Contextualized Graph Representations for Generating Counter-Narratives against Hate Speech

Selene Baez Santamaria[1], Helena Gómez-Adorno[2], Ilia Markov[1]

[1]CLTL, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
{s.baezsantamaria, i.markov}@vu.nl
[2]IIMAS, Universidad Nacional Autónoma de México, Mexico City, Mexico
helena.gomez@iimas.unam.mx

Abstract

Hate speech (HS) is a widely acknowledged societal problem with potentially grave effects on vulnerable individuals and minority groups. Developing counter-narratives (CNs) that confront biases and stereotypes driving hateful narratives is considered an impactful strategy. Current automatic methods focus on isolated utterances to detect and react to hateful content online, often omitting the conversational context where HS naturally occurs. In this work, we explore strategies for the incorporation of conversational history for CN generation, comparing text and graphical representations with varying degrees of context. Overall, automatic and human evaluations show that 1) contextualized representations are comparable to those of isolated utterances, and 2) models based on graph representations outperform text representations, thus opening new research directions for future work.

Offensive Content Warning: This paper contains offensive language that some readers may find distressing.

1 Introduction

Hate speech (HS) is a widespread problem in society with severe repercussions at both personal and societal levels. At the individual level, it can lead to severe psychological and emotional impacts on individuals who are targeted, e.g., fear of becoming the target of physical violence (Saresma et al., 2021) or increased suicide rates (Hinduja and Patchin, 2010). At a systemic level, it can create a feedback loop between offline violence and online HS (Siegel, 2020), e.g., encouraging school violence (Hinduja and Patchin, 2007). The pervasiveness of HS in the digital age makes its countering a pressing issue (Gagliardone et al., 2015).

Several approaches exist to counter online HS, including legal sanctions, content regulation, and counter-speech (Donzelli, 2021). The latter consists in responding to HS through counter-narratives (CNs), which are non-negative responses to HS, targeting extreme statements through fact-bound arguments or alternative perspectives (Benesch, 2014). CNs aim to promote understanding between individuals and are regarded as an approach that does not pose normative or censoring issues (Donzelli, 2021). Furthermore, CNs can play a social role in educating those exposed to biased narratives, providing evidence-based responses and exposing misinformation (Pariser, 2011).

CNs naturally appear online and are typically authored by community members addressing the phenomenon of hate speech or carefully curated by a mediating party like NGO operators. Significant cognitive effort and time investment are needed to create CNs manually, which is not feasible at large scale (Schieb and Preuss, 2016). This became the primary motivation for exploring automatic methods for CN generation, a task initially proposed by Qian et al. (2019). As some of these automatic methods rely on data-hungry approaches, research efforts have been directed at providing high-quality datasets containing manually created CNs. Their provenance ranges from mined instances on social media platforms like Twitter (Mathew et al., 2018), Gab or Reddit (Qian et al., 2019), or instances manually created by trained NGO operators (Chung et al., 2019).

Currently, automatically generated counter-narratives tend to be generic and often fail to engage users effectively (Zheng et al., 2023). As Doğanç and Markov (2023) observe, current automatic CN generation methods are ineffective in persuading authors against their expressed biases due to a lack of personalization and contextualization. This is partially because authority is insufficient to change the course of the conversation; instead, establishing psychological and epistemic common ground is required (McGowan, 2018).

Findings of the Association for Computational Linguistics: EACL 2024, pages 7664–7674
November 12-16, 2024

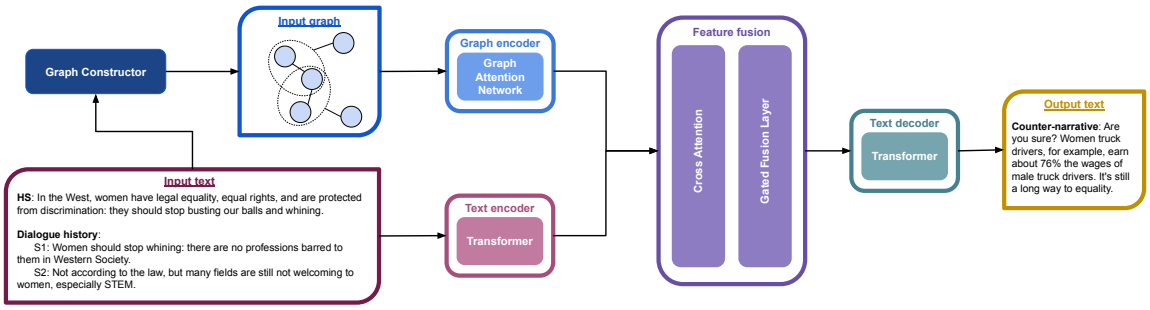

Figure 1: System architecture. From left to right: 1) The text representation is translated into a graph representation. 2) The graph and text inputs get encoded in parallel to generate vector representations. 3) The encoded representations are aligned using a feature fusion mechanism. 4) These aligned features are passed to a text decoder to generate output text.

In this work, we aim to automatically generate *counter-narratives contextualized in dialogue* that include conversational context and history for better customization of CNs. To do so, we propose an approach that leverages the use of *graphical models* to depict the dialogue history as the context in which HS arises, beyond relying on text representations. These graphical representations can make implicit aspects of HS explicit, fostering a better understanding of the cognitive and social mechanisms underlying hateful content.[1]

2 Related work

In this section, we begin by describing the research on counter-narrative generation, with a particular focus on the role of dialogue context. We then shift our attention to the use of graph representations for textual data.

2.1 Counter-narrative generation in dialogue

Extensive research has utilized conversational context for detecting the (fine-grained) types (Pavlopoulos et al., 2020; Menini et al., 2021) and targets (Markov and Daelemans, 2022) of HS, some focusing on the usage of social networks (Mishra et al., 2018) and other graph representations (Mishra et al., 2019). However, these techniques have not been well studied for CN generation (Qian et al., 2019). Some efforts have focused on creating datasets with discourse annotations from Reddit (Hassan and Alikhani, 2023) and Youtube (Mathew et al., 2019), yet these datasets still consist only of HS-CN pairs. Closer work involves intervening during conversation (de los Riscos and D'Haro, 2021), however these systems require a HS classification model and thus do not tackle the task of CN generation in an isolated manner. All and all, when the dialogue history is taken into account, it is usually represented as text, which is typically encoded as a vector (Bonaldi et al., 2024). To the best of our knowledge, there is no work that explores alternative representations of the dialogue context, such as graph representations, for CN generation.

2.2 Graph representations

Several methods for representing text as graphs focus on either syntax or semantics. Syntactic parsing techniques, such as dependency parsing, constituency parsing, and syntactic integrated graphs (Gómez-Adorno et al., 2016), capture structural and hierarchical aspects of language. In contrast, semantic techniques like Frame Semantics (Fillmore, 1976) and Abstract Meaning Representations (AMR) (Fillmore, 1976) describe events or situations conceptually, relying on catalogues of possible frames or abstract representations. While this is effective for parsing individual pieces of factual text, they may fall short in handling discourse-level information, especially in subjective contexts such as detecting and withstanding HS, where open-world graph constructions are potentially better suited.

As an example, Yao et al. (2023) propose a Chain-Of-Thought representation (GoT) to approximate the non-linear reasoning process that humans are capable of. They test their framework on a Question Answering task, both with text only and in multimodal scenarios, using a two-stage framework to generate in-between rationales followed by the final answer.

[1]All code and data available at: https://github.com/selBaez/graph-based-hs-cn

Alternatively, Baez Santamaria et al. (2021) propose to use episodic Knowledge Graphs (eKGs) to represent the content, form and context of dyadic dialogues. Using RDF technologies, these graphs contain subgraphs relating to: (i) `Ontology`: the world model, (ii) `Claims`: the atomic pieces of knowledge, (iii) `Instances`: the individual entities in claims and their inter-claim connections, (iv) `Perspectives`: the viewpoint of the source regarding a claim, (v) `Interactions`: the conversational provenance of each claim. As such, these graphs also serve as a model for the Theory of Mind, as the information incoming from each speaker is maintained and interpreted separately.

3 Research framework

The research questions guiding this work are:

- What is the role of discourse context in counter-narrative generation?
- To what extent can graph representations aid in generating contextualized counter-narratives in a conversational context?

To answer the stated research questions, we generate CNs under several conditions. We first compare the generated CNs utilizing only the targeted HS utterance against the ones utilizing the previous dialogue history. Then, we compare the generated CNs using only text against the ones relying on graph representations.

3.1 Methodology

In this paper, we transform a text-based dialogue dataset into different graph representations, resulting in data structures that include dialogue context information. We employ a graph-based text generation architecture (Yao et al., 2023), originally tailored to produce rationales in a reasoning task but here adapted to the CN generation task. The system's pipeline, as shown in Figure 1, primarily consists of four steps:

1. Graph construction from text;
2. Encoding of the text and graph modalities;
3. Alignment of embeddings across modalities;
4. Text decoder for language generation.

.

3.1.1 Graph construction

Various types of graph representations can be used with the proposed architecture. In this work, we use two graph representations: 1) Graph of Thought (GoT) and 2) Episodic Knowledge Graph (eKG). GoT captures the content of a dialogue, while the eKG additionally captures the dialogue's structure, including speaker identity and utterance sequence. Hereby, we describe how these graph representations are constructed. For a grounded example with corresponding graph visualizations of the representations used in this work, see Appendix A.

Graph of Thought We create Graph of Thought (GoT) representations using the code provided by Yao et al. (2023)[2], keeping the maximum number of nodes to 100. Following Yao et al. (2023), we extract triples using Stanford Open Information Extraction (OpenIE) framework (Angeli et al., 2015) and cluster the nodes referring to the same mentions utilizing the Extract-Clustering Coreference (ECC) mechanism. As a result, these graphs represent the joint information in the communications between HS and CN authors.

Episodic Knowledge Graphs We create Episodic Knowledge Graph (eKG) representations using the RDF graph generation package[3], with OIE as the base triple extractor. These graphs are usually bigger, but as computational reasons require us to keep the maximum number of nodes comparable to GoT[4], we only keep the `Claims`, `Perspectives`, and `Interactions` subgraphs. We further remove triples stating `rdf:type` and `rdfs:label`. Consequently, eKGs represent the exchange of subjective information in the communication between HS and CN authors.

3.1.2 Encoders

The input text gets encoded by extracting the hidden states of the last layer of the T5 model (Raffel et al., 2020), specifically the FLAN-Alpaca checkpoint[5]. As shown in Figure 1, the prompt is formatted with the HS to be addressed, preceded by the dialogue history.

[2] https://github.com/Zoeyyao27/Graph-of-Thought

[3] https://github.com/leolani/cltl-knowledgerepresentation

[4] The adjacency matrix used as input for the encoder has to have fixed dimensions.

[5] https://huggingface.co/declare-lab/flan-alpaca-base

Table 1: Results from automatic evaluation. Models: **txt (NC)** - text only without context, **txt** - text only with context, **GoT (NC)** - text and GoT graph without context, **GoT** - text and GoT graph with context, **eKG** - text and eKG graph with context. For all metrics, higher is better, except for Repetition Rate and Toxicity where lower is better. Underlined denotes the best scores across all metrics, while * denotes the best scores per modality (text vs graph).

	BLEU1	BLEU4	ROUGEL	METEOR	GLEU	RR	FRE	SentSim	BLEURT	CS	Tox	Div	Nov
txt (NC)	13.57*	0.51	17.10	20.59*	6.62	22.63	50.72*	51.77	-52.64	41.68	35.38*	52.95	66.45*
txt	13.54	0.87*	17.20*	20.24	6.71*	22.47*	49.71	52.11*	-52.00*	46.99*	36.14	57.31*	66.10
GoT (NC)	14.34*	1.03	17.64	20.58	6.99*	22.40*	50.44	51.08	-50.09*	51.72*	33.31*	53.99	66.58*
GoT	13.60	1.10*	17.58	20.98	6.96	23.76	53.45*	53.68	-52.67	41.98	40.09	55.53	66.16
eKG	13.82	0.80	17.65*	21.97*	6.93	24.38	53.15	54.10*	-50.60	42.91	40.95	57.29*	67.38

In parallel, the graph gets encoded using the Graph Attention Network (Veličković et al., 2018; Chen and Yang, 2021). The node embeddings are encoded by the same T5 text encoder, where the text representation of the graph consists of the concatenated triples, having each triple element separated by special <s> </s> tokens.

3.1.3 CN generation

The encoded representations are aligned using a crossed attention layer followed by a gated fusion attention layer. The output of this is fed into the T5 text decoder to generate CNs.

3.2 Evaluation

To verify the method's effectiveness, we carry out standard Natural Language Generation (NLG) automatic evaluation. Following Saha et al. (2024a), we compute BLEU1 (Papineni et al., 2002), BLEU4, ROUGEL (Lin, 2004), METEOR (Banerjee and Lavie, 2005), GLEU (Wu et al., 2016), Repetition Rate (RR), and Flesch Reading Ease (FRE) (Farr et al., 1951). We also use pre-trained models specialized on NLG evaluation: Sentence Similarity (SentSim)[6], BLEURT (Sellam et al., 2020), Counterspeech (CS) (Saha et al., 2024a), and Toxicity (Tox) (Mathew et al., 2021). Finally, we compute metrics for novelty (Nov) and diversity (Div) (Wang and Wan, 2018).

While human evaluation of CN remains difficult due to the subjectivity of the task (Khurana et al., 2022), in this work we follow the criteria from Bengoetxea et al. (2024a) and adapt it to a contextualized setting. Five different aspects were annotated: Relatedness, Specificity, Richness, Coherence, and Grammaticality. We additionally include Effectiveness since it is particularly important for efficient CNs (Benesch, 2014). The scores range between 1-5. Six annotators participated in the evaluation task (see Appendix D), where the HS to be addressed and the dialogue history were provided, followed by 6 shuffled CNs, including those generated by text-only models, graph-based models (with and without context), and the ground data. 10% of the test set was randomly selected for human evaluation, resulting in 30 dialogue instances.

4 Experimental setup

Data We work with the DIALOCONAN dataset (Bonaldi et al., 2022), consisting of 3,059 dialogues between a HS author and a CN author. The average length of these dialogues is 5.43 turns. We divide the data into 80/10/10 for train/dev/test, resulting in 2,447/306/306 instances, respectively. We consider the last CN as the ground truth (avg. length=159.8), the last HS as the one to be addressed (avg. length=91.1), and the rest as discourse context (avg. length=393.4, 3.4 turns).

Models We train five models: two using only the text representations (with and without context), and three including also the graph representations (GoT or eKG). For the ones including graph representations, we tested two models using the GoT graph format (with and without context) and one model using the eKG graph format (with context)[7].

The training details are provided in Appendix B, while an example of the responses produced by each model is shown in Appendix C.

5 Results

We present automatic evaluation in Table 1. The results are in line with recent papers on CN gen-

[6] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[7] Note that eKGs are inherently contextualized representations and thus cannot exclude the dialogue history.

eration, with BLEU scores ranging from 11-16, ROUGE-L scores ranging from 16-18, and GLEU scores ranging from 5-12 (Bengoetxea et al., 2024b; Saha et al., 2024b; Doğanç and Markov, 2023).

Table 2: Results from human evaluation. Abbreviations are provided in the caption of Table 1. Scores range from 1 (low quality) to 5 (high quality).

	Relatedness	Specificity	Richness	Coherence	Gramaticality	Effectiveness
gold	4.20	4.08	4.30	4.68	4.94	3.94
txt (NC)	3.62	3.30	3.28	3.60*	4.70*	2.54
txt	3.66*	3.44*	3.28	3.58	4.38	2.54
GoT (NC)	3.48	3.08	3.46	4.00*	4.72*	2.71
GoT	4.02*	3.70*	3.50*	3.96	4.32	2.82*
eKG	3.66	3.40	3.36	3.64	4.32	2.58

Human evaluation was conducted according to the guidelines provided in Appendix D, and the results are reported in Table 2. The inter-annotator agreement is 37.68% (Average Pairwise Percent Agreement), indicating slight agreement and highlighting the inherent subjectivity of the task (Plank, 2022).

Firstly, regarding graph representations, we observe that models based on graphs outperform those using only text representations, as evidenced by both the automatic metrics and human evaluation.

The benefits of encoding contextual information are less pronounced. In the text modality, automatic metrics hint at better performance for the contextualized models (9 out of 13 metrics); yet this is not reflected in the human evaluation. In the graph modalities, the contextualized representations outperform the non-contextualized ones in the majority of automatic (7 out of 13) and human metrics (4 out of 6). Furthermore, we observe a slight preference for GoT over eKG, which is not a graph model specifically tailored to represent dialogue discourse.

Finally, we note that the non-contextualized GoT model increases the counterspeech strength and reduces the average toxicity of the generated CNs. This improvement is not reflected in the effectiveness score in human annotation. The discrepancy may indicate a mismatch between automatic and human evaluation in contextualized settings. Annotators were explicitly asked to consider the previous dialogue history in their judgments, however the automatic metrics only evaluate a given CN independently of its context (see Limitations).

6 Conclusion

In this work, we explored the usage of graphs for encoding discourse contextual information for CN generation, providing initial evidence of the benefits of contextualized graph representations for this task. Further research could investigate the impact of specific graph representation models (i.e., other syntactic or semantic types of graphs) for capturing context in the task of CN generation. Furthermore, future work could involve the intersection with knowledge-grounded CN generation, using knowledge graph repositories like Wikidata for supporting factually-enhanced CNs, potentially further improving their effectiveness.

Limitations

In this study, we use the DIALOCONAN dataset, which provides a dialogue context and a HS utterance to respond to. It is assumed that the dialogues provided are coherent and fluent, thus giving an opportunity to explore contextualized HS and CNs. However, the dataset was created by grouping HS-CN pairs targeting the same vulnerable group, and then asking annotators to make the dialogue more coherent. This artificial dialogue creation may create a reality-gap with dialogues encountered in the wild. Hence, the dataset's structure could potentially limit the generalizability of this work's conclusions regarding the effectiveness of contextualized models.

Ethics Statement

The models described in this research were developed with the purpose of combating online hate speech. However, we acknowledge that their misuse in unintended contexts could cause the outputs to be inappropriate or even exacerbate harm. Hence, the application of such technology should be approached with caution in real-world contexts and only deployed after a thorough testing, as it may be misused for purposes like censorship.

We recognize that the proposed methodology might inadvertently introduce or amplify biases. Consequently, we strongly recommend manual refinement and analysis of the generated counter-narratives to identify and correct biased outputs before implementing fully automated methods in real-world situations. For this reason, we included

human oversight in the review process when manually evaluating the generated counter-narratives.

In our human evaluation, we included a diverse group of people, aiming to represent a wide range of perspectives. Regardless of the efforts for diversity, we recognize that this annotator pool is not exhaustive and some minorities, which are more exposed to hateful content, might not be represented.

Furthermore, we acknowledge that the model's performance and output depend on the data it is trained on. We use a publicly available dataset developed by the scientific community relying on counter-narratives created by trained NGO operators with the aim to mitigate and limit the spread of hateful content online.

Acknowledgements

This research was funded by the Vrije Universiteit Amsterdam and the Netherlands Organisation for Scientific Research (NWO) via the *Spinoza* grant (SPI 63-260) awarded to Piek Vossen, and the *Hybrid Intelligence Centre* via the Zwaartekracht grant (024.004.022). We thank SURF (www.surf.nl) for the support in using the National Supercomputer Snellius.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. 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 344–354, Beijing, China. Association for Computational Linguistics.

Selene Baez Santamaria, Thomas Baier, Taewoon Kim, Lea Krause, Jaap Kruijt, and Piek Vossen. 2021. EMISSOR: A platform for capturing multimodal interactions as episodic memories and interpretations with situated scenario-based ontological references. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 56–77, Groningen, Netherlands (Online). Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Susan Benesch. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024a. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia. ELRA and ICCL.

Jaione Bengoetxea, Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. 2024b. Basque and Spanish counter narrative generation: Data creation and evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia. ELRA and ICCL.

Helena Bonaldi, Yi-Ling Chung, Gavin Abercrombie, and Marco Guerini. 2024. NLP for counterspeech against hate: A survey and how-to guide. *Computing Research Repository*, arXiv:2403.20103.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Agustín Manuel de los Riscos and Luis Fernando D'Haro. 2021. ToxicBot: A conversational agent to fight online hate speech. *Conversational Dialogue Systems for the Next Decade*, pages 15–30.

Mekselina Doğanç and Ilia Markov. 2023. From generic to personalized: Investigating strategies for generating targeted counter narratives against hate speech. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 1–12, Prague, Czechia. Association for Computational Linguistics.

Silvia Donzelli. 2021. Countering harmful speech online. (In)effective strategies and the duty to counterspeak. *Phenomenology and Mind*, (20):76–87.

James N. Farr, James J. Jenkins, and Donald G. Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Origins and Evolution of Language and Speech*, 280(1):20–32.

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

Helena Gómez-Adorno, Grigori Sidorov, David Pinto, Darnes Vilariño, and Alexander Gelbukh. 2016. Automatic authorship detection using textual patterns extracted from integrated syntactic graphs. *Sensors*, 16(9):1374.

Sabit Hassan and Malihe Alikhani. 2023. DisCGen: A framework for discourse-informed counterspeech generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 420–429, Nusa Dua, Bali. Association for Computational Linguistics.

Sameer Hinduja and Justin W. Patchin. 2007. Offline consequences of online victimization: School violence and delinquency. *Journal of School Violence*, 6:89–112.

Sameer Hinduja and Justin W Patchin. 2010. Bullying, cyberbullying, and suicide. *Archives of suicide research*, 14(3):206–221.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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

Ilia Markov and Walter Daelemans. 2022. The role of context in detecting the target of hate speech. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 37–42, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on Twitter. *Computing Research Repository*, arXiv:1812.02712.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Mary Kate McGowan. 2018. Responding to harmful speech: The more speech response, counter speech, and the complexity of language use 1. In *Voicing Dissent*, pages 182–199. Routledge.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection. *Computing Research Repository*, arXiv:2103.14916.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2018. Author profiling for abuse detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1088–1098, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive Language Detection with Graph Convolutional Networks. 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 2145–2150, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin Group.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305, Online. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1).

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024a. On zero-shot counterspeech generation by LLMs. *Computing Research Repository*, arXiv:2403.14938.

Punyajoy Saha, Aalok Agrawal, Abhik Jana, Chris Biemann, and Animesh Mukherjee. 2024b. On zero-shot counterspeech generation by LLMs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12443–12454, Torino, Italia. ELRA and ICCL.

Tuija Saresma, Sanna Karkulehto, and Piia Varis. 2021. Gendered violence online: Hate speech as an intersection of misogyny and racism. In *Violence, gender and affect*. Palgrave Macmillan.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on Facebook. In *66th ICA Annual Conference*, pages 1–23, Fukuoka, Japan.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 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.

Alexandra A. Siegel. 2020. *Online Hate Speech*, page 56–88. SSRC Anxieties of Democracy. Cambridge University Press.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations ICLR 2018*, Vancouver, BC, Canada. OpenReview.net.

Ke Wang and Xiaojun Wan. 2018. SentiGAN: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Computing Research Repository*, arXiv:1609.08144.

Yao Yao, Zuchao Li, and Hai Zhao. 2023. Beyond chain-of-thought, effective graph-of-thought reasoning in large language models. *Computing Research Repository*, arXiv:2305.16582.

Yi Zheng, Björn Ross, and Walid Magdy. 2023. What makes good counterspeech? a comparison of generation approaches and evaluation metrics. In *Proceedings of the 1st Workshop on CounterSpeech for Online Abuse (CS4OA)*, pages 62–71, Prague, Czechia. Association for Computational Linguistics.

A Graph examples

Listing 1 shows a randomly chosen example dialogue containing a HS utterance to be addressed. Figure 2 shows the three graph representations investigated in this paper. Note that the Graph of Thought (GoT) model focuses on the content of the dialogue (with or without dialogue history), while episodic Knowledge Graphs (eKG) also encode discourse information, such as the order of the claims mentioned (blue nodes) and their speaker provenance (purple nodes).

```
HS: "If you left 1000 men on an island
    with 1000 women who were men but now
     according to morons on twitter are
    now women the islands would become
    extinct in a generation.",

CN: "If you put 1000 men on an island
    with 1000 cisgender women with
    reproductive health conditions that
    prevent them from getting pregnant,
    the population would still become
    extinct in a generation. Are you
    saying women who can't have children
     aren't women?"

HS: "But they've got a womb. What I mean
     is if everyone woke up and decided
    to be a different gender then we'd
    all die out."
```

Listing 1: Example of dialogue containing hate speech.

B Training details

We restrict the input length to 512 tokens and the output length to 256 tokens, based on the dataset analysis. All models were trained for 50 epochs, with a learning rate of $5e-5$ and a batch size of 8. All experiments were run three times. We utilize 4 NVIDIA A100 GPUs with 25G of memory. The experiments were performed on the Dutch National Supercomputer Snellius.

C Counter-narrative examples

Hereby, we provide a randomly chosen example of a dialogue in the DIALOCONAN dataset. The dialogue history is shown in Listing 2 and the generated CNs are shown in Listing 3.

```
HS: "We need to eradicate all those
    refugees coming to our land."

CN: "We should rather show compassion
    and sympathy to people who are
    fleeing persecution."

HS: "They flee here only to destroy our
    culture, we have to stop immigrants
    ."

CN: "Immigrants are part of our culture,
     so it wouldn't exist like we know
    it if we stopped accepting them."

HS: "We cannot keep letting in hundreds
    of thousands of immigrants! This is
    not what the American dream is about
    ."
```

Listing 2: Example of contextual dialogue history.

The generated counter-narrative responses per model, along with the ground truth provided in the dataset, are presented below.

```
# GOLD -
    Well, I hate to break it to you but
    this is EXACTLY what the American
    dream is about: to strive to improve
     your life.

# GoT -
    The American dream is about the
    integration of people from different
     backgrounds, and the acceptance of
    them is the key to the success of
    our country.

# GoT (NC) -
    The American dream is about a free
    and open society, and the fact that
    we are all in this together makes it
     a reality.

# eKG -
    The American dream is about the
    opportunity to live our lives in a
    safe and welcoming country. We
    should not restrict the right of
    immigrants to come to our country.

# txt (NC) -
    The American dream is about equality
     and inclusion, not discrimination
    and violence.

# txt -
    The American dream is about the
    equality of the people, not the
    letting in of people.
```

Listing 3: Examples of counter-narrative responses.

D Annotation details

D.1 Annotators pool

We recruited volunteer annotators from different geographical regions, including 2 European, 2 North American (Mexican), 1 African, and 1 Middle Eastern), genders (4 male, 2 female). All annotators have native-level proficiency in English and are familiar with the task of annotating hate speech (HS) and counter-narratives (CNs). We disclosed to all annotators that they might be exposed to trigger-

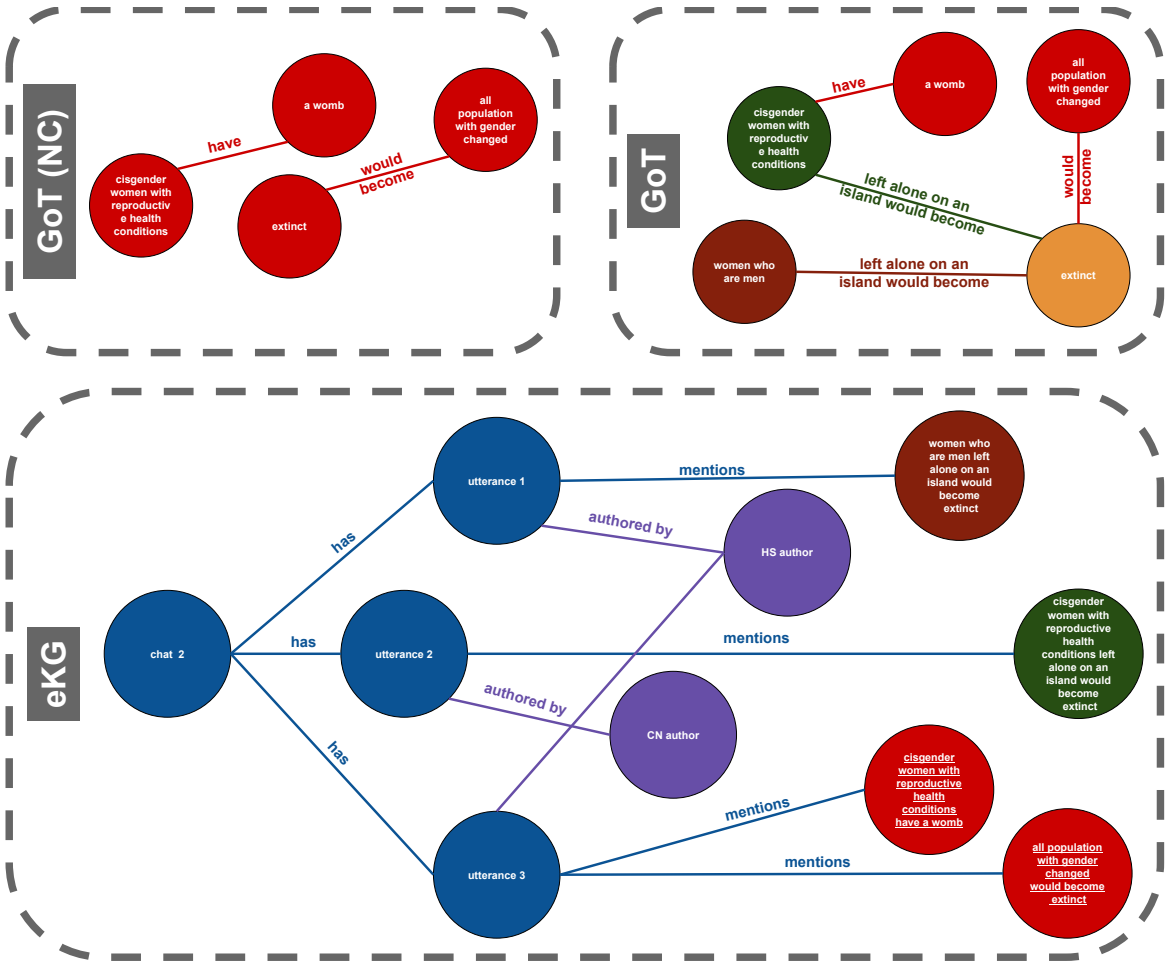

Figure 2: Graph visualization for the three specific representations explored: **GoT (NC)** - text and GoT graph without context, **GoT** - text and GoT graph with context, **eKG** - text and eKG graph with context. In bright red are the triple elements related to the target HS, in dark green are the triple elements related to the previous CN utterances, in dark red are the triple elements related to the previous HS utterances, in blue are the triple elements related to the discourse structure information, and in purple are the triple elements related to the speaker information.

og_idx	id	text	Relatedness	Specificity	Richness	Coherence	Grammatically	Persuasiveness
		Evaluate on a scale from 1 to 5	It measures how relevant the CN is for the HS that it is responding to, given the dialogue context. Reduce points for topics that are missing or for unrelated topics brought up.	It states whether the CN is rather generic or specific for the given dialogue/HS it is responding to, thus replying to the question "can it be used for another completely different dialogue/HS or not?"	In terms of language and vocabulary, it measures whether the CNs are simple or rather complex. Maximum score for a domain specific and versatile vocabulary. Reduce points for common/generic words.	It tells us whether the sentences make sense together, and if all ideas are clear and can be easily understood. Reduce points for incoherence with the dialogue history.	It measures the grammatical correctness of the CNs. Reduce points for repetitions and redundancy within sentences, across sentences and at the dialogue level.	It considers the persuasiveness of the CNs in combating the targeted HS, given the dialogue history. Responds to the question, how effective would this CN be in responding to HS given the dialogue history?
Dialogue history:								
HS: Jews are a disease like COVID that is killing Europeans. CN: I am not sure what you're talking about. The disease is mostly due to contact between people, Jews aren't to blame. They're not a disease.								
HS to address:								
HS: Even if they're not a disease, they are a threat to Europe, they are actually a problem.								
114	7898	Jews are not a threat to Europe, they are just a group of people who have contributed greatly to our country.	3	3	3	4	5	3
114	6470	The European Convention on Human Rights states that the right to life and to life in the public domain is protected by the law. The European Convention on Human Rights states that the right to life and to life in the public domain is protected by the law.	2	1	5	2	1	3
114	2616	Jews are not a threat to Europe. They are just a part of our society.	4	3	3	4	5	2
114	6661	Jews are not a threat to Europe, nor are they a problem. Why should we let this Anti-Semitism run in our country? It's time we start accepting all the people that make up our society, including Jews.	3	3	5	4	5	5
114	9182	Jews are not a threat to Europe. They are just a group of people who have been part of our society for centuries. They are just like us.	4	4	4	4	5	4
114	1275	The EU welcomes people from all over the world, including those from different backgrounds. We must respect them and respect their rights.	3	1	3	3	4	2

Figure 3: Example dialogue annotation provided to annotators.

ing language and ensured they understood and felt comfortable to stop the task at any point.

D.2 Annotation guidelines

The annotators were provided with instructions on the annotation task and a description of the evaluation criteria, as well as an example dialogue and the reasoning behind the given scores. The exam-

ple annotation is shown in Figure 3. The criteria descriptions were as follows:

- **Relatedness**: It measures how relevant the CN is for the HS that it is responding to, given the dialogue context. Reduce points for topics that are missing or for unrelated topics brought up.
- **Specificity**: It states whether the CN is rather generic or specific for the given dialogue/HS it is responding to, thus replying to the question "can it be used for another completely different dialogue/HS or not?"
- **Richness**: In terms of language and vocabulary, it measures whether the CNs are simple or rather complex. Maximum score for a domain specific and versatile vocabulary. Reduce points for common/generic words.
- **Coherence**: It tells us whether the sentences make sense together, and if all ideas are clear and can be easily understood. Reduce points for incoherence with the dialogue history.
- **Gramaticality**: It measures the grammatical correctness of the CNs. Reduce points for repetitions and redundancy within sentences, across sentences and at the dialogue level.
- **Effectiveness**: It considers the persuasiveness of the CNs in combating the targeted HS, given the dialogue history. Responds to the question, how effective would this CN be in responding to HS given the dialogue history?