CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech

Yi-Ling Chung[1,2], **Elizaveta Kuzmenko**[2], **Serra Sinem Tekiroğlu**[1], **and Marco Guerini**[1]

[1]Fondazione Bruno Kessler, Via Sommarive 18, Povo, Trento, Italy
ychung@fbk.eu, tekiroglu@fbk.eu, guerini@fbk.eu
[2]University of Trento, Italy
elizaveta.kuzmenko@studenti.unitn.it

Abstract

Although there is an unprecedented effort to provide adequate responses in terms of laws and policies to hate content on social media platforms, dealing with hatred online is still a tough problem. Tackling hate speech in the standard way of content deletion or user suspension may be charged with censorship and overblocking. One alternate strategy, that has received little attention so far by the research community, is to actually oppose hate content with counter-narratives (i.e. informed textual responses). In this paper, we describe the creation of the first large-scale, multilingual, expert-based dataset of hate speech/counter-narrative pairs. This dataset has been built with the effort of more than 100 operators from three different NGOs that applied their training and expertise to the task. Together with the collected data we also provide additional annotations about expert demographics, hate and response type, and data augmentation through translation and paraphrasing. Finally, we provide initial experiments to assess the quality of our data.

1 Introduction

Together with the rapid growth of social media platforms, the amount of user-generated content is steadily increasing. At the same time, abusive and offensive language can spread quickly and is difficult to monitor. Defining hate speech is challenging for the broadness and the nuances in cultures and languages. For instance, according to UNESCO hate speech refers to "expressions that advocate incitement to harm based upon the targets being identified with a certain social or demographic group" (Gagliardone et al., 2015).

Victims of hate speech are usually targeted because of various aspects such as gender, race, religion, sexual orientation, physical appearance. For example, Sentence 1 shows explicit hostility towards a specific group with no reasons explained[1].

(1) I hate Muslims. They should not exist.

Online hate speech can deepen prejudice and stereotypes (Citron and Norton, 2011) and bystanders may receive false messages and consider them correct. Although Social Media Platforms (SMP) and governmental organizations have elicited unprecedented attention to take adequate actions against hate speech by implementing laws and policies (Gagliardone et al., 2015), they do not seem to achieve the desired effect, since hate content is continuously evolving and adapting, making its identification a tough problem (Davidson et al., 2017).

The standard approach used on SMPs to prevent hate spreading is the suspension of user accounts or deletion of hate comments, while trying to weigh the right to freedom of speech. Another strategy, which has received little attention so far, is to use counter-narratives. A counter-narrative (sometimes called counter-comment or counter-speech) is a response that provides non-negative feedback through fact-bound arguments and is considered as the most effective approach to withstand hate speech (Benesch, 2014; Schieb and Preuss, 2016). In fact, it preserves the right to freedom of speech, counters stereotypes and misleading information with credible evidence. It can also alter the viewpoints of haters and bystanders, by encouraging the exchange of opinions and mutual understanding, and can help de-escalating the conversation. A counter-narrative such as the one in Sentence 2 is a non-negative, appropriate response to Sentence 1, while the one in 3 is not, since it escalates the conversation.

[1]It is crucial to note that this paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 2819–2829
Florence, Italy, July 28 - August 2, 2019.

(2) Muslims are human too. People can choose their own religion.

(3) You are truly one stupid backwards thinking idiot to believe negativity about Islam.

In this respect, some NGOs are tackling hatred online by training operators to monitor SMPs and to produce appropriate counter-narratives when necessary. Still, manual intervention against hate speech is a toil of Sisyphus, and automatizing the countering procedure would increase the efficacy and effectiveness of hate countering (Munger, 2017).

As a first step in the above direction, we have nichesourced the collection of a dataset of counter-narratives to 3 different NGOs. Nichesourcing is a specific form of outsourcing that harnesses the computational efforts from niche groups of experts rather than the 'faceless crowd' (De Boer et al., 2012). Nichesourcing combines the strengths of the crowd with those of professionals (De Boer et al., 2012; Oosterman et al., 2014). In our case we organized several data collection sessions with NGO operators, who are trained experts, specialized in writing counter-narratives that are meant to fight hatred and de-escalate the conversation. In this way we build the first large-scale, multilingual, publicly available, expert-based dataset of hate speech/counter-narrative pairs for English, French and Italian, focusing on the hate phenomenon of *Islamophobia*. The construction of this dataset involved more than 100 operators and more than 500 person-hours of data collection. After the data collection phase, we hired three non-expert annotators, that performed additional tasks that did not require specific domain expertise (200 person-hours of work): paraphrase original hate content to augment the number of pairs per language, annotate hate content sub-topic and counter-narrative type, translate content from Italian and French to English to have parallel data across languages. This additional annotation grants that the dataset can be used for several NLP tasks related to hate speech.

The remainder of the paper is structured as follows. First, we briefly discuss related work on hate speech in Section 2. Then, in Section 3, we introduce our CONAN dataset and some descriptive statistics, followed by a quantitative and qualitative analysis on our dataset in Section 4. We conclude with our future works in Section 5.

2 Related Work

With regard to hatred online, we will focus on three research aspects about the phenomenon, i.e. (i) publicly available datasets, (ii) methodologies for detecting hate speech, (iii) seminal works that focus on countering hate speech.

Hate datasets. Several hate speech datasets are publicly available, usually including a binary annotation, i.e. whether the content is hateful or not (Reynolds et al., 2011; Rafiq et al., 2015; Hosseinmardi et al., 2015; de Gibert et al., 2018; ElSherief et al., 2018). Also, several shared tasks have released their datasets for hate speech detection in different languages. For instance, there is the German abusive language identification on SMPs at Germeval (Bai et al., 2018), or the hate speech and misogyny identification for Italian at EVALITA (Del Vigna et al., 2017; Fersini et al., 2018) and for Spanish at IberEval (Ahluwalia et al., 2018; Shushkevich and Cardiff, 2018). Bilingual hate speech datasets are also available for Spanish and English (Pamungkas et al., 2018).

Waseem and Hovy (2016) released 16k annotated tweets containing 3 offense types: sexist, racist and neither. Ross et al. (2017) first released a German hate speech dataset of 541 tweets targeting refugee crisis and then offered insights for the improvement on hate speech detection by providing multiple labels for each hate speech.

It should be noted that, due to the copyright limitations, usually hate speech datasets are distributed as a list of tweet IDs making them ephemeral and prone to data loss (Klubička and Fernández, 2018). For this reason, Sprugnoli et al. (2018) created a multi-turn annotated WhatsApp dataset for Italian on Cyberbullying, using simulation session with teenagers to overcome the data collection/loss problem.

Hate detection. Several works have investigated online English hate speech detection and the types of hate speech. Owing to the availability of current datasets, researchers often use supervised-approaches to tackle hate speech detection on SMPs including blogs (Warner and Hirschberg, 2012; Djuric et al., 2015; Gitari et al., 2015), Twitter (Xiang et al., 2012; Silva et al., 2016; Mathew et al., 2018a), Facebook (Del Vigna et al., 2017), and Instagram (Zhong et al., 2016). The predominant approaches are to build a classifier trained on various features derived from lexical resources

(Gitari et al., 2015; Burnap and Williams, 2015, 2016), n-grams (Sood et al., 2012; Nobata et al., 2016) and knowledge base (Dinakar et al., 2012), or to utilize deep neural networks (Mehdad and Tetreault, 2016; Badjatiya et al., 2017). In addition, other approaches have been proposed to detect subcategories of hate speech such as anti-black (Kwok and Wang, 2013) and racist (Badjatiya et al., 2017). Silva et al. (2016) studied the prevalent hate categories and targets on Twitter and Whisper, but limited hate speech only to the form of *I <intensity> <user intent> <any word>*. A comprehensive overview of recent approaches on hate speech detection using NLP can be found in (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018).

Hate countering. Lastly, we should mention that a very limited number of studies have been conducted on counter-narratives (Benesch, 2014; Schieb and Preuss, 2016; Ernst et al., 2017; Mathew et al., 2018b). Mathew et al. (2018b) collected Youtube comments that contain counter-narratives to YouTube videos of hatred. Schieb and Preuss (2016) studied the effectiveness of counter-narrative on Facebook via a simulation model. The study of Wright et al. (2017) shows that some arguments among strangers induce favorable changes in discourse and attitudes. To our knowledge, there exists only one very recent seminal work (Mathew et al., 2018a), focusing on the idea of collecting hate message/counter-narrative pairs from Twitter. They used a simple pattern in the form *(I <hate> <category>)* to first extract hate tweets and then manually annotate counter-narratives found in the responses. Still, there are several shortcomings of their approach: (i) this dataset already lost more that 60% of the pairs in a small time interval (content deletion) since only tweet IDs are distributed, (ii) it is only in English language, (iii) the dataset was collected from a specific template which limits the coverage of hate speech, and (iv) many of these answers come from ordinary web users and contain -for example- offensive text, that do not meet the de-escalation intent of NGOs and the standards/quality of their operators' responses.

Considering the aforementioned works, we can reasonably state that no suitable corpora of counter-narratives is available for our purposes, especially because the natural 'countering' data that can be found on SMP – such as example 3 – often does not meet the required standards. For this reason we decided to build CONAN, a dataset of COunter NArratives through Nichesourcing.

3 CONAN Dataset

In this section, we describe the characteristics that we intend our dataset to posses, the nichesourcing methodology we employed to collect the data and the further expansion of the dataset together with the annotation procedures. Moreover, we give some descriptive statistics and analysis for the collected data. CONAN can be downloaded at the following link `https://github.com/marcoguerini/CONAN`.

3.1 Fundamentals of the Dataset

Considering the shortcomings of the existing datasets and our aim to provide a reliable resource to the research community, we want CONAN to comply with the following characteristics:

Copy-free data. We want to provide a dataset that is not ephemeral, by releasing only copy-free textual data that can be directly exploited by researches without data loss across time, as originally pointed out in (Klubička and Fernández, 2018).

Multilingual data. Our dataset is produced as a multilingual resource to allow for cross lingual studies and approaches. In particular, it contains hate speech/counter-narrative pairs for English, French, and Italian.

Expert-based data. The hate speech/counter-narrative pairs have been collected through nichesourcing to three different NGOs from United Kingdom, France and Italy. Therefore, both the responses and the hate speech itself are expert-based and composed by operators, specifically trained to oppose online hate speech.

Protecting operator's identity. We aim to create a secure dataset that will not disclose the identity of operators in order to protect them against being tracked and attacked online by hate spreaders. This might be the case if we were to collect their real SMP activities, following a procedure similar to the one in Mathew et al. (2018a). Therefore our data collection was based on simulated SMP activity.

Demographic-based metadata. Demographic-based NLP can be used for several tasks, such as characterizing personal linguistic styles (Johannsen et al., 2015; Hovy and Spruit, 2016; van der Goot et al., 2018; DellOrletta and Nissim, 2018), improving text classification (Mandel et al., 2012; Volkova et al., 2013; Hovy, 2015), or personalizing conversational agents (Qiu and Benbasat, 2010; Mazaré et al., 2018a). In this work, we collect demographic information of participants; i.e. gender, age, and education level, to provide data for counter-narrative personalization.

3.2 Dataset Collection

We have followed the same data collection procedure for each language to grant the same conditions and comparability of the results. The data collection has been conducted along the following steps:

1. Hate speech collection. For each language we asked two native speaker experts (NGO trainers) to write around 50 prototypical islamophobic short hate texts. This step was used to ensure that: (i) the sample uniformly covers the typical 'arguments' against Islam as much as possible, (ii) we can distribute to the NLP community the original hate speech as well as its counter-narrative.

2. Preparation of data collection forms. We prepared three online forms (one per language) with the same instructions for the operators translated in the corresponding language. For each language, we prepared 2 types of forms: in the first users can respond to hate text prepared by NGO trainers, in the second users can write their own hate text and counter-narratives at the same time. In each form operators were first asked to anonymously provide their demographic profile including age, gender, and education level; secondly to compose up to 5 counter-narratives for each hate text.

3. Counter-narrative instructions. The operators were already trained to follow the guidelines of the NGOs for creating proper counter-narratives. Such guidelines are highly consistent across languages and across NGOs, and are similar to those in 'Get the Trolls Out' project[2]. These guidelines emphasize using fact-bounded information and non-offensive language in order to avoid escalating the discussion as outlined in Table 1. Furthermore, for our specific data collection task, operators were asked to follow their intuitions without over-thinking and to compose reasonable responses. The motivation for this instruction was to collect as much and as diverse data as possible, since for current AI technologies (such as deep learning approaches) quantity and quality are of paramount importance and few perfect examples do not provide enough generalization evidence. Other than this instruction and the fact of using a form – instead of responding on a SMP – operators carried out their normal counter messaging activities.

4. Data collection sessions. For each language, we performed three data collection sessions on different days. Each session lasted roughly three hours[3] and had a variable number of operators – usually around 20 (depending on their availability). Operators are different from NGO trainers and might change across sessions. Operators were gathered in the same room (NGO premises) with a computer, and received a brief introduction from the NGO trainer. This introduction was about our specific counter-narrative collection task, as described above. A sample of the collected data for the three languages is given in Table 2.

3.3 Dataset Augmentation and Annotation

After the data collection phase, we hired three non-expert annotators, that performed additional work that did not require specific domain expertise. Their work amounted to roughly 200 hours. In particular they were asked to (i) paraphrase original hate content to augment the number of pairs per language, (ii) annotate hate speech subtopics and counter-narrative types (iii) translate content from French and Italian to English to have parallel data across languages. To guarantee data quality, after the annotation and the augmentation phase, a validation procedure has been conducted by NGO trainers on the newly generated data for their specific language.

Paraphrasing for augmenting data pairs. Recent deep learning approaches are data hungry, and data augmentation is a way to mitigate the problem. For instance, to improve text classification performance for sexism, new tweets are generated by replacing words in original tweets with synonyms from ConceptNet (Sharifirad et al.,

[2]http://stoppinghate.getthetrollsout.org/

[3]Duration was decided during a pilot session: after three hours operators were no longer effective in producing counter-narratives, since it is a cognitively fatiguing task.

Don't be abusive Before submitting a response, make sure the response does not spread any hate, bigotry, prejudice or illegal content. We want to maintain the conversations peaceful and not to degenerate into a conflict. We are talking about people not categories.
Think about the objectives Before writing a response, think about the effect it may create and the one you want to obtain. Paying attention to the objectives will help use proper words.
Call for influential users Enlisting influential supporters (civic leaders, politicians, subject experts) will help bring attention and deepen the effect to counter-narrative.
Use credible evidence The information in hate speech may be confusing and misleading. Ask for clarification when necessary. Counter it with credible evidence and use reliable sources.
Think about the tone We can demonstrate understanding and support to those who might be attacked. Be careful of using sarcasm, humour, parody and satire. We can use them, if we are able to master it as they run the danger of being antagonistic.

Table 1: Counter-narrative guidelines adopted by the three NGOs.

Hate Speech	Counter-Narrative
Every Muslim is a potential terrorist.	Every Muslim is also a potential peacemaker, doctor, philanthropist... What's your point?
I musulmani violentano le nostre donne e vanno castrati.	Buonasera, se ho capito bene, lei mi sta dicendo che tutti gli uomini adulti di fede islamica presenti in Italia, hanno violentato, violentano o violenteranno le donne italiane? Può portare dei dati a supporto della sua tesi?
Muslims rape our women and must be neutered.	*Good evening, if I understood correctly, you're telling me that all adult men of Islamic faith present in Italy, raped, rape or are going to rape Italian women? Can you bring in data in support of your statement?*
Le voile est contraire à la laïcité.	Bien au contraire la laïcité permet à tout citoyen de vivre librement sa confession.
The veil is contrary to secularism.	*On the contrary, secularism allows every citizen to freely profess his faith.*

Table 2: Example pairs for the three languages, along with English translations.

2018). Other examples of data augmentation strategies are back translation (Sennrich et al., 2016) and gold standard repetition (Chatterjee et al., 2017) that have been used in sequence-to-sequence Machine Translation. In all these tasks, adding the synthetic pairs to the original data always results in significant improvements in the performance.

In line with the idea of artificially augmenting pairs, and since in our dataset we have many responses for few hate speeches, we produced two manual paraphrases of each hate speech and paired them with the counter-narratives of the original one. Therefore we increased the number of our pairs by three times in each language.

Counter-narrative type annotation. In this task, we asked the annotators to label each counter-narrative with types. Based on the counter-narrative classes proposed by (Benesch et al., 2016; Mathew et al., 2018b), we defined the following set of types: PRESENTATION OF FACTS, POINTING OUT HYPOCRISY OR CONTRADICTION, WARNING OF CONSEQUENCES, AFFILIATION, POSITIVE TONE, NEGATIVE TONE, HUMOR, COUNTER-QUESTIONS, OTHER. With respect to the original guidelines, we added a new type of counter-narrative called COUNTER-QUESTIONS to cover expressions/replies using a question that can be thought-provoking or asking for more evidence from the hate speaker. In fact, a preliminary analysis showed that this category is quite frequent among operator responses. Finally, each counter-narrative can be labeled with more than one type, thus making the annotation more fine-grained.

Two annotators per language annotated all the counter-narratives independently. A reconciliation

phase was then performed for the disagreement cases.

Hate speech sub-topic annotation. We labeled sub-topics of hate content to have an annotation that can be used both for fine grained hate speech classification, and for exploring the correlation between hate sub-topics and counter-narrative types. The following sub-topics are determined for the annotation based on the guidelines used by NGOs to identify hate messages (mostly consistent across languages): CULTURE, criticizing Islamic culture or particular aspects such as religious events or clothes; ECONOMICS, hate statements about Muslims taking European workplaces or not contributing economically to the society; CRIMES, hate statements about Muslims committing actions against the law; RAPISM, a very frequent topic in hate speech, for this reason it has been isolated from the previous category; TERRORISM, accusing Muslims of being terrorists, killers, preparing attacks; WOMEN OPPRESSION, criticizing Muslims for their behavior against women; HISTORY, stating that we should hate Muslims because of historical events; OTHER/GENERIC, everything that does not fall into the above categories.

As before, two annotators per language annotated all the material. Also in this annotation task, a reconciliation phase was performed for the disagreement cases.

Parallel corpus of language pairs. To allow studying cross-language approaches to counter-narratives and more generally to increase language portability, we also translated the French and the Italian pairs (i.e. hate speech and counter-narratives) to English. Similar motivations can be found in using zero-short learning to translate between unseen language pairs during training (Johnson et al., 2017). With parallel corpora we can exploit cross-lingual word embeddings to enable knowledge transfer between languages (Schuster et al., 2018).

3.4 Dataset Statistics

In total we had more than 500 hours of data collection with NGOs, where we collected 4078 hate speech/counter-narrative pairs; specifically, 1288 pairs for English, 1719 pairs for French, and 1071 pairs for Italian. At least 111 operators participated in the 9 data collection sessions and each counter-narrative needed about 8 minutes on average to be composed. The paraphrasing of hate messages and the translation of French and Italian pairs to English brought the total number of pairs to more than 15 thousand. Regarding the token length of counter-narratives, we observe that there is a consistency across the three languages with 14 tokens on average for French, and 21 for Italian and English. Considering counter-narrative length in terms of characters, only a small portion (2% for English, 1% for French, and 5% for Italian) contains more than 280 characters, which is the character limit per message in Twitter, one of the key SMPs for hate speech research. Further details on the dataset can be found in Table 3.

	English	French	Italian
original pairs	1288	1719	1071
augmen. pairs	2576	3438	2142
transl. pairs	2790	-	-
total pairs	6654	5157	3213
HS	136	50	62
CN_per_HS$_{\mu}$	9.47	34.38	17.27
CN_per_HS$_{sd}$	7.56	53.86	26.48
HS vocabulary	947	193	343
HS+aug. vocab.	1631	333	790
CN vocabulary	3556	4018	3728
HS words	2950	434	751
HS+aug. words	9770	1172	2633
CN words	27677	23730	23129
HS_words$_{\mu}$	21.69	8.68	12.11
HS_words$_{sd}$	10.29	4.02	6.69
HS+aug._words$_{\mu}$	18.72	5.31	14.16
HS+aug._words$_{sd}$	10.05	4.73	7.65
CN_words$_{\mu}$	21.49	13.80	21.60
CN_words$_{sd}$	11.06	11.44	12.42

Table 3: Main statistics of the dataset. HS stands for Hate Speech, CN stands for Counter-Narrative.

Regarding demographics, the majority of responses were written by operators that held a bachelor's or a higher degree (95% for English, 65% for French, and 69% for Italian). As it is shown in Table 4, there is a good balance in responses with regard to declared gender, with a slight predominance of counter-narratives written by female operators in English and Italian (53 and 55 per cent respectively) while a slight predominance of counter-narratives written by male operators is present in French (61%). Finally, the predominant age bin is 21-30 for English and Italian,

while for French is in the range 31-40.

	EN	FR	IT
< high school	-	5%	14%
high school	-	14%	10%
< university	5%	16%	6%
bachelor	51%	17%	34%
master	44%	35%	30%
PhD	-	13%	5%
female	53%	39%	55%
male	47%	61%	45%
<= 20	-	-	15%
21 - 30	74%	15%	42%
31 - 40	-	51%	7%
41 - 50	18%	20%	15%
51 - 60	-	11%	16%
> 60	8%	3%	5%

Table 4: Demographic profile of the operators.

Type	EN	FR	IT
affiliation	1	4	1
consequences	0	1	0
denouncing	19	18	13
facts	38	37	47
humor	8	6	5
hypocrisy	16	14	10
negative	0	0	0
other	0	4	1
positive	6	5	7
question	12	11	16

Table 5: Counter-narrative type distribution over the three languages (% over the total number of labels).

Considering the annotation tasks, we give the distribution of counter-narrative types per language in Table 5. As can be seen in the table, there is a consistency across the languages such that FACTS, QUESTION, DENOUNCING, and HYPOCRISY are the most frequent counter-narrative types. Before the reconciliation phase, the agreement between the annotators was moderate: Cohen's Kappa[4] 0.55 over the three languages. This can be partially explained by the complexity of the messages, that often fall under more than one category (two labels were assigned in more than 50% of the cases). On the other hand, for hate speech sub-topic annotation, the agreement between the annotators was very high even before the reconciliation phase (Cohen's Kappa 0.92 over the three languages). A possible reason is that such messages represent short and prototypical hate arguments, as explicitly requested to the NGO trainers. In fact, the vast majority has only one label. In Table 6, we give a distribution of hate speech sub-topics per language. As can be observed in the table, the labels are distributed quite evenly among sub-topics and across languages - in particular, CULTURE, ISLAMIZATION, GENERIC, and TERRORISM are the most frequent sub-topics.

[4]Computed using Mezzich's methodology to account for possible multiple labels that can be assigned to a text by each annotator (Mezzich et al., 1981).

Type	EN	FR	IT
crimes	10	0	7
culture	30	26	11
economics	4	1	8
generic	20	27	8
islamization	11	7	36
rapism	15	0	7
terrorism	6	14	19
women	4	25	4

Table 6: hate speech sub-topic type distribution over the three languages (% over the total number of labels).

4 Evaluation

In order to assess the quality of our dataset, we ran a series of preliminary experiments that involved three annotators to judge hate speech/counter-narrative pairs along a yes/no dimension.

Augmentation reliability. The first experiment was meant to assess how natural a pair is when coupling a counter-narrative with the manual paraphrase of the original hate speech it refers to. We administered 120 pairs to the subjects to be evaluated: 20 were kept as they are so to have an upper bound representing ORIGINAL pairs. In 50 pairs we replaced the hate speech with a PARAPHRASE, and in the 50 remaining pairs, we randomly matched a hate speech with a counter-narrative from another hate speech (UNRELATED baseline). Results show that 85% of the times in the ORIGINAL condition hate speech and counter-narrative were considered as clearly tied, followed by the 74% of times by PARAPHRASE condition, and only 4% of the UNRELATED baseline, this difference is statistically significant with $p < .001$ (w.r.t. χ^2 test). This indicates that the quality of augmented pairs is almost as good as the one of original pairs.

Augmentation for counter-narrative selection. Once we assessed the quality of augmented pairs, we focused on the possible contribution of the paraphrases also in standard information retrieval approaches that have been used as baselines in dialogue systems (Lowe et al., 2015; Mazaré et al., 2018b). We first collected a small sample of natural/real hate speech from Twitter using relevant keywords (such as "stop Islam") and manually selected those that were effectively hate speeches. We then compared 2 tf-idf response retrieval models by calculating the tf-idf matrix using the following document variants: (i) hate speech and counter-narrative response, (ii) hate speech, its 2 paraphrases, and counter-narrative response. The final response for a given sample tweet is calculated by finding the highest score among the cosine similarities between the tf-idf vectors of the sample and all the documents in a model.

For each of the 100 natural hate tweets, we then provided 2 answers (one per approach) selected from our English database. Annotators were then asked to evaluate the responses with respect to their relevancy/relatedness to the given tweet. Results show that introducing the augmented data as a part of the tf-idf model provides 9% absolute increase in the percentage of the agreed 'very relevant' responses, i.e. from 18% to 27% - this difference is statistically significant with $p < .01$ (w.r.t. χ^2 test). This result is especially encouraging since it shows that the augmented data can be helpful in improving even a basic automatic counter-narrative selection model.

Impact of Demographics. The final experiment was designed to assess whether demographic information can have a beneficial effect on the task of counter-narrative selection/production. In this experiment, we selected a subsample of 230 pairs from our dataset written by 4 male and 4 female operators that were controlled for age (i.e. same age range). We then presented our subjects with each pair in isolation and asked them to state whether they would definitely use that particular counter-narrative for that hate speech or not. Note that, in this case, we did not ask whether the counter-narrative was relevant, but if they would use that given counter-narrative text to answer the paired hate speech. The results show that in the SAMEGENDER configuration (gender declared by the operator who wrote the message and gender declared by the annotator are the same), the appreciation was expressed 47% of the times, while it decreases to 32% in the DIFFERENTGENDER configuration (gender declared by the operator who wrote the message and gender declared by the annotator are different). This difference is statistically significant with $p < .001$ (w.r.t. χ^2 test), indicating that even if operators were following the same guidelines and were instructed on the same possible arguments to build counter-narratives, there is still an effect of their gender on the produced text, and this effect contributes to the counter-narrative preference in a SAMEGENDER configuration.

5 Conclusion

As online hate content rises massively, responding to it with counter-narratives as a combating strategy draws the attention of international organizations. Although a fast and effective responding mechanism can benefit from an automatic generation system, the lack of large datasets of appropriate counter-narratives hinders tackling the problem through supervised approaches such as deep learning. In this paper, we described CONAN: the first large-scale, multilingual, and expert-based hate speech/counter-narrative dataset for English, French, and Italian. The dataset consists of 4078 pairs over the 3 languages. Together with the collected data we also provided several types of metadata: expert demographics, hate speech sub-topic and counter-narrative type. Finally, we expanded the dataset through translation and paraphrasing.

As future work, we intend to continue collecting more data for Islam and to include other hate targets such as migrants or LGBT+, in order to put the dataset at the service of other organizations and further research. Moreover, as a future direction, we want to utilize CONAN dataset to develop a counter-narrative generation tool that can support NGOs in fighting hate speech online, considering counter-narrative type as an input feature.

Acknowledgments

This work was partly supported by the HATEMETER project within the EU Rights, Equality and Citizenship Programme 2014-2020. We are grateful to the following NGOs and all annotators for their help: Stop Hate UK, Collectif Contre l'Islamophobie en France, Amnesty International (Italian Section - Task force hate speech).

References

Resham Ahluwalia, Evgeniia Shcherbinina, Edward Callow, Anderson Nascimento, and Martine De Cock. 2018. Detecting misogynous tweets. *Proc. of IberEval*, 2150:242–248.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee.

Xiaoyu Bai, Flavio Merenda, Claudia Zaghi, Tommaso Caselli, and Malvina Nissim. 2018. Rug at germeval: Detecting offensive speech in german social media. In *14th Conference on Natural Language Processing KONVENS 2018*.

Susan Benesch. 2014. Countering dangerous speech: new ideas for genocide prevention. *Washington, DC: US Holocaust Memorial Museum*.

Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. 2016. Counterspeech on twitter: A field study. *Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field- study/*.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data Science*, 5(1):11.

Rajen Chatterjee, Matteo Negri, Marco Turchi, Marcello Federico, Lucia Specia, and Frédéric Blain. 2017. Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168.

Danielle Keats Citron and Helen Norton. 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Victor De Boer, Michiel Hildebrand, Lora Aroyo, Pieter De Leenheer, Chris Dijkshoorn, Binyam Tesfa, and Guus Schreiber. 2012. Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 16–20. Springer.

Fabio Del Vigna, Andrea Cimino, Felice DellOrletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook.

Felice DellOrletta and Malvina Nissim. 2018. Overview of the evalita 2018 cross-genre gender prediction (gxg) task. *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy. CEUR. org*.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Mai ElSherief, Shirin Nilizadeh, Dana Nguyen, Giovanni Vigna, and Elizabeth Belding. 2018. Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.

Julian Ernst, Josephine B Schmitt, Diana Rieger, Ann Kristin Beier, Peter Vorderer, Gary Bente, and Hans-Joachim Roth. 2017. Hate beneath the counter speech? a qualitative content analysis of user comments on youtube related to counter speech videos. *Journal for Deradicalization*, (10):1–49.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA18), Turin, Italy. CEUR. org*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. 2015. *Countering online hate speech*. Unesco Publishing.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Njagi Dennis Gitari, Zhang Zuping, Hanyurwimfura Damien, and Jun Long. 2015. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. *arXiv preprint arXiv:1805.03122*.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Dirk Hovy. 2015. Demographic factors improve classification performance. 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)*, volume 1, pages 752–762.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 591–598.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. 2015. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2017. Googles multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Filip Klubička and Raquel Fernández. 2018. Examining a hate speech corpus for hate speech detection and popularity prediction. In *LREC*.

Irene Kwok and Yuzhou Wang. 2013. Locate the hate: Detecting tweets against blacks. In *Twenty-seventh AAAI conference on artificial intelligence*.

Ryan Lowe, Nissan Pow, Iulian V Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, page 285.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. 2012. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the Second Workshop on Language in Social Media*, pages 27–36. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. 2018a. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Binny Mathew, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherje. 2018b. Thou shalt not hate: Countering online hate speech. *arXiv preprint arXiv:1808.04409*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018a. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018b. Training millions of personalized dialogue agents. In *EMNLP*.

Yashar Mehdad and Joel Tetreault. 2016. Do characters abuse more than words? In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 299–303.

Juan E Mezzich, Helena C Kraemer, David RL Worthington, and Gerald A Coffman. 1981. Assessment of agreement among several raters formulating multiple diagnoses. *Journal of psychiatric research*, 16(1):29–39.

Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th international conference on world wide web*, pages 145–153. International World Wide Web Conferences Steering Committee.

Jasper Oosterman, Alessandro Bozzon, Geert-Jan Houben, Archana Nottamkandath, Chris Dijkshoorn, Lora Aroyo, Mieke HR Leyssen, and Myriam C Traub. 2014. Crowd vs. experts: nichesourcing for knowledge intensive tasks in cultural heritage. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 567–568. ACM.

Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. 14-exlab@ unito for ami at ibereval2018: Exploiting lexical knowledge for detecting misogyny in english and spanish tweets. In *3rd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2018*, volume 2150, pages 234–241. CEUR-WS.

Lingyun Qiu and Izak Benbasat. 2010. A study of demographic embodiments of product recommendation agents in electronic commerce. *International Journal of Human-Computer Studies*, 68(10):669–688.

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 617–622. ACM.

Kelly Reynolds, April Kontostathis, and Lynne Edwards. 2011. Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine learning and applications and workshops*, volume 2, pages 241–244. IEEE.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at Fukuoka, Japan*, pages 1–23.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. *EMNLP 2018*, page 107.

Elena Shushkevich and John Cardiff. 2018. Classifying misogynistic tweets using a blended model: The ami shared task in ibereval 2018. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain*.

Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690.

Sara Owsley Sood, Elizabeth F Churchill, and Judd Antin. 2012. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology*, 63(2):270–285.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a whatsapp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Lucas Wright, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Susan Benesch. 2017. Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Guang Xiang, Bin Fan, Ling Wang, Jason Hong, and Carolyn Rose. 2012. Detecting offensive tweets via topical feature discovery over a large scale twitter corpus. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1980–1984. ACM.

Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958.