratives from scratch.

6.3. Content analysis and activity logs

Since previous evaluations suggest that using the counter-narrative tool may be less time-consuming than manually writing messages from scratch, we added a final evaluation to measure this aspect. For this reason, we kept the logs of operator activities during the deployment

---

Table 4
Statistics of 6 indicators from additional questionnaire on CN⁻, CN⁺ and CN⁻/⁺ conditions.

	CN⁻				CN⁺				CN⁻/⁺			
	Mean	Median	SD	N	Mean	Median	SD	N	Mean	Median	SD	N
Monitor	7.33	8.00	1.62	55	7.48	8.00	1.56	49	7.78	8.00	0.97	9
Analysis	6.84	7.00	2.03	55	7.29	7.00	1.47	49	7.33	8.00	1.41	9
Support	7.33	8.00	1.88	55	7.46	8.00	1.86	49	8.00	8.00	0.87	9
Ease of use	7.51	8.00	1.93	55	7.35	8.00	1.87	49	7.78	8.00	1.48	9
Look	7.51	8.00	1.88	55	7.60	8.00	1.97	49	8.00	8.00	1.22	9
Suggest	–	–	–	–	6.85	7.00	2.06	49	6.89	7.00	2.15	9

Table 5
Time required for writing CN (Modified CN Based on Suggestions v.s. New CN from Scratch) across 3 countries. Time is expressed in minutes.

	CN⁻ cond.			CN⁺ cond.		
	UK	FR	IT	UK	FR	IT
modified_CN$_\mu$	–	–	–	4.95	2.13	4.30
new_CN$_\mu$	8	8	8	4.21	3.62	4.98

days, and in particular we recorded the time needed in writing new responses per CN suggestion. The assumption is that with the CN suggestion tool the time needed for composing new or modifying responses would be shorter compared to the one without suggestion tool.

Therefore we recorded how many suggested messages were further modified and used by operators, and how many were written by operators and saved through the platform after seeing the CN suggestions. In total, using the platform we collected 999 modified messages and 335 new counter-narratives across three countries. As it is shown in Table 5, for the three data collections there is a significant time decrease by half after the suggestion tool is introduced (the time needed for writing a CN from scratch without the CN suggestion tool was estimated in a separate session for each country). This result indicates that the suggestion tool we provided is able to offer suitable counter-narrative candidates and meets the goal of increasing efficiency of the operators. In fact, with this design, operators can utilize their time to refine responses on top of suggestions and construct more diverse counter-narratives. We further differentiated the time required for modifying counter-narrative suggestions and writing new ones from scratch. For both France and Italy, operators spent less time on modification than on writing new counter-narratives. This implies that our suggestion tool is of good quality and can provide proper counter-narratives after small adaptation. The fact that operators from UK spent slightly less time writing new counter-narratives than adapting the suggested ones may depend on the fact that this user group was the only one including some professional operators, and in general volunteers from this specific NGO receive more training than the others. This may explain why UK operators need to personalize more the suggested counter-narratives, and are faster in writing their own.

7. Limitations

This work presents some limitations that we will possibly try to address in the future. First of all, we chose to build our counter-narrative tool around a retrieval-based suggestion engine, that produces fixed suggestions for similar hate messages. In this regard, users were instructed to paraphrase the suggested counter-narratives in order to increase the variability of the messages posted online. It may be possible to adopt an approach generating more diverse counter-narratives by taking advantage of recent advances in neural text generation [49]. However, with such approaches it is difficult to control the fluency and the grammaticality of the generated messages, and we avoided them in the first experiments because badly written sentences may discourage operators from using the tool, or lead them to correct mistakes instead of focusing on the best counter-narrative strategy. As a next step, we

will compare the quality of the suggested messages using a retrieval-based and a neural generation approach, measuring also the time spent to paraphrase or correct a counter-narrative.

Second, while we believe that the findings on the counter-narrative suggestion tool could apply to hate speech in general, we tested it only on the specific topic of islamophobia, because it was part of the planned activities of the three NGOs involved in the evaluation. On this topic, it was rather straightforward to identify hashtags to be monitored, as well as to collect examples of hate messages and counter-narratives. It is possible that hate messages with a different target, for example misogynistic ones, are less associated with hashtags, or are expressed in the more subtle form of microaggressions. It may, therefore, be more difficult to implement an easy yet effective counter-narrative strategy like the one we introduce in this paper.

8. Conclusions

As digital technologies have been deeply woven into our social lives and hate content spreads rapidly, approaches to deal with this issue at scale are very much needed, especially by organizations that fight hate content online but rely on a limited number of volunteers and operators. To this purpose, we have developed an online platform that monitors islamophobic messages on Twitter and supports NGO operators in quickly responding to hate messages with the help of a counter-narrative tool. We evaluated in particular the contribution of the counter-narrative suggestion tool for English, French, and Italian in addressing islamophobic messages. We performed our evaluation involving more than 100 NGO operators from three countries, using two questionnaires, a set of open-ended questions and analysing the user logs. The evaluation shows that operators have a positive attitude towards the adoption of our platform, finding it both effective and easy-to-use. The counter-narrative tool proved to efficiently support operators in responding to hate messages, also reducing the time needed for the task. Our work gives new insight into the adoption of digital tools to empower NGOs in countering online hate messages.

CRediT authorship contribution statement

Yi-Ling Chung: Data collection, Curation and analysis, Writing - original draft, Writing - review & editing, Software. **Serra Sinem Tekiroğlu:** Data collection, Curation and analysis, Software, Writing - original draft. **Sara Tonelli:** Investigation, Writing - original draft, Writing - review & editing, Supervision, Project administration, Resources. **Marco Guerini:** Data collection, Experimental design, Investigation, Writing - original draft, Writing - review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] M. Mondal, L.A. Silva, F. Benevenuto, A measurement study of hate speech in social media, in: Proceedings of the 28th ACM Conference on Hypertext and Social Media, HT '17, ACM, New York, NY, USA, 2017, pp. 85–94, http://dx.doi.org/10.1145/3078714.3078723.

[2] C. Aguilera-Carnerero, A.H. Azeez, 'Islamonausea, not islamophobia': The many faces of cyber hate speech, J. Arab & Muslim Media Res. 9 (1) (2016) 21–40.

[3] I. Awan, Islamophobia on social media: A qualitative analysis of the facebook's walls of hate, Int. J. Cyber Criminol. 10 (1) (2016).

[4] N. Hall, Hate Crime, Willan, 2013.

[5] I. Gagliardone, D. Gal, T. Alves, G. Martinez, Countering online hate speech, Unesco Publishing, 2015.

[6] S. Benesch, Countering dangerous speech: New ideas for genocide prevention, United States Holocaust Memorial Museum, Washington, DC, 2014.

[7] C. Schieb, M. Preuss, Governing hate speech by means of counterspeech on Facebook, in: 66th Ica Annual Conference, At Fukuoka, Japan, 2016, pp. 1–23.

[8] M. Stella, M. Cristoforetti, M. De Domenico, Influence of augmented humans in online interactions during voting events, PLoS One 14 (5) (2019) e0214210.

[9] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M. Rangel Pardo, P. Rosso, M. Sanguinetti, Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63, http://dx.doi.org/10.18653/v1/S19-2007, URL https://www.aclweb.org/anthology/S19-2007.

[10] J.M. Struß, M. Siegel, J. Ruppenhofer, M. Wiegand, M. Klenner, Overview of germeval task 2, 2019 shared task on the identification of offensive language, in: Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019, Erlangen, Germany, October 9-11, 2019, 2019, URL https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/germeval/GermEvalSharedTask2019Iggsa.pdf.

[11] M. Zampieri, P. Nakov, S. Rosenthal, P. Atanasova, G. Karadzhov, H. Mubarak, L. Derczynski, Z. Pitenis, Ç. Çöltekin, Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020), in: A. Herbelot, X. Zhu, A. Palmer, N. Schneider, J. May, E. Shutova (Eds.), Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (Online), December 12-13, 2020, International Committee for Computational Linguistics, 2020, pp. 1425–1447, URL https://www.aclweb.org/anthology/2020.semeval-1.188/.

[12] M. Sanguinetti, G. Comandini, E.D. Nuovo, S. Frenda, M. Stranisci, C. Bosco, T. Caselli, V. Patti, I. Russo, Haspeede 2 @ EVALITA2020: overview of the EVALITA 2020 hate speech detection task, in: V. Basile, D. Croce, M.D. Maro, L.C. Passaro (Eds.), Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Online event, December 17th, 2020, Vol. 2765 of CEUR Workshop Proceedings, CEUR-WS.org, 2020, URL http://ceur-ws.org/Vol-2765/paper162.pdf.

[13] J. Qian, A. Bethke, Y. Liu, E. Belding, W.Y. Wang, A benchmark dataset for learning to intervene in online hate speech, 2019, ArXiv Preprint ArXiv:1909.04251.

[14] S.S. Tekiroglu, Y.-L. Chung, M. Guerini, Generating counter narratives against online hate speech: Data and strategies, 2020, ArXiv Preprint ArXiv:2004.04216.

[15] M.J. Riedl, G.M. Masullo, K.N. Whipple, The downsides of digital labor: Exploring the toll incivility takes on online comment moderators, Comput. Hum. Behav. 107 (2020) 106262.

[16] B. Dang, M.J. Riedl, M. Lease, But who protects the moderators? the case of crowdsourced image moderation, 2018, ArXiv Preprint ArXiv:1804.10999.

[17] S. Karunakaran, R. Ramakrishan, Testing stylistic interventions to reduce emotional impact of content moderation workers, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7, 2019, pp. 50–58.

[18] N. Sambuli, K. Awori, Monitoring online dangerous speech in Kenya. Insights from the umati project, in: V.L. R., I. Wall (Eds.), Communications Technology and Humanitarian Delivery: Challenges and Opportunities for Security Risk Management, European Interagency Security Forum, 2014.

[19] P. Burnap, O. Rana, M. Williams, W. Housley, A. Edwards, J. Morgan, L. Sloan, J. Conejero, COSMOS: Towards an integrated and scalable service for analysing social media on demand, Int. J. Parallel Emergent Distrib. Syst. 30 (2) (2015) 80–100.

[20] S. Benesch, C. Buerger, T. Glavinic, S. Manion, Dangerous speech: A practical guide, dangerous speech project, 2018, URL https://dangerousspeech.org/guide/.

[21] A.T. Capozzi, M. Lai, V. Basile, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, C. Musto, M. Polignano, et al., Computational linguistics against hate: Hate speech detection and visualization on social media in the" Contro L'odio" project, in: 6th Italian Conference on Computational Linguistics, CLiC-It 2019, Vol. 2481, CEUR-WS, 2019, pp. 1–6.

[22] S. Menini, G. Moretti, M. Corazza, E. Cabrio, S. Tonelli, S. Villata, A system to monitor cyberbullying based on message classification and social network analysis, in: Proceedings of the Third Workshop on Abusive Language Online, 2019, pp. 105–110.

[23] D. Paschalides, D. Stephanidis, A. Andreou, K. Orphanou, G. Pallis, M.D. Dikaiakos, E. Markatos, MANDOLA: A big-data processing and visualization platform for monitoring and detecting online hate speech, ACM Trans. Internet Technol. (TOIT) 20 (2) (2020) 1–21.

[24] A.T.E. Capozzi, V. Patti, G. Ruffo, C. Bosco, A data viz platform as a support to study, analyze and understand the hate speech phenomenon, in: Proceedings of the 2nd International Conference on Web Studies, WS.2 2018, ACM, New York, NY, USA, 2018, pp. 28–35, http://dx.doi.org/10.1145/3240431.3240437, URL http://doi.acm.org/10.1145/3240431.3240437.

[25] A. Ben-David, A. Matamoros-Fernandez, Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in Spain, Int. J. Commun. 10 (2016) 1167–1193, URL https://eprints.qut.edu.au/101369/.

[26] P. Burnap, M.L. Williams, Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making, Policy & Internet 7 (2) (2015) 223–242, http://dx.doi.org/10.1002/poi3.85.

[27] D. Paschalides, D. Stefanidis, M. Hernando, G. Pallis, M. Dikaiakos, Deliverable 3.1: Mandola monitoring dashboard, tech. rep., MANDOLA EU REC Project, 2016.

[28] J. Bartlett, A. Krasodomski-Jones, Counter-speech examining content that challenges extremism online, demos, 2015.

[29] R.D. Richards, C. Calvert, Counterspeech 2000: A new look at the old remedy for bad speech, BYU L. Rev. (2000) 553.

[30] L. Wright, D. Ruths, K.P. Dillon, H.M. Saleem, S. Benesch, Vectors for counter-speech on twitter, in: Proceedings of the First Workshop on Abusive Language Online, 2017, pp. 57–62.

[31] J. Ernst, J.B. Schmitt, D. Rieger, A.K. Beier, P. Vorderer, G. Bente, H.-J. Roth, Hate beneath the counter speech? A qualitative content analysis of user comments on youtube related to counter speech videos, J. Deradicalization (10) (2017) 1–49.

[32] B. Mathew, H. Tharad, S. Rajgaria, P. Singhania, S.K. Maity, P. Goyal, A. Mukherje, Thou shalt not hate: Countering online hate speech, 2018, ArXiv Preprint ArXiv:1808.04409.

[33] T. Silverman, C.J. Stewart, J. Birdwell, Z. Amanullah, The impact of counter-narratives, Inst. Strateg. Dialogue (2016) 1–54.

[34] K. Munger, Tweetment effects on the tweeted: Experimentally reducing racist harassment, Political Behav. 39 (3) (2017) 629–649.

[35] Y.-L. Chung, E. Kuzmenko, S.S. Tekiroglu, M. Guerini, Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, 2019, pp. 2819–2829.

[36] B. Mathew, N. Kumar, P. Goyal, A. Mukherjee, et al., Analyzing the hate and counter speech accounts on Twitter, 2018, ArXiv Preprint ArXiv:1812.02712.

[37] F. Moretti, Distant Reading, Verso, London, 2013.

[38] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11–21.

[39] L. Shang, Z. Lu, H. Li, Neural responding machine for short-text conversation, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 1577–1586.

[40] O. Vinyals, Q.V. Le, A neural conversational model, CoRR abs/1506.05869, 2015, ArXiv Preprint arXiv:1506.05869.

[41] D.-T. Le, C.T. Nguyen, K.A. Nguyen, Dave the debater: a retrieval-based and generative argumentative dialogue agent, in: Proceedings of the 5th Workshop on Argument Mining, 2018, pp. 121–130.

[42] X. Hua, Z. Hu, L. Wang, Argument generation with retrieval, planning, and realization, 2019, ArXiv Preprint ArXiv:1906.03717.

[43] A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, B. Dolan, A neural network approach to context-sensitive generation of conversational responses, in: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2015, pp. 196–205.

[44] I.V. Serban, A. Sordoni, Y. Bengio, A. Courville, J. Pineau, Building end-to-end dialogue systems using generative hierarchical neural network models, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.

[45] J. Li, M. Galley, C. Brockett, J. Gao, B. Dolan, A diversity-promoting objective function for neural conversation models, in: Proceedings of NAACL-HLT, 2016, pp. 110–119.

[46] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: Symposium of the Austrian HCI and Usability Engineering Group, Springer, 2008, pp. 63–76.

[47] D. Kahneman, J.L. Knetsch, Valuing public goods: the purchase of moral satisfaction, J. Environ. Econ. Manag. 22 (1) (1992) 57–70.

[48] W.H. Desvousges, F. Reed Johnson, R.W. Dunford, K. Nicole Wilson, K.J. Boyle, Measuring natural resource damages with contingent valuation, in: Contingent Valuation: A Critical Assessment, Emerald Group Publishing Limited, 1993, pp. 91–164.

[49] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog 1 (8) (2019).