Italian Counter Narrative Generation to Fight Online Hate Speech

Yi-Ling Chung
University of Trento
Fondazione Bruno Kessler
ychung@fbk.eu

Serra Sinem Tekiroğlu
Fondazione Bruno Kessler
tekiroglu@fbk.eu

Marco Guerini
Fondazione Bruno Kessler
guerini@fbk.eu

Abstract

English. Counter Narratives are textual responses meant to withstand online hatred and prevent its spreading. The use of neural architectures for the generation of Counter Narratives (CNs) is beginning to be investigated by the NLP community. Still, the efforts were solely targeting English. In this paper, we try to fill the gap for Italian, studying how to implement CN generation approaches effectively. We experiment with an existing dataset of CNs and a novel language model, recently released for Italian, under several configurations, including zero and few shot learning. Results show that even for under-resourced languages, data augmentation strategies paired with large unsupervised LMs can held promising results.

Italiano. *Le Contro Narrative sono risposte testuali volte a contrastare l'odio online e a prevenirne la diffusione. La comunità di NLP ha iniziato a studiare l'uso di architetture neurali per la generazione di CN. Tuttavia, gli sforzi sono stati rivolti esclusivamente all'inglese. In questo lavoro, cerchiamo di colmare la lacuna per l'italiano, mostrando come implementare efficacemente approcci di generazione di CN. Sperimentiamo con un dataset esistente di CN e un modello del linguaggio per l'italiano recentemente rilasciato, in diverse configurazioni, tra cui zero e few shot learning. I risultati mostrano che anche per lingue con poche risorse, strategie di data augmentation abbinate a potenti modelli del linguaggio possono offrire risultati promettenti.*

1 Introduction

The rise of online Hate Speech (HS) brings along the need for combating strategies as it can trigger harmful psychological effects on the target groups and more crimes against them. While research studies have been widely focusing on hate speech detection methodologies for social media platforms (Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018), a recent line of research has taken the problem a step further by addressing the automatic generation of counter responses, aka counter narratives (Qian et al., 2019; Tekiroğlu et al., 2020), in order to assist non-governmental organizations in their real-world online hatred combating efforts. An example of HS along with a possible CN are shown below:

HS: *Gli arabi sono tutti terroristi e vogliono conquistarci con la violenza e le bombe. Bisogna rispondere con il napalm.* [Arabs are all terrorists and they want to conquer us with violence and bombs. We must respond with napalm.]

CN: *Essere di origine araba non significa essere terroristi, evitiamo generalizzazioni che portano solo ad altro odio.* [Being of Arab descent does not mean being a terrorist, let's avoid generalizations that only lead to more hatred.]

Despite the encouraging results of the counter narrative generation task, experiments have been limited to English due to the scarcity of hate speech / counter narrative data in other languages. In this paper, we investigate counter narrative generation for Italian as a case study where zero or only a small amount of task specific in-language data is available. We first explore the portability of generation across languages, considering that recent neural machine translation (NMT) systems have shown outstanding performances. We pro-

pose utilizing off-the-shelf NMT models to synthesize silver data from other languages, and fine-tuning GePpeTto (Mattei et al., 2020), a recently developed GPT-2 based language model for Italian, on the silver data. We then examine the effect of combining silver with gold data on CN generation by experimenting with various gold data sizes. Our findings show that a proper combination of silver and gold data while fine-tuning LMs can drastically reduce the need for expert-annotator effort on target languages.

2 Related Work

In this section we briefly recap relevant works for our counter narrative generation task, including the problem of online hatred recognition, effectiveness of approaches to hatred intervention, methodologies for generating counter-arguments, and text generation for low-resourced languages.

Hate problem. A wealth of work has investigated online hateful content, aiming at creating datasets for hate speech identification (Warner and Hirschberg, 2012; Burnap and Williams, 2015; Silva et al., 2016). For instance, there are datasets collected from Facebook (Kumar et al., 2018), forums (Silva et al., 2016; de Gibert et al., 2018), and Twitter (Silva et al., 2016; Waseem and Hovy, 2016). Hate speech detection tasks are available at IberEval (Fersini et al., 2018) for Spanish and EVALITA (Del Vigna12 et al., 2017; Fersini et al., 2018) for Italian.

Hate countering. Counter narratives can be used as an effective approach to moderate hateful content on social media platforms such as Twitter (Munger, 2017; Wright et al., 2017), Youtube (Ernst et al., 2017; Mathew et al., 2019) and Facebook (Schieb and Preuss, 2016). Previous studies on hate countering cover several aspects of CNs. For example: defining counter narratives (Benesch et al., 2016), studying their effectiveness (Schieb and Preuss, 2016; Silverman et al., 2016; Ernst et al., 2017; Munger, 2017; Wright et al., 2017), linguistically characterizing online counter narrative accounts (Mathew et al., 2018), creating real or simulated CN datasets (Mathew et al., 2019; Chung et al., 2019; Qian et al., 2019; Tekiroğlu et al., 2020), and neural approaches to CN generation (Qian et al., 2019; Tekiroğlu et al., 2020).

Counter-argument Generation. This task share the same abstract goal as CN generation - i.e. to produce the opposite or alternate stance of a statement. Previous works adopted sequence-to-sequence architectures to generate arguments (Rakshit et al., 2019; Hua et al., 2019; Rach et al., 2018; Le et al., 2018) targeting specific domains in which massive discussion is available, such as politics (Hua et al., 2019; Hua and Wang, 2018; Le et al., 2018), and economy (Le et al., 2018; Wachsmuth et al., 2018).

NLG for under-resourced languages. In spite of several studies addressing NLG, only a few have investigated the generation for languages other than English. For instance, there is the porting of SimpleNLG API (Gatt and Reiter, 2009) to Dutch (de Jong and Theune, 2018) and Italian (Mazzei et al., 2016), or Bilingual generation via combining NMT and Generative Adversarial Networks (Rashid et al., 2019).

3 Italian Counter Narrative Generation

Our main goal is to determine a methodology for Italian counter narrative generation considering the scarcity of gold standard data for training. Accordingly, we hypothesize that the availability of a decent amount of silver data can provide a kick-start for the generative models. Therefore, we resort to data augmentation through translation with the help of the existing datasets of hate speech / counter narrative pairs in other languages. For translation setting, we use DeepL[1], an off-the-shelf and well-performing MT system, to translate data from other languages to Italian. The translated pairs are used for fine-tuning a large Italian pre-trained generative model, i.e. GePpeTto, along with the original Italian gold standard pairs.

4 Dataset

For our study, we use CONAN dataset (Chung et al., 2019), which is a niche-sourced hate-countering dataset that consists of HS/CN pairs focusing on Islamophobia. The dataset provides pairs in English, French, and Italian, collected with the help of operators from three European NGOs specialized in online hate countering. Each pair in CONAN can either be an original or one of the 2 paraphrases of an original pair. In the experiments, we used the following splits:

1. 2142 pairs (original IT pairs and 1 IT paraphrase pair) as a training set made of gold

[1] https://www.deepl.com/translator

standard data.

2. 5996 pairs as a training set made of silver data obtained by automatically translating FR and EN pairs to IT.
3. 1071 pairs (the rest of the IT paraphrased pairs) are kept for testing purposes.

5 Models

In order to inspect how Italian CN generation can be accomplished under different resource conditions, we test the effect of using (i) silver data, (ii) gold standard data, and (iii) their combination. In particular we experiment with the following configurations on which `GePpeTto` is fine-tuned:

GP-trans. `GePpeTto` is fine-tuned on the silver data obtained by translating EN and FR pairs to IT using DeepL. This configuration represents the worst case scenario, where no HS/CN pair is available in the target language, and corresponds to a zero-shot learning setting.

Gp-ita. We fine-tune `GePpeTto` on all the original IT pairs in CONAN. This represents our practical best-case scenario, despite the fact that more pairs might provide better results.

GP-hybrid. We conjecture that introducing even a small amount of gold standard examples can help LMs adapt to the domain-specific idiosyncrasies. Moreover, we inspect how generation performance varies with the size of gold standard data provided. In this regard, we conduct a second phase of fine-tuning on top of the GP-trans model using 100, 300, 500, 800, and full IT pairs of CONAN. Therefore, we can represent various intermediate conditions of few-shot learning where few to several pairs for the target language are available. Thus, we assess how much the pretraining with the silver data helps to reduce the amount of gold standard data needed to reach a proper generation performance.

5.1 Training Details

For all the experiments, we have used `GePpeTto` as the pretrained Italian language model adopted from HuggingFace's Transformers library[2] and fine-tuned our models on a single K80 GPU using a batch size of 2048 tokens. The training pairs are represented as $[HS\_start\_token]$ HS $[CN\_start\_token]$ CN $[CN\_end\_token]$. The hyperparameter tuning details are provided in the following. At test time, we employed nucleus sampling with a p value of 0.9 for the generation of CNs. Conditioned on HSs, the generated sequence of text tagged with $[CN\_start\_token]$ CN $[CN\_end\_token]$ is selected as output.

[2]https://github.com/huggingface/transformers

Training Epochs We have empirically chosen 5 epochs for training for all the configurations, tuned from $\{2, 3$ and $5\}$ on test set. Preliminary experiments show that while lower number of epochs grant higher novelty in the output, they also came at the cost of lower BLEU scores. A further manual evaluation confirmed that the generation with 5 epoch provides more suitable responses.

Learning rate Once defining the epochs, we experimented with different learning rates of [1,2,5]e-5 and chose 5e-5 for the best performing setting - preliminary experiments show that while producing less novel and slightly more repeated text, the learning rate of 5e-5 consistently has better results in terms of BLEU and ROUGE scores.

Fine-tuning steps. In case where multiple datasets (silver and gold standard) were used, we followed a multi-step fine-tuning procedure by first using the silver and then the gold standard dataset. Gururangan et al. (2020) showed that task-adaptive pretraining using curated datasets from a dataset with similar distribution with the end task, provides significant improvements. Our fine-tuning schema follows this finding by first fine-tuning `GePpeTto` with the silver data as the task adaptive pretraining with an augmented dataset. Our preliminary experiments confirmed that adapting fine-tuned models towards the language characteristics of the target corpus is more effective than mixing silver and gold data together in a single fine-tuning procedure.

5.2 Evaluation

For our experiments we report word-overlap metrics BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to evaluate the CN generation on the gold standard test set. As for the generation quality, we compute Repetition Rate (Bertoldi et al., 2013) and Novelty (Wang and Wan, 2018) to assess how *Diverse* a response is with reference to the given HS and how *Novel* the generation is concerning the training data.

We also conduct a human evaluation to compare the generation quality of the configurations based on 3 criteria: **(i) Suitableness.** How suitable the given CN is as a response for the input HS. **(ii) Specificity.** How specific the given CN is as a response. This metric is used to discern suitable responses that are nonetheless very generic. **(iii) Grammaticality.** How grammatically correct the given CN is. All scores were in a scale from 1 to 5.

6 Results and Discussion

Model comparison. Results in Table 1 show that using the silver data (GP-trans) provides a viable step towards a proper model. When gold standard data is also available (GP-hybrid), we obtain better quantitative performance in terms of BLEU and ROUGE scores in comparison to the best case scenario (GP-ita). Furthermore, mixing the silver translation and the Italian gold standard data (GP-hybrid) yields better performances also in terms of output diversity (RR 11.7 vs 12.8). On the contrary, the most novel output is obtained by GP-trans, which can be expected since EN and FR pairs usually have slightly different focus on the topic of Islamophobia (topics and tropes can vary across nations and cultures). In Table 2 we provide few examples of generated CNs.

Learning Curve Discussion. As can be seen in Figure 1, even 100 Italian pairs are enough to dramatically improve the performances of `GePpeTto` on the task of CN generation over the baseline GP-trans. If we continue fine-tuning GP-trans with more and more Italian pairs, soon we are able to outperform also GP-ita. The number of examples required to obtain a new state of the art CN generation in Italian comes within 200 and 300, which reduces the required amount of gold standard data by around 80%. Therefore, it becomes clear that a good NMT model can be of fundamental help while porting the generation task to new languages, especially if few or no gold standard examples are available in the target language. Considering the fact that the counter narrative data collection is an expert-based task requiring costly human effort (Chung et al., 2019), decreasing the required amount of expert data can be of remarkable importance for low-resource languages.

Human Evaluation. As annotators, we employed 2 Italian native speakers that are expert

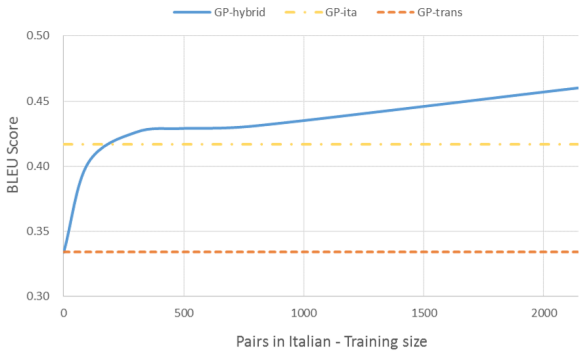

Figure 1: Learning Curve of GP-hybrid model while Italian pairs being added. GP-hybrid performance with no examples is shown as GP-trans.

in counter narrative production. The annotators were instructed in assessing CN *suitableness*, *specificity*, and *grammaticality* with respect to the paired hate speech. During training, we explained what a common and suitable counter narrative is, and then asked them to intuitively evaluate the generation without overthinking. We further presented 20 examples of HS/CN pairs to demonstrate the appropriate evaluation. In order to avoid comparison or primacy/recency effects, we have presented 20 random pairs from each condition to each annotator as a single randomized file and asked them to evaluate each counter narrative with respect to the 3 criteria. The results presented in Table 3 show that all models reach very high levels of grammaticality; most of the sentences were completely grammatical and few ungrammatical ones were due to dangling sentences. Moreover, using silver data alone can already provide a performance lower than but close to the GP-ita case for Suitableness and Specificity. Finally, fine-tuning GP-trans further using gold standard data (GP-hybrid) provides the most suitable and the least generic responses among the 3 models in line with their performance ranking of automatic metrics.

7 Conclusion and Future Work

Counter narrative generation using neural architectures is beginning to be studied for hatred intervention. In this paper, we presented the first attempt of CN generation for Italian, investigating several variations of generation when gold data is limited or not available. Our experiments reveal that with simple data augmentation strategies paired with powerful LMs can bring promising

Model	$BLEU_s$	$BLEU_c$	ROUGE1	ROUGE2	ROUGEL	RR	Novelty
GP-trans	0.327	0.048	0.236	0.061	0.196	12.795	0.785
GP-ita	0.417	0.231	0.343	0.187	0.305	12.870	0.561
GP-hybrid	0.460	0.287	0.380	0.234	0.344	11.752	0.522

Table 1: Quantitative results of fine-tuned models. BLEU scores are reported at sentence-level ($BLEU_s$) and corpus-level ($BLEU_c$).

GP-trans: *La maggior parte dei pedofili nel Regno Unito sono uomini bianchi, quindi secondo la vostra logica il mondo sarebbe un posto migliore senza uomini bianchi?* [Most pedophiles in the UK are white men, so by your logic would the world be a better place without white men?]
GP-ita: *Non esiste nessuna invasione islamica dell'occidente. La popolazione musulmana in Italia non supera il 5%.* [There is no Islamic invasion of the West. The Muslim population in Italy does not exceed 5%.]
GP-hybrid: *Non si può ritenere una invasione quando il numero di musulmani in Italia è inferiore addirittura al 5%.* [It cannot be considered an invasion when the number of Muslims in Italy is less than even 5%.]
Gold: *Salve, in realtà il numero di musulmani in Italia è pari a poco meno il 5% della popolazione totale: ritiene davvero che si possa parlare di sostituzione?* [Hello, in reality the number of Muslims in Italy is equal to a little less than 5% of the total population: do you really think that we can talk about substitution?]

Table 2: Sample CN generations along with EN translation. GP-trans generation is grammatically correct but focused on the UK/FR scenario. Instead, GP-ita and GP-hybrid can mimic gold arguments with novel and diverse wording.

results for under-resourced languages. As future work we plan to explore methods for improving specificity via the use of diverse types of knowledge (e.g., facts, events and named entities) from external resources. This would also mitigate the problem of the model possibly generating hallucinated content (i.e. non-existent or wrong content), see for example (Maynez et al., 2020; Nie et al., 2019). Finally, we plan to apply this approach to other hate phenomena such as antisemitism, homophobia, and misogyny.

Model	Suitable	Specific	Grammar
GP-trans	2.47	2.20	4.52
GP-ita	2.78	2.32	4.72
GP-hybrid	2.82	2.57	4.40

Table 3: Human evaluation results.

References

[Benesch et al.2016] Susan Benesch, D Ruths, KP Dillon, HM Saleem, and L Wright. 2016. Considerations for successful counterspeech.

[Bertoldi et al.2013] Nicola Bertoldi, Mauro Cettolo, and Marcello Federico. 2013. Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

[Burnap and Williams2015] 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.

[Chung et al.2019] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. 2019. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829.

[de Gibert et al.2018] Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.

[de Jong and Theune2018] Ruud de Jong and Mariët Theune. 2018. Going dutch: Creating simplenlg-nl. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 73–78.

[Del Vigna12 et al.2017] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

[Ernst et al.2017] 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.

[Fersini et al.2018] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

[Fortuna and Nunes2018] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

[Gatt and Reiter2009] Albert Gatt and Ehud Reiter. 2009. Simplenlg: A realisation engine for practical applications. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG 2009)*, pages 90–93.

[Gururangan et al.2020] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[Hua and Wang2018] Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. *arXiv preprint arXiv:1805.10254*.

[Hua et al.2019] Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. *arXiv preprint arXiv:1906.03717*.

[Kumar et al.2018] Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

[Le et al.2018] Dieu-Thu Le, Cam Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

[Lin2004] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

[Mathew et al.2019] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

[Mattei et al.2020] Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model.

[Maynez et al.2020] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

[Mazzei et al.2016] Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th international natural language generation conference*, pages 184–192.

[Munger2017] Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

[Nie et al.2019] Feng Nie, Jin-Ge Yao, Jinpeng Wang, Rong Pan, and Chin-Yew Lin. 2019. A simple recipe towards reducing hallucination in neural surface realisation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2673–2679, Florence, Italy, July. Association for Computational Linguistics.

[Papineni et al.2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[Qian et al.2019] Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

[Rach et al.2018] Niklas Rach, Saskia Langhammer, Wolfgang Minker, and Stefan Ultes. 2018. Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*.

[Rakshit et al.2019] Geetanjali Rakshit, Kevin K Bowden, Lena Reed, Amita Misra, and Marilyn Walker. 2019. Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*, pages 45–52. Springer.

[Rashid et al.2019] Ahmad Rashid, Alan Do-Omri, Md Haidar, Qun Liu, Mehdi Rezagholizadeh, et al. 2019. Bilingual-gan: A step towards parallel text generation. *arXiv preprint arXiv:1904.04742*.

[Schieb and Preuss2016] 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.

[Schmidt and Wiegand2017] 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.

[Silva et al.2016] Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. *arXiv preprint arXiv:1603.07709*.

[Silverman et al.2016] Tanya Silverman, Christopher J Stewart, Jonathan Birdwell, and Zahed Amanullah. 2016. The impact of counter-narratives. *Institute for Strategic Dialogue*, pages 1–54.

[Tekiroğlu et al.2020] Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[Wachsmuth et al.2018] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251.

[Wang and Wan2018] Ke Wang and Xiaojun Wan. 2018. Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

[Warner and Hirschberg2012] 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.

[Waseem and Hovy2016] 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.

[Wright et al.2017] 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.