UNIVERSITÀ
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

ICT International Doctoral School

Counter Narrative Generation for Fighting Online Hate Speech

Yi-Ling Chung

Advisor:

Marco Guerini Fondazione Bruno Kessler

Thesis Committee:

Malvina Nissim University of Groningen

Viviana Patti University of Turin

April, 2022

Trento, Italy

Abstract

Studies on online hate speech have mostly focused on the automated detection of harmful messages. Still, tackling hate speech in the standard way of content moderation may be charged with censorship and overblocking. One alternate strategy, that has received little attention so far, is to directly intervene in the discussion with counter narratives (i.e., textual responses meant to withstand hatred and prevent its spreading). While recent advances in NLP can help in automatizing counter narrative generation by using pre-trained language models, challenges such as lack of high-quality data, generic generation, hallucination and multilinguality must be addressed. This dissertation focuses on automatic and effective hate mitigation targeting Islamophobia by data collection and counter narrative generation. Firstly, we tackle the problem of data scarcity. We present CONAN, the first large-scale and expert-based hate countering dataset for English, Italian and French. Then, we present an author-reviewer approach that can create automatically high-quality data while reducing human effort. Secondly, we develop models to generate counter narratives focusing on informative and multilingual responses. We introduce a knowledge-driven pipeline that can produce suitable and informative English counter narratives while avoiding hallucination phenomena. We address multilinguality by presenting approaches to counter narrative generation for Italian, and characterizing the effect of data size and of data quality on model performance. Thirdly, we present an extensive evaluation of automatic counter narrative generation embedded in a platform that NGO operators can use to monitor social media data and counter hate speech. Results show an increased efficiency and effectiveness of operators' activities. We conclude by discussing our contributions and future research directions on building models for hate countering.

Keywords

[Hate Speech Countering, Counter Narrative Generation, Multilinguality, Knowledge-Grounded Generation, Counter Narrative Generation Evaluation]

Acknowledgments

This thesis is dedicated to my dad, a great man who has tried to give me the whole world and basically never said no to my requests.

I would never imagine one day I could complete a PhD. Undoubtedly, this dissertation cannot be accomplished without all the people who supported and helped me, especially when I got injured.

I want to express my gratitude to my advisor, Marco Guerini, for his guidance, support, and patience. Marco is a fun, thoughtful and empathetic man who always creates a positive environment that stimulates insightful discussion. Most importantly, thank you for encouraging me to pursue my interests, for brainstorming with me, for pointing out my weakness, for helping me improve my writing, and for persisting in exploring even when experimental results are not so exciting. I appreciate that Marco introduces me to counter narratives, which I did not know anything about prior to my PhD journey. Also, I am grateful for having met Serra Sinem Tekiroğlu who is a great collaborator and friend. Thank you for all technical and emotional support as well as for being critical that helps refine this research.

Special thanks to Rodrigo Agerri for hosting me at the University of the Basque Country during my internship and for sharing different perspectives to approach multilingual counter narrative generation. I want to thank reviewers, Malvina Nissim and Viviana Patti, for their valuable comments to improve this thesis. I would give special thanks to all contributors and NGO operators who provided high-quality data and evaluated model generation in this thesis.

I have been fortunate to meet colleagues and friends in FBK and Trento: Samuel, Enrica, Silvia, Alina, Marco, Sara, Vevake, Begoña, Milene, Federica, Burcu, Nandu, Elham, Aniruddha, Sudipan and people whose names I cannot recall. Lastly, I want to thank my family and Xabier, who are always there for me, cheer me up, and support all the decisions I made, even when I lost my temper.

Table of Contents

Abstract **iii**

Acknowledgments **v**

Table of Contents **vii**

List of Tables **xii**

List of Figures **xvii**

1 Introduction **1**

1.1 Research Problems 4

1.2 Contributions 6

1.3 Ethical Considerations 7

1.4 Structure of the Thesis 8

2 Background **9**

2.1 The Problem of Online Hate 9

2.1.1 Hate Speech Terminology and Typology 10

2.2 Standard Approach to Hate Moderation 11

2.2.1 Content Moderation of Hate Speech 12

2.2.2 Hate Speech Datasets 13

2.2.3 Approaches to Hate Speech Detection 15

2.3 Neural Language Models . . . 16
2.3.1 Sequence-to-Sequence Models . . . 16
2.3.2 Encoder-Decoder Architecture . . . 17
2.3.3 Transformer Model . . . 18
2.3.3.1 Self-Attention . . . 19
2.3.3.2 Positional Embeddings . . . 20
2.3.4 Pre-Trained Language Models . . . 21
2.4 Natural Language Generation . . . 24
2.4.1 Counter Argument Generation . . . 24
2.4.2 Knowledge-Grounded Generation . . . 25
2.4.3 NLG for Under-Resourced Languages . . . 26
2.5 Evaluation Metrics . . . 26

3 Strategies for Hate Countering Dataset Collection 29
3.1 Introduction . . . 29
3.2 Related Work . . . 30
3.3 Strategies of Counter Narrative Collection . . . 30
3.4 Creation of Hate Countering Dataset through Nichesourcing . . . 32
3.4.1 Fundamentals of CONAN . . . 32
3.4.2 Data Collection Methods . . . 33
3.4.3 Dataset Augmentation and Annotation . . . 36
3.4.4 Dataset Statistics . . . 38
3.4.5 Evaluation . . . 41
3.5 Creation of Hate Countering Dataset through Author-Reviewer Architecture 44
3.5.1 Characteristics of the Datasets . . . 44
3.5.2 CN Generation through Author-Reviewer Architecture . . . 47
3.5.3 The Author . . . 49

3.5.3.1 Generation Models 49
3.5.3.2 Baselines 50
3.5.3.3 Results 50
3.5.4 The Reviewer 52
3.5.4.1 Machine Reviewer Experiment 52
3.5.4.2 Non-Expert Reviewer Experiment 54
3.5.4.3 Expert Reviewer Experiments 55
3.5.5 Author-Reviewer Architecture Discussion 57
3.5.6 Generation Examples 59
3.6 Conclusion 63

4 Knowledge-Guided Counter Narrative Generation 65
4.1 Introduction 65
4.2 Counter Arguments versus Counter Narratives 67
4.3 HS-CN Dataset 68
4.4 Knowledge-Guided Counter Narrative Generation Architecture 68
4.5 Knowledge Retrieval Module 69
4.5.1 Knowledge Repository 69
4.5.2 Query Construction 70
4.5.3 Knowledge Sentence Selection 72
4.6 Counter Narrative Generation Module 73
4.6.1 Experiments 73
4.6.1.1 Models 73
4.6.2 Results for the Silver Knowledge Test Set 74
4.6.3 Results for the Gold Knowledge Test Set 77
4.7 Discussion 82
4.8 Conclusion 82

4.9 Generation examples 84

5 Multilinguality for Counter Narrative Generation 89

5.1 Introduction 89

5.2 Counter Narrative Generation for Under-Resourced Settings 90

5.3 Experiments 91

5.3.1 Dataset 91

5.3.2 Models 91

5.3.3 Training Details 92

5.3.4 Evaluation Metrics 93

5.4 Results 93

5.5 Conclusion 96

5.6 Generation Examples 97

6 Counter Narrative Generation Evaluation 101

6.1 Introduction 101

6.2 Hate Countering Platform 102

6.2.1 Monitoring Functionalities 102

6.2.2 Counter Narrative Suggestion Approach 104

6.3 Experiment Setup 107

6.4 Questionnaires 109

6.4.1 UEQ 109

6.4.2 Additional Questionnaire and Open-ended Questions 109

6.5 Data Analysis 110

6.5.1 UEQ Analysis 110

6.5.2 Additional Questionnaire and Open Questions Analysis 114

6.5.3 Content Analysis and Activity Logs 115

6.6 Conclusion 117

7 Conclusions **119**

7.1 Challenges and Future Directions 120

References **123**

A Additional Materials **151**

A.1 Preliminary Analysis on Decoding Methods for XNLG. 151

A.2 Usability Exercise Instructions 153

List of Tables

1.1 An example hate speech with appropriate counter narratives (CN1 and CN2) and an inappropriate one (CN3) containing abusive language. 3

3.1 Counter narrative guidelines adopted by the three NGOs. 35

3.2 Example pairs for the three languages, along with English translations. . . 36

3.3 Example paraphrases of a hate speech. 37

3.4 Main statistics of the dataset. 39

3.5 Demographic profile of the operators. 40

3.6 Counter narrative type distribution over the three languages (% over the total number of labels). 40

3.7 Hate speech sub-topic type distribution over the three languages (% over the total number of labels). 41

3.8 Diversity analysis of the three datasets. Semantic diversity is reported in terms of CN type percentages, Lexical diversity in terms of Repetition Rate (RR - average over 5 shuffles). 45

3.9 Some examples of the categories relevant to our analysis. Hostile from CRAWL dataset, Denouncing from CROWD, Fact (other) from NICHE. . 46

3.10 Comparison of different approaches proposed in the literature according to the main characteristics required for the CN dataset. 47

3.11 Evaluation results of the best author configurations with different datasets. Novelty is computed w.r.t. to the corresponding training set, RR in the produced test output. 51

3.12 F1, Precision and Recall results for the two main classifier configurations we tested. 54

3.13 Percentage of filtered pairs according to various filtering conditions. 55

3.14 Results for CN collection under various configurations. RR for 'no suggestion' is computed on NICHE dataset and the time needed is the one reported in Section 3.4. Time is expressed in seconds. Pairs_{selec} indicates the percentage of original author pairs that have been passed to the expert for reviewing, Pairs_{final} indicates the percentage of selected pairs that have been accepted or modified by the expert themselves. Crowd_{time} is computed considering that annotators gave a score every 35 seconds, and we required two judgments per pair. 56

3.15 Randomly sampled CNs generated from GPT-2 and TRF models trained on CROWD dataset. 59

3.16 Randomly sampled CNs generated from GPT-2 and TRF models trained on NICHE dataset. 60

3.17 Randomly sampled CNs generated from GPT-2 and TRF models trained on NICHE dataset. 61

3.18 Randomly sampled CNs, generated from GPT-2, filtered from various reviewer configurations. 62

4.1 Hallucinated CN generated by GPT-2 that is fluent and credible (according to Office for National Statistics, the Muslim population is just above 3M). 66

4.2 Examples of KN retrieved using (i) extracted queries from HS (Q_{hs}), (ii) generated queries (Q_{gen}), and (iii) extracted queries from both HS and CN ($Q_{hs \cup cn}$). 72

4.3 Results of CN generation with silver knowledge. We report novelty (Nov.), RR, BLEU-2 (B-2), ROUGE-L (R-L), KN overlap with generation and the average amount of words (#W.) and sentences (#S.) per generation. . . . 76

4.4 Results of unique HS test set (U) for novelty and diversity. For comparison, we keep the results of whole test set (W). 78

4.5 Results of CN generation with gold knowledge in-target and cross-target test sets. 79

4.6 Samples of cross-domain generation. 79
4.7 Annotation examples with score explanations. 81
4.8 Human evaluation results of CN generation. 81
4.9 Sample of in-domain generation with KN retrieved using Q_{hs}. 84
4.10 Sample of in-domain generation with KN retrieved using $Q_{hs \cup gen}$. 85
4.11 Sample of in-domain generation with KN retrieved using $Q_{hs \cup cn}$. 86
4.12 Sample of cross-domain CN generation with gold knowledge. 87

5.1 Quantitative results of fine-tuned models. 94
5.2 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. 94
5.3 Human evaluation results. 96
5.4 Sample CN generations along with EN translation. 97
5.5 Sample CN generations along with EN translation. 98
5.6 Sample CN generations along with EN translation. 99
5.7 Sample CN generations along with EN translation. 100

6.1 Islamophobic messages monitored between October 2018 and May 2019 in English, French and Italian. 103
6.2 Overview of the tasks for CN^- and CN^+ conditions. 110
6.3 UEQ scales with and without counter narrative suggestion tool (CN^+ and CN^- respectively). $CN^{-/+}$ denotes CN^- receiving suggestion tool in a post-hoc experiment. 111
6.4 Pragmatic and hedonic quality. 112
6.5 Benchmark intervals for the UEQ scores (per scale). 112
6.6 Statistics of 6 indicators from additional questionnaire on CN^-, CN^+ and $CN^{-/+}$ conditions (Med. denotes median.). 114

6.7 Time required for writing CN (modified CN based on suggestions vs. new CN from scratch) across 3 countries. Time is expressed in minutes. 117

A.1 Human evaluation on CN generation using various decoding methods. . . . 152

A.2 Instructions for usability exercises. 153

List of Figures

2.1 A simplified encoder-decoder model reads an HS source sentence and produces a CN as the target sentence. 18

2.2 The architecture of the Transformer (figure source: Vaswani et al. (2017)). 19

3.1 Screenshot of data collection form. 35

3.2 The author-reviewer configuration. The author module produces HS-CN candidates and the reviewer(s) filter them. Finally, an NGO operator validates and eventually post-edits the filtered candidates. 48

4.1 Architecture of knowledge grounded generation with extracted (green solid arrow) and generative (dotted arrow) queries (topical phrases) that are exploited to retrieve relevant knowledge. The knowledge sentences extracted together with input HS are fed to CN generation. 69

5.1 Learning curve of GP-hybrid model while Italian pairs being added. GP-hybrid performance with no examples is shown as GP-trans. 95

6.1 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 direclty obtain counter narratives to that particular tweet. 104

6.2 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. 105

6.3 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. 105

6.4 On top the hate content "Every Muslim is a potential terrorist". Below a list of possible counter narratives automatically suggested for that hate content. 106

6.5 Statistics on CN^{-}, CN^{+} and $CN^{-/+}$ conditions, compared with the benchmark for UEQ scales. 113

Chapter 1

Introduction

Social media have become the main place for information exchange and opinionated discussion. While a huge number of users passively receive and digest news and messages from social media, some can actively post inflammatory comments while remaining anonymous (Aguilera-Carnerero and Azeez, 2016). In this context, online Social Media Platforms (SMPs) can be used maliciously as a tool for spreading hate speech, appealing to a wide audience, and marginalizing specific communities (Awan, 2016; Hall, 2013). Consequently, hate speech (including offensive language and abusive language) can proliferate quickly and is difficult to regulate. As an example, a recent survey shows that in the UK 30-40% of people have experienced online abuse, and 10-20% of people have been targeted by hateful content (Vidgen et al., 2019).

In this respect, inaction against hate speech allows for deepening prejudice and stereotypes, while bystanders may receive false messages and consider them correct (Citron and Norton, 2011). Victims of hate speech are likely to fear for their physical safety and experience psychological illness such as anxiety, self-blame, and suicidal thoughts (Jay, 2009). Thus, this scenario has motivated international authorities in the search for effective ways of combating online hatred. For instance, several countries constrain hate speech and discrimination by enforcing laws and sanctions such as fines and imprisonment (Waldron, 2012). In parallel, the approach to hatred intervention adopted by SMPs is content moderation. Such approach relies on the suspension of user accounts and deletion of hate messages, while trying to weigh the right to freedom of expression.

Still, these containment strategies 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). An alternative and preferable strategy is to use **counter**

narratives (or counter comments or counter speech). They are polite and non-aggressive responses to hate speech, targeting and counteracting extreme statements (Benesch, 2014; Schieb and Preuss, 2016). A Counter Narrative (CN) is direct feedback in hate discussion and is considered as an effective tool since it: preserves the right to freedom of speech, rectifies stereotypes and misleading information with credible evidence, can alter the viewpoints of haters and bystanders, and encourages the exchange of opinions among possible bystanders to change their perspectives (Benesch, 2014; Schieb and Preuss, 2016; Mathew et al., 2019; Stroud and Cox, 2018). It can also help mutual understanding and de-escalate the conversation. By emphasizing equality, inclusivity, and diversity using sympathetic approaches in counter narratives, we can change the Internet sphere and guide audiences to withstand hateful content.

In practice, not all responses to a hate message can be seen as 'good' counter narratives (Silverman et al., 2016; Tuck and Silverman, 2016). Following guidelines[1] provided by Non-Governmental Organizations (NGOs), a good counter narrative is a response that provides non-negative feedback with fact-bound arguments or alternative viewpoints, in order to avoid escalating the hatred in the discussion. Prior work on Twitter shows that effective counter narratives, inducing favorable impact on hate users, often use strategies such as warning of consequences, empathy, affiliation, and humor (Benesch et al., 2016b). In Table 1.1 we show an example Hate Speech (HS) with appropriate counter narratives (CN1 and CN2) and an inappropriate one (CN3)[2]. Although both examples CN1 and CN2 are appropriate responses, example CN2 is more effective as it is tailored to the HS and demonstrates contextualized and persuasive statements with supporting facts, knowledge, and logical reasoning (Habernal and Gurevych, 2016). On the contrary, example CN1 is a generic response that can be applied to any hate speech and CN3 is not appropriate since it contains abusive language.

Recently, several NGOs have trained operators to monitor SMPs and to compose effective counter narratives when necessary. Still, hate countering faces a number of challenges. Firstly, given the sheer amount of hate produced on a daily basis on SMPs, manual intervention via counter narratives is not scalable. To compose a counter narrative, an expert operator usually will go through reading hate messages, searching for and filtering related information online, and then composing a counter narrative. The whole process may take up to several minutes (Chung et al., 2021b). Advances in Natural Language

[1]For instance: `https://getthetrollsout.org/stoppinghate`

[2]It is crucial to note that this thesis contains examples of language which can be sensitive or offensive to some readers. They do not represent the view of the author.

HS: *The world would be a better place without Muslims. They are only killing and raping our children*
CN1: *Evidence for this? This is not true. How can you say this about an entire faith?*
CN2: *On the contrary, most children abuse is operated by people they know: a relative, family friends, sports coach, someone in a position of trust and authority. Besides, Muslims help people - A Muslim woman rushed to help the victims of a triple stabbing in Manchester on New Year's Eve.*
CN3: *You are truly one stupid backwards thinking idiot to comment on Muslims like that.*

Table 1.1: An example hate speech with appropriate counter narratives (CN1 and CN2) and an inappropriate one (CN3) containing abusive language.

Processing (NLP) can be used to assist NGO operators in composing counter narratives, thus decreasing the time and effort in fighting hatred online. Nevertheless, automatic CN generation requires a large amount of data and no data or tools is available for such purpose. Secondly, generating reasonable and effective counter narratives is challenging for machines (Zellers et al., 2019; Solaiman et al., 2019), since it requires several skills that are not fully developed yet: language comprehension, the storage and recall of relevant background knowledge (including facts and credible evidence), logical reasoning and language generation. Though recent approaches of sequence-to-sequence models (Cho et al., 2014; Sukhbaatar et al., 2015) have produced quite fluent responses in dialog systems (Ghazvininejad et al., 2018; Le et al., 2018) and question answering (Das et al., 2017; Kumar et al., 2019), such models have difficulties in maneuvering memory and knowledge, leading often to content hallucination (Zellers et al., 2019; Solaiman et al., 2019), and generic and incoherent responses (Li et al., 2016; Dinan et al., 2019; Ghazvininejad et al., 2018). For instance, example CN1 that is more generic and easier to generate may not be as effective as example CN2, since hate countering involves varieties of topics, targets and arguments. Thirdly, while models that can handle multiple languages are becoming increasingly important, related work focuses on English due to the limited data availability for most languages to obtain good performance (Rashid et al., 2019).

These challenges motivate this thesis to explore methods for generating effective counter narratives in a multilingual setting using NLP techniques to aid NGO operators. This thesis first tackles the problem of limited training examples available by experimenting with methods for creating high-quality data. Second, we aim to broaden counter narrative generation by infusing external unstructured knowledge to Language Models

(LMs), enabling the selection of new information and relevant facts. Third, we explore cross-lingual transfer methods and the use of pre-trained LMs to generate counter narratives in multiple languages. Ultimately, we examine the effectiveness of providing counter narrative suggestion in assisting NGOs via an extensive evaluation campaign.

1.1 Research Problems

This thesis investigates the automation of counter narrative generation concerning data collection, knowledge-grounded generation, low-resource languages, multilingual generation, and evaluation:

Q1. How can we obtain automatically high-quality training examples? Manual hate mitigation via counter narrative writing is difficult to scale up since substantial effort from experts is required. The data needed for developing Natural Language Generation (NLG) systems, that can be employed to help NGO operators, is not currently available. For this reason, we describe the creation of the first multilingual and high-quality hate countering dataset based on a niche-sourcing methodology in Chapter 3. Based on this manually crafted dataset, we explore a new approach to obtaining data: synthetic data generation in an author-reviewer framework, in which an author is tasked with text generation and a reviewer filters the produced output.

Q2. What makes a counter narrative effective and how can we produce such counter narratives? In Chapter 4, we consider the studies from two fields closely related, argumentation theory and hate countering, for producing successful and effective counter narratives. Then, we probe automated mechanisms producing such counter narratives with supporting and fact-bound arguments, exploiting information retrieval and language generation. From a modeling perspective, we examine empirically the potential of pre-trained and non-pretrained language models, under in-domain and cross-domain configurations.

Q3. How can we transfer the learned knowledge from high-resource language to low-resource languages? The adaptation of language models across languages is key for developing applications targeting multilingual online hate speech, particularly when gold data is limited or not available. Data augmentation has been highly beneficial to improve classification and machine translation performance. As a result, in Chapter 5 we explore data augmentation strategies through translation to alleviate the demand for gold standard data for training. Our approach is heavily based on machine translation systems

and pre-trained language models. We further investigate various learning configurations, including zero and few shot learning.

Q4. What are the effectiveness and efficacy of employing counter narrative generation systems? At the time of this work, no research has addressed the use of counter narrative generation tools in real scenarios. To understand and make progress in this direction, in Chapter 6 we present a retrieval-based counter narrative suggestion tool and assess its effectiveness for daily activities of NGO operators to curb hatred online. The performance of the suggestion tool is evaluated by more than 100 expert operators across 3 different countries through an extensive evaluation that encompasses open-ended questions, a user experience questionnaire, and time measurement.

1.2 Contributions

The specific contributions in this thesis with referenced publications include:

- Different methodologies for the creation of hate countering datasets, providing answers to questions such as how to automatically create high-quality datasets and how to identify counter narratives and their types.

  - Chung, Y.-L., Kuzmenko, E., Tekiroğlu, S. S., and Guerini, M. (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, Florence, Italy. Association for Computational Linguistics
  - Tekiroğlu, S. S., Chung, Y.-L., and Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics

- An approach to generating informative counter narratives using external unstructured knowledge.

  - Chung, Y.-L., Tekiroğlu, S. S., and Guerini, M. (2021a). Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics

- An approach to multilingual counter narrative generation for under-resourced languages when training data is limited or not available.

  - Chung, Y.-L., Tekiroğlu, S. S., and Guerini, M. (2020). Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Online

- An evaluation campaign on employing counter narrative suggestion tool for helping NGOs counter online hatred in terms of efficiency and effectiveness.

  - Chung, Y.-L., Tekiroğlu, S. S., Tonelli, S., and Guerini, M. (2021b). Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150

1.3 Ethical Considerations

While we believe that counter narratives are a better tool than content moderation in fighting hate speech (e.g. they do not hinder freedom of speech), still the automatic generation of CNs should be taken with care. Since this thesis aims at presenting methodologies for data collection and production of counter narratives through neural approaches, several ethical implications should be considered.

First, neural models can be vulnerable and exhibit prompted or unprompted toxic degeneration (Gehman et al., 2020). As an example, Microsoft's chatbot Tay learned to produce offensive content within 24 hours of its release[3]. To mitigate this issue, possible solutions include (1) our approach integrating in the pipeline a classifier or human reviewers for validation and possible post-editing, (2) detoxification techniques for controllable generation methods (Gehman et al., 2020), and (3) deletion of undesirable content from the corpora used for training the LMs we use (Raffel et al., 2020), even if the appropriate criteria for such purpose are still investigated.

Second, Neural Machine Translation (NMT) and NLG systems tend to hallucinate and produce misleading statements (Zellers et al., 2019; Solaiman et al., 2019). While our approach to knowledge-bound CN generation reduces the risks of content hallucination, an additional step, where the accuracy of the generated text is checked against the provided knowledge (Nie et al., 2019; Dušek and Kasner, 2020), would provide further robustness to the system.

Third, language models may still induce unintended social biases. Controlling this phenomenon is challenging, often requiring carefully defined notions of social biases and their sources. Prior work moderated this issue by measuring/promoting fairness in models and data employed (Blodgett et al., 2020), also designing bias triggers (Sheng et al., 2020) or regularization methods (Bordia and Bowman, 2019; Corbett-Davies et al., 2017) for controllable bias.

To sum up, though some additional automated techniques may help in maintaining generation quality, human evaluation should always be considered as the foremost solution, at least for delicate tasks such as 'real' hate countering on social media platforms. For this reason, we advocate that generation systems should be used as a suggestion tool for NGO operators, to make their countering work more effective. In this way, there is always a "human moderator" taking the final decision (Chung et al., 2019).

[3] https://en.wikipedia.org/wiki/Tay_(bot)

1.4 Structure of the Thesis

This thesis is organized in 7 chapters:

Chapter 2 discusses online hate speech phenomenon in general and methods for automatic hatred moderation by neural networks, including sequence-to-sequence models, Transformer and pre-trained language models. It then focuses on natural language generation using neural networks, in connection with the approaches proposed in the following chapters. Finally, it concludes by introducing metrics for evaluating the performance of generation models.

Chapter 3 introduces our multilingual expert-based dataset for hate countering, targeting English, Italian, and French, and probes methodologies to automatically produce training examples. It further discusses the pros and cons of methodologies for dataset collection.

Chapter 4 introduces a knowledge-driven approach to generating informative counter narratives. Experimental results are analyzed in in-domain and zero-shot cross-domain configurations, and comparisons are provided for multiple models followed by a human evaluation.

Chapter 5 proposes a method that extends the portability of generation across languages using data augmentation. The approach is evaluated under several configurations, including zero and few shot learning.

Chapter 6 presents an approach for evaluating the efficacy and effectiveness of producing counter narratives to assist NGOs in combating online hate speech. Experimental settings are evaluated in a multilingual perspective for English, Italian, and French.

Chapter 7 concludes by summarizing the contributions of the thesis, discussing the remaining open challenges and potential directions for future work.

Chapter 2

Background

This chapter provides the background material needed for the methodologies explored in this thesis. We begin by introducing the problem of online hate speech, and the definition of hateful content along with their key aspects in Section 2.1. We then discuss in Section 2.2 the standard approaches to hate mitigation, hate speech datasets used for detection, and the techniques for hate speech detection. We further review neural language models in Section 2.3 and how they are used for natural language generation task in Section 2.4. Finally we introduce the metrics employed for evaluating generated text in Section 2.5.

2.1 The Problem of Online Hate

On social media platforms various forms of online hate are commonly identified, such as cyberbullying (Xu et al., 2012) and racism (Waseem and Hovy, 2016). According to a survey conducted by the Pew Research Center, more than 60% of adults under thirty have been the victims of harassment online (Vogels, 2021). Hate incidents online are increasing and receiving more attention across countries. The COVID-19 pandemic has further stirred up hatred, both online and offline, peculiarly against East Asian with reports of insults and physical abuse (Tahmasbi et al., 2021).

Victims of hate speech are usually targeted because of various aspects such as race, gender, age, religion, sexual orientation, and physical appearance. Online hate speech can deepen prejudice and stereotypes and bystanders may receive false messages and consider them correct (Citron and Norton, 2011).

Addressing online hate is an urgent but challenging task. Several reasons can ac-

count for the difficulties. First, abusive language spreads quickly and it is difficult to monitor. The possibility of staying anonymous online stimulates haters to spread inflammatory messages without considering consequences (Rainie et al., 2017). Second, language changes over time and on the Internet it presents distinct characteristics in vocabulary and word-formation than offline. Undoubtedly addressing all variations is hard.

2.1.1 Hate Speech Terminology and Typology

The first step to combat hate speech is clarifying terminology. Unfortunately, the boundary that sets hate (or abuse) from non-hate is controversial, subjective and under debate, even when a definition is provided (Waseem, 2016; Ross et al., 2016). In fact, the definition of hate speech content in international human rights law varies in scope, culture, language, and severity. In this thesis we adopt the definition by UNESCO that characterizes hate speech as "expressions that advocate incitement to harm based upon the target's being identified with a certain social or demographic group" (Gagliardone et al., 2015). For instance, Sentence 1 shows explicit hostility towards a specific group with no reasons explained.

(1) *Muslims are a problem in my opinion. They do not want to assimilate. They demand that everyone shows respect for Islam, but do they respect our laws? One day they will oppress us with the Sharia and polygamy.*

Several umbrella terms are used in NLP research community when tackling online hate (Fortuna and Nunes, 2018), such as offensive language (Razavi et al., 2010), hate speech (Warner and Hirschberg, 2012; Burnap and Williams, 2015; Silva et al., 2016; Djuric et al., 2015), and cyberbullying (Xu et al., 2012; Hosseinmardi et al., 2015; Zhong et al., 2016). We believe that understanding the similarities and differences between terminologies can help shape the scope and goals of research. The terms and the typology delineating hate phenomenon in prior work are summarized below.

One clear attribute to identify hate speech is the presence of a *specific* target, i.e., attacking a specific individual, group or entity, or it targets a *generalized other* (Waseem, 2016; Fortuna and Nunes, 2018). Commonly cyberbullying and trolling are directed towards a specific individual and community, respectively (Waseem, 2016; Zampieri et al., 2019), while hate speech has been understood as specific (Zampieri et al., 2019; Sigurbergsson and Derczynski, 2020), or both directed and generalized (Burnap and Williams, 2015; Waseem and Hovy, 2016; Davidson et al., 2017). On the other hand, offensive language

is captured as a more general speech expressing profanity (e.g., non-acceptable language) or insult (Sigurbergsson and Derczynski, 2020). Indeed, hate speech can be classified as a subset of offensive language (Sigurbergsson and Derczynski, 2020). Other typical aspects separating hate speech from offensive language include if (i) violence or hate is incited; (ii) the speech aims at attacking or diminishing (Fortuna and Nunes, 2018).

We should also mention a relatively narrow but still important category of hate phenomenon, called dangerous speech. It refers to speech that "has a reasonable chance of catalyzing or amplifying violence by one group against another" (Benesch, 2014). Distinct from hate speech, dangerous speech is not considered as a crime. For this reason, companies are not obliged to regulate such speech on SMPs, making it a challenging task. In this regard, counter narratives are particularly encouraged and helpful to make audiences more resistant to dangerous speech (Benesch, 2014).

2.2 Standard Approach to Hate Moderation

Many international institutions and countries have taken actions to monitor, restrict, and counter hate speech. As early of 1966, the International Covenant on Civil and Political Rights monitored by United Nations states that "Any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law[1]." The convention on the Elimination of All Forms of Racial Discrimination, adopted by United Nations, condemns, prohibits, and outlaws discrimination in any form[2]. From the legal perspective, hate speech is regulated at least to some extent in European Union and other countries including Australia, Denmark, France, Germany, India, Italy, South Africa, Sweden, and New Zealand (Waldron, 2012). The C.O.N.T.A.C.T. project[3] supported by European Union aims to identify and react to hate crime and extreme speech in collaboration with experienced NGOs. Conversely, in the United States hate speech is generally not regulated by laws with the exception of actual threats or incitement to violence, as the freedom of speech is protected by the First Amendment (Ben-David and Fernández, 2016).

From a computational point of view, NLP can be helpful for identifying hate speech, monitoring media, and providing counter measures to create a safe and open Web. In the following sections, we first discuss the conventional approaches to mitigating online

[1] https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx
[2] https://www.ohchr.org/en/professionalinterest/pages/cerd.aspx
[3] https://reportinghate.eu/en/

hatred, then the datasets and techniques for hate detection.

2.2.1 Content Moderation of Hate Speech

On social media, the most common actions against online hatred is through content moderation that seeks to remove "problematic" user accounts or to screen "sensitive" user-generated content from posts, images, videos, or hashtags (Roberts, 2016; Gerrard, 2018). Often the "sensitive" content is flagged by moderators that can be employees, actual users of a platform or algorithms, following the hate speech policy implemented by individual social media companies. For instance, in supporting the European Commission, social media companies signed a Code of Conduct on combating online hate speech[4]. This Code mandate staff training and accommodates internal protocols for responding to hateful content within 24 hours. Companies committed to such convention include Facebook, YouTube, Twitter, Instagram, Snapchat, and many others.

Content moderation seems to diminish hate speech fast at scale. As an example, between January and March in 2021 Facebook removed 96.7% of problematic content flagged by its employees before users reported it on the site[5]. Even so, many have criticized this approach and its actual impact, calling for alternative approaches. One major critique concerns the violation of the free speech. If no suitable Code of Conduct is available, social media companies can act like gatekeepers of information access by controlling and influencing what can be kept on the Internet.

Next, Codes of Conduct on hate regulation are different based on the companies and regions they are developed. Moderation processes typically lack transparency, bringing the risk of selective speech and over-censorship. In addition, what comment can be acceptable remains under debate and is often bound by political pressure and financial decisions (Caplan, 2018). More importantly, taking hate speech down from SMPs does not mean undoing the harm and removing it from community. Instead, it limits the opportunity to counteract possible harm with non-negative, objective and informed arguments (i.e., counter narratives, counter measure, or counter speech).

Another concern relates to the quality and the transparency of techniques and human moderators. Social network platforms develop hate detection models using algorithms to identify content containing slurs and harmful speech. Nevertheless, automated content filters can have drawbacks. For instance, the performance of a model is highly correlated

[4] https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937

[5] https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/#data

to the data employed for training in terms of quantity and quality (e.g., annotations and biases). Furthermore, existing models do not take into consideration the usage of context, a key to determine if a speech is harmful, and it is challenging to incorporate contextual cues into models (Wilson and Land, 2020). As for human moderators, their decisions on content moderation tend to be subjective (Caplan, 2018). When facing large amounts of online hate speech, workers need to review a post in brevity and they may not read the complete thread of a post or discussion; hence, rushing into a decision without considering the circumstances. Moreover, repeatedly exposed to a negative and stressful environment reading hostile speech, moderators can experience mental health issues such as depression and anxiety.

2.2.2 Hate Speech Datasets

Manually scrutinizing all flagged content is very time consuming, and automating hate identification can facilitate hate regulation. Myriad datasets have been created to address the task of online hatred detection, reflecting the variation of hate phenomena, as well as motivations and annotation frameworks behind research. A few studies have attempted to survey systematically resources for training hate detection systems at scale, and have highlighted practices for improvement (Fortuna and Nunes, 2018; Vidgen and Derczynski, 2021; Poletto et al., 2021). In this section, we briefly review existing datasets on the basis of the following dimensions: language, hate target, data source, and granularity of taxonomies.

- **Language.** Unsurprisingly, English is the most represented language among the hate speech datasets released. As hate phenomenon becomes prominent globally, research community has started creating datasets for languages other than English, such as Arabic, French, Hindi, Italian, Danish, and Turkish. Several shared tasks have released their datasets for hate speech detection in European languages. For instance, there are the multilingual offensive language identification in social media at OffensEval under SemEval (Zampieri et al., 2020), the German abusive language identification on SMPs at Germeval (Bai et al., 2018), and the misogyny identification for Italian at EVALITA (Del Vigna et al., 2017; Fersini et al., 2018a) and for Spanish at IberEval (Ahluwalia et al., 2018; Shushkevich and Cardiff, 2018). Multilingual hate speech datasets are also available for Spanish and English at IberEval (Pamungkas et al., 2018) and HateEval (Basile et al., 2019), and for English, Bengali, and Hindi at TRAC-2020 (Kumar et al., 2020).

- **Hate target.** Identifying the hate targets is a straightforward way to characterize hate datasets. This refers to the attributes (e.g., gender, race, religion, sexual orientation, physical appearance) of hate content targets. The boundary between hate identities can be vague and prone to subjectivity based on the social characteristics of annotators (Vidgen and Derczynski, 2021; Poletto et al., 2021).

  Commonly, datasets publicly available cover a wide range of hate phenomena including anti-gypsism (Sanguinetti et al., 2018), homophobia (de Pelle and Moreira, 2017; Gomez et al., 2020), Islamophobia (Sanguinetti et al., 2018; Chung et al., 2019), misogyny (or sexism) (Waseem and Hovy, 2016; Waseem, 2016; Jha and Mamidi, 2017; de Pelle and Moreira, 2017; Bosco et al., 2018; Fersini et al., 2018b; Rezvan et al., 2018; Mathur et al., 2018; Gomez et al., 2020), politics-based attack (Bosco et al., 2018; Rezvan et al., 2018), racism (Waseem and Hovy, 2016; Waseem, 2016; de Pelle and Moreira, 2017; Bosco et al., 2018; Rezvan et al., 2018; Gomez et al., 2020), and religion-based attack (de Pelle and Moreira, 2017; Bosco et al., 2018; Gomez et al., 2020). Noticeably, in terms of the research quantity, Islamophobia have far been explored in comparison to misogyny and racism.

  In addition to directing at single class, datasets annotated with multiple types of abuse are also available. Ousidhoum et al. (2019) released 13k annotated tweets directed at many targets such as women, Asians, Muslims, refugees, and immigrants in general. For non-English dataset, as an example, Ross et al. (2016) 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.

- **Data source.** The sources of datasets refer to where the content is collected. Most datasets are retrieved from SMPs such as Facebook (Del Vigna et al., 2017; Bosco et al., 2018; Kumar et al., 2018; Mandl et al., 2019; Mossie and Wang, 2020), Instagram (Corazza et al., 2019), Twitter (Silva et al., 2016; Waseem and Hovy, 2016; Waseem, 2016; Ross et al., 2016; Bosco et al., 2018; Mandl et al., 2019), WhatsApp (Sprugnoli et al., 2018), Whisper (Silva et al., 2016), and Youtube (Hammer, 2016; Alakrot et al., 2018; Kumar Sharma et al., 2018). Other resources include Gab (Qian et al., 2019), Reddit (Olteanu et al., 2018; Qian et al., 2019), Stormfront (de Gibert et al., 2018), Wikipedia (Wulczyn et al., 2017), and World of Warcraft (Bretschneider and Peters, 2016).

  Twitter is the dominant source of hate speech, given the easily accessible API required for data retrieval and relatively liberal terms and conditions for research com-

munities (Cihon and Yasseri, 2016; Plaza-del-Arco et al., 2021). The unbalanced retrieval source, hindering the diversity in demographic and linguistic practices, can bring challenges for building robust systems (Vidgen and Derczynski, 2021). 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 Fernandez, 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.

- **Granularity of taxonomies.** Another factor to distinguish datasets is through what information it covers, i.e., a binary annotation (hate speech or not hate speech) (Reynolds et al., 2011; Rafiq et al., 2015; Hosseinmardi et al., 2015; de Gibert et al., 2018; ElSherief et al., 2018) or a multi-class annotation (hate speech, offensive, neither offensive nor hate speech) (Davidson et al., 2017). The number of classes for multi-class annotation varies depending on the reach of the research: the scale of abuse (e.g., overtly aggressive, covertly aggressive, not aggressive (Kumar et al., 2018)), or the type of targets (e.g., sexism, racism, neither (Waseem and Hovy, 2016)).

2.2.3 Approaches to Hate Speech Detection

Several works have investigated online hate speech detection and hate speech type classification. 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), Support Vector Machines (Karan and Šnajder, 2018; Fortuna et al., 2021) and knowledge bases (Dinakar et al., 2012), or to utilize deep neural networks such as Recurrent Neural Networks (Mehdad and Tetreault, 2016) and Long Short-Term Memory (LSTM) networks (Badjatiya et al., 2017; Arango et al., 2020). More recently, the advancements in pre-trained language models (see Section 2.3 for further details) have led to state-of-the-art results in many natural language tasks including hate speech detection. These language models, pre-trained on domain-general text and fine-tuned on target data, have superseded the employment of traditional machine learning models. For instance, the top performing systems of offensive language identification at OffensEval-2020 (SemEval-2020 Task 12), across five languages, utilize pre-trained models such as ALBERT (Lan et al., 2020), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020), and BERT (Devlin et al., 2019).

Concerning the levels of classification, prior work has contributed to offensive language identification (Wang et al., 2020), classification of offense types (Pham-Hong and Chokshi, 2020), and offense target identification (Pàmies et al., 2020). Other approaches have also 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).

2.3 Neural Language Models

In this section, we discuss neural language modeling and different architectures of its type, namely sequence-to-sequence models, encoder-decoder architecture, Transformer model, and pre-trained language models. A *language model* determines the probability of a sequence of words.

Given a set of tokens $W = \{w_1, ..., w_n\}$ in a text, a language model assigns probability $P(W)$ to the sequence W following the Equation 2.1:

$$P(W) = \prod_{i=1}^{n} p(w_i \mid w_1, ..., w_{i-1}) \tag{2.1}$$

This probability can be used to predict the next word based on the conditional distribution of previous tokens in the sequence. When employing a neural network for modeling such probabilities, it is called *neural language modeling*. Language modeling is essential for numerous NLP tasks such as machine translation, part-of-speech tagging and dialogue systems.

2.3.1 Sequence-to-Sequence Models

A sequence-to-sequence (seq2seq) model takes a sequence as input and transforms it into another sequence as output. This can be done by connecting two networks *encoder* and *decoder*, in which the former reads a variable-length input and learns its latent representation, whereas the latter constructs a variable-length output sequence based on the representation. Seq2seq models not only gained success for machine translation (Sutskever

et al., 2014), but also achieved strong performance on sequence prediction tasks such as conversational models (Vinyals and Le, 2015) and speech recognition (Prabhavalkar et al., 2017). We start our discussion on *encoder-decoder* architecture and then dive into *Transformer* architecture making use of *attention*.

2.3.2 Encoder-Decoder Architecture

Encoder-decoder models learn the mapping between an input sequence X and an output sequence Y. The length of both input and output can be different. Various neural networks can be utilized as encoder/decoder. Take counter narrative generation as an example (see Figure 2.1). Given a source *HS*: "*Islam*", "*is*", "*a*", "*disease*", ".", an encoder first encodes this sentence into a fixed-shape vector hidden state, then a decoder uses this vector representation to compute the output, e.g., "*This*", "*is*", "*an*", "*irrational*", "*comment*", ".". These models process a sequence on token-level one by one consecutively. Consequently, the order of a sequence can be represented using *time step* $t \in [1, \mathrm{T}]$, where T marks the length of the computation. An encoder-decoder model based on recurrent neural network (Elman, 1990) is formalized as:

$$
\begin{aligned}
\mathbf{h_t} &= f^{(e)}(\mathbf{w_t}, \mathbf{h_{t-1}}) && t = 1, ..., T_x \\
\mathbf{c} &= g(\mathbf{h_1}, ..., \mathbf{h_{T_x}}) \\
\mathbf{z_1} &= \mathbf{c} \\
\mathbf{z_t} &= f^{(d)}(\mathbf{y_{t-1}}, \mathbf{c}, \mathbf{z_{t-1}}) && t = 1, ..., T_y \\
\mathbf{y_t} &= \phi(\mathbf{z_t})
\end{aligned}
\tag{2.2}
$$

At each time step t, the encoder f^e computes the latent representation (i.e., embedding) of current token $\mathbf{w_t}$ and the hidden state of previous step as the current hidden step $\mathbf{h_t}$. Once all tokens are fed to the encoder, the hidden state $\mathbf{c}$ representing the entire source input is generated as context. Conditioned on the context variable $\mathbf{c}$, the decoder $f^{(d)}$ transforms the previous hidden state $\mathbf{z_{t-1}}$ and the previously generated token $\mathbf{y_{t-1}}$ into $\mathbf{z_t}$. For inference at time $t = 1$, only the context variable $\mathbf{c}$ is fed to the decoder. To predict the next word occurrence y_t at time step t, an activation function ϕ (e.g., softmax) converts the hidden state z_t into conditional probability distribution over all possible target tokens. The approach of modeling the state z_t based on the previously generated token y_{t-1} is called **auto-regressive generation** (Graves, 2013).

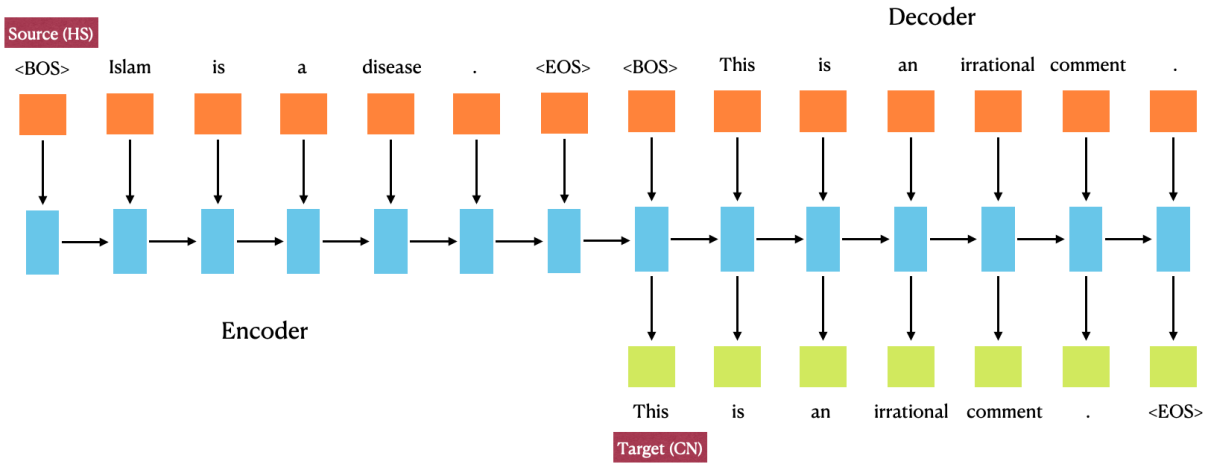

Figure 2.1: A simplified encoder-decoder model reads an HS source sentence and produces a CN as the target sentence.

2.3.3 Transformer Model

Transformer architecture has shown to outperform recurrent language models across numerous tasks (Vaswani et al., 2017). It is also the backbone of many pre-trained language models that will be covered in Section 2.3.4. As discussed in section 2.3.2, recurrence-based encoder-decoder architectures encode the input sentence sequentially and the hidden representations are dependent on the history of words read previously in the computation. This poses a challenge to parallel computing, particularly when dealing with long-range dependencies. In this context, Transformer is introduced to mitigate such circumstances by replacing the recurrent layers with self-attention layers for input and output representation (Vaswani et al., 2017). The success of the Transformer model comes from *positional encoding* and *self-attention*.

The architecture of the Transformer is illustrated in Figure 2.2. As depicted in the Figure, the Transformer comprises an encoder and a decoder, in which each of them is a stack of multiple identical layers based on self-attention. Each encoder layer consists of a multi-head self-attention mechanism followed by a fully-connected feed-forward network. Similarly, a decoder layer consists of the sub-layers in an encoder layer with an additional masked multi-head self-attention layer to ensure that at inference time the decoder only attends to outputs (i.e., positions) that have been generated. Both of the input and output sequences are embedded using positional encoding.

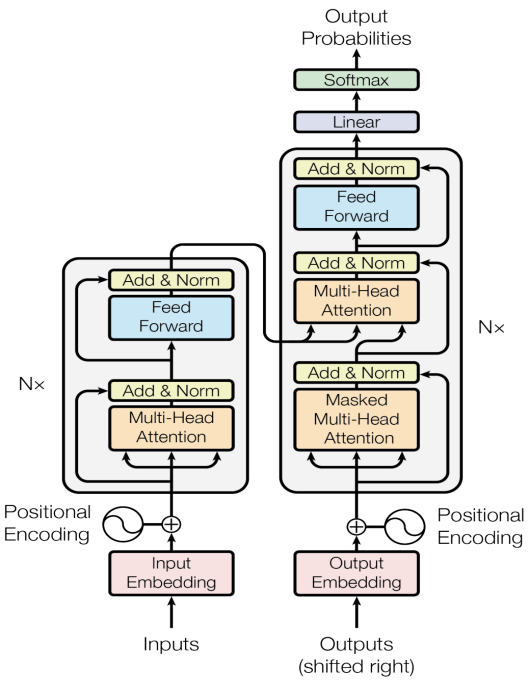

Figure 2.2: The architecture of the Transformer (figure source: Vaswani et al. (2017)).

2.3.3.1 Self-Attention

Attention mechanism is introduced in machine translation to encode the long sequence efficiently. Originally it is implemented in the decoder, signaling models which position to focus on in the input sequence without the need of compressing all information into a fixed-length vector (Bahdanau et al., 2015). Taking a step further, self-attention reflects the relations between different words (positions) in the same sequence. Self-attention determines the amount of focus to allocate on other positions of the same sequence when encoding a token at a specific position.

In Transformer, each word has a *Query* vector, a *Key* vector, and a *Value* vector that are used to compute *Scaled Dot-Product Attention*, where d_k is the dimension of queries and keys, and d_v the dimension of values. The attention function calculates a score as a scaled dot product of the query Q and the corresponding key K. Then, it adds to this score a scaling factor $\frac{1}{\sqrt{d_k}}$ followed by a softmax to produce weights on the values V. Formally, the attention operation is defined as,

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.3}$$

Multi-Head Self-Attention. In comparison with performing a single self-attention layer, the implementation of multi-head attention can increase the computational efficiency and provide multiple representation subspaces. The multi-head attention comprises h different self-attention layers in parallel, each one is called a *head* H that can learn independent linear projection with its query, key, and value using attention pooling. The outputs of each head are further concatenated and transformed via another linear projection as the final attention value. Formally, multi-head attention is defined as:

$$\begin{aligned} Multihead(Q, K, V) &= Concat(H_1, H_2, \ldots, H_i)W^O \qquad i = 1, \ldots, h \\ H_i &= Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2.4}$$

where W_i^Q, W_i^K, W_i^V denotes the learned parameter matrices for each attention head. The advantages of using multi-head attention are two-fold. Firstly, it allows for highly-efficient implementation through simultaneous computation of attention pooling and the dimension minimization of each head. Thus, computing h attention heads consumes a similar cost to performing single-head attention. Secondly, it offers multiple representation sub-spaces that project distinct behaviors derived from the same attention mechanism with different sets of query, key, and value weight metrics learned. Multi heads can increase the attention span to wider range of sequences. This behavior is beneficial, for instance, when dealing with long-distance and short-distance dependencies. More attention heads are considered as advantageous for pushing the boundaries of model performance. The original Transformer implementation contains 8 attention heads.

2.3.3.2 Positional Embeddings

Since the Transformer is based on attention mechanism in which each token in the sequence is processed independently, the sequential or temporal aspect of language is not reflected while it supports parallel computing. To fill this gap, positional embeddings are concatenated with the input encoding vectors to capture the absolute or relative position of words in a sequence. The positional encodings are calculated applying sine and cosine functions as follows:

$$PE_{(pos,2i)} = sin(\frac{pos}{10000^{2i/d}})$$
$$PE_{(pos,2i+1)} = cos(\frac{pos}{10000^{2i/d}}) \tag{2.5}$$

where *pos* and i are the position in the sequence and the position along the embedding vector dimension, respectively, and d is the dimension of the model.

2.3.4 Pre-Trained Language Models

Prior to pre-trained language models, most NLP systems rely on task-specific architectures trained on limited labeled instances to learn both the language and the task knowledge of interests. These task-specific architectures require domain expertise while often failing at generalizing to unseen tasks. In contrast, pre-trained language models, generally based on Transformer architecture, such as ALBERT (Lan et al., 2020), BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and T5 (Raffel et al., 2020) have dominated many NLP tasks like text classification (Piskorski et al., 2020) and natural language understanding (Liu et al., 2019). Such models benefit from self-supervised learning and transfer learning (Kalyan et al., 2021). The typical setup of pre-trained models is designed as follows. A model is firstly pre-trained on vast volumes of unlabeled text data using a desired language modeling loss to learn language representation. Next, the pre-trained model is fine-tuned on target-domain labelled data using cross-entropy loss to transfer the knowledge learned to downstream tasks. Unlike prior work (e.g., models relied on word embeddings), pre-trained language models require less feature engineering and annotated data while demonstrating better generalization ability (Kalyan et al., 2021). Whereas pre-training such models requires high computation power and sophisticated model architectures, once trained, they can be applied easily to downstream tasks through fine-tuning. Fine-tuning is comparatively inexpensive when contrasted with pre-training; hence, making it a popular option for NLP from a practical perspective (Ruder, 2021). Below we overview the pre-trained models that are exploited in this thesis, i.e., BERT, ALBERT, and GPT-2.

- **BERT.** BERT model is developed based on a series of Transformer encoder layers (Transformer Blocks). Each layer consists of two sub-layers: multi-head self-attention and a position-wise fully connected feed-forward network. The BERT

model takes as input a token sequence flowing through the Transformer encoder layer. A special token *[CLS]* is prepended to the beginning of every sequence, representing the sentence-level understanding for classification tasks. Sentences are differentiated by (1) adding a separation token *[SEP]* between sentences, and (2) learning an embedding for every token marking which sentence it belongs to. The authors provided two model sizes:

  - BERT_{base} with 108M parameters, 12 layers of Transformer blocks, a hidden size of 768, and 12 self-attention heads.
  - BERT_{large} with 334M parameters, 24 layers of Transformer blocks, a hidden size of 1024, and 16 self-attention heads.

  BERT is pre-trained on BooksCorpus (Zhu et al., 2015) and English Wikipedia utilizing two pre-training tasks, Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). MLM learns to predict the masked word by randomly masking 15% of the input tokens with token *[MASK]* and then predicting the masked word. Unlike unidirectional language models (e.g., GPT (Radford et al., 2018)) that attend only to previous tokens (typically in a left-to-right order), MLM aims at learning bidirectional representations leveraging both left and right context, which is beneficial for question answering (Devlin et al., 2019). In addition to randomly masking input tokens, the authors also attempt to replace a chosen word with another random token for improving model performance. NSP aims at capturing the relationships between sentences by predicting if two sentences appear consecutively. BERT is commonly used for sentiment classification, named entity recognition, and question answering tasks.

- **ALBERT.** ALBERT (aka A Lite BERT) is an improved version of BERT that has fewer parameters but achieves better results, notably on GLUE (Wang et al., 2018) and SQuAD Rajpurkar et al. (2016) benchmarks. Given the huge structure of BERT, scaling it becomes costly and demands longer training time. In order to increase the training efficiency, the authors propose three modifications: factorized embedding parameterization to reduce the size of embedding space, cross-layer parameter sharing to stabilize the network, and inter-sentence coherence loss to address discourse-level consistency (Lan et al., 2020). In terms of model size, ALBERT has much fewer parameters than BERT. As an example, the biggest model ALBERT-xxlarge has 70% of BERT_{large}'s parameters, with 235M parameters, 12 layers of Transformer blocks, and a hidden size of 4096.

- **GPT-2.** Generative Pre-trained Transformer (GPT) models have overwhelmingly taken the NLP research by achieving good results on various down-stream tasks, strikingly in text generation, even under zero-shot setting without explicit supervised adaptation. They allow for learning a generative language model using task conditioning and zero shot transfer. Such models require few training samples to learn task representations and achieve performance comparable to or better than models trained in a supervised manner.

  After the first GPT (GPT-1) was introduced (Radford et al., 2018), two more models were developed to improve model performance by enlarging GPT-1 architecture, and by training with massive datasets and more sophisticated learning frameworks such as multitask and meta-learning settings (Radford et al., 2019; Brown et al., 2020). Here we focus on the pre-trained model used in this thesis, i.e., GPT-2 (Radford et al., 2019).

  GPT-2 is based on decoder-only Transformer blocks and comes with 4 model sizes. The core of GPT-2 lies in an unsupervised pre-training phase, followed by a supervised fine-tuning stage. The unsupervised pre-training phase is to learn a high-capacity language model using a large corpus, and the supervised fine-tuning stage adapts the model learned to a target task. The objective function for learning to perform a task (e.g., language modeling) can be framed as approximating the conditional distribution $p(\textit{output} \mid \textit{input})$. Instead of learning a new model for every single task, conditioning the probability of output on the desired task along with input enables multitask learning with the same dataset. This task conditioning has moved GPT-2 towards a general system excelling at a wide range of tasks under zero-shot scenarios. For pre-training a general system that can handle various tasks, a language modeling loss applied to maximize the likelihood is shown below:

$$L_1(W) = \sum_i logP(w_i \mid w_{i-k}, ..., w_{i-1}, T; \Theta) \tag{2.6}$$

  where w_i is a token in a dataset W, k is the size of context window, T is task conditioning, and Θ is the learnable parameter for training the conditional probability P via stochastic gradient descent. Task conditioning can be natural language instructions or training examples showing the expected outputs. For instance, the training inputs for machine translation task can be structured as *"translate from english to french,* $\langle$*english text*$\rangle$*,* $\langle$*french text*$\rangle$*"*. Similar training strategies can be

found in Raffel et al. (2020) where various NLP tasks are treated as a text-to-text problem.

2.4 Natural Language Generation

Having laid the groundwork of neural language models, we discuss how it is applied to natural language generation research, especially focusing on three areas that can aid counter narrative generation, i.e., counter argument generation, knowledge-grounded generation, and NLG for under-resourced languages.

2.4.1 Counter Argument Generation

Counter argument generation to some extent share the same goal as counter narrative generation - to produce the opposite or alternate stance of a statement - and hence we review its relevant work. A limited number of studies have investigated argument generation due to the challenges in text generation. These studies usually focus on specific domains for the large volume and coverage of discussion, 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). Most of existing work on argument generation deals with rule-based systems and information retrieval systems. Rule-based systems use operators that follow the guidelines from argumentation theory (Reed et al., 1996; Carenini and Moore, 2000). The goal of the retrieval-based approaches is to select the most relevant arguments (Sato et al., 2015) given a topic and the context of a conversation. While retrieval approaches can produce semantically related and controllable responses to a user's message, they cannot generate new content.

Owing to the recent progress on machine learning, several studies have investigated using encoder-decoder architectures to generate arguments (Rakshit et al., 2019; Hua et al., 2019; Rach et al., 2018; Le et al., 2018). The closest work to CN generation is by Hua et al. (2019), where an argument retrieval model and an argument generation model are combined for general counter argument generation incorporating knowledge. In their study, a search engine is built based on Wikipedia and news articles. Given an input statement, an argument retrieval model selects a set of key concepts as talking points from the search engine. A text planning decoder then takes the talking points as inputs to produce sentence representations which will be used to generate counter arguments via

a content realization decoder.

Despite of the similarity to counter argument generation, counter narrative generation faces peculiar difficulties such as the absence in HS of explicit or well-structured 'arguments' (e.g., HS "*Islam is a disease*" vs. an argument "*Renewable energy is a better option for replacing fossil fuels than unsustainable and expensive energy like nuclear power.*") and the limited amount of data available for training.

2.4.2 Knowledge-Grounded Generation

Effective counter narratives often require fact-bound arguments. Towards this goal, we seek to employ knowledge to enrich counter narratives with facts and related knowledge. There is growing interest in exploiting external knowledge to generate more informative and versatile responses for various applications such as dialog systems (Hua et al., 2019; Ghazvininejad et al., 2018; He et al., 2017a; Moon et al., 2019; Parthasarathi and Pineau, 2018; Young et al., 2018), question answering (Das et al., 2017; Miller et al., 2016; Saha et al., 2019), and sentiment analysis (Poria et al., 2015).

Regarding the knowledge sources, previous knowledge-grounded conversation models have injected guidance through topic information (Fan et al., 2019; Hua et al., 2019; Wang et al., 2019a), structured knowledge graphs (Zhou et al., 2018; Zhu et al., 2017), unstructured texts (Dinan et al., 2019; Ghazvininejad et al., 2018; Hua et al., 2019) and incorporation of the last two (Parthasarathi and Pineau, 2018) using attention mechanism (Moon et al., 2019; Zhou et al., 2018), copy mechanism (He et al., 2017b) or memory networks (Sukhbaatar et al., 2015) to attend to external knowledge.

In Wang et al. (2019a), a topic-guided generation model is proposed by jointly training a topic model and a sequence model, where the former learns to capture semantic meanings across documents, and the latter generates a response conditioned on the topic assigned. The work of Yao et al. (2019) proposes a plan-and-write hierarchical framework for story generation given a generated title, enabling dynamic and static schemas. Specifically, a story planning module generates plots (i.e. a sequence of words) and a surface realization produces a story constrained on the plots. Ghazvininejad et al. (2018) use an encoder-decoder architecture, together with a modified memory network modeling external facts from Foursquare, to handle social conversations. Similarly, Parthasarathi and Pineau (2018) incorporate Wikipedia summaries and NELL knowledge base to enhance next utterance prediction.

2.4.3 NLG for Under-Resourced Languages

Hatred online has increased worldwide and can be found in all languages. Therefore, it is worthwhile to investigate counter narrative generation in languages other than English. Still, there has been little research on approaches for generating text in multiple languages or 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).

Rashid et al. (2019) introduce Bilingual-GAN that can generate bilingual text concurrently by learning a shared latent space between two languages and then producing a parallel text from the space. While their system can generate fluent parallel texts in a supervised scenario, and even some parallel texts in unsupervised models without parallel corpora, such approach requires large dataset that is not feasible for low-resource scenarios including counter narrative generation.

Recently, multilingual pre-trained language models are started to be investigated, aiming to scale up generalization across language and cross-lingual representation on several NLP tasks, particularly for low-resource languages. Examples of pre-trained multilingual language models, generally covering a wide range of language families in a single model, are XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), multilingual BERT (Devlin et al., 2019), mBART (Liu et al., 2020), and XNLG (Chi et al., 2020). Monolingual pre-trained language models specifically trained for languages other than English have also been established. For instance, there are Bert-based LMs for French (Le et al., 2020), Spanish (Canete et al., 2020) and Dutch (Delobelle et al., 2020), and GPT-2 for Italian (De Mattei et al., 2020).

2.5 Evaluation Metrics

In this section, we review the commonly adopted metrics such as BLEU, ROUGE, novelty, and Repetition Rate used to assess NLG model performance.

- **BLEU.** BLEU (Papineni et al., 2002) assesses the generation quality based on word overlap. It computes a modified n-gram precision and then applies a brevity penalty to penalize short generations. The precision measures the similarity between the reference and the generation through unique n-gram matches. BLEU is a standard

automatic metric for machine translation, but is also employed for NLG even it is criticised (Marie et al., 2021).

- **ROUGE.** Similar to BLEU, ROUGE (Lin, 2004) computes the word-overlapping between the ground-truth and the generated outputs. The metric is commonly used for summarization task and comes with different options based on the calculation of the overlap: n-gram, word sequences, and word pairs. For instance, ROUGE-L measures the longest common subsequence overlapping, ROUGE-N assesses the number of matched n-grams, and ROUGE-S considers skip-bigram co-occurrence statistics.

- **Novelty.** As shown in Wang and Wan (2018), the novelty of a generation regarding the training data can be captured with Jaccard similarity function, which measures dissimilarity between two sample sets. Jaccard similarity is formalized as the size of the intersection divided by the size of the union of two sets (Jaccard, 1912). We include this measurement since we want to generate new outputs instead of copying sentences as in information retrieval based systems.

- **Repetition rate.** The Repetition Rate (RR) evaluates the repetitiveness of a collection of texts, by considering the rate of non-singleton n-gram types it contains (Bertoldi et al., 2013; Cettolo et al., 2014). It indicates the capability of a system to produce responses with high lexical diversity. The higher the score, the more repeated the text.

Chapter 3

Strategies for Hate Countering Dataset Collection

3.1 Introduction

As established in Chapter 1, many NGOs specialized in hate countering are already training operators to compose counter narratives. Since manual intervention against hate speech is not scalable, automatizing the countering procedure would increase the efficacy and effectiveness of hate countering (Munger, 2017). Therefore, data-driven NLG approaches are beginning to be investigated to assist NGO operators in writing CNs. At the time of this work, the research community had solely begun to understand counter narratives on SMPs, and no high-quality data was available for automating counter narrative generation. The lack of high-quality data hinders the development of effective generation systems to hate countering. As a necessary first step, we propose diverse CN collection strategies to fill the gap.

In this chapter, we first briefly review the related work on hate countering via counter narratives and its effectiveness (Section 3.2). Then we present our collection of the first multilingual hate countering dataset nichesourced to experts (Section 3.4). Based on such dataset, we further experiment with various approaches, leveraging language models, automatic classifiers, experts and non-experts for data generation (Section 3.5).

3.2 Related Work

Several CN methods to counter hatred are delineated and tested on social media platforms including Twitter (Munger, 2017; Wright et al., 2017), Facebook (Schieb and Preuss, 2016), and Youtube (Ernst et al., 2017; Mathew et al., 2019). Previous studies on hate countering cover several aspects of CNs; for instance: defining counter narratives (Benesch, 2014; Benesch et al., 2016a), studying their effectiveness (Schieb and Preuss, 2016; Silverman et al., 2016; Ernst et al., 2017; Munger, 2017; Wright et al., 2017), and linguistically characterizing online counter narrative accounts (Mathew et al., 2018).

Mathew et al. (2019) 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, focusing on the idea of collecting hate message/counter narrative pairs from Twitter (Mathew et al., 2018). 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 than 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 developing effective generation systems, especially because the natural 'countering' data that can be found on SMP – such as the ones containing abusive language as example CN3 in Chapter 1 – often does not meet the required standards, and those from crowdsourcing are not of good quality.

3.3 Strategies of Counter Narrative Collection

We first review two prototypical strategies (crawling and crowdsourcing) that have been presented recently, and then discuss our two newly proposed approaches (nichesourcing

and author-reviewer) to collect HS-CN pairs.

Crawling (CRAWL). Mathew et al. (2018) focus on the intuition that counter narratives can be found on SMPs as responses to hateful expressions. Their approach is a mix of automatic HS collection via linguistic patterns, and a manual annotation of replies to check if they are responses that counter the original hate content. Thus, all the material collected is made of *natural/real* occurrences of HS-CN pairs. Similarly, Mathew et al. (2019) collected counter narratives on SMP. However, the majority type of counter narrative collected uses hostile language, which is explicitly discouraged by NGOs working on hatred intervention.

Crowdsourcing (CROWD). Qian et al. (2019) propose that once a list of HS is collected from SMPs and manually annotated, we can briefly instruct crowd-workers (non-expert) to write possible responses to such hate content. In this case, the content is obtained in controlled settings as opposed to crawling approaches.

Nichesourcing (NICHE). We propose outsourcing and collecting CNs under controlled settings (introduced in Section 3.4). In this nichesourcing, the CNs are written by NGO operators, i.e., persons specifically trained to constrain online hatred via textual responses that can be considered as experts in CN production. While such approach produces expert-based CNs, automation strategies are desirable and scalable.

Author-Reviewer (HYBRID). In order to address the above shortcomings, we employ an author-reviewer architecture to collect CNs. In particular, a hybrid approach is adopted: an author is tasked with CN generation and a reviewer can be a human or a trained classifier that filters the produced text. Such approach can be easily and flexibly extended, for instance, with a validation/post-editing phase performed by professionals over the filtered text, allowing for quality data. This work is introduced in Section 3.5.

3.4 Creation of Hate Countering Dataset through Nichesourcing

This section presents **CONAN**, a dataset[1] of COunter NArratives through Nichesourcing 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). Considering that composing counter narratives requires expert knowledge, nichesourcing methodology can fulfill our expectations on hate countering datasets that provide a tailored and non-negative counter narrative to a given hate speech. For such purpose, we organized several data collection sessions with NGO operators, who are trained experts, specialized in writing counter narratives that are meant to curb hatred and de-escalate the conversation. In this way we build the first large-scale, multilingual, publicly available, and 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 to perform additional tasks that did not require specific domain expertise (in total 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, and 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, that we will discuss in the following chapters.

3.4.1 Fundamentals of CONAN

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.

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

[1]CONAN can be downloaded at the following link: `https://github.com/marcoguerini/CONAN`

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 Fernandez (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. (2018). 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; Dell'Orletta 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., 2018). In this work, we collect demographic information of participants, i.e., gender, age, and education level, to provide data for counter narrative personalization.

3.4.2 Data Collection Methods

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. A screenshot of the form used is displayed in Figure 3.1.

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 3.1. Furthermore, for our specific data collection task, operators were asked to follow their intuitions without over-thinking and to compose reasonable responses (e.g., *Please try to write 5 responses to the hate message provided. Write the responses following your intuition without thinking too much: they do not need to be perfect, but 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 3.2.

[2] `https://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.

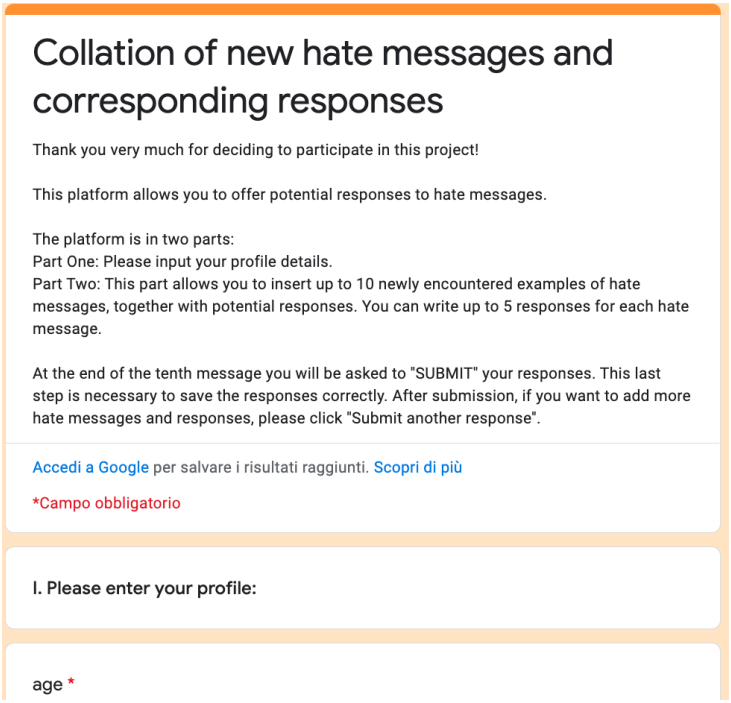

Figure 3.1: Screenshot of data collection form.

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 3.1: 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 3.2: Example pairs for the three languages, along with English translations.

3.4.3 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 sub-topics and counter narrative types, and (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., 2018). Other examples of data augmentation strategies are back-translation (Sennrich et al., 2016)

HS: According to a recent OFSTED report, a school in Birmingham is still segregating girls and boys, despite a High Court ruling in 2017 that this is unlawful.
Paraphrase 1: In Birmingham there is a school where girls and boys are separated even if it isn't legal.
Paraphrase 2: A school in Birmingham is still separating girls from boys. This information come from a OFSTED report. A High court said in 2017 that this type of action is illegal.

Table 3.3: Example paraphrases of a hate speech.

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 with small effect. Two paraphrase examples of a hate speech are presented in Table 3.3.

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., 2016b; Mathew et al., 2019), 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 follow-

ing 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 (discussed in Chapter 5), 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-shot 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., 2019).

3.4.4 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. On average 9, 34, and 17 responses per hate speech are collected for English, French, and Italian, respectively. 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 15,000. 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 3.4.

	English	French	Italian
original pairs	1288	1719	1071
augmented pairs	2576	3438	2142
translated 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+augmented vocabulary	1631	333	790
CN vocabulary	3556	4018	3728
HS words	2950	434	751
HS+augmented 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+augmented_words$_{\mu}$	18.72	5.31	14.16
HS+augmented_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 3.4: Main statistics of the dataset.

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 3.5, 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.

Considering the annotation tasks, we give the distribution of counter narrative types per language in Table 3.6. As can be seen in the table, there is a consistency across the languages such that Facts, Question, Denouncing, and hypocrisy are the most

	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 3.5: 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 3.6: Counter narrative type distribution over the three languages (% over the total number of labels).

frequent counter narrative types. Before the reconciliation phase, the agreement between

the annotators was moderate: Cohen's Kappa[4] was 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 3.7, 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, History, Generic, and Terrorism are the most frequent sub-topics.

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

Table 3.7: Hate speech sub-topic type distribution over the three languages (% over the total number of labels).

3.4.5 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

[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).

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., 2018). 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 effective 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 notably 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.

3.5 Creation of Hate Countering Dataset through Author-Reviewer Architecture

In the previous section we presented the dataset creation nichesourced to professionals. Despite obtaining diverse counter narratives, our analysis shows that on average an expert annotator spends 8 minutes to compose a counter narrative. In this respect, we can conclude that composing counter narratives for data collection is slow and cannot address effectively online hate speech that is colossal and varied, even when employing experts. For this reason, we ask, how can we obtain high-quality counter narratives while reducing manual efforts from experts.

To answer this question, we first diagnose data collection strategies depending on the two main requirements that CN datasets must meet: (i) data quantity and (ii) data quality. Finding the right trade-off between the two is in fact a key element for an effective automatic CN generation. To our understanding none of the collection strategies presented so far is able to fulfill this requirement. Thus, we test several *hybrid* strategies to collect data, by combining niche-sourcing, crowd-sourcing, and synthetic data generation obtained by fine-tuning deep neural architectures specifically developed for NLG tasks, such as GPT-2 (Radford et al., 2019). We propose using an author-reviewer framework in which an author is tasked with text generation and a reviewer can be a human or a classifier model that filters the produced output. Finally, a validation/post-editing phase is conducted with NGO operators over the filtered data. Our findings show that this framework is scalable allowing to obtain datasets that are suitable in terms of diversity, novelty, and quantity.

3.5.1 Characteristics of the Datasets

Regardless of the HS-CN collection strategy, datasets must meet two criteria: *quality* and *quantity*. While *quantity* has a straightforward interpretation, we propose that data *quality* should be decomposed into *conformity* (to NGOs guidelines) and *diversity* (lexical & semantic). Additionally, HS-CN datasets should not be ephemeral, which is a structural problem with crawled data since, due to copyright limitations, datasets are usually distributed as a list of tweet IDs (Klubička and Fernandez, 2018). By generating the data through crowdsourcing or nichesourcing, the problem is avoided.

Quantity. While the CRAWL dataset is very small and ephemeral, representing more

a proof of concept than an actual dataset, the CROWD dataset involved more than 900 workers to produce $\approx 41K$ CNs. On the other hand, the NICHE dataset is constructed by the participation of ≈ 100 expert-operators to obtain $\approx 4K$ pairs (in three languages) and resorted to HS paraphrasing and pair translation to obtain the final $\approx 14K$ HS-CN pairs. Evidently, employing non-experts, e.g, crowdworkers or annotators, is preferable in terms of data quantity.

Quality. In terms of quality, we consider that diversity is of paramount importance, since verbatim repetition of arguments can become detrimental for operator credibility and for the CN intervention itself. Following Li et al. (2016), we distinguish between (i) *lexical diversity* and (ii) *semantic diversity*. While lexical diversity focuses on the diversity in surface realization of CNs and can be captured by word overlapping metrics, semantic diversity focuses on the meaning and is harder to be captured, as in the case of CNs with similar meaning but different wordings (e.g.,"*Any source?*" vs. "*Do you have a link?*").

(i) Semantic Diversity & Conformity. To model semantic diversity and conformity, we focus on the CN 'argument' types that are present in various datasets. Argument types are useful in assessing content richness (Hua et al., 2019). In a preliminary analysis, CROWD CNs are observed to be simpler and mainly focus on 'denouncing' the use of profanity while NICHE CNs are found richer with a higher variety of arguments. On the other hand, CRAWL CNs can cover diverse arguments to a certain extent while being highly prone to contain profanities. To perform a quantitative comparison, we randomly sampled 100 pairs from each dataset and annotated them according to the CN types presented by Benesch et al. (2016b), which is the most comprehensive CN schema to our knowledge.

	CRAWL	CROWD	NICHE
Hostile	50	0	0
Denouncing	16	76	10
Den.+Oth.	0	10	9
Other	34	14	81
RR	3.16	4.83	2.72

Table 3.8: Diversity analysis of the three datasets. Semantic diversity is reported in terms of CN type percentages, Lexical diversity in terms of Repetition Rate (RR - average over 5 shuffles).

The results are reported in Table 3.8. For the sake of conciseness we focus on the *hostile*, *denouncing*, and *consequences* classes, giving *other* to all remaining types (including the *fact* class). Clearly, CRAWL does not meet the conformity standards of CNs considering the vast amount of *hostile* responses (50%), still granting a certain amount of type variety (*other:* 34%). Contrarily, CROWD conforms to the CN standards (*hostile:* 0%), yet mostly focuses on pure *denouncing* (76%) or denouncing with simple arguments (10%). The class *other* (14%) consists of almost only simple arguments, such as "*All religions deserve tolerance*". In NICHE instead, arguments are generally and expectedly more complex and articulated, and represent the vast majority of cases (81%). Few examples of CN types are given in Table 3.9.

Hostile	"Hell is where u belong! Stupid f***t... go hang yourself!!"
Denouncing	"The N word is unacceptable. Please refrain from future use."
Fact	"The majority of sexual assaults are committed by a family member, friend, or partner of the victim, and only 12% of convicted rapists are Muslim. It is not the religion, its the individuals, whether they're Muslim or not."

Table 3.9: Some examples of the categories relevant to our analysis. Hostile from CRAWL dataset, Denouncing from CROWD, Fact (other) from NICHE.

(ii) Lexical Diversity. The Repetition Rate (RR) is used to measure the repetitiveness of a collection of texts (Cettolo et al., 2014; Bertoldi et al., 2013). We utilize RR instead of the simple count of distinct ngrams (Xu et al., 2018; Li et al., 2016) or the standard type/token ratio (Richards, 1987) since it allows us to compare corpora of diverse sizes by averaging the statistics collected on a sliding window of 1000 words. Since CROWD and NICHE contain repeated CNs for different HSs[5], we first removed repeated CNs and then applied a shuffling procedure to avoid that CNs that are answering to the same HS (so more likely to contain repetitions) appear close together. Results in Table 3.8 show that NICHE is the dataset with more lexical diversity (lower RR), followed by CRAWL and CROWD.

Discussion. We can reasonably conclude that: (i) crawling, as presented in (Mathew et al., 2018, 2019), is not a mature procedure yet for CN collection, even if it is promising, (ii) nichesourcing is the one producing the best and most diverse material by far, however it

[5]While this is an explicit data augmentation choice in NICHE, for CROWD it seems to derive from writing the same CNs for similar HSs by crowd-workers.

is also the most challenging to implement considering the difficulty of making agreements with NGOs specialized in CN creation and it does not provide sufficient amount of data. (iii) On the contrary, CROWD seems to be the only one that can grant the amount of data that is needed for deep learning approaches, but contains more simple and stereotyped arguments. A summary of the pros and cons of each collection approach is presented in Table 3.10.

	Quantity	Quality		Non-ephemerality
		Conformity	Diversity	
Crawl.	✓	-	✓	-
Crowd.	✓	✓	-	✓
Niche.	-	✓	✓	✓

Table 3.10: Comparison of different approaches proposed in the literature according to the main characteristics required for the CN dataset.

3.5.2 CN Generation through Author-Reviewer Architecture

Since none of the aforementioned approaches alone can be decisive for creating proper CN datasets, we propose a novel framework that combines crowdsourcing and nichesourcing to obtain new quality data while reducing collection cost/effort. The key elements of this combination are: (i) there must be an external element in the framework that produces HS-CN candidates, (ii) non-experts should pre-filter the material to be presented/validated by experts. Thus, we settle on the author-reviewer modular architecture (Oberlander and Brew, 2000; Manurung et al., 2008). In this architecture the author has the task of generating a text that conveys the correct propositional content (a CN), whereas the reviewer must ensure that the author's output satisfies certain properties. The reviewer finally evaluates the text viability and picks the ones to present to the NGO operators for final validation/post-editing.

The author-reviewer architecture that we propose differs from the previous studies in two respects: (i) it is used for data collection rather than for NLG, (ii) we modified the original configuration by adding a human "reviewer" and a final post-editing step.

We first tested four different author configurations, then three reviewer configurations keeping the best author configuration constant. A representation of the architecture is shown in Figure 3.2.

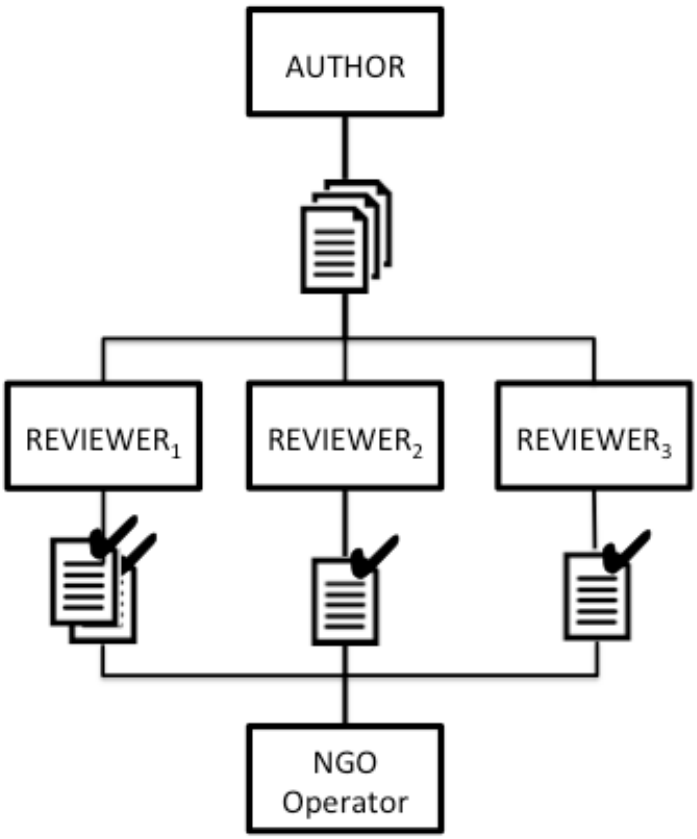

Figure 3.2: The author-reviewer configuration. The author module produces HS-CN candidates and the reviewer(s) filter them. Finally, an NGO operator validates and eventually post-edits the filtered candidates.

3.5.3 The Author

In order to obtain competent models that can provide automatic counter narrative hints and suggestions to NGO operators, we have to overcome the data bottleneck/limitations, i.e., either the limited amount of training data in NICHE or its repetitiveness in CROWD, especially for using neural NLP approaches. Since pre-trained language models have achieved promising results when fine-tuned on challenging generation tasks such as chit-chat dialog (Wolf et al., 2018; Golovanov et al., 2019), we propose using a recent large-scale language model GPT-2 (Radford et al., 2019).

We used the GPT-2 medium model, which contains 345 million parameters, with 24 layers, 16 attention heads, and hidden state size of 1024. We fine-tuned two models with GPT-2, one on NICHE and one on CROWD datasets for counter narrative generation.

NICHE - Training and test data. We have split the dataset into 5366 pairs of HS-CN for training and the rest (1288 pairs) for testing. In particular, the original HS-CN pairs, one HS paraphrase, and the pairs translated from FR and IT were kept for training while the other HS paraphrases were used for testing.

CROWD - Training and test data. Although the CROWD dataset was created for dialogue level HS-CN, we could extract HS-CN pairs by selecting the dialogues in which only 1 utterance was labeled as HS. Therefore, we could guarantee that the crowd-produced CNs are exactly for the labeled utterance. We then applied a 80/20 training and test split, obtaining 26320 and 6337 pairs, respectively.

3.5.3.1 Generation Models

We fine-tuned GPT-2[6], with a batch size of 1024 tokens and a learning rate of 2e-5. The training pairs are represented as:

$$[HS_start_token]\ HS\ [HS_end_token]\ [CN_start_token]\ CN\ [CN_end_token].$$

While we empirically selected model checkpoint at the 3600^{th} step of fine-tuning with NICHE dataset, with CROWD dataset we selected the checkpoint at the 5000^{th} step[7]. After fine-tuning the models the generation of CNs for the test HSs has been performed using Nucleus Sampling (Holtzman et al., 2020) with a p value of 0.9, which provides an enhanced diversity on the generation in comparison to the likelihood maximization

[6]We adopted the fine-tuning implementation from `https://github.com/nshepperd/gpt-2`

[7]If not explicitly specified, all experiments in this thesis have been conducted on a Nvidia Tesla V100 GPU.

decoding methods while preserving the coherency by truncating the less reliable tail of the distribution. At the test time, the input HSs are fed into models as conditions, which are used as the initial contexts while sampling the next tokens. Given an input HS, the models produce a chunk of text which is a list of HS-CN pairs of which the first sequence marked with $[CN_start_token]$ CN $[CN_end_token]$ is the generated output.

3.5.3.2 Baselines

In addition to the fine-tuned GPT-2 models, we also evaluate two baseline models. Considering the benefits of the Transformer architectures on parallelization and learning long-term dependencies over recurrent models (Vaswani et al., 2017), we have implemented the baseline models using Transformer (TRF) architecture. The models have been trained similar to the base model described by Vaswani et al. (2017) with 6 Transformer layers, batch size of 64, 100 epochs, 4000 warmup steps, input/output dimension of 512, 8 attention heads, inner-layer dimension of 2048, and drop-out rate of 0.1. We used Nucleus Sampling also for the baselines with a p value of 0.9 during decoding.

In brief, we have trained four different configurations/models as authors:

1. $\mathbf{TRF}_{crowd}$: baseline on CROWD dataset
2. $\mathbf{GPT}_{crowd}$: fine-tuned GPT-2 on CROWD dataset
3. $\mathbf{TRF}_{niche}$: baseline on NICHE dataset
4. $\mathbf{GPT}_{niche}$: fine-tuned GPT-2 on NICHE dataset

3.5.3.3 Results

We report both standard metrics (BLEU (Papineni et al., 2002), BertScore (Zhang* et al., 2020)) concerning the lexical and semantic generation performances and a specific *Diversity* metric (RR) (Cettolo et al., 2014) regarding the generation quality. As a second quality metric, we report *Novelty* (Wang and Wan, 2018). While diversity is used to measure the ability of the model to produce diverse/varied responses with respect to the given input HS, novelty is used to measure how different the generated sequences are with regard to the training corpus (Wang and Wan, 2018).

Results of the author model experiments are shown in Table 3.11. In terms of BLEU and BertScore, baseline models yield a better performance. However, a few peculiarities of

Author	RR	Novelty	BLEU	BertScore
TRF_{crowd}	8.93	0.04	0.305	0.485
GPT_{crowd}	5.89	0.46	0.270	0.482
TRF_{niche}	4.89	0.10	0.569	0.457
GPT_{niche}	3.23	0.70	0.316	0.445

Table 3.11: Evaluation results of the best author configurations with different datasets. Novelty is computed w.r.t. to the corresponding training set, RR in the produced test output.

CN generation task and the experiment settings hinder the direct and objective comparison of the presented scores among the models. First, gathering a finite set of all possible counter narratives for a given hate speech is a highly unrealistic target. Therefore, we have only a sample of proper CNs for each HS, which is a possible explanation of very low scores using the standard metrics. Second, the train-test splits of NICHE dataset contain same CNs since the splitting has been done using one paraphrase for each HS and its all original CNs, while CROWD train-test splits have a similar property since an exact same CN can be found for many different HSs. Consequently, the non-pretrained Transformer models, which are more prone to generating an exact sequence of text from the training set, show a relatively better performance with the standard metrics in comparison to the advanced pre-trained models. Some randomly sampled CNs, generated by the various author configurations, are provided in Table 3.15–3.17 in Section 3.5.6.

Regarding the generation quality, we observe that baseline models cannot achieve the diversity achieved by GPT-2 models in terms of RR – both for NICHE and CROWD (4.89 vs 3.23, and 8.93 vs. 5.89). Moreover, GPT-2 provides an impressive boost in novelty (0.04 vs 0.46 and 0.10 vs 0.70). Among the GPT-2 models, the quality scores (in terms of RR and novelty) of the CNs generated by GPT_{niche} are better in comparison to those generated with GPT_{crowd}.

With regard to the overall results, GPT_{niche} is the most promising configuration to be employed as author. In fact, we observed that, after the output CN, the over-generated chunk of text consists of semantically coherent brand-new HS-CN pairs, marked with proper HS/CN start and end tokens consistent with the training data representation. Therefore, on top of CN generation for a given HS, we can also take advantage of the over-generation capabilities of GPT-2, so that the author module can continuously output

plausible HS-CN pairs without the need to provide the HS to generate the CN response. This expedient allows us to avoid the ephemerality problem for HS collection as well.

To generate HS-CN pairs with the author module, we basically exploited the model test setting and conditioned the fine-tuned model with each HS in the NICHE test-set. After removing the CN output for the test HS, we could obtain new pairs of HS-CN. In this way, we generate 2700 HS-CN pairs that we used for our reviewer-configuration experiments.

3.5.4 The Reviewer

The task of the reviewer is a sentence-level Confidence Estimation (CE) similar to the one of Machine Translation (Blatz et al., 2004). In this task, the reviewer must decide whether the author output is correct/suitable for a given source text, i.e., a hate speech. Consistently with the MT scenario, one application of CE is filtering candidates for possible human post-editing, which is conducted by the NGO operator by validating the CN. We tested three reviewer configurations:

1. **machine reviewer**: Filtering is done by a classifier neural-architecture before operator validation.
2. **non-expert reviewer**: Author output is filtered by human reviewers, then validated by operators.
3. **expert reviewer**: Author output is directly presented to NGO operators.

3.5.4.1 Machine Reviewer Experiment

As the machine reviewer we implemented 2 neural classifiers tasked with assessing whether the given HS-CN is a proper data pair. The two models are based on BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2020) architectures.

Training data. We created a balanced dataset with 1373 positive and 1373 negative examples for training purposes. The positive pairs come both from NICHE dataset and from the examples annotated in the human reviewer setting (Reviewer$_{\geq 2}$). The negative pairs consist of the examples annotated in the human reviewer setting, in the 'at least one 0' bin. In addition, 50 random HSs from NICHE-training are utilized with verbatim repetition as HS-HS to discourage the same text for both HS and CN in a pair, and 50

random HSs are paired with other random HSs simulating the condition of inappropriate CNs with hateful text.

Test data. We collected a balanced test set, with 101 positive and 101 negative pairs. Both positive and negative examples are created replicating the non-expert reviewer annotation described in Section 3.5.4.2 for new CN generation with NICHE test set by using the author model GPT_{niche}.

Models. For the first model, we follow the standard sentence-pair classification fine-tuning schema of the original BERT study. First, the input HS-CN is represented as $[CLS]$ HS_tokens $[SEP]$ CN_tokens $[SEP]$ and fed into BERT. By using the final hidden state of the first token [CLS] as the input, originally denoted as $C \in R^H$, we obtain a fixed-dimensional pooled representation of the input sequence. Then, a classification layer is added with the parameter matrix $W \in R^{KH}$, where K denotes the number of labels, i.e., 2 for HS-CN classification. The cross-entropy loss has been used during the fine-tuning.

We have conducted a hyperparameter tuning phase with a grid-search over the batch sizes 16 and 32, the learning rates [4,3,2,1]e-5 and the number of epochs in the range of 3 to 8. We obtained the best model by fine-tuning uncased BERT-large, with a learning rate of 1e-5, batch size of 16, and after 6 epochs at the 1029^{th} step on a single GPU.

The second model is built by fine-tuning ALBERT, which shows better performance than BERT on inter-sentence coherence prediction by using a sentence-order prediction loss instead of next-sentence prediction. In sentence-order prediction loss, while the positive examples are created similar to BERT by using the consecutive sentences within the same document, the negative examples are created by swapping sentences, which leads the model to capture the discourse-level coherence properties better (Lan et al., 2020). This objective is particularly suitable for HS-CN pair classification task, since HS and CN order and their coherence are crucial for our task. We fine-tuned ALBERT similarly to BERT model, by adding a classification layer on top of it. We applied the same grid-search that we used for BERT model to fine-tune ALBERT-xxlarge which contains 235M parameters. We saved a checkpoint at every 200 steps and finally, obtained the best model by using the learning rate of 1e-5, the batch size of 16, and at the 1200^{th} step.[8]

Metrics. To find the best model for machine reviewer, we compared BERT and ALBERT models over the test set. Although it seems more intuitive to focus on precision since we

[8]All the experiments have been conducted on a single GeForce RTX 2080 Ti GPU. Only the ALBERT classifier model has been trained with 8 TPU cores on Google Cloud.

search for an effective filtering over many possible solutions, we observed that a model with a very high precision tends to overfit on generic responses, such as "*Evidence please?*". Therefore, we aim to keep the balance between the precision and recall and we opted for F1 score for model selection. We report the best configurations for each model in Table 3.12, and the percentage of filtered pairs in Table 3.13. ALBERT classifier outperformed BERT model in all three metrics; F1, Precision, and Recall. Considering 6% of absolute F1 score improvement with respect to BERT model, we employed ALBERT model as the Machine Reviewer.

$\text{Reviewer}_{machine}$	F1	Precision	Recall
ALBERT	0.73	0.74	0.73
BERT	0.67	0.69	0.65

Table 3.12: F1, Precision and Recall results for the two main classifier configurations we tested.

3.5.4.2 Non-Expert Reviewer Experiment

In this section we describe the annotation procedure for the non-expert reviewer configuration.

Setup. We administered the generated 2700 HS-CN pairs to three non-expert annotators, and instructed them to evaluate each pair in terms of CN 'suitableness' with regard to the corresponding hate speech.

Instructions. We briefly described what an appropriate and suitable CN is, then we instructed them not to overthink during the evaluation, but to give a score based on their intuition. We also provided a list of 20 HS-CN pairs exemplifying the proper evaluation.

Measurement. We opted for a scale of 0-3, rather than a CE binary response, since it allows us to study various thresholds for better data selection. In particular, the meanings of the scores are as follows: 0 is not suitable; 1 is suitable with small modifications, such as grammar or semantic; 2 is suitable; and 3 is extremely good as a CN. We also ask to discard the pairs in which the hate speech was not well formed. For each pair we gathered two annotator scores.

Filtered Data. After the non-expert evaluation, we applied two different thresholds to obtain the pairs to be presented to the expert operators: (i) at least a score of 2 by both

annotators (Reviewer$_{\geq 2}$) yielding high-quality data where no post-editing is necessary, (ii) at least a score 1 by both annotators (Reviewer$_{\geq 1}$) providing reasonable quality with a possible need for post-editing.

The statistics reported in Table 3.13 show that high-quality pairs (Reviewer$_{\geq 2}$) account for only a small fraction (10%) of the produced data and only one third was of reasonable quality (Reviewer$_{\geq 1}$), while the vast majority was discarded. Some randomly selected filtered pairs are provided in Table 3.18 in Section 3.5.6.

Threshold	count	Percentage
Reviewer$_{\geq 2}$	276	10.0%
Reviewer$_{\geq 1}$	902	32.6%
at least one 0	1723	62.2%
bad HS	145	5.2%
Reviewer$_{machine}$	-	40.2%

Table 3.13: Percentage of filtered pairs according to various filtering conditions.

3.5.4.3 Expert Reviewer Experiments

To verify that the author-reviewer approach can boost HS-CN data collection, we run an experiment with 5 expert operators from an NGO. We compared the filtering strategies to reveal the best depending on several metrics.

Within Subject Design. We administered lists of HS-CN pairs to 5 operators from each filtering condition, and instructed them to evaluate/modify each pair in terms of 'suitableness' of the CN to the corresponding HS.

Instructions. For each HS-CN pair, we asked the operators: a) if the CN is a "perfect answer", to validate it without any modification, b) if the CN is not perfect, but a good answer can be obtained with some editing, to modify it, c) if the CN is completely irrelevant and/or needs to be completely rewritten to fit the given HS, to discard it.

Measurement. The main goal of our effort is to reduce the time needed by experts to produce training data for automatic CN generation. Therefore the primary evaluation measure is the average time needed to obtain a proper pair. The other measurements of interest are Diversity and Novelty, to understand how the reviewing procedure can affect the variability of the obtained pairs.

Procedure and material. We gave the instructions along with a list of 20 HS-CN exemplar pairs for each condition (i.e., Reviewer$_{\geq 1}$, $_{\geq 2}$, $_{machine}$, $_{expert}$). The condition order was randomized to avoid primacy effect. In total, each NGO operator evaluated 80 pairs. Pairs were sampled from the pool of 2700 pairs described before (apart from the automatic filtering condition). To guarantee that the sample was representative of the corresponding condition, we performed a stratified sampling and avoided repeating pairs across subjects.

Results and Discussion. As it is shown in Table 6.7, there is a substantial decrease in data collection time (NGO$_{time}$) when automatic generation mechanisms are introduced (no suggestion vs. Reviewer$_{expert}$). If crowd filtering is applied (Reviewer$_{\geq 1}$, $_{\geq 2}$), the amount of time can be further reduced, and the more stringent the filtering criterion, the higher the time saved. Conversely, the more stringent the filtering criterion, the higher the time to obtain a filtered pair from non-expert annotators (CROWD$_{time}$). For instance to obtain a single pair with at least a score of 2 by both annotators, 700 sec (around 12 min) are needed on average (only 10% of examples are in ≥ 2 condition). Results indicate that providing an automatic generation tool meets the first goal of increasing efficiency of the operators in data collection.

Approach	NGO$_{time}$	Crowd$_{time}$	RR	Novelty	Pairs$_{selec}$	Pairs$_{final}$
no suggestion	480	-	2.72	-	-	-
Reviewer$_{expert}$	76	-	3.56	0.73	100%	45%
Reviewer$_{\geq 1}$	72	215	4.31	0.70	33%	54%
Reviewer$_{machine}$	68	-	4.48	0.68	40%	63%
Reviewer$_{\geq 2}$	49	703	5.70	0.65	10%	72%

Table 3.14: Results for CN collection under various configurations. RR for 'no suggestion' is computed on NICHE dataset and the time needed is the one reported in Section 3.4. Time is expressed in seconds. Pairs$_{selec}$ indicates the percentage of original author pairs that have been passed to the expert for reviewing, Pairs$_{final}$ indicates the percentage of selected pairs that have been accepted or modified by the expert themselves. Crowd$_{time}$ is computed considering that annotators gave a score every 35 seconds, and we required two judgments per pair.

Regarding diversity and novelty metrics, pre-filtering author's output (Reviewer$_{\geq 1}$, $_{\geq 2}$ and $_{machine}$) has a negative impact: the more stringent the filtering condition the higher the RR and the lower the novelty of the filtered CNs. We performed some manual analysis

of the selected CNs and we observed that notably for the Reviewer$_{\geq 2}$ case (which was the most problematic in terms of RR and novelty) there was a significantly higher ratio of "generic" responses, such as "*This is not true.*" or "*How can you say this about an entire faith?*" , for which reviewers' agreement is easier to attain. Therefore, the higher agreement on the generic CNs reveals itself as a negative impact in the diversity and novelty metrics. Conversely, the percentage of pre-filtered pairs that are accepted by the expert increases with the filtering condition becoming more stringent, the baseline being 45% for the Reviewer$_{expert}$ condition.

As for the amount of operators' effort, we observed a slight decrease in HTER[9] with the increase of pre-filtering conditions, indicating an improvement in the quality of candidates. However, HTER scores were all between 0.1 and 0.2, much below the 0.4 acceptability threshold defined by Turchi et al. (2013), indicating that operators modified CNs only if "easily" amendable. Finally, we observe that despite reducing the ouput diversity and novelty, the reduction of expert effort by Reviewer$_{\geq 2}$ in terms of the percentage of the obtained pairs is not attainable by a machine yet. On the other hand, automatic filtering (Reviewer$_{machine}$) is a viable solution since (i) it helps the NGO operators save time better than human filter $\geq$1, (ii) it preserves diversity and novelty better than Reviewer$_{\geq 2}$ and in line with Reviewer$_{\geq 1}$.

3.5.5 Author-Reviewer Architecture Discussion

We presented a set of experiments on how to reduce data collection effort, using a mix of several strategies. We analyzed the existing hate countering datasets on the basis of two principles, data quantity and data quality. Our results show that the datasets collected by existing approaches do not satisfy both requirements at the same time. To effectively and efficiently obtain varied and novel data, we first proposed the generation of silver counter narratives – using large scale unsupervised language models – then a filtering stage by crowd-workers and finally an expert validation/post-editing. We also showed promising results obtained by replacing crowd-filtering with an automatic classifier.

Altogether, our findings suggest that there are different strategies for collecting HS-CN pairs depending on the amount of experts available. When a significant number of experts is available, nichesourcing the task to experts is recommended for granting the quality of data needed for neural approaches. Alternatively, a hybrid approach is practical, if the

[9]HTER is a measure used to compute the post-editing effort of annotators at sentence level translations for correcting the output of automatic machine translation systems. (Specia and Farzindar, 2010).

amount of experts is limited, by generating data with a strong fine-tuned language model along with a post-editing phase via experts. For an extreme scenario where experts are greatly insufficient for the demand, it is helpful adding a pre-filtering step before post-editing phase via non-expert human or classifier model. As a final remark, we believe that the proposed framework can be useful for other NLG tasks such as paraphrase generation or text simplification.

3.5.6 Generation Examples

HS	I love the Pres also he is a real man ... not a commie cunt like the last muslim in there.
GPT-2	Use of the cword here is totally unacceptable because it is hurtful and offensive.
	That is very interesting, but they already have that kind of man who is always saying something different. No need for your comment.
	Hi there, please refrain from using derogatory comments in the thread. They are hurtful and unwanted. If you continue Admin will be alerted.
TRF	Please refrain from using derogatory terms for females.
	It's really hard to respect someone who uses racial slurs in an attempt at comedy.
	Please don't use hateful words, please use your words or else removal will take place.
HS	What cha doin America? Oh...just waiting to hang the cunt the Muslim nigga.
GPT-2	Name calling speaks volumes about one's character.
	To act like those of a race are the same is to admit to your narrow mindedness.
	Hey there, racist and derogatory views are hurtful and uncalled for. Please refrain from this type of behavior or you will be reported. Thx.
TRF	It doesn't make sense to name call someone you've never met based on what the media says.
	I'd be more inclined to listen to your side if you didn't use that slur.
	Please muslims are not so two as a high everyone has no need to get an ignorant little elected official further.
HS	The muslim nigger terrorist obama needs death the hussein needs it's limbs ripped off and his tongue seared with a hot poker.
GPT-2	You don't know that, but I do.
	I understand your point but please watch the language.
	You are throwing a lot of hatred and anti-Semite in to this, which is not good.
TRF	I'd be inclined to agree with you if you didn't use slurs.
	Your reply isn't need, not big or clearer. Think before you type next time please.
	All religions, ethnicities, and social groups deserve tolerance.

Table 3.15: Randomly sampled CNs generated from GPT-2 and TRF models trained on CROWD dataset.

HS	White women converted to Islam are the saddest and most selfish persons that you can encounter. They praise happily Islam beliefs and become part of a sisterhood of tollerance and understanding, but they do not know the situation of women in Pakistan and other islamic states.
GPT-2	Do you know that in many Muslim-majority countries, women's rights are recognized and actively promoted? Laws are passed that guarantee the right to have children, and foreign charities work to ensure that women have access to quality education and health services.
	Are you a believing woman who has known and respected great men like Mohammed, Albert and Winston Churchill. You have forgotten about those who sacrificed so much for your benefit.
	If white women who love Islam were doing such things why would their profile be so public?
TRF	Honestly, i see around many more muslim women without burqa or veil than women dressed like europeans.
	Are you sure? muslim nobel prize winners would suggest otherwise.
	Sounds like you are generalizing what a very small percentage of extremists do to an entire religion.

Table 3.16: Randomly sampled CNs generated from GPT-2 and TRF models trained on NICHE dataset.

HS	We are every day in constant danger, because Muslims can blow us all up. For this fact we should put all Muslims in a hole in the ground.
GPT-2	If you had said this about all the non-Muslim groups who may espouse hateful or violent views, how would you feel about being labelled a bigot?
	Apart from being offensive and ignorant, is also the belief that Muslims only create problems. In fact, many Muslims are working to solve some of these social problems.
	Muslims have never sought to destroy our world, rather they have offered their goods and services in return for peace and unity. Nostalgia may have something to offer us, but not for taking over countries and nations.
TRF	People who commit offenses are punished by law. Muslims are not all terrorists. muslims are not trying to generalize on their own community, indeed anything. If someone would like you, but you are peaceful muslim countries like which, please.
	No, a person can be immigrants for multiple allegiances and responsibilies.
	Why do you think that muslim criminals will receive different treatment or even privileges?
HS	Koran is a book that should be burnt and mosques should be demolished. If we do all these things, all problems in our community will be solved.
GPT-2	All book burning is prohibited under the Universal Declaration of Human Rights, which says: ' Everyone has the right to book his escape from cruelty and oppression wherever it may be found.'?
	Yes, because this book encourages people to respect for life and to respect God. Glad you like it!
	Praise be to God that we have these humble requests. Can you give us some figures on the number of people who would like to migrate to your country if you could only send em home?
TRF	If you dislike diversity so much , why do not you go off and live in a country where there are no muslims, only people like you?
	Mgm fgm have been in britain for hundreds of years. it is never acceptable but it does not come from just one country or ideology.
	The idea of blow someone up is itself crazy.

Table 3.17: Randomly sampled CNs generated from GPT-2 and TRF models trained on NICHE dataset.

HS	Girls and boys are brainwashed by the same people.
No filter	That's why most religions end well.
$\text{Reviewer}_{\geq 1}$	Mindless Hate and fear mongering divide and conquer.
$\text{Reviewer}_{\geq 2}$	Mindless Hate and fear mongering do not help anyone.
HS	Everything will be better if Muslims did not exist. Their only contribute to our society is violence.
No filter	Did you know that the greatest child abusers in British history were white men? That is if they weren't SO EXTREMELY LIE!?!?!
$\text{Reviewer}_{\geq 1}$	Why would you say that? I would say that all the people using social media and whatever to vent their anger and frustration are clearly not serious about their religion. Some may be, and are angry, but that does not mean they are responsible for anything.
	The world would be a better place without forced baptisms and child rapists.
	What exactly are you suggesting?
$\text{Reviewer}_{\geq 2}$	The world would actually be a very dark place without Muslims who contribute a lot to our society. How about a place of peace, tolerance and understanding for all?
HS	Europe is civilised, Muslims should not stay there.
No filter	That is a misunderstanding. On the one hand, Europe is a civilized place, on the other, Muslims cannot excel at anything.
$\text{Reviewer}_{\geq 2}$	How can you say that about an entire faith of 1.6 billion people?
$\text{Reviewer}_{machine}$	Muslims have been a part of the European society for centuries, they have been a part of its institutions since the middle ages, they have been active in politics during the 19^{th} century and have been crucial in the development of Europe during the last decades.
HS	Multiculturalism has brought us nothing but disaster.
$\text{Reviewer}_{machine}$	The multiethnic society has produced many smart and talented people, who have gone on to work in prominent public offices. No one is asking us to sacrifice our culture or our identity for the greater good of the community.

Table 3.18: Randomly sampled CNs, generated from GPT-2, filtered from various reviewer configurations.

3.6 Conclusion

In Chapter 1, we identified our first research question on obtaining high-quality training examples to oppose hate content with counter narratives. 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.

As a necessary first step, in this chapter we described the creation of the first large-scale, multilingual, expert-based dataset of hate speech/counter narrative pairs. This dataset was built with the effort of more than 100 operators from three different NGOs that applied their training and expertise to the task. In total, the dataset consists of more than 15 thousand pairs over the 3 languages for English, French, and Italian. Together with the collected data we also provided several types of metadata about expert demographics, hate and response type, and data augmentation through translation and paraphrasing. Finally, we provided initial experiments to assess the quality of our data.

Grounded on this dataset, we further presented a study on how to collect counter narratives to hate effectively, employing large scale unsupervised language models such as GPT-2 for the generation of silver data, and the best annotation strategies/neural architectures that can be used for data filtering before expert validation/post-editing.

Chapter 4

Knowledge-Guided Counter Narrative Generation

4.1 Introduction

In Chapter 3, we discussed the properties of desired data quality in terms of diversity (lexical and semantic diversity) and conformity to NGO guidelines. We now turn to the generation of counter narratives.

Understanding effective counter narratives is essential to make progress towards persuasive online hate countering via automation. However, identifying effective CNs is hard and remains an open question. This is because the field is still young and counter narratives vary considerably in tone and communicative strategies (Benesch et al., 2016b). In this chapter, we focus on effective counter narratives based on facts and argumentation. Indeed, an effective counter narrative can refute false information in hate speech by supplying facts from credible sources and adequate reasoning (Benesch, 2014; Tuck and Silverman, 2016; Benesch et al., 2016a). From an argumentation's perspective, a convincing argument is often backed up and reasoned logically through facts and examples (Habernal and Gurevych, 2016). In this context, we take these remarks into consideration when developing systems for counter narrative generation. The CN2 in Table 1.1 in Chapter 1 exemplifies an effective counter narrative.

Recently, studies have started to focus on strategies for the automatic generation of counter narratives in an end-to-end fashion (Chung et al., 2019; Qian et al., 2019; Tekiroğlu et al., 2020; Chung et al., 2020; Zhu and Bhat, 2021). However, the seminal studies,

based on training models using only the HS-CN data samples, do not directly address the required properties for proper and effective counter narratives such as providing credible evidence and knowledge with facts, statistics or examples (e.g., social movements). While such models possess knowledge to a certain extent through utilizing existing generative language models such as GPT-2 (Radford et al., 2019), this knowledge (e.g., events or news) will quickly get out-dated or not be able to handle new unseen events, arguments and facts. Furthermore, such models are able to create plausible but not necessarily true arguments - a problem known as content hallucination - (Zellers et al., 2019; Solaiman et al., 2019). A hallucinated example of a fact-bound CN, generated by GPT-2 model fine-tuned as done in Section 3.5, is provided in Table 4.1. Therefore, in this chapter we investigate the generation of knowledge-bound counter narratives, which had never been studied yet to the best of our knowledge. We hypothesized that knowledge infusion can not only leverage generating informative counter narratives but also handle hate speech from unseen domains (i.e., hate targets) without further training (e.g., train on Islamophobia and test on Antisemitism).

HS: *Muslims are invading our country*
CN: *Actually, there were 16,938,000 Muslims in the UK in 2016, so if you exclude London, that is actually increasing by 2%, which doesn't seem very significant.*

Table 4.1: Hallucinated CN generated by GPT-2 that is fluent and credible (according to Office for National Statistics, the Muslim population is just above 3M).

To this end, we explore methodologies to generate informative counter narratives using external knowledge. In particular, we propose an extension of knowledge-grounded generation approaches by adopting an intermediate step where we generate keyphrases for retrieving needed knowledge. We first train a counter narrative keyphrase generator, then the generated keyphrases are employed for selecting relevant knowledge sentences (Section 4.5). Finally, pre-trained LMs are fine-tuned on the relevant knowledge sentences, together with the HS input, to produce knowledge-augmented counter narratives (Section 4.6). Our extensive experiments on counter narrative generation, including both automatic and expert evaluation, demonstrate that the presented approach produces more specific and tailored responses both for in-domain and zero-shot cross-domain configurations as compared to other approaches, such as fine-tuning standard LMs for the task without the use of external knowledge (Section 4.6.2 and 4.6.3).

Our summary findings are that: (i) external knowledge can boost informative CN gen-

eration, (ii) keyphrase generation improves the quality of retrieved documents, (iii) silver knowledge is utilizable for the task when no gold knowledge is available, (iv) knowledge-bound models are advantageous while tackling zero-shot cross-domain generation particularly if (v) using injection of knowledge in large pre-trained LMs.

4.2 Counter Arguments versus Counter Narratives

While counter arguments and counter narratives have a similar purpose - to present an alternative stance to a statement - a counter argument is not necessarily a counter narrative, and vice versa. Counter narratives face peculiar difficulties such as the absence in hate speech of explicit or well-structured 'arguments' (e.g., "*Islam is a disease*") and the limited amount of data available for training. The main differences between the two are briefly discussed below, and point for the development of specific generation approaches for counter narratives.

Presence of a counter argument. A counter argument is typically defined as an argument (equivalent to a claim or standpoint) contradicting the initial statement with well-structured 'arguments', supporting evidence or reasoning (Khishfe, 2012; Schiappa and Nordin, 2013; Stab et al., 2018). The presence of supported evidence or reasoning is vital for counter-argumentation and argumentative information retrieval, regardless of it being implicit or explicit (e.g., Statement: *Nuclear energy should be used to replace fossil fuels.* v.s. Counter argument: *Renewable energy is a better option for replacing fossil fuels than unsustainable and expensive energy like nuclear power.*). In contrast, counter narratives can be formulated without using arguments. For example, a counter narrative can be without supporting evidence (e.g., Hate speech: *Islam is a plague.* v.s. Counter narrative: *How can you use such language to describe a religion of 1.6 billion people?*).

Non-negative arguments. Following guidelines[1] generally adopted by NGOs, an appropriate counter narrative should not contain hostile language, prejudice, and unlawful content. Using negative tones as responses may risk escalating the conversation and engaging in hateful sentiment. On the contrary, counter arguments are not bound by such constraints.

The closest work to ours is counter argument generation with external knowledge augmentation by Hua et al. (2019). Our approach differs from theirs in three aspects: (i) we explore generating queries to extract knowledge for grounding CN with, (ii) pre-trained

[1]As an example: `https://getthetrollsout.org/stoppinghate`

generative models are utilized for leveraging the knowledge present, (iii) our approach requires less manipulation over knowledge.

4.3 HS-CN Dataset

To the best of our knowledge, there is no high-quality hate speech - counter narrative dataset available yet where CNs are explicitly paired with relevant knowledge. Since constructing such dataset with a decent-size would be too costly[2], we resort to a "reverse-engineering" strategy such that we automatically paired relevant knowledge with an already existing high-quality counter narrative dataset. We chose CONAN (see Section 3.4), which is the best and most diverse material among the other CN datasets (Tekiroğlu et al., 2020). We split the English data including translated pairs into 4069/1288/1288 samples for train/dev/test set.

4.4 Knowledge-Guided Counter Narrative Generation Architecture

Our architecture, illustrated in Figure 4.1, consists of a knowledge retrieval module that retrieves sentence-level relevant knowledge, and a generation module that generates a counter narrative. Specifically, the knowledge retrieval module first prepares variants of a query Q for a given hate speech (HS) using two strategies: query extraction (Q_{hs}) and automatic query generation (Q_{gen}). Then, the obtained queries are employed to search for relevant knowledge articles via a search engine. Finally, it uses a sentence selector to filter and rank the most relevant sentences as the relevant knowledge (KN) from the retrieved articles. For the counter narrative generation module, we fine-tuned several LMs that take a HS and the ranked knowledge sentences KN as input and output a corresponding counter narrative.

[2]While obtaining access to a pool of trained NGO operators is possible but also very complicated, keeping track of their search activity and the material they used during CN production would require even longer and more complex data collection sessions than the one described in the previous chapter to collect CONAN.

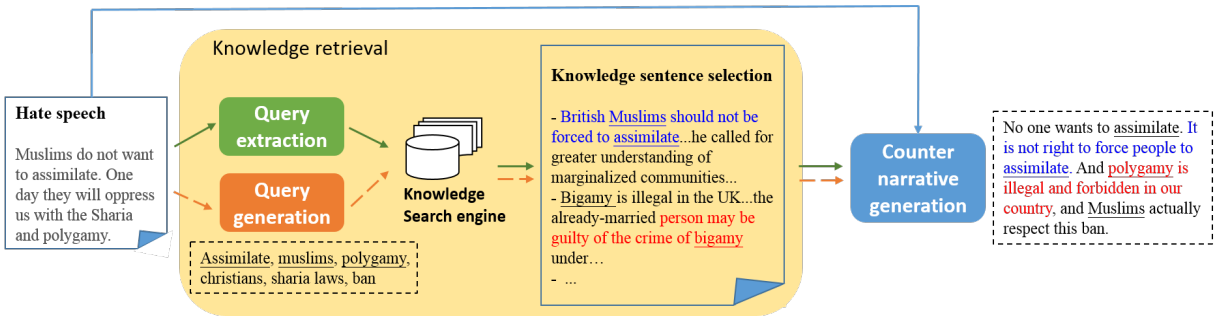

Figure 4.1: Architecture of knowledge grounded generation with extracted (green solid arrow) and generative (dotted arrow) queries (topical phrases) that are exploited to retrieve relevant knowledge. The knowledge sentences extracted together with input HS are fed to CN generation.

4.5 Knowledge Retrieval Module

The knowledge retrieval module in the architecture incorporates a knowledge repository, query construction sub-module, and a knowledge sentence selection sub-module that are described in the following sections.

4.5.1 Knowledge Repository

Previous approaches on introducing external knowledge for dialog generation have exploited unstructured (Dinan et al., 2019; Hua et al., 2019) and structured knowledge (Zhou et al., 2018). Since no structured knowledge is available for the hate speech domain, we rely on unstructured textual knowledge in the format of articles, which allows for updating the knowledge repository easily. Considering that the proliferation of HS is also triggered with target-related events (e.g., terrorist attacks), being able to update the knowledge, such as news articles, would let us produce proper CNs that contain the latest statistics or evidence from the current events.

We include Newsroom (Grusky et al., 2018) and WikiText-103 (Merity et al., 2017) to our knowledge repository. WikiText-103 is a large collection of 28,595 full Wikipedia articles covering over 103 million words. Newsroom consists of 1.3 million articles extracted from major news publications between 1998 and 2017, featuring over 6.9 million words.

4.5.2 Query Construction

To construct comprehensive and proper queries to search for relevant knowledge for the HS/CN pairs, we applied two strategies: (i) query extraction and (ii) query generation. In both strategies, the query is composed of keyphrases that can be defined as the important and topical phrases from a text (Turney, 2000).

Query extraction. We extracted keyphrases from CONAN dataset using Keyphrase Digger (Moretti et al., 2015), a multilingual keyphrase extraction system that uses statistical measures and linguistic information, and is proven to be one of the best systems for unsupervised settings[3]. Following the knowledge retrieval strategy using input argument by Hua et al. (2019) for counter argument generation, we first obtained the HS keyphrases to construct the initial query Q_{hs}. However, HSs from CONAN mostly contain hateful and simplistic phrases in comparison to the input arguments used by Hua et al. (2019) that can be rich in content[4]. Therefore, in the HS-CN scenario, we hypothesize that the keyphrases from Q_{hs} alone would not be sufficient for relevant knowledge search, in particular for mapping the knowledge onto training data.

To this end, we also extracted keyphrases from CN together with HS to increase the possibility that the retrieved knowledge sentences contain pieces of information found in the ground truth. Hence, the second query $Q_{hs \cup cn}$ contains CN keyphrases for the relevancy to the target CN and HS keyphrases for preserving the hate context.

Analysis on keyphrase extraction configurations. We conducted a preliminary manual analysis to determine the best keyphrase extraction configurations. In particular, we randomly sampled 48 pairs of hate speech and counter narrative from CONAN dataset and extracted the keyphrases. Then, we retrieved the KN with the queries Q_{hs}, Q_{cn} and $Q_{hs \cup cn}$. For each condition, we asked an annotator, expert in CN writing, to assign a score for the KN retrieved based on the relevance to a given hate speech in the scale of 1 to 5; 1 meaning no-relevance, and 5 perfect relevance. As a result, we have noticed that Q_{cn} is the worst condition, i.e., non-optimal, having an average score of 2.30. The analysis shows that it causes the loss of context related to HS, bringing information mostly from whole another topic. For instance, especially when CNs are rather generic, often, no lexical hint can be found related to the topic of Islamophobia (e.g., "*Do you have*

[3]Keyphrase Digger is a new implementation of KX (Pianta and Tonelli, 2010) with several improvements, and it was ranked the best performing unsupervised system on task 5 of SemEval 2010 evaluation campaign (Moretti et al., 2015).

[4]e.g., an argument "*A universal basic income will help the labor market adapt to inevitable disruptions caused by advancements in automation and artificial intelligence.*" vs. HS "*Islam is a disease*".

any proof?"). Indeed, Q_{hs} provides an apparently better average score (3.46) since it provides a better context to search for. However, as expected, the best score (3.77) has been obtained through $Q_{hs \cup cn}$, i.e, optimal, verifying our hypothesis of utilizing both HS and CN keyphrases for training.

Query generation. Since the best query configuration $Q_{hs \cup cn}$ cannot be available at test time, we need a way to obtain keyphrases that serve as CN cues for searching knowledge sentences during the CN generation. To this end, we built a query generation model that takes HS as input and outputs a list of CN keyphrases, which is then used as Q_{gen}. Our aim is to obtain an approximation of $Q_{hs \cup cn}$ via $Q_{hs \cup gen}$ at the test time.

The model is trained using Transformer (Vaswani et al., 2017) architecture as it has obtained state-of-the-art performance for generation tasks (Dinan et al., 2019; Ghazvininejad et al., 2018). For the training data, we used CONAN dataset and discarded the CNs that are less than 10 words, since they are usually generic, poor in terms of argumentative content and cannot provide a meaningful search (e.g., "*No they are not - prove this?*", "*What does that even mean?*", "*Any evidence?*"). Accordingly, we kept 3906/1257/1257 instances for train/dev/test set. The train set includes the pairs marked as original in the dataset, and all translated pairs from French and Italian; the dev set consists of one paraphrase of each original HS and its CNs; and the test set contains the rest of the paraphrased HSs. The training inputs are represented as

$$HS\ [HS_end_token]\ KP\ [KP_end_token]$$

where KP is the list of keyphrases extracted from the gold CN.

The model has been trained following the configuration of the base model in Vaswani et al. (2017): with 6 Transformer layers, 8 Transformer heads, embedding size of 512, hidden size of 2048, dropout rate of 0.1, batch size of 64 for 100 epochs. The training time lasted around 7 hours. For decoding, we used nucleus sampling (Holtzman et al., 2020) with a p value of 0.9.

We report keyphrase generation results in terms of BLEU and ROUGE against the keyphrases extracted from gold test CNs. We obtained a score of 0.162 for BLEU-2 and a score of 0.353 for ROUGE-L. Although both scores can be considered as low, this is due to the open ended nature of the set of possible CN keyphrases for a given HS. Example queries for a single pair, extracted from its HS, its CN, and generated with the keyphrase generation model are shown in Table 4.2.

HS	CN	Query	Knowledge sentences (KN)
Islam is a disease.	Like Christianity or any other, islam is a religion of tolerance. Disease does not discriminate on religious grounds.	Q_{hs}: islam, disease	(i) Do Muslims want to heal from the disease...? (ii) Being infected by religious extremism is like being infected by a disease...
		Q_{gen}: islamic law, god, christians	(i) Islamic law is to create an environment...submission to God. (ii) Certain areas of the Muslim world have always been home to large populations of Christians...
		$Q_{hs \cup cn}$: islam, disease, tolerance, christianity, discriminating	(i) Islam is a 1400-year-old religion that preaches tolerance...like Christianity and... (ii) Disease does not discriminate...on religious grounds...

Table 4.2: Examples of KN retrieved using (i) extracted queries from HS (Q_{hs}), (ii) generated queries (Q_{gen}), and (iii) extracted queries from both HS and CN ($Q_{hs \cup cn}$).

4.5.3 Knowledge Sentence Selection

We use Solr[5] to index the articles and retrieve those relevant to a given query based on the similarity between the articles and the query using BM25 (Robertson et al., 1995). Once the queries have been obtained either through extraction or generation, they are presented to Solr for retrieving the 25 top-ranked articles. Next, we used spaCy sentence segmentation[6] to split an article into sentences. Similar to Zhang et al. (2020), given a query Q we score each sentence x_i in the set of articles D independently, using ROUGEL-F1 (Lin, 2004) as in Equation 4.1.

$$s_i = rouge(x_i, Q), \forall i \in D \tag{4.1}$$

In the final step, we distilled the knowledge by keeping the top 5 knowledge sentences that have the highest scores among 25 top-ranked articles. Instead of a more stringent filtering, such setting has been applied to grant a better variety of source articles and corresponding distilled sentences. We refer to such automatically associated sentences as "silver knowledge".

[5] https://lucene.apache.org/solr/

[6] https://spacy.io/universe/project/spacy-sentence-segmenter

4.6 Counter Narrative Generation Module

Large pre-trained LMs require less amount of high-quality data to be fine-tuned on downstream tasks while providing strong performances and they already store large amount of factual and commonsense knowledge from their training data (Petroni et al., 2019). To this respect, we built the following models: (1) **GPT-2**$_{KN}$, obtained by fine-tuning GPT-2 on CONAN data paired with KN; (2) **GPT-2**$_{KN,MT}$, by fine-tuning GPT-2$_{KN}$ in a multi-task learning fashion for learning to distinguish counter narratives from hate speech as next utterances; (3) **XNLG** (Chi et al., 2020) for its ability to copy information to the output (in our case the retrieved KN to be copied to the CN). We expect all three models to attend over the HS and retrieve KN and look for the relevant snippets to be recovered while generating a CN.

4.6.1 Experiments

4.6.1.1 Models

The training HS-CN pairs are represented as:

$$HS\ [HS_end_token]\ KN\ [KN_end_token]\ \ CN\ [CN_end_token].$$

Each model is trained with $Q_{hs \cup cn}$ and then tested on Q_{hs}, Q_{gen}, and $Q_{hs \cup gen}$. We also tested the models with $Q_{hs \cup cn}$ to define an oracle scenario with an upperbound performance when the data can only be paired with silver knowledge.

1. **GPT-2**$_{KN}$**.** We fine-tuned the GPT-2[7] medium model for 3 epochs with a batch size of 2048 tokens. We used Adam optimizer with a learning rate of 5e-5. At inference time, responses were generated employing nucleus sampling with a p value of 0.9, conditioned on HSs and corresponding KN.

2. **GPT-2**$_{KN,MT}$**.** Since we noticed that GPT-2 occasionally produces responses that contain fragments of abusive language, we combined the language modeling objective with a next-sentence prediction objective for fine-tuning GPT-2 in a multi-task setting, inspired by Wolf et al. (2018). Next-sentence prediction adopts a linear classification layer added to the last layer of the Transformer language model and then applies a cross-entropy loss to classify a proper next response to the input HS from 2 distractors randomly selected from HS. We used Adam optimizer with a

[7]We adopted the implementation from `https://github.com/huggingface/transformers`

learning rate of 5e-5 and empirically fine-tuned it for 1 epoch and the same sampling strategy as GPT-2_{KN} has been applied.

3. **XNLG** is a pre-trained Transformer-based language model trained on Wikipedia dumps with two relevant objectives for our task: to obtain contextual representations and to recover a given input. We fine-tuned XNLG[8] for generating counter narratives on all layers with a batch size of 10 for 100 epochs. We used Adam optimizer with a learning rate of 1e-4. We tokenized and removed accent from the entire dataset and applied the same BPE codes used by Chi et al. (2020). For KN and CN we kept the first 256 tokens, while setting the HS to 70 tokens, which is the maximum length of hate speech in the dataset. We experimented with various decoding methods and adopted beam search with a beam-width of 3 for the best performing setting (Details on the selection of best decoding methods are provided in Appendix A.1).

Baselines used for comparison are: (1) non-pretrained **Transformer** without knowledge using the same hyper-parameters as keyphrase generation model; (2) **GPT-2** without knowledge following the same configuration as GPT-2_{KN}; (3) **Candela** (Hua et al., 2019), an LSTM-based state-of-the-art knowledge-driven architecture for argument generation. Since CONAN is relatively small, we hypothesize that a pre-training procedure[9] on data from a similar task (argument generation) can be beneficial for generalization and porting knowledge. Thus, we first pre-trained Candela architecture on argument generation dataset (Hua et al., 2019), following the configuration described in the paper. We then fine-tuned the model for 20 epochs on CONAN with KN using Q_{hs} as it is done in the original setting of Candela.

4.6.2 Results for the Silver Knowledge Test Set

We report BLEU-2 (B-2) and ROUGE-L (R-L) scores for all proposed models and baselines in Table 4.3 on the test split of CONAN that we automatically paired with silver knowledge using various queries. We also measure the capability of each model to produce *novel* responses with respect to the training data by Jaccard similarity (Wang and Wan, 2018), and *diverse* responses for the given input by repetition rate (RR) (Cettolo et al., 2014).

[8] https://github.com/CZWin32768/XNLG

[9] We also trained Candela from scratch on CONAN but decided not to proceed with this setting for the poor performance.

Among our models, GPT-2_{KN} yields the highest B-2 and XNLG the highest novelty, diversity, and R-L. The notably improved novelty achieved by knowledge-grounded models indicates the benefits of adding knowledge on producing CNs, in comparison to the baselines - particularly Transformer TRF. On the other hand, the quantitative performance of XNLG does not reflect its true performance in terms of quality. A quick glance at the output CNs showed that XNLG model copies almost everything from KN to the output instead of a proper CN generation, increasing the novelty and diversity scores. The issue can easily be observed from the average numbers of words and sentences in the XNLG output in comparison to the outputs of the other models presented in Table 4.3. GPT-$2_{KN,MT}$ falls behind among our models in terms of RR, B-2, and R-L, still providing a competitive novelty. Regarding Candela, while it obtained similar performances to our models in terms of R-L and B-2, the generation is repetitive and less novel.

As for the testing with different query types, $Q_{hs \cup gen}$ induces more novel responses than $Q_{hs \cup cn}$ and Q_{hs}. While XNLG yields the highest novelty with Q_{hs} (0.824), it can be explained again with the problem of copying the whole KN, which is more varied due to the less restrictive search using only HS. Conversely, XNLG obtained the highest repetition rate with Q_{hs}, since the models exploit the same knowledge for multiple test samples due to the repeated HSs with different CNs in the test set.

The oracle query $Q_{hs \cup cn}$, in which we deliberately provide the best knowledge possible through the keyphrases coming also from the gold CN, yields the best R-L scores among the query variations of knowledge-grounded models. Among all the models, $Q_{hs \cup cn}$ also leads to the best B-2 through GPT-2_{KN}, and the best R-L through XNLG, as we have anticipated. Finally, $Q_{hs \cup gen}$ outperforms Q_{hs} and Q_{gen} over most metrics, hinting at the advantages of using generated queries together with hate context for silver-knowledge retrieval.

Unique HS test set analysis. Taking into consideration the design choices and the various phenomena in our study, we have further conducted complementary experiments on unique HS test set. Since in our test set, in line with CONAN, a HS can be paired with more than one CN, Q_{hs} would retrieve the same KN for all the target CNs of the same input HS. Contrarily, we obtain a different set of KN using queries Q_{gen}, $Q_{hs \cup gen}$ and $Q_{hs \cup cn}$ for each target CN. Accordingly, we also report an evaluation on unique HS-CN pairs, where a single target CN has been randomly chosen for each HS, among all query types.

The unique test set is comprised of each unique HS and one randomly selected CN

Models	Nov.	RR	B-2	R-L	#W.	#S.	KN overlap (ngram) 1	2	3
without KN									
TRF	0.467	7.72	0.082	0.094	21.47	1.70	-	-	-
GPT-2	0.688	9.04	0.045	0.100	15.95	1.35	-	-	-
Train$_{cn}$	-	3.91	-	-	21.79	1.87	0.307	0.054	0.016
with KN									
Candela (Q_{hs})	0.692	21.87	0.040	0.098	23.85	2.47	0.173	0.008	0.001
GPT-2$_{KN}$									
w/ Q_{hs}	0.723	8.13	0.082	0.094	15.60	1.32	0.258	0.023	0.008
w/ Q_{gen}	0.728	7.48	0.067	0.091	12.75	1.17	0.260	0.050	0.019
w/ $Q_{hs \cup gen}$	0.735	6.30	0.085	0.103	15.35	1.59	0.358	0.068	0.024
w/ $Q_{hs \cup cn}$	0.727	7.17	**0.166**	0.110	13.10	1.16	0.282	0.058	0.022
GPT-2$_{KN,MT}$									
w/ Q_{hs}	0.744	11.69	0.050	0.090	13.35	1.17	0.269	0.049	0.017
w/ Q_{gen}	0.731	10.37	0.052	0.092	13.34	1.14	0.253	0.044	0.017
w/ $Q_{hs \cup gen}$	0.747	7.59	0.091	0.090	16.91	1.26	0.269	0.033	0.009
w/ $Q_{hs \cup cn}$	0.731	9.56	0.048	0.107	13.05	1.13	0.276	0.057	0.023
XNLG									
w/ Q_{hs}	**0.824**	14.42	0.073	0.084	55.51	3.71	0.841	0.650	0.558
w/ Q_{gen}	0.819	6.88	0.097	0.084	55.64	3.64	0.849	0.656	0.558
w/ $Q_{hs \cup gen}$	0.812	6.98	0.074	0.089	57.58	3.00	0.828	0.579	0.475
w/ $Q_{hs \cup cn}$	0.819	**5.69**	0.076	**0.116**	55.69	3.42	0.840	0.631	0.529

Table 4.3: Results of CN generation with silver knowledge. We report novelty (Nov.), RR, BLEU-2 (B-2), ROUGE-L (R-L), KN overlap with generation and the average amount of words (#W.) and sentences (#S.) per generation.

among its CNs in CONAN dataset. This adjustment lets us perform a fairer comparison among query configurations in particular for Q_{hs} with models employing beam search (i.e., XNLG and Candela).

The results are given in Table 4.4. For XNLG and GPT-2$_{KN,MT}$, we observed an increase in the novelty and diversity with Q_{hs} and Q_{gen} on the unique HS set over the

whole test set. As for GPT-2_{KN}, while diversity improves for all query configurations, we did not observe an increase in novelty through Q_{hs}. For Candela, while novelty also increases, the diversity does not improve.

4.6.3 Results for the Gold Knowledge Test Set

To isolate the effect of the knowledge retrieval strategies from the knowledge-grounded generation performances, we conducted a second evaluation on a newly crafted test set paired with gold standard knowledge. In this evaluation, in addition to stereotypical islamophobic in-domain (i.e., in-target) scenario, we also explore the effect of knowledge infusion on cross-domain (i.e., cross-target) CN generation under zero-shot setting. We hypothesize that having a system trained to make use of substandard silver knowledge to generate proper CNs for a given context, could be robust to cross-domain zero-shot conditions. Therefore, we organized a data collection session with an expert operator in writing CNs. In this session, 50 islamophobic HSs randomly sampled from CONAN and 144 new cross-target HSs (covering misogyny, antisemitism, racism, and homophobia) are provided along with the knowledge retrieved by $Q_{hs \cup cn}$ queries. The expert is tasked with composing a suitable CN using the corresponding knowledge as much as possible. Thus, we could obtain a gold test set[10], in which the input knowledge can certainly be found in the CNs.

We tested all models with gold knowledge in-domain and cross-domain test cases. Results are given in Table 4.5. For in-domain scenario, as we have anticipated, knowledge grounded models yield better performances in B-2 and R-L in comparison to the silver knowledge test setting. Especially with the striking jump in the performance of GPT-2_{KN}, we can confirm the proper infusion of the given knowledge to the generated CNs. As for cross-domain tests, GPT-2_{KN} still yields better performance than baselines while the performance for all models (except for XNLG) drops due to unseen events during training. All GPT-2 variations present better diversity performances on the cross-domain setting as compared to both in-domain and silver-knowledge settings. Regardless of domains, XNLG yields fallaciously high scores due to its extensive copying. A cross-target generation from the models can be seen in Table 4.6. More examples of in-/cross-domain generations from all the models are provided in Table 4.9–4.12 in Section 4.9.

Human evaluation. We further resort to human evaluation to assess the final generation quality of each model. For this reason we perform human evaluation of generation using

[10]We release the gold test set at `https://github.com/marcoguerini/CONAN`

Models	Novelty (W/U)	RR (W/U)
without knowledge		
TRF	0.467/0.457	7.72/6.97
GPT-2	0.688/0.675	9.04/7.79
Train_{cn}	-	3.91/3.31
with knowledge		
Candela (Q_{hs})	0.692/0.697	21.87/21.97
GPT-2$_{KN}$		
w/ Q_{hs}	0.723/0.719	8.13/7.97
w/ Q_{gen}	0.728/0.720	7.48/6.34
w/ $Q_{hs \cup gen}$	0.735/0.740	6.30/5.81
w/ $Q_{hs \cup cn}$	0.727/0.731	7.17/6.16
GPT-2$_{KN,MT}$		
w/ Q_{hs}	0.744/0.748	11.69/10.24
w/ Q_{gen}	0.731/0.750	10.37/9.25
w/ $Q_{hs \cup gen}$	0.747/0.747	7.59/8.41
w/ $Q_{hs \cup cn}$	0.731/0.728	9.56/11.08
XNLG		
w/ Q_{hs}	**0.824**/**0.828**	14.42/6.22
w/ Q_{gen}	0.819/0.821	6.88/**4.03**
w/ $Q_{hs \cup gen}$	0.812/0.817	6.98/5.19
w/ $Q_{hs \cup cn}$	0.819/0.816	**5.69**/4.64

Table 4.4: Results of unique HS test set (U) for novelty and diversity. For comparison, we keep the results of whole test set (W).

	in-domain				cross-domain			
	Nov.	RR	B-2	R-L	Nov.	RR	B-2	R-L
without knowledge								
TRF	0.30	7.57	0.014	0.10	0.46	8.62	0.015	0.08
GPT-2	0.72	8.53	0.020	0.11	0.72	8.01	0.022	0.09
with knowledge								
Candela	0.69	19.31	0.072	0.10	0.70	22.22	0.022	0.09
GPT-2_{KN}	0.71	**6.85**	0.201	0.19	0.75	**6.33**	0.041	0.19
GPT-$2_{KN,MT}$	**0.85**	11.55	0.066	0.12	**0.86**	10.38	0.022	0.11
XNLG	0.83	6.94	**0.256**	**0.33**	0.84	8.15	**0.291**	**0.35**

Table 4.5: Results of CN generation with gold knowledge in-target and cross-target test sets.

HS: Jews are different from non-Jews not because of their beliefs, but because of their physical nature. “Jewishness” and malevolence is inherited through blood.
Human: Today one of the biggest concerns is the reinvention of clinical concepts of race, based on inaccurate generalizations. The notion that the superiority of the white race was part of the natural order was deviously reinforced by the rise of modern ”scientific” racism in the late 19th century. Despite the strong objections of many of our contemporaries, this idea of a ”jewish race” is still going around.
GPT-2: Muslims are not different from non-Muslims because of their faith.
GPT-2_{KN}: I think that the notion of race is a generalization based on ignorance.
GPT-$2_{KN,MT}$: All that is wrong is that we cannot accept that violence is necessary.
XNLG: today , one of the biggest concerns is the reinvention of clinical concepts of race , based on inaccurate generalizations about the prejudice to certain physical characteristics and civilizations . despite the strong objections of many contemporary contemporaries , the notion of fixed ” race ” - packages of physical and behavioral characteristics

Table 4.6: Samples of cross-domain generation.

gold knowledge, to rule out the effect of possible noise in the knowledge that may result from the retrieval process.

Our models were evaluated by 3 annotators who are already experienced, and specif-

ically trained, in reading hateful content and writing CNs for online hate countering. The annotators are instructed to assess all generated pairs in gold knowledge test sets in terms of *suitableness* to the HS, *informativeness*, and *intra-coherence* of CN regardless of HS (annotation instruction is provided below). Each score is on a scale of 1 (the least) to 5 (the most). To avoid possible bias and hints towards models, we normalized the pairs (e.g., lowercase and space between words and punctuation) and divided them into 3 partitions of randomized files, one for each expert. Each expert was given 388 pairs, resulting in a total of 1164 pairs for evaluation. To avoid excessive workload, annotators were allowed to complete the task over multiple sessions at their preference.

Human evaluation instruction. Three experts are tasked with CN generation evaluation. We briefed the annotators on the aim of the evaluation, and then we instructed them to evaluate the generation with intuition. The generation is evaluated in terms of 3 aspects: (i) **suitableness:** How suitable the counter narrative is under evaluation for its hate speech, i.e., coherence and relevance of a counter narrative with regard to the hate speech. (ii) **informativeness:** How informative/specific the counter narrative is as a response for its hate speech. (iii) **intra-coherence:** How coherent the counter narrative is under evaluation as a response, i.e., internal coherence of a counter narrative regardless of its hate speech. In Table 4.7 we give the annotation examples with the score explanations provided to the annotators.

Human evaluation results. Results are reported in Table 4.8. We also computed Kendall's Tau-b (Kendall, 1938) to measure the annotators' agreement towards the model ranking for each aspect. The high correlations indicate a strong concordance among the annotators (threshold tau-b > 0.35). Regardless of domains, annotators consider XNLG generations as the most informative and GPT-2_{KN} generations as the most suitable. TRF yields a reasonable suitableness and coherence since it tends to memorize the training CNs, almost behaving like a retrieval system on the human responses, i.e., the examples present in the training data. However, such behavior can be fatal in cross-domain settings. Candela fails to generate suitable cross-domain CNs despite preserving the intra-CN coherence. While GPT-2 and GPT-2_{KN} generations are found almost equally coherent, the lower suitableness and informativeness of GPT-2 output (2.26 and 1.92) for cross-domain as compared to GPT-2_{KN} (2.51 and 2.29) encourages the grounding CNs in knowledge.

HS: Muslims do not want to assimilate. They demand that everyone shows respect for Islam, but they don't respect our laws. One day they will oppress us with the Sharia and polygamy.
CN: Do you really believe that they are a problem? I have a job. **evaluation**: suit.: 3, info.: 1, cohe.: 1 **Note**: The first sentence is relevant but not the second one. There is no coherence between 2 sentences.
CN: As far as they do not force us to follow sharia laws, we should not be concerned with them following their laws. **evaluation**: suit.: 5, info.: 3, cohe.: 5 **Note**: The CN is an appropriate and coherent response, but not providing a substantial amount of information or facts.
CN: Muslims should not be forced to assimilate, since it is not right and no one wants that. And polygamy is illegal and forbidden in UK and Muslims actually respect this ban. **evaluation**: suit.: 5, info.: 5, cohe.: 5 **Note**: The whole CN is perfectly suitable for the given HS. It is highly informative with the facts and internally coherent.

Table 4.7: Annotation examples with score explanations.

	in-domain			cross-domain		
	suit.	info.	cohe.	suit.	info.	cohe.
without knowledge						
TRF	2.65	2.25	3.39	1.47	2.09	3.45
GPT-2	2.67	2.16	4.10	2.26	1.92	**4.24**
with knowledge						
Candela	2.41	2.25	3.14	1.42	2.09	3.40
GPT-2_{KN}	**3.02**	2.35	**4.33**	**2.51**	2.29	4.21
GPT-$2_{KN,MT}$	1.76	1.65	3.73	2.03	1.76	3.88
XNLG	1.43	**3.88**	2.12	1.88	**4.10**	2.79
Kendall's tau-b	0.82	0.69	0.82	0.51	0.91	0.73

Table 4.8: Human evaluation results of CN generation.

4.7 Discussion

Our findings suggest that a large pre-trained LM with knowledge injection is preferred to alleviate the demand for gold data and improves in-/cross-domain generations. GPT-2_{KN} outperforming GPT-2, which becomes more clear with every increase in the quality of provided KN (i.e., from silver Q_{gen} to silver $Q_{hs \cup gen}$ to gold in-domain), indicates that our model trained on silver knowledge learns to grasp relevant KN to the given HS instead of random copying. As for GPT-$2_{KN,MT}$, since the model is trained to distinguish CN from HS, certain typical counter responses are prone to appear. Through a manual analysis, we confirmed that its responses tend to be simpler (e.g., negation: "*terrorism is* ***not*** *a religion.*", "*It's* ***not*** *true...*") than the responses of other GPT-2 variant models. Future work can explore alternative multitask learning frameworks to improve the generation.

From a practical perspective, we can conclude that fine-tuning a pre-trained language model for conditional generation, utilizing silver-knowledge sentences retrieved through a query generation model is the most promising approach for knowledge-bound CN generation. In fact, (i) among various models we tried, GPT-2 with knowledge infusion is shown to produce the most informative and suitable output. (ii) At deployment time, using input context together with generated keyphrases becomes prominent with more novel and diverse CNs and (iii) for new hate-targets in a zero-shot scenario, the generation model can produce coherent and suitable CNs without retraining if a reasonable set of input knowledge is provided. Moreover, this approach can be applied to other scenarios, such as dialogue response generation, since it does not require any extra data collection or annotation.

4.8 Conclusion

In this chapter, we have proposed methods for improving counter narrative generation by incorporating external knowledge retrieved through extracted and generated keyphrases. To this end, CONAN pairs were augmented with relevant knowledge to train systems that are able to produce suitable and informative arguments. Our experiments on in-/cross-domain generation indicate that the generated responses can meet these desiderata. Future work can test other query generation approaches (e.g., exploit human-crafted queries in an interactive setup or the use of pre-trained LMs) to improve knowledge selection and to test other architectures for the final generation step.

One limitation of our approach is that we do not model directly the degree of hallucinations. Future work could inspect automatic methods to identify, quantify and characterize hallucinated text - for instance through measuring semantic correctness using natural language inference (Dušek and Kasner, 2020) - in order to improve model performance. Future work can also explore techniques for improving specificity via the use of diverse types of knowledge (e.g., facts, events and named entities) in knowledge graphs from external resources. This would also mitigate the problem of the model possibly generating hallucinated content (i.e. non-existent or wrong content) (Maynez et al., 2020; Nie et al., 2019).

4.9 Generation examples

HS: It seems that I cannot define Islam as a cult which promote wars and conquers host nations and legitimate this by its scriptures. Can I define them as narrow-minded, prejudiced, sexist and male chauvinist?
Human: Everyone has the right of free speech, while it does not mean anyone can use it to hurt people. Plus, based on what you critisize Islam? I recommend you to read something about true Islam and Muslims. You would find that they are no difference from us.
TRF: Islam is all about the religions and tolerance insert quote from koran. Migrants is not evil?
GPT-2: It seems to me that this is the antithesis of what Islam teaches.
KN Q_{hs}: Hakim al-Mutairi, secretary-general of the emirate's Umma (Islamic Nation) party, urged the pope to apologise immediately "to the Muslim world for his calumnies against the Prophet Muhammad and Islam".... Instead of bringing Islam and Christianity closer, he is straining relations between the two religions". In the present political atmosphere such views can be exploited by those who are trying to malign Muslims and Islam. **Candela:** I have no accountability. I have no problem with those who claim to belong to, but i do not have the right to practice whatever religion. **GPT-2$_{KN}$:** The Qur'an clearly teaches peace, love and tolerance towards everyone, including the non-Muslims. The only thing that Islam does not have is a message of conquest and war. **GPT-2$_{KN,MT}$:** Islam is a religion founded on the principle that you must believe in the good will of all people. **XNLG:** Extremist muslim national party secretary, hamid al mutairi, urged the pope to define immediately to the muslim world for his calumies against the prophet muhammad and islam. Pope benedict xvi was full of hostility and hatred against muslims and islam. In the present political atmosphere such views can be exploited by those who bring

Table 4.9: Sample of in-domain generation with KN retrieved using Q_{hs}.

HS: It seems that I cannot define Islam as a cult which promote wars and conquers host nations and legitimate this by its scriptures. Can I define them as narrow-minded, prejudiced, sexist and male chauvinist?
Human: Everyone has the right of free speech, while it does not mean anyone can use it to hurt people. Plus, based on what you critisize Islam? I recommend you to read something about true Islam and Muslims. You would find that they are no difference from us.
TRF: Islam is all about the religions and tolerance insert quote from koran. Migrants is not evil?
GPT-2: It seems to me that this is the antithesis of what Islam teaches.
KN $Q_{hs\cup gen}$: Islam, as practiced by the vast majority of people, is a peaceful religion, a religion that respects others. "The large majority of Muslims tinker," said Franck Fregosi, a sociologist who has written extensively on Islam in Europe.... **GPT-2$_{KN}$:** Hi, I do not understand the problem you have. The vast majority of Muslims do not advocate wars and conquer the world. They practice a peaceful religion that respects others. **GPT-2$_{KN,MT}$:** It seems that i have just come across the iota of muslims who have been insulted, persecuted, killed, raped, harassed, murdered, killed by the muslims in europe. **XNLG:** It seems that islam, as practiced by the vast majority of people, is a peaceful religion, a religion that respects others. The large majority of muslims tinker, as franck fregosi, who has written extensively on islam in europe. It seems that it does not define islam as an inclusion and respect for all of the religion

Table 4.10: Sample of in-domain generation with KN retrieved using $Q_{hs\cup gen}$.

HS: It seems that I cannot define Islam as a cult which promote wars and conquers host nations and legitimate this by its scriptures. Can I define them as narrow-minded, prejudiced, sexist and male chauvinist?
Human: Everyone has the right of free speech, while it does not mean anyone can use it to hurt people. Plus, based on what you critisize Islam? I recommend you to read something about true Islam and Muslims. You would find that they are no difference from us.
TRF: Islam is all about the religions and tolerance insert quote from koran. Migrants is not evil?
GPT-2: It seems to me that this is the antithesis of what Islam teaches.
KN $Q_{hs\cup cn}$: No doubt, it is the legitimate right of Muslims to build a community center near Ground Zero. So please stop conflating criticism of Islam with discrimination of Muslims. As a nation of free people, we can never defeat Islam but we can make war against terrorists and governments that support them. a "moderate Islamic countries" when all unanimously reject the Universal Declaration of Human Rights and support The Cairo Declaration of Human Rights in Islam. **GPT-2$_{KN}$:** The Islamic countries do not promote wars and conquer host nations. Islam is not a cult but a peaceful religion which respects all peoples. **GPT-2$_{KN,MT}$:** The vast majority of men in the world are not able to articulate the spiritual values of islam. **XNLG:** It seems that it is the legitimate right of muslims to build a community center near zero ground, so please confine criticism of islam with discrimination against muslims. As a nation of free people, we can never defeat islam, but we can make war against terrorists and governments that support them. My comment gives a lot of information about what islam and muslims

Table 4.11: Sample of in-domain generation with KN retrieved using $Q_{hs\cup cn}$.

HS: Jews are Christ Killers.
Human: While the Catholic Church has for decades taught that Jews weren't collectively responsible, the argument laid out by the Pope Benedict XVI was a landmark statement. In his book, he not only confirms church teaching refuting the decide charge, he also explains biblically and theologically why there is no basis in Scripture for the argument that the Jewish people as a whole were responsible for Jesus' death.
TRF: Why do you consider muslims like beasts? Did they do something wrong? If you take into consideration the crimes committed by muslims, why do you distinguish the muslim criminals from the non muslim? They are all criminals and all of them are punished.
GPT-2: I know many Muslims who are peaceful and respectful and do not kill.
Candela: Do you have proof? Where did you get this information from? I do not see the link between islamism and the news of the muslim community.
GPT-2$_{KN}$: No, the crucifixion of Jesus Christ is not a source for anti-Semitism. The Jewish people were not responsible for the death of Jesus.
GPT-2$_{KN,MT}$: On the basis of christianity, as christianity is not a religion, jews are not responsible for their own death.
XNLG: In "jesus of nazareth-part ii" excerpts the church's teaching that refuted the deicide charge of the jewish people, as well as theologically explains one of the most controversial issues in christianity.
KN: Pope Benedict XVI has made a sweeping exoneration of the Jewish people for the death of Jesus Christ. While the Catholic Church has for five decades taught that Jews weren't collectively responsible, Jewish scholars said... would help fight anti-Semitism today. In "Jesus of Nazareth-Part II" excerpts released Wednesday, Benedict explains biblically and theologically why there is no basis in Scripture for the argument that the Jewish people as a whole were responsible for Jesus' death.

Table 4.12: Sample of cross-domain CN generation with gold knowledge.

Chapter 5

Multilinguality for Counter Narrative Generation

5.1 Introduction

As highlighted in Chapter 1, approaches to countering online hatred from a multilingual perspective at scale are very much needed, since hate speech spreads rapidly in many countries. In this respect, to restrain online hatred via counter narrative generation, a large amount of training instances is indispensable for developing systems achieving reasonable performance. Unfortunately, this is not attainable for most languages, since producing counter narratives requires domain experts. In Chapter 3 we showed that manually creating high-quality counter narratives cannot address the dissemination of hate speech at scale. While our experiments demonstrate that it is feasible to automatize CN collection via an author-reviewer framework, a reasonable amount of data is still required. For the aforementioned reasons, despite the encouraging results of the counter narrative generation task, experiments have been limited to English due to the scarcity of HS/CN data in other languages.

In this context, it is natural to ask how can we transfer the generation capability of a model across languages, in particular by leveraging high-resource languages such as English. To answer this question, we resort to data augmentation via translation. Data augmentation aims at increasing the variety of training examples without deliberately collecting new data and has been explored in various NLP tasks (Feng et al., 2021).

As a first step towards multilinguality, 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 hypothesized that data augmentation via translation can not only overcome the problem that stems from the lack of data for low-resource languages but also facilitate models' generalization capabilities.

In this chapter, we first explore the automatic portability of data across languages, considering that recent neural machine translation systems have shown outstanding performances. In particular, we explore data augmentation utilizing off-the-shelf NMT models to synthesize silver data translating from English and French to Italian, and fine-tuning GePpeTto (De 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.

5.2 Counter Narrative Generation for Under-Resourced Settings

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 of CONAN hate speech/counter narrative pairs in other languages. For translation setting, we use DeepL[1], an off-the-shelf and well-performing NMT system, to translate data from other languages to Italian. The translated pairs are then used for fine-tuning a large Italian pre-trained generative model, i.e., GePpeTto, along with the original Italian gold standard pairs.

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

5.3 Experiments

5.3.1 Dataset

Since automating the creation of counter narratives is still a novel task, very limited amount of data is available, especially for languages other than English. In our case, we use CONAN, which is the only existing dataset targeting hate speech for Italian. Specifically, we prepared the following splits:

1. 2142 pairs (original Italian pairs and 1 Italian paraphrase pair) as a training set made of gold standard data.
2. 5996 pairs as a training set made of silver data obtained by automatically translating French and English pairs to Italian.
3. 1071 pairs (the rest of the Italian paraphrased pairs) are kept for testing purposes.

5.3.2 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 English and French pairs to Italian 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 Italian 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 Italian 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 pre-training with the silver data helps to reduce the amount of gold standard data needed to reach a proper generation performance.

5.3.3 Training Details

For all the experiments, we have used GePpeTto as the pre-trained 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.

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.

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

5.3.4 Evaluation Metrics

Following the experimental setting in Chapter 4, we report BLEU and ROUGE to evaluate the CN generation on the gold standard test set. As for the generation quality, we compute repetition rate and novelty 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 are in a scale from 1 to 5.

5.4 Results

Model comparison. Results in Table 5.1 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 English and French pairs usually have slightly different focus on the topic of Islamophobia (topics and tropes can vary across nations and cultures). In Table 5.2 we provide few examples of generated CNs. As shown in the table, GP-trans generation is grammatically correct but focused on the UK/FR scenario coming from the training data. On the contrary, GP-ita and GP-hybrid can mimic gold arguments with novel and diverse wording. More generation examples can be found in Table 5.4-5.7 in Section 5.6.

Learning curve discussion. As can be seen in Figure 5.1, 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

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

Table 5.1: Quantitative results of fine-tuned models.

HS: *Dovremmo scagliarci su di voi con delle bombe, siete spregevoli. Smettetela maleodoranti.* [We should throw bombs on you, you are despicable. Stop smelling it.]
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 5.2: 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.

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.

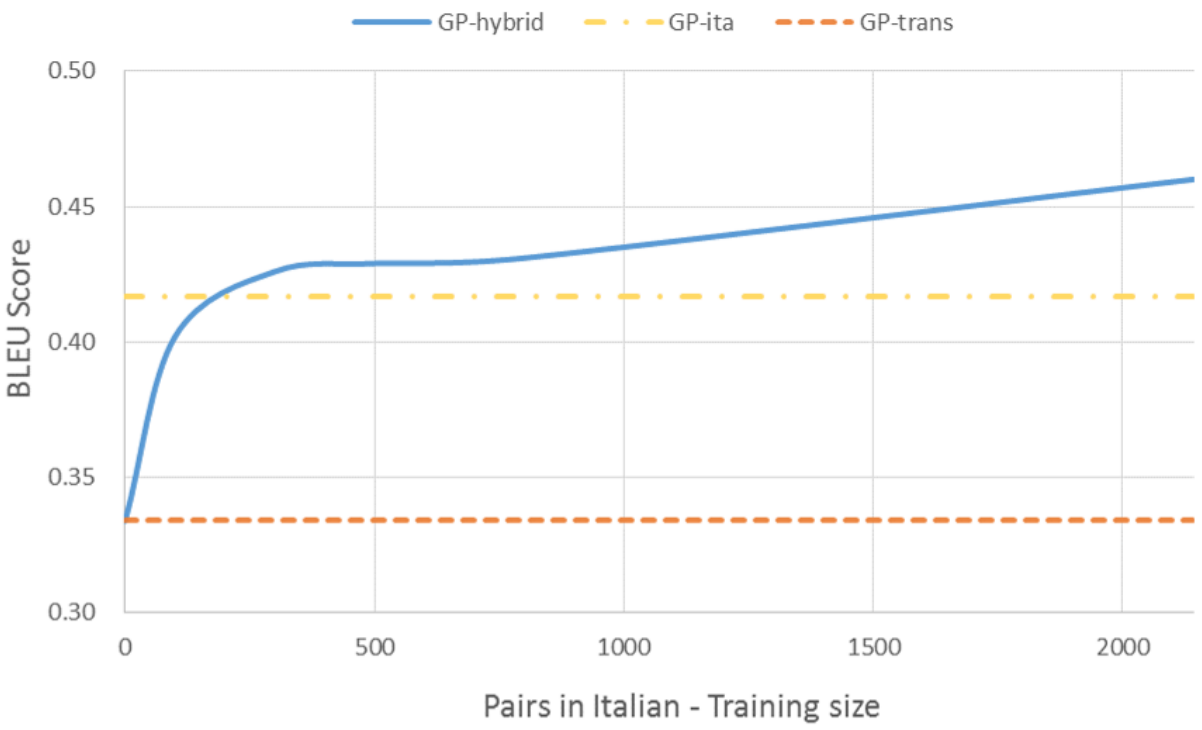

Figure 5.1: Learning curve of GP-hybrid model while Italian pairs being added. GP-hybrid performance with no examples is shown as GP-trans.

Human evaluation. As annotators, we employed 2 Italian native speakers that are expert 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 5.3 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.

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 5.3: Human evaluation results.

5.5 Conclusion

In this chapter, we presented simple strategies for low-resource counter narrative generation. We investigate approaches to effective multilingual CN generation, focusing on Italian as a case study. We experiment with several variations of generation under zero and few shot scenarios employing a novel language model recently released for Italian. Our experiments reveal that with simple data augmentation strategies paired with powerful LMs can bring promising results for under-resourced languages.

Our approach is based on having a large pre-trained LM for the target language (GePpeTto in our case). However, such models are not always available for every language. Future work can explore approaches utilizing multilingual generative models that cover massive amount of languages, such as mT5 (Xue et al., 2021) and mBart (Liu et al., 2020). Alternatively, translation at deployment time offers a practical way to alleviate the lack of data in target languages. This can be done by, for instance, translating a text input in a target language into a temporary language for which LMs are available to generate a response, and then translating back this response to the target language as final output.

5.6 Generation Examples

HS: *L'islam alimenta la violenza e non segue la pace.* [Islam fuels violence and does not follow peace.]
GP-trans: *L'Islam non sostiene la violenza e non la legittima i delinquenti.* [Islam does not support violence and does not legitimize offenders.]
GP-ita: *L'islam non è una religione di pace, anzi è impegnata a combattere il terrorismo in nome di Dio.* [Islam is not a religion of peace, rather it is committed to fighting terrorism in the name of God.]
GP-hybrid: *L'islam non è una religione che professa la violenza. I musulmani hanno il diritto di professare la propria religione.* [Islam is not a religion that professes violence. Muslims have the right to profess their religion.]
Gold: *Vorrei farle notare che Islam e Isis sono differenti. L'islam è una religione pacifica, mentre l'Isis è un'organizzazione militare e terroristica che sostiene il fondamentalismo islamico.* [I would like to point out that Islam and Isis are different. Islam is a peaceful religion, while Isis is a military and terrorist organization that supports Islamic fundamentalism.]

Table 5.4: Sample CN generations along with EN translation.

HS: *Tutti i musulmani sono terroristi che cercano di invaderci con le armi. La soluzione è attacarli a nostra volta con la violenza.* [All Muslims are terrorists who try to invade us with weapons. The solution is to attack them in turn with violence.]
GP-trans: *Il terrorismo non si limita a una religione in particolare.* [Terrorism is not limited to one particular religion.]
GP-ita: *L'odio contro l'occidente è un problema ma non ha nulla a che vedere con la religione.* [Hatred of the West is a problem but it has nothing to do with religion.]
GP-hybrid: *Il terrorismo non ha fede e nazionalita. E' una devianza che si impossessa di singoli individui.* [Terrorism has no faith and no nationality. It is a deviance that takes possession of single individuals.]
Gold: *Buonasera, le principali vittime degli attentati terroristici sono proprio gli abitanti di paesi arabi. Per la maggior parte sono, dunque, vittime e non terroristi.* [Good evening, the main victims of the terrorist attacks are precisely the inhabitants of Arab countries. For the most part, therefore, they are victims and not terrorists.]

Table 5.5: Sample CN generations along with EN translation.

HS: *Noi li ospitiamo nel nostro paese, forniamo un aiuto economico e loro ci uccidono: sono da considerarsi più simili agli animali che alle persone.* [We host them in our country, we provide financial help and they kill us: they are to be considered more like animals than people.]
GP-trans: *Cosa suggerisci esattamente?* [What exactly do you suggest?]
GP-ita: *Salve, in realtà la percezione del fenomeno è differente a seconda che si tratti di animali o di persone con una loro vita o appartenenza religiosa.* [Hi, in reality the perception of the phenomenon is different depending on whether it is animals or people with their own life or religious affiliation.]
GP-hybrid: *Nessuno vuole imporre nulla a nessuno. Gli animali non hanno il diritto di avere un luogo di culto.* [Nobody wants to impose anything on anyone. Animals do not have the right to have a place of worship.]
Gold: *Guardi che le vittime non siamo solo noi, ma lo sono tutti coloro che sono costretti a lasciare il proprio paese e a cercare qui riparo e rifugio, inoltre mi sembra che lei stia enormemente generalizzando.* [Look that the victims are not just us, but all those who are forced to leave their country and seek shelter and refuge here, moreover it seems to me that you are enormously generalizing.]

Table 5.6: Sample CN generations along with EN translation.

HS: *L'islam tra pochi anni prenderà il controllo costringendo tutte quelle buoniste che li sostengono, per seguire la moda, ad indossare il velo!* [Islam in a few years will take control by forcing all those do-gooders who support them, to follow the fashion, to wear the veil!]
GP-trans: *La parola stessa islam, che significa "resa", viene dall'arabo salam, o pace - l'islam non è una religione violenta* [The word Islam itself, which means "surrender", comes from the Arabic salam, or peace - Islam is not a violent religion]
GP-ita: *L'islam è una religione pacifica, non fanatismo. I musulmani non sono tutti fanatici, alcuni lo sono.* [Islam is a peaceful religion, not fanaticism. Muslims are not all fanatics, some are.]
GP-hybrid: *L'islam non è una religione intollerante. I musulmani hanno solo il diritto di professare la loro religione.* [Islam is not an intolerant religion. Muslims have only the right to profess their religion.]
Gold: *Il velo non è obbligatorio nella maggioranza dei paesi musulmani. Figuriamoci in paesi con una tradizione del tutto diversa.* [The veil is not mandatory in most Muslim countries. Let alone in countries with a completely different tradition.]

Table 5.7: Sample CN generations along with EN translation.

Chapter 6

Counter Narrative Generation Evaluation

6.1 Introduction

As discussed in Chapter 1, manual intervention strategy via CN writings cannot govern the severe magnitude of online hate, since it would become too costly and time-consuming, even for expert operators in NGOs. In Chapter 3, we present data and strategies to hatred countering. Since that work, various methodologies have explored approaches to automation of CN writing (Qian et al., 2019), addressing various aspects of generation including informativeness (see Chapter 4), multilinguality (see Chapter 5), and grammaticality and diversity (Zhu and Bhat, 2021).

However, no previous work has addressed automatic creation of hate countering messages through an interactive and real-time interface, also evaluating the effectiveness and efficiency of providing counter narrative suggestions to assist in NGO operators' activities against online hate. The expected benefit of employing such tools lies in reducing the time and effort of NGO operators in responding to offensive posts, thus yielding more systematic and sustainable impact against online hate speech. Furthermore, automation of hatred countering 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 (Riedl et al., 2020). The issue of moderators' well-being has been tackled in the past via automatic image modification, for example image blurring (Dang et al., 2018) or grayscaling

(Karunakaran and Ramakrishan, 2019).

To fill this gap, in this chapter we evaluate the effectiveness of providing counter narrative suggestions on a novel platform that enhances NGO operators' online hate monitoring activities (similar to the concept of 'digital augmentation' presented in Stella et al. (2019)). Specifically, we build a counter narrative suggestion tool to counter online hatred in English, French, and Italian, focusing on Islamophobic messages. Given a hate speech on SMPs, users can rely on the suggestion tool to automatically compose suitable responses to hate content (Section 6.2). The platform was tested with 112 NGO operators across 3 different countries (Section 6.3) through an extensive qualitative (Section 6.4) and quantitative evaluation that includes open-ended questions, a user experience questionnaire, and time measurement. The experimentation shows that the platform can effectively empower operators in the fundamental task of hate countering (Section 6.5).

6.2 Hate Countering Platform

We present in this section the online platform that we employed and its main functionality: 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% reported messages are deleted compared to 71.7% Facebook, and Instagram on average[1]. 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.

6.2.1 Monitoring Functionalities

The monitoring functionalities of the platform cover both real-time Twitter activity and past messages, and were developed within the HATEMETER[2] project. The analyses are

[1] http://scan-project.eu/wp-content/uploads/sCAN_monitoring_report2_final.pdf

[2] http://hatemeter.eu/

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[3] 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 6.1, showing the number of tweets containing at least one islamophobic term or hashtag, divided into original messages, retweets or replies for each language.

	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
IT	35,569	27,562	213,713	286,844

Table 6.1: Islamophobic messages monitored between October 2018 and May 2019 in English, French and Italian.

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 Figure 6.1. Hashtag monitoring has been included in the platform as a straightforward way to track which semantic domains revolve around islamophobic messages and how they change over time. By monitoring hashtags it is possible to understand, for instance, whether some external events have an impact on online discourse against Muslims, and how islamophobic content is framed.

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 time window of 10 days and a predefined hashtag/keyword of interest, and to display the online activity around this term, such as the hashtags that most co-occurred with the one of interest (Figure 6.2) and the most popular messages containing it. It is also possible to

[3] `https://developer.twitter.com/en/docs/api-reference-index`

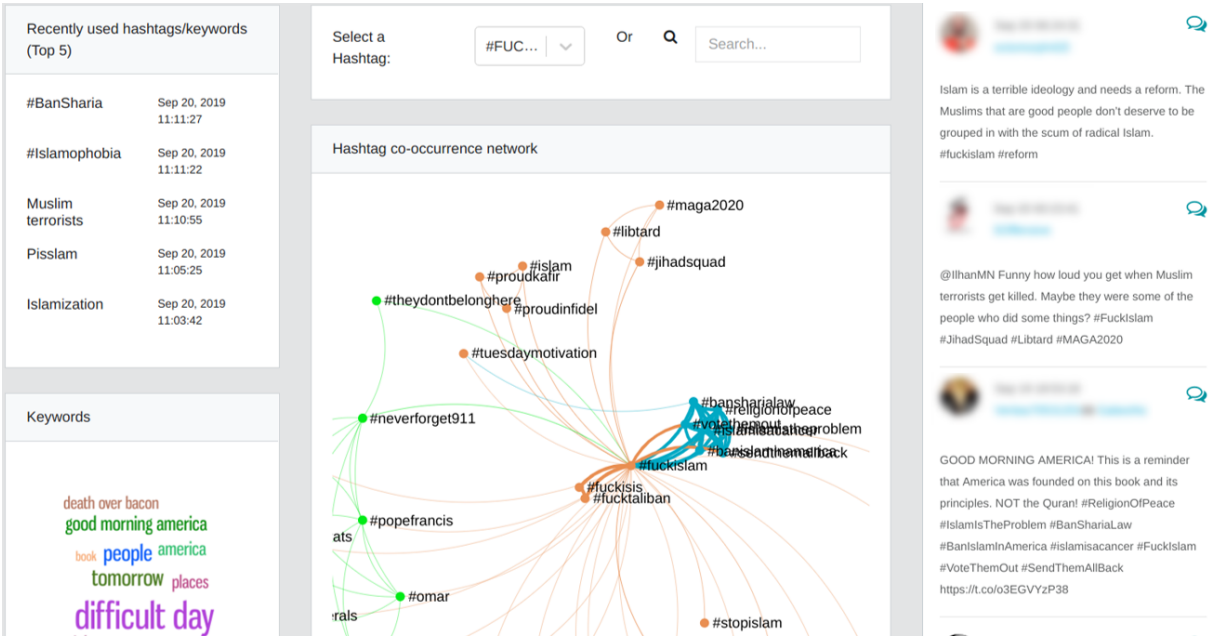

Figure 6.1: Screenshot of the real-time monitoring view: after selecting the #FUCKISLAM hashtag, the network of co-occurring hashtags is displayed, together with the most recent tweets containing it. A small bubble button can be clicked to direclty obtain counter narratives to that particular tweet.

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 (Figure 6.3), so that operators can use this information to connect peaks with past events occurring on specific dates.

6.2.2 Counter Narrative Suggestion Approach

The main hate countering feature consists in assisting operators in writing counter narratives based on automatically generated suggestions. There is a specific tab of the interface that can be either activated from each hate tweet (as shown in Figure 6.1) 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 operators must copy-paste the hate content of interest. Figure 6.4 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 dialog icon on the right, the platform will output a list of counter narrative sugges-

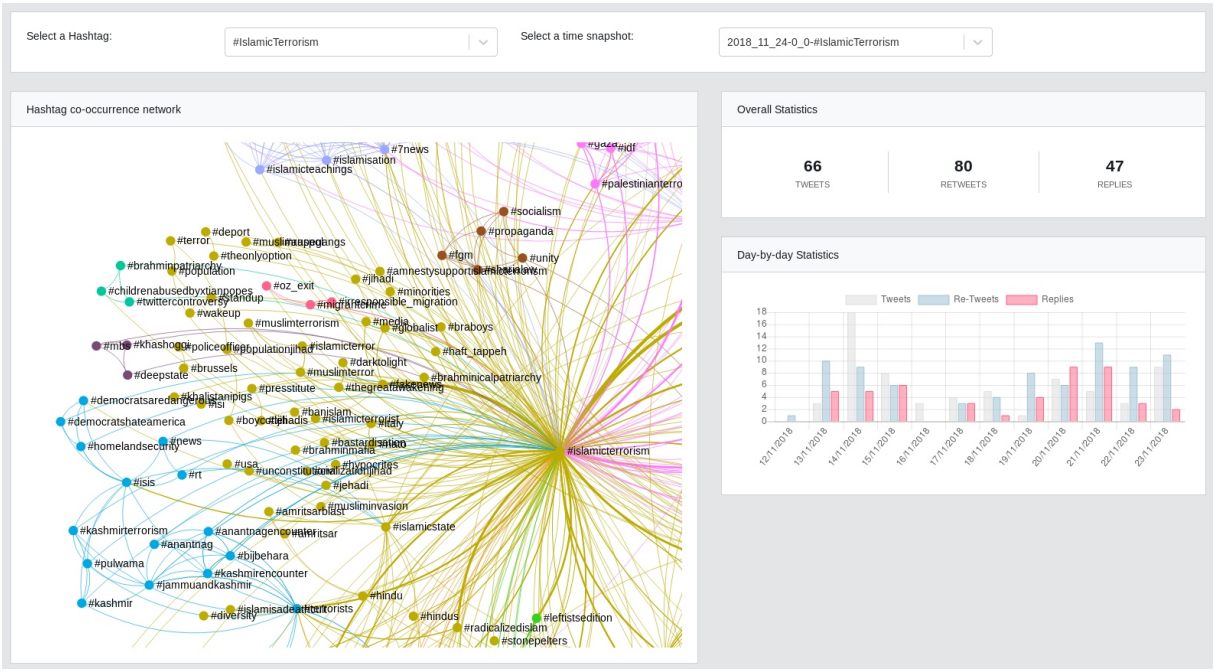

Figure 6.2: Screenshot of the monitoring view for past tweets: after selecting the #IslamicTerrorism hashtag and the time slot, the network of co-occurring hashtags is displayed, together with overall statistics on the use of the hashtag in the period of interest.

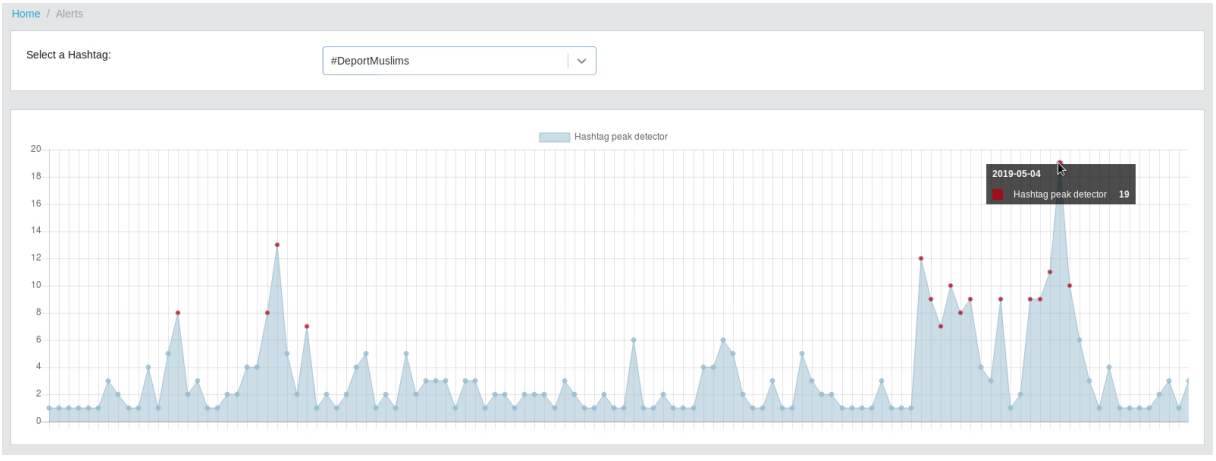

Figure 6.3: Screenshot of the peak detector view: after selecting the #DeportMuslims hashtag, the number of occurrences for the given hashtag on each monitoring day is reported. Peaks are highlighted in red, with the associated date.

tions 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.

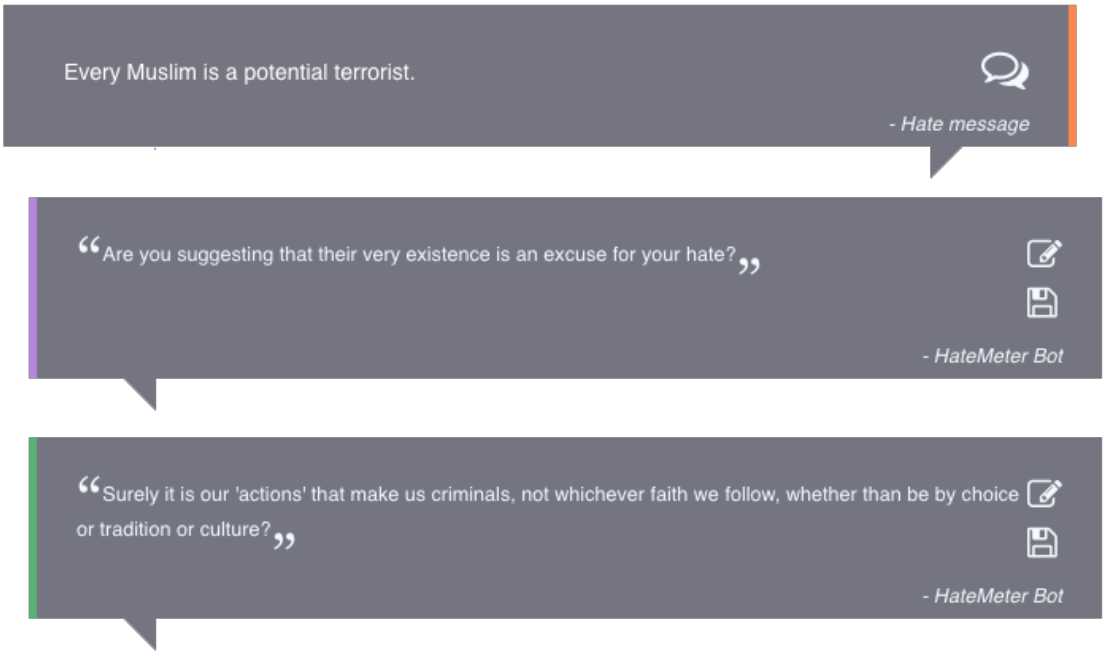

Figure 6.4: On top the hate content “Every Muslim is a potential terrorist”. Below a list of possible counter narratives automatically suggested for that hate content.

For implementing the counter narrative suggestion tool, we have adopted a data-driven approach and we used CONAN hate-message/counter-narrative pairs. As a suggestion methodology, we have implemented a tf-idf information retrieval model (Jones, 1972) 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 with regard to 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 (Shang et al., 2015; Radford et al., 2019) and argument generation (Hua et al., 2019) techniques have also been proposed recently, they still have limitations that might lead to substandard generations. Foremost, neural models can be uncontrollable and produce outputs containing undesirable biases and toxic language (Gehman et al., 2020; Liang et al., 2021). Substandard CN examples generating by LMs can be found in Chapter 3 and 4. Furthermore, these models tend to produce responses that are generic and not engaging (Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016), or to overproduce negations (Hua et al., 2019) that would be impractical to effectively counter online hatred. Last but not least, LMs are prone to hallucination so they might provide a CN that seems grammatically and semantically correct, still being misleading given it contains made up information. Additionally, 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, Hua et al. (2019) discusses that automated generation systems have not been able to fully capture the organization and complexity of human arguments yet. Considering the aforementioned issues, 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. We adopt an information retrieval strategy, which guarantees informative, controllable and well-organized responses, at the expense of novelty or full relevance.

6.3 Experiment Setup

The goal of our experiment is to evaluate the effect of using the suggestion tool in daily activities of NGO operators to restrain hatred online. Since the counter narrative activity has always been carried out manually by operators, 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, and usability (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). We summarize the tasks required for the CN^+ and CN^- conditions in Table 6.2.

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 curb Islamophobia online and undergo specific training to comply with NGO code of conduct and fairness when it comes to online behavior. 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 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 exercises were meant to let operators inspect the functionalities of the platform. Four dimensions of hate monitoring functionalities are included in the exercises: hate distribution, hate timeline, islamisation analysis and recent trends. The instruction for each exercise is given in Appendix A.2. For the CN^+ condition, 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. Furthermore, operators were explicitly asked to modify the suggested CNs if possible.

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.

6.4 Questionnaires

To evaluate the operators' experience with the platform, two questionnaires were used: the User Experience Questionnaire (UEQ) (Laugwitz et al., 2008) and an additional questionnaire paired with some qualitative questions specifically developed for the hate countering task of our interface.

6.4.1 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 are 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.

6.4.2 Additional Questionnaire and Open-ended Questions

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

- 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 3 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.

Tasks	**CN$^-$**	**CN$^+$**
Usability tasks		
Hate distribution	yes	yes
Hate timeline	yes	yes
Islamisation analysis	yes	yes
Recent trends	yes	yes
CN writing	yes	yes
Questionnaires		
UEQ	yes	yes
Additional Questionnaire		
Support	yes	yes
Ease of use	yes	yes
Look	yes	yes
Suggestion	no	yes

Table 6.2: Overview of the tasks for CN$^-$ and CN$^+$ conditions.

6.5 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.

6.5.1 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 6.3, 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

[4]`https://www.ueq-online.org`

	$\mathbf{CN^-}$		$\mathbf{CN^+}$		$\mathbf{CN^{-/+}}$	
Scale	*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

Table 6.3: UEQ scales with and without counter narrative suggestion tool (CN^+ and CN^- respectively). $CN^{-/+}$ denotes CN^- receiving suggestion tool in a post-hoc experiment.

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.

With regards to counter narratives, operators reported higher evaluation for the condition where the suggested responses provided by 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 retrieval-based automated suggestion mechanism is that the collected responses (i.e. CNs written by operators in our experiments) are primed by the suggested responses. The collected responses can be lexically or semantically 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.

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 our goal

	CN$^-$	**CN$^+$**	**CN$^{-/+}$**
Scale	*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

Table 6.4: Pragmatic and hedonic quality.

of countering hate speech. In Table 6.4, 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 understand how our UEQ scores compared to those of other products. The benchmark dataset involves 18,483 participants judging 401 products, including web pages, social media, and business software. As shown in Table 6.5, 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.

categories	**evaluation results**
excellent	$90\% < \text{results} \leq 100\%$
good	$75\% < \text{results} \leq 90\%$
above average	$50\% < \text{results} \leq 75\%$
below average	$25\% < \text{results} \leq 50\%$
bad	$\text{results} \leq 25\%$

Table 6.5: Benchmark intervals for the UEQ scores (per scale).

The benchmark chart is displayed in Figure 6.5. 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

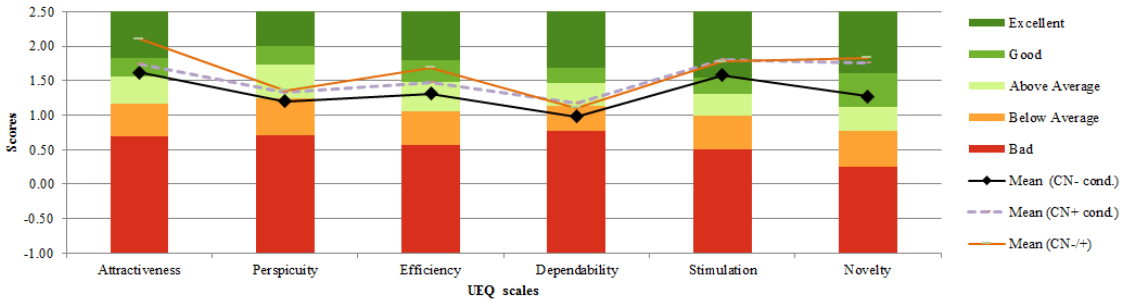

Figure 6.5: Statistics on CN^{-}, CN^{+} and $CN^{-/+}$ conditions, compared with the benchmark for UEQ scales.

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 (randomly selected) 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 (Kahneman and Knetsch, 1992; Desvousges et al., 1993) 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 6.3 and 6.4, and also in Figure 6.5. 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 appear even more clearly.

	CN$^-$				**CN$^+$**				**CN$^{-/+}$**			
	Mean	*Med.*	*SD*	*N*	*Mean*	*Med.*	*SD*	*N*	*Mean*	*Med.*	*SD*	*N*
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 6.6: Statistics of 6 indicators from additional questionnaire on CN$^-$, CN$^+$ and CN$^{-/+}$ conditions (Med. denotes median.).

6.5.2 Additional Questionnaire and Open Questions Analysis

We performed an additional qualitative evaluation to assess how the platform can be used to 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 6.6. 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.

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. 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.

Ease of use. As for the ease of use, participants commented that the experience in using the platform is very intuitive and user-friendly. 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.

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 even more to guide the interpretation of the results. In the CN^+ condition, no specific comment concerned the counter narrative interface alone.

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.

6.5.3 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 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. This implies that our suggestion tool based on tf-idf has around 75% accuracy, i.e. in three cases out of four it was helping operators in their countering activity by providing relevant suggestions. If we focus on relevance of suggestions, using the cosine similarity between actual HS and suggested CNs, we found that the score of modified suggestions is almost double than the similarity of suggestions that have been discarded (0.13, 0.12, 0.06 vs. 0.07, 0.07, 0.03 for EN, IT, FR respectively). This implies that there is definitely a relevance phenomenon in the selection process. Finally, as it is shown in Table 6.7, for the three data collections there is expectedly a significant time decrease by half after the suggestion

tool is introduced. The numbers indicate the average time required to complete the task (including reading CN suggestions, for the CN^+ condition).

For CN^- condition, we estimated the time needed for writing a CN from scratch without the CN suggestion tool in a separate session for each country. We divided the total recorded time by the total number of CNs collected, resulting in on average 8 minutes per CN, which is similar to the one reported in our prior study on counter narrative data collection through NGO operators (see Section 3.4). This time includes possible searching for useful information on the web and structuring a proper argument before writing the corresponding counter narrative.

For the CN^+ condition, the time needed to modify a suggestion or to write a new CN is similar[5]. Since operators in CN^+ condition composed new CNs after seeing suggestions, we hypothesised that these "prompts" reduced the time required to search for useful information and digest it as compared to CN^- condition. We performed also an additional analysis, computing the HTER (Specia and Farzindar, 2010), for the modified CNs. We found that the average HTER is above 0.4 for all languages, which represents an empirical threshold above which correction or rewriting from scratch are almost the same (Turchi et al., 2013). This explains why the two conditions have similar timing. 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. This efficiency is primarily obtained by providing hints and arguments to operators rather than simply allowing them to save writing time. In fact, CN^+ condition is always better than CN^- for both modified and new CN configurations with almost comparable time. In fact, with this design, operators can utilize their time to refine responses on top of suggestions and construct more diverse counter narratives. Considering the different countries, we observe that 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 personalise more the suggested counter narratives, and are faster in writing their own.

[5]The suggested CNs in the database are generally short and easily understandable with 2 to 3 sentences, so the time for reading them is almost negligible as compared to the time for writing or searching for useful information.

	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

Table 6.7: Time required for writing CN (modified CN based on suggestions vs. new CN from scratch) across 3 countries. Time is expressed in minutes.

6.6 Conclusion

In this chapter, we described a retrieval-based CN suggestion tool on an online platform that monitors islamophobic messages on Twitter and supports NGO operators in quickly responding to hate messages. We evaluated in particular the contribution of the counter narrative suggestion tool for English, French, and Italian in addressing islamophobic discourse. 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.

We chose to build our counter narrative tool around a retrieval-based suggestion engine, that produces fixed suggestions for similar hate messages. It may be possible to adopt approaches as presented in Chapter 3 and 4 for generating more diverse counter narratives. Future work could 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.

Finally, while our results obtained by testing with 112 operators on 6 different deployment days demonstrate that CN tool is helpful for countering online hate messages, we recognize that a long-term evaluation is required through daily activities, to ensure the effectiveness and usability of the platform. Future work can extend the assessment period and evaluate the actual activity at scale.

Chapter 7

Conclusions

As digital technologies have been deeply woven into our social lives, hateful content has found fertile ground for spreading rapidly and approaches for dealing with it are urgently needed. To aid NGOs relying on a limited number of volunteers for countering hatred at scale, we have focused on recent advances in NLP to automatize counter narrative production.

In Chapter 1, we identified four main challenges in NLP-based hate countering: the lack of data, generic generation, content hallucination and multilinguality. To fill the gap, in this thesis we have performed study on effective CN generation from a multilingual perspective for English, Italian and French, addressing the hate phenomenon of Islamophobia.

Overall, this thesis provides 4 main contributions. The first contribution, described in Chapter 3, is a new dataset and strategies for hate countering. We started by gathering the first large-scale and multilingual hate-speech/counter-narrative dataset CONAN, written by NGO operators, for English, French, and Italian. 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. Based on this dataset, we then presented an author-reviewer framework to generate high-quality data relying on substantially less expert-annotator effort. Through a detailed evaluation, we identified best practices for data collection based on the amount of experts available.

Next, in Chapter 4 we established the first complete knowledge-bound counter narrative generation pipeline, grounded in an external knowledge repository that can provide

more informative content to withstand online hatred. Together with our approach, we presented a series of experiments that show its feasibility to produce suitable and informative counter narratives in in-domain and cross-domain settings.

The third contribution, described in Chapter 5, extended CN generation for Italian, studying approaches to generating effectively CNs in languages other than English. We experimented with CONAN dataset presented in Chapter 3 and a novel generative language model, recently released for Italian, under several configurations, including zero and few shot learning. Results indicate that even for under-resourced languages, data augmentation strategies paired with large unsupervised LMs hold promising results.

Finally, in Chapter 6, we introduced a novel tool that provides counter narrative suggestions to NGO operators and evaluate it. Our platform aims at increasing the efficiency and effectiveness of operators' activities against Islamophobia. We tested the platform with more than one hundred NGO operators in three countries through qualitative and quantitative evaluation. Results indicate that NGOs favour the platform solution with the suggestion tool, and that the time required to produce counter narratives significantly decreases.

7.1 Challenges and Future Directions

Taking a wider view, this thesis brings several steps towards online safety and provides an overview of counter narrative generation using NLP. Though we focused on data and strategies for counter narrative generation targeting informativeness and multilinguality, many research challenges remain open.

While we believe our approaches to counter narrative generation could apply to hate speech in general, in this thesis 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 experiments in the HATEMETER project. It is possible that hate messages with a different target, for example misogynistic ones, 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 ones we introduce in this thesis. Future work can continue collecting data for Islam and include other hate targets such as migrants, people of colors, Jews or LGBT+, in order to put the dataset at the service of other organizations and further research. Understanding what counter narrative types are commonly employed by NGO operators for various hate targets is also a potential approach.

With the rise of online hate speech, methods that can generalize out of domain would be helpful for producing counter narratives involving new topics and events. In Chapter 4, we showed that grounding CN generation in knowledge text can facilitate handling cross-target hateful content without re-training. Recent work investigates methodologies for the detection of hallucination by fine-tuning pre-trained language models on synthetic hallucinated data (Zhou et al., 2021). Future work can apply such hallucination detection method as an indicator to either discard unfaithful generations or replace them with faithful and informative ones.

In Chapter 5, we focused on Italian CN generation representing low-resource scenarios. In practice, investigating model performance for other languages would also be necessary to advance multilingual counter narrative generation. For instance, it is desirable to understand whether our approach of data augmentation via translation can contribute the same amount of effects for all languages.

Another important yet unexplored research line is to model personalization of CN generation according to haters' characteristics and personality traits. Recently, personalized persuasion has become of great interest in dialogue systems (Wang et al., 2019b) and argument generation (Lukin et al., 2017), demonstrating various styles of persuasions based on, for instance, modality and layout (Berkovsky et al., 2012). In this context, we argue that personalized CNs can encourage engagement in discussion as well as strength the impact of counter narratives on haters and bystanders. Thus, future work can tailor CNs to different individuals, conditioned on relevant aspects such as demographic data (Wang et al., 2019b) and prior belief (Durmus and Cardie, 2018). In the same vein, we can also incorporate contextual information into CN generation by considering hate speech types and counter narrative types as input features.

Lastly, assessing CN generation is complex and challenging due to the lack of clear evaluation criteria and robust evaluation techniques. The characteristics of a good CN have been put forward in prior work (Benesch et al., 2016a; Munger, 2017; Hangartner et al., 2021). Overall, we consider good CNs to be effective, credible and can convince audience to stop spreading hatred. In this thesis, we assessed the effectiveness of CNs via two aspects, automatic metrics and human evaluation. Automatic metrics, such as BLEU and ROUGE, are quantitatively scalable methods comparing the word overlap between model outputs and gold standard. However, these metrics can only infer model performance according to references provided (e.g., dependent heavily on exact word usage and word order) and gathering an exhaustive list of all appropriate CNs is not feasible. For this reason, such metrics cannot capture properly model performance, particularly for

open-ended tasks (Liu et al., 2016; Novikova et al., 2017) including CN generation. As a result, human evaluation is heavily employed in open-ended NLG tasks. Despite of being trusted and high-performing, human evaluation has inherent limitations such as it is costly, tricky (e.g., evaluator biases and question formatting), time-consuming (both in terms of evaluation and moderator training), and can be inconsistent and inflict psychological harm on the moderators. The effectiveness of CN generations should be also carefully investigated "in-the-wild" to understand its social media impact, reach of content, and the dynamics of hateful content and CN. To our knowledge, no work has examined this line of research yet. While it is beyond the scope of this thesis, we identify several aspects for characterising how CNs can alter discourse and the behaviour/beliefs of audience (e.g., hate speakers and bystanders) on social media. Intuitively, it can be investigated via explicit measurements such as number of likes, views, and replies received after CNs are posted (Mathew et al., 2018, 2019). Qualitatively, surveys (Krosnick and Abelson, 1992) can be used to explore the changes in attitude and behaviours, for instance before and after receiving CN treatment. Together, a vital future direction is design metrics that are scalable, high-performing, robust, and transparent - for instance BLEURT (Sellam et al., 2020) and MAUVE (Pillutla et al., 2021).

While we specifically investigated counter narratives in this thesis, online hate speech still presents substantial challenges. We hope that the data and approaches proposed are beneficial not only for hate mitigation but also for other tasks for social good. We also hope this thesis can provide insights and motivate future work that considers counter narratives for online hate countering.

Bibliography

Aguilera-Carnerero, C. and Azeez, A. H. (2016). 'islamonausea, not islamophobia': The many faces of cyber hate speech. *Journal of Arab & Muslim Media Research*, 9(1):21–40.

Ahluwalia, R., Shcherbinina, E., Callow, E., Nascimento, A., and De Cock, M. (2018). Detecting misogynous tweets. *Proc. of IberEval*, 2150:242–248.

Alakrot, A., Murray, L., and Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in arabic. *Procedia Computer Science*, 142:174–181. Arabic Computational Linguistics.

Arango, A., Pérez, J., and Poblete, B. (2020). Hate speech detection is not as easy as you may think: A closer look at model validation (extended version). *Information Systems*, page 101584.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the facebook's walls of hate. *International Journal of Cyber Criminology*, 10(1).

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (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.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Bai, X., Merenda, F., Zaghi, C., Caselli, T., and Nissim, M. (2018). Rug at germeval: Detecting offensive speech in german social media. In *14th Conference on Natural Language Processing KONVENS 2018*.

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., and Sanguinetti, M. (2019). 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ben-David, A. and Fernández, A. M. (2016). Hate speech and covert discrimination on social media: Monitoring the facebook pages of extreme-right political parties in spain. *International Journal of Communication*, 10:27.

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

Benesch, S., Ruths, D., Dillon, K. P., Saleem, H. M., and Wright, L. (2016a). Considerations for successful counterspeech. *Dangerous Speech Project*.

Benesch, S., Ruths, D., P Dillon, K., Mohammad Saleem, H., and Wright, L. (2016b). Counterspeech on twitter: A field study. *Dangerous Speech Project. Available at: https://dangerousspeech.org/counterspeech-on-twitter-a-field-study/*.

Berkovsky, S., Freyne, J., and Oinas-Kukkonen, H. (2012). Influencing individually: Fusing personalization and persuasion. *ACM Trans. Interact. Intell. Syst.*, 2(2).

Bertoldi, N., Cettolo, M., and Federico, M. (2013). Cache-based online adaptation for machine translation enhanced computer assisted translation. In *MT-Summit*, pages 35–42.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., and Ueffing, N. (2004). Confidence estimation for machine translation. In *Coling 2004: Proceedings of the 20th international conference on computational linguistics*, pages 315–321.

Blodgett, S. L., Barocas, S., Daumé III, H., and Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Bordia, S. and Bowman, S. R. (2019). Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter*

of the Association for Computational Linguistics: Student Research Workshop, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosco, C., Felice, D., Poletto, F., Sanguinetti, M., and Maurizio, T. (2018). Overview of the evalita 2018 hate speech detection task. In *EVALITA Evaluation of NLP and Speech Tools for Italian: Proceedings of the Final Workshop*, volume 2263, pages 1–9. CEUR.

Bretschneider, U. and Peters, R. (2016). Detecting cyberbullying in online communities.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Burnap, P. and Williams, M. L. (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.

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

Canete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., and Pérez, J. (2020). Spanish pre-trained bert model and evaluation data. *Pml4dc at iclr*, 2020:2020.

Caplan, R. (2018). Content or context moderation? artisanal, community-reliant, and industrial approaches. *Data & Society*.

Carenini, G. and Moore, J. D. (2000). A strategy for generating evaluative arguments. In *Proceedings of the first international conference on Natural language generation-Volume 14*, pages 47–54. Association for Computational Linguistics.

Cettolo, M., Bertoldi, N., and Federico, M. (2014). The repetition rate of text as a predictor of the effectiveness of machine translation adaptation. In *Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas (AMTA 2014)*, pages 166–179.

Chatterjee, R., Negri, M., Turchi, M., Federico, M., Specia, L., and Blain, F. (2017). Guiding neural machine translation decoding with external knowledge. In *Proceedings of the Second Conference on Machine Translation*, pages 157–168.

Chi, Z., Dong, L., Wei, F., Wang, W., Mao, X., and Huang, H. (2020). Cross-lingual natural language generation via pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7570–7577. AAAI Press.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Chung, Y.-L., Kuzmenko, E., Tekiroğlu, S. S., and Guerini, M. (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, Florence, Italy. Association for Computational Linguistics.

Chung, Y.-L., Tekiroğlu, S. S., and Guerini, M. (2020). Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Online.

Chung, Y.-L., Tekiroğlu, S. S., and Guerini, M. (2021a). Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Chung, Y.-L., Tekiroğlu, S. S., Tonelli, S., and Guerini, M. (2021b). Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.

Cihon, P. and Yasseri, T. (2016). A biased review of biases in Twitter studies on political collective action. *Frontiers in Physics*, 4:34.

Citron, D. K. and Norton, H. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Corazza, M., Menini, S., Cabrio, E., Tonelli, S., and Villata, S. (2019). Cross-Platform Evaluation for Italian Hate Speech Detection. In *CLiC-it 2019 - 6th Annual Conference of the Italian Association for Computational Linguistics*, Bari, Italy.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 797–806, New York, NY, USA. Association for Computing Machinery.

Dang, B., Riedl, M. J., and Lease, M. (2018). But who protects the moderators? the case of crowdsourced image moderation. *arXiv preprint arXiv:1804.10999*.

Das, R., Zaheer, M., Reddy, S., and McCallum, A. (2017). Question answering on knowledge bases and text using universal schema and memory networks. In Barzilay, R. and Kan, M., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 358–365, Vancouver, Canada. Association for Computational Linguistics.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515.

De Boer, V., Hildebrand, M., Aroyo, L., De Leenheer, P., Dijkshoorn, C., Tesfa, B., and Schreiber, G. (2012). Nichesourcing: harnessing the power of crowds of experts. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 16–20. Springer.

de Gibert, O., Perez, N., García-Pablos, A., and Cuadros, M. (2018). Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

de Jong, R. and Theune, M. (2018). Going dutch: Creating simplenlg-nl. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 73–78.

De Mattei, L., Cafagna, M., Dell'Orletta, F., Nissim, M., and Guerini, M. (2020). Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy.

de Pelle, R. and Moreira, V. (2017). Offensive comments in the brazilian web: a dataset and baseline results. In *Anais do VI Brazilian Workshop on Social Network Analysis and Mining*, Porto Alegre, RS, Brasil. SBC.

Del Vigna, F., Cimino, A., Dell'Orletta, F., Petrocchi, M., and Tesconi, M. (2017). Hate me, hate me not: Hate speech detection on facebook.

Dell'Orletta, F. and Nissim, M. (2018). Overview of the EVALITA 2018 cross-genre gender prediction (gxg) task. In *EVALITA@CLiC-it*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Delobelle, P., Winters, T., and Berendt, B. (2020). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Desvousges, W. H., Reed Johnson, F., Dunford, R. W., Nicole Wilson, K., and Boyle, K. J. (1993). Measuring natural resource damages with contingent valuation. In *Contingent valuation: A critical assessment*, pages 91–164. Emerald Group Publishing Limited.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):18.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations*, New Orleans, LA.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015). Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30. ACM.

Durmus, E. and Cardie, C. (2018). Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*

(Long Papers), pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Dušek, O. and Kasner, Z. (2020). Evaluating semantic accuracy of data-to-text generation with natural language inference. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 131–137, Dublin, Ireland. Association for Computational Linguistics.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2):179–211.

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., and Belding, E. (2018). Peer to peer hate: Hate speech instigators and their targets. In *Twelfth International AAAI Conference on Web and Social Media*.

Ernst, J., Schmitt, J. B., Rieger, D., Beier, A. K., Vorderer, P., Bente, G., and Roth, H.-J. (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.

Fan, A., Lewis, M., and Dauphin, Y. (2019). Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660, Florence, Italy. Association for Computational Linguistics.

Feng, S. Y., Gangal, V., Wei, J., Chandar, S., Vosoughi, S., Mitamura, T., and Hovy, E. (2021). A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Fersini, E., Nozza, D., and Rosso, P. (2018a). Overview of the evalita 2018 task on automatic misogyny identification (AMI). In *EVALITA@CLiC-it*, volume 2263 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Fersini, E., Rosso, P., and Anzovino, M. (2018b). Overview of the task on automatic misogyny identification at ibereval 2018. In *IberEval@SEPLN*, volume 2150 of *CEUR Workshop Proceedings*, pages 214–228. CEUR-WS.org.

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

Fortuna, P., Soler-Company, J., and Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, 58(3):102524.

Gagliardone, I., Gal, D., Alves, T., and Martinez, G. (2015). *Countering online hate speech*. Unesco Publishing.

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

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. (2020). RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, 20(12):4492–4511.

Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M. (2018). A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10(4):215–230.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T. (2019). Large-scale transfer learning for natural language generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6053–6058.

Gomez, R., Gibert, J., Gomez, L., and Karatzas, D. (2020). Exploring hate speech detection in multimodal publications. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1459–1467, Los Alamitos, CA, USA. IEEE Computer Society.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Grusky, M., Naaman, M., and Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (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*.

Habernal, I. and Gurevych, I. (2016). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Hall, N. (2013). *Hate crime*. Taylor Francis, 2nd edition.

Hammer, H. L. (2016). Automatic detection of hateful comments in online discussion. In *International Conference on Industrial Networks and Intelligent Systems*, pages 164–173. Springer.

Hangartner, D., Gennaro, G., Alasiri, S., Bahrich, N., Bornhoft, A., Boucher, J., Demirci, B. B., Derksen, L., Hall, A., Jochum, M., Munoz, M. M., Richter, M., Vogel, F., Wittwer, S., Wüthrich, F., Gilardi, F., and Donnay, K. (2021). Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

He, H., Balakrishnan, A., Eric, M., and Liang, P. (2017a). Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1766–1776, Vancouver, Canada. Association for Computational Linguistics.

He, S., Liu, C., Liu, K., and Zhao, J. (2017b). Generating natural answers by incorporating copying and retrieving mechanisms in sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–208.

Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2020). The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Hovy, D. (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.

Hovy, D. and Spruit, S. L. (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.

Hua, X., Hu, Z., and Wang, L. (2019). Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Hua, X. and Wang, L. (2018). Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

Jaccard, P. (1912). The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50.

Jay, T. (2009). Do offensive words harm people? *Psychology, public policy, and law*, 15(2):81–101.

Jha, A. and Mamidi, R. (2017). When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16.

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

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

Kahneman, D. and Knetsch, J. L. (1992). Valuing public goods: the purchase of moral satisfaction. *Journal of environmental economics and management*, 22(1):57–70.

Kalyan, K. S., Rajasekharan, A., and Sangeetha, S. (2021). Ammus: A survey of transformer-based pretrained models in natural language processing. *arXiv preprint arXiv:2108.05542*.

Karan, M. and Šnajder, J. (2018). Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Karunakaran, S. and Ramakrishan, R. (2019). Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 50–58.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Khishfe, R. (2012). Relationship between nature of science understandings and argumentation skills: A role for counterargument and contextual factors. *Journal of Research in Science Teaching*, 49(4):489–514.

Klubička, F. and Fernandez, R. (2018). Examining a hate speech corpus for hate speech detection and popularity prediction. In Branco, A., Calzolari, N., and Choukri, K., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Krosnick, J. A. and Abelson, R. P. (1992). The case for measuring attitude strength in surveys. *Questions about questions: Inquiries into the cognitive bases of surveys*, pages 177–203.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2018). Benchmarking aggression identification in social media. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 1–11.

Kumar, R., Ojha, A. K., Malmasi, S., and Zampieri, M. (2020). Evaluating aggression identification in social media. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 1–5, Marseille, France. European Language Resources Association (ELRA).

Kumar, V., Joshi, N., Mukherjee, A., Ramakrishnan, G., and Jyothi, P. (2019). Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy. Association for Computational Linguistics.

Kumar Sharma, H., Kshitiz, K., and Shailendra (2018). NLP and machine learning techniques for detecting insulting comments on social networking platforms. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 265–272.

Kwok, I. and Wang, Y. (2013). Locate the hate: Detecting tweets against blacks.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2020). Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Laugwitz, B., Held, T., and Schrepp, M. (2008). Construction and evaluation of a user experience questionnaire. In *Symposium of the Austrian HCI and Usability Engineering Group*, pages 63–76. Springer.

Le, D.-T., Nguyen, C. T., and Nguyen, K. A. (2018). Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016). A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Liang, P. P., Wu, C., Morency, L.-P., and Salakhutdinov, R. (2021). Towards understanding and mitigating social biases in language models. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lowe, R., Pow, N., Serban, I. V., and Pineau, J. (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.

Lukin, S., Anand, P., Walker, M., and Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B., and Rodrigue, J. (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.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., and Patel, A. (2019). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, page 14–17, New York, NY, USA. Association for Computing Machinery.

Manurung, R., Ritchie, G., and Thompson, H. (2008). An implementation of a flexible author-reviewer model of generation using genetic algorithms. In *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation*, pages 272–281, The University of the Philippines Visayas Cebu College, Cebu City, Philippines. De La Salle University, Manila, Philippines.

Marie, B., Fujita, A., and Rubino, R. (2021). Scientific credibility of machine translation research: A meta-evaluation of 769 papers. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7297–7306, Online. Association for Computational Linguistics.

Mathew, B., Kumar, N., Goyal, P., Mukherjee, A., et al. (2018). Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhania, P., Maity, S. K., Goyal, P., and Mukherjee, A. (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.

Mathur, P., Sawhney, R., Ayyar, M., and Shah, R. (2018). Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of*

the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.

Mazaré, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779, Brussels, Belgium. Association for Computational Linguistics.

Mazzei, A., Battaglino, C., and Bosco, C. (2016). Simplenlg-it: adapting simplenlg to italian. In *Proceedings of the 9th international natural language generation conference*, pages 184–192.

Mehdad, Y. and Tetreault, J. (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.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2017). Pointer sentinel mixture models. *International Conference on Learning Representations*.

Mezzich, J. E., Kraemer, H. C., Worthington, D. R., and Coffman, G. A. (1981). Assessment of agreement among several raters formulating multiple diagnoses. *Journal of psychiatric research*, 16(1):29–39.

Miller, A., Fisch, A., Dodge, J., Karimi, A.-H., Bordes, A., and Weston, J. (2016). Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Moon, S., Shah, P., Kumar, A., and Subba, R. (2019). Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854.

Moretti, G., Sprugnoli, R., and Tonelli, S. (2015). Digging in the dirt: Extracting keyphrases from texts with kd. In *Proceedings of the 2nd Italian Conference on Computational Linguistics*, Trento, Italy.

Mossie, Z. and Wang, J.-H. (2020). Vulnerable community identification using hate speech detection on social media. *Inf. Process. Manage.*, 57(3).

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

Nie, F., Yao, J.-G., Wang, J., Pan, R., and Lin, C.-Y. (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. Association for Computational Linguistics.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (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.

Novikova, J., Dušek, O., Cercas Curry, A., and Rieser, V. (2017). Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Oberlander, J. and Brew, C. (2000). Stochastic text generation. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 358(1769):1373–1387.

Olteanu, A., Castillo, C., Boy, J., and Varshney, K. (2018). The effect of extremist violence on hateful speech online. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Oosterman, J., Bozzon, A., Houben, G.-J., Nottamkandath, A., Dijkshoorn, C., Aroyo, L., Leyssen, M. H., and Traub, M. C. (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.

Ousidhoum, N., Lin, Z., Zhang, H., Song, Y., and Yeung, D.-Y. (2019). Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Pàmies, M., Öhman, E., Kajava, K., and Tiedemann, J. (2020). LT@Helsinki at SemEval-2020 task 12: Multilingual or language-specific BERT? In *Proceedings of the Fourteenth*

Workshop on Semantic Evaluation, pages 1569–1575, Barcelona (online). International Committee for Computational Linguistics.

Pamungkas, E. W., Cignarella, A. T., Basile, V., Patti, V., 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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Parthasarathi, P. and Pineau, J. (2018). Extending neural generative conversational model using external knowledge sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 690–695, Brussels, Belgium. Association for Computational Linguistics.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Pham-Hong, B.-T. and Chokshi, S. (2020). PGSG at SemEval-2020 task 12: BERT-LSTM with tweets' pretrained model and noisy student training method. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2111–2116, Barcelona (online). International Committee for Computational Linguistics.

Pianta, E. and Tonelli, S. (2010). KX: A flexible system for keyphrase eXtraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 170–173, Uppsala, Sweden. Association for Computational Linguistics.

Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., and Harchaoui, Z. (2021). Mauve: Measuring the gap between neural text and human text using divergence frontiers. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4816–4828. Curran Associates, Inc.

Piskorski, J., Haneczok, J., and Jacquet, G. (2020). New benchmark corpus and models for fine-grained event classification: To BERT or not to BERT? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6663–6678, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Plaza-del-Arco, F. M., Halat, S., Padó, S., and Klinger, R. (2021). Multi-Task Learning with Sentiment, Emotion, and Target Detection to Recognize Hate Speech and Offensive Language. *arXiv e-prints*, page arXiv:2109.10255.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Poria, S., Cambria, E., Gelbukh, A., Bisio, F., and Hussain, A. (2015). Sentiment data flow analysis by means of dynamic linguistic patterns. *IEEE Computational Intelligence Magazine*, 10(4):26–36.

Prabhavalkar, R., Rao, K., Sainath, T., Li, B., Johnson, L. M., and Jaitly, N. (2017). A comparison of sequence-to-sequence models for speech recognition. In *INTERSPEECH*.

Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (2019). A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

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

Rach, N., Langhammer, S., Minker, W., and Ultes, S. (2018). Utilizing argument mining techniques for argumentative dialogue systems. In *Proceedings of the 9th International Workshop On Spoken Dialogue Systems (IWSDS)*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding with unsupervised learning.

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

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rafiq, R. I., Hosseinmardi, H., Han, R., Lv, Q., Mishra, S., and Mattson, S. A. (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.

Rainie, H., Anderson, J. Q., and Albright, J. (2017). *The future of free speech, trolls, anonymity and fake news online*. Pew Research Center Washington, DC.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rakshit, G., Bowden, K. K., Reed, L., Misra, A., and Walker, M. (2019). Debbie, the debate bot of the future. In *Advanced Social Interaction with Agents*, pages 45–52. Springer.

Rashid, A., Do-Omri, A., Haidar, M. A., Liu, Q., and Rezagholizadeh, M. (2019). Bilingual-GAN: A step towards parallel text generation. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 55–64, Minneapolis, Minnesota. Association for Computational Linguistics.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010). Offensive language detection using multi-level classification. In *Proceedings of the 23rd Canadian Conference on Advances in Artificial Intelligence*, AI'10, page 16–27, Berlin, Heidelberg. Springer-Verlag.

Reed, C., Long, D., and Fox, M. (1996). An architecture for argumentative dialogue planning. In *International Conference on Formal and Applied Practical Reasoning*, pages 555–566. Springer.

Reynolds, K., Kontostathis, A., and Edwards, L. (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.

Rezvan, M., Shekarpour, S., Balasuriya, L., Thirunarayan, K., Shalin, V. L., and Sheth, A. (2018). A quality type-aware annotated corpus and lexicon for harassment research. In *Proceedings of the 10th ACM Conference on Web Science*, WebSci '18, page 33–36, New York, NY, USA. Association for Computing Machinery.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Riedl, M. J., Masullo, G. M., and Whipple, K. N. (2020). The downsides of digital labor: Exploring the toll incivility takes on online comment moderators. *Computers in Human Behavior*, 107:106262.

Roberts, S. T. (2016). *Commercial Content Moderation: Digital Laborers' Dirty Work*, chapter The Intersectional Internet: Race, Sex, Class and Culture Online, pages 147–160. New York: Peter Lang.

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., Gatford, M., et al. (1995). Okapi at trec-3. *Nist Special Publication*.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2016). Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In Beißwenger, M., Wojatzki, M., and Zesch, T., editors, *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9, Bochum.

Ruder, S. (2021). Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`.

Saha, A., Ansari, G. A., Laddha, A., Sankaranarayanan, K., and Chakrabarti, S. (2019). Complex program induction for querying knowledge bases in the absence of gold programs. *Transactions of the Association for Computational Linguistics*, 7:185–200.

Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., and Stranisci, M. (2018). An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sato, M., Yanai, K., Miyoshi, T., Yanase, T., Iwayama, M., Sun, Q., and Niwa, Y. (2015). End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114.

Schiappa, E. and Nordin, J. P. (2013). *Argumentation: Keeping Faith with ReasonKeeping Faith With Reason*. Pearson UK.

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

Schmidt, A. and Wiegand, M. (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.

Schuster, S., Gupta, S., Shah, R., and Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

Sellam, T., Das, D., and Parikh, A. (2020). BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (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, Berlin, Germany. Association for Computational Linguistics.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Shang, L., Lu, Z., and Li, H. (2015). 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)*, pages 1577–1586.

Sharifirad, S., Jafarpour, B., and Matwin, S. (2018). Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. *EMNLP 2018*, page 107.

Sheng, E., Chang, K.-W., Natarajan, P., and Peng, N. (2020). Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational*

Linguistics: EMNLP 2020, pages 3239–3254, Online. Association for Computational Linguistics.

Shushkevich, E. and Cardiff, J. (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.*

Sigurbergsson, G. I. and Derczynski, L. (2020). Offensive language and hate speech detection for Danish. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3498–3508, Marseille, France. European Language Resources Association.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016). Analyzing the targets of hate in online social media. In *ICWSM*, pages 687–690.

Silverman, T., Stewart, C. J., Birdwell, J., and Amanullah, Z. (2016). The impact of counter-narratives. *Institute for Strategic Dialogue, London. https://www. strategicdialogue. org/wp-content/uploads/2016/08/Impact-of-Counter-Narratives_ONLINE. pdf–73.*

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., and Wang, J. (2019). Release strategies and the social impacts of language models. *arXiv:1908.09203.*

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

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). 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*, pages 196–205.

Specia, L. and Farzindar, A. (2010). Estimating machine translation post-editing effort with HTER. In *Proceedings of the Second Joint EM+/CNGL Workshop: Bringing MT to the User: Research on Integrating MT in the Translation Industry*, pages 33–43, Denver, Colorado, USA. Association for Machine Translation in the Americas.

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

Stab, C., Miller, T., Schiller, B., Rai, P., and Gurevych, I. (2018). Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Stella, M., Cristoforetti, M., and De Domenico, M. (2019). Influence of augmented humans in online interactions during voting events. *PLOS ONE*, 14(5):1–16.

Stroud, S. R. and Cox, W. (2018). The varieties of feminist counterspeech in the misogynistic online world. In *Mediating Misogyny*, pages 293–310. Springer.

Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2440–2448, Cambridge, MA, USA. MIT Press.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Tahmasbi, F., Schild, L., Ling, C., Blackburn, J., Stringhini, G., Zhang, Y., and Zannettou, S. (2021). "go eat a bat, chang!": On the emergence of sinophobic behavior on web communities in the face of covid-19. In *Proceedings of the Web Conference 2021*, WWW '21, page 1122–1133, New York, NY, USA. Association for Computing Machinery.

Tekiroğlu, S. S., Chung, Y.-L., and Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.

Tuck, H. and Silverman, T. (2016). *The counter-narrative handbook*. Institute for Strategic Dialogue.

Turchi, M., Negri, M., and Federico, M. (2013). Coping with the subjectivity of human judgements in MT quality estimation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 240–251, Sofia, Bulgaria. Association for Computational Linguistics.

Turney, P. D. (2000). Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336.

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., and Plank, B. (2018). Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Vidgen, B. and Derczynski, L. (2021). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Vidgen, B., Margetts, H., and Harris, A. (2019). How much online abuse is there. *Alan Turing Institute*.

Vinyals, O. and Le, Q. V. (2015). A neural conversational model. In *ICML Deep Learning Workshop*.

Vogels, E. A. (2021). The state of online harassment.

Volkova, S., Wilson, T., and Yarowsky, D. (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.

Wachsmuth, H., Syed, S., and Stein, B. (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.

Waldron, J. (2012). *The harm in hate speech*. Harvard University Press.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, K. and Wan, X. (2018). Sentigan: Generating sentimental texts via mixture adversarial networks. In *IJCAI*, pages 4446–4452.

Wang, S., Liu, J., Ouyang, X., and Sun, Y. (2020). Galileo at SemEval-2020 task 12: Multi-lingual learning for offensive language identification using pre-trained language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1448–1455, Barcelona (online). International Committee for Computational Linguistics.

Wang, W., Gan, Z., Xu, H., Zhang, R., Wang, G., Shen, D., Chen, C., and Carin, L. (2019a). Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., and Yu, Z. (2019b). Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Warner, W. and Hirschberg, J. (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.

Waseem, Z. (2016). Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Waseem, Z. and Hovy, D. (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, San Diego, California. Association for Computational Linguistics.

Wilson, R. A. and Land, M. K. (2020). Hate speech on social media: Content moderation in context. *Connecticut Law Review*, 52:1029.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2018). Transfertransfo: A transfer learning approach for neural network based conversational agents. *NeurIPS 2018 workshop on Conversational AI: "Today's Practice and Tomorrow's Potential"*.

Wright, L., Ruths, D., Dillon, K. P., Saleem, H. M., and Benesch, S. (2017). Vectors for counterspeech on twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 57–62.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Xu, J., Ren, X., Lin, J., and Sun, X. (2018). Diversity-promoting GAN: A cross-entropy based generative adversarial network for diversified text generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. Association for Computational Linguistics.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012). Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666. Association for Computational Linguistics.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., and Yan, R. (2019). Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Young, T., Cambria, E., Chaturvedi, I., Zhou, H., Biswas, S., and Huang, M. (2018). Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., and Çöltekin, Ç. (2020). SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. In *Advances in Neural Information Processing Systems*, pages 9051–9062.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., and Caragea, C. (2016). Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, pages 3952–3958.

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzmán, F., Zettlemoyer, L., and Ghazvininejad, M. (2021). Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

Zhu, W. and Bhat, S. (2021). Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Zhu, W., Mo, K., Zhang, Y., Zhu, Z., Peng, X., and Yang, Q. (2017). Flexible end-to-end dialogue system for knowledge grounded conversation. *CoRR*, abs/1709.04264.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Appendix A

Additional Materials

A.1 Preliminary Analysis on Decoding Methods for XNLG.

To find a suitable decoding method for our task, we generated CNs with 3 candidate settings: beam search with a beam-width of 3 and top-k sampling with a k value of 8 and 10. For each setting we utilized KN retrieved with both non-optimal (Q_{cn}) and optimal ($Q_{hs \cup cn}$) queries. Then we sampled 120 HS-CN pairs and served them to three experts in CN writing for evaluating the generation on a scale of 1 (the worst) to 5 (the best) in terms of suitableness and informativeness. Suitableness measures if the generated CN is relevant to the HS and informativeness evaluates the amount of information (e.g. statistics and facts) enclosed in the counter narrative.

The results reported in Table A.1 reveal a clear difference between beam search and top-k sampling regardless of KN being optimal or non-optimal. In a manual investigation, we observed that the generation using both beam search and top-k sampling generally can copy some pieces of information from the given KN, while top-k seems to replace part of the text with slightly relevant and uncommon words. Hence, copying the right knowledge pieces through the decoding strategy is a key factor instead of diverging from the knowledge solely for the sake of lexical diversity. Therefore, based on the results, we adopt beam search with a beam-width of 3, which is shown to be the most suitable and informative, for decoding method in our experiments.

Decoding methods	Suitableness	Informativeness
Non-optimal knowledge		
Beam-3	1.950	2.325
Topk-8	1.275	1.775
Topk-10	1.625	2.100
Optimal knowledge		
Beam-3	2.325	2.450
Topk-8	1.975	2.175
Topk-10	2.050	2.325

Table A.1: Human evaluation on CN generation using various decoding methods.

A.2 Usability Exercise Instructions

Task	Instruction
Hate distribution	Please choose two hashtags and a temporal framework of your own choice. Now analyse the network of hate speakers. What is their shape? What is the distribution of hate speech? Are there main nodes (i.e. influencers) which lead the flow of hate speech? Is the net wide or narrow? In your opinion, what network is the most dangerous in terms of anti-Muslim hate speech and why?
Hate timeline	In the "Hashtag trends" section, please first select one hashtag of your own choice and, subsequently, a temporal window from the drop-down menu (snapshot). Then maintain the same hashtag, changing the snapshot. How do trends vary across time with regard to that specific hashtag? Does the shape of the network change?
Islamisation analysis	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
Recent trends	Please select one of the most recent Islamophobic hashtags and analyse the context in which it is used: does it refer to the national or the international context? What political actors are involved? Are the hate speakers who are active closely affiliated to any particular political party?

Table A.2: Instructions for usability exercises.