FC-CONAN: An Exhaustively Paired Dataset for Robust Evaluation of Retrieval Systems

Juan Junqueras, Florian Boudin, May Myo Zin, Nguyen Ha Thanh, Wachara Fungwacharakorn, Damián Ariel Furman, Akiko Aizawa, Ken Satoh

FC-CONAN: An Exhaustively Paired Dataset for Robust Evaluation of Retrieval Systems

Juan Junqueras[1*], **Florian Boudin**[2], **May Myo Zin**[3], **Nguyen Ha Thanh**[3,4], **Wachara Fungwacharakorn**[3], **Damián Ariel Furman**[1], **Akiko Aizawa**[4], **Ken Satoh**[3,4]

[1]Universidad de Buenos Aires, FCEyN, Departamento de Computación, Buenos Aires, Argentina

[2]JFLI, CNRS, Nantes University, Nantes, France

[3]Center for Juris-Informatics, ROIS-DS, Tokyo, Japan

[4]National Institute of Informatics (NII), Tokyo, Japan

Abstract

Hate speech (HS) is a critical issue in online discourse, and one promising strategy to counter it is through the use of counter-narratives (CNs). Datasets linking HS with CNs are essential for advancing counterspeech research. However, even flagship resources like CONAN (Chung et al., 2019) annotate only a sparse subset of all possible HS–CN pairs, limiting evaluation. We introduce **FC-CONAN** (Fully Connected CONAN), the first dataset created by exhaustively considering all combinations of 45 English HS messages and 129 CNs. A two-stage annotation process involving nine annotators and four validators produces four partitions—Diamond, Gold, Silver, and Bronze—that balance reliability and scale. None of the labeled pairs overlap with CONAN, uncovering hundreds of previously unlabelled positives. FC-CONAN enables more faithful evaluation of counterspeech retrieval systems and facilitates detailed error analysis. The dataset is publicly available[1][2].

Keywords: hate speech, counter-narrative, exhaustive annotation, fully paired dataset, dataset creation, benchmark creation, quality-graded partition, label sparsity, lower-bound bias, retrieval-system evaluation, recommender systems, information retrieval, bias in evaluation, evaluation metrics, annotation quality, argumentation, counterspeech, natural language processing (NLP).

1 Introduction

Disclaimer. This paper quotes hate speech verbatim for research purposes; some readers may find the language offensive.

Many Natural Language Processing (NLP) datasets consist of paired sentences, such as questions and answers (Rajpurkar et al., 2016), paraphrases (Dolan and Brockett, 2005), entailment (Bowman et al., 2015), translation (Koehn, 2005), and dialog (Li et al., 2017). While some datasets allow a single sentence to link with multiple others, such as CONAN (Chung et al., 2019), most cover only a fraction of all possible combinations. Exhaustive annotation is rarely attempted due to combinatorial growth and cost, so unlabeled pairs remain ambiguous—often reflecting oversight rather than a true absence of relation. This incompleteness is especially problematic for recommendation tasks, where metrics can severely underestimate system performance (§4.3).

This challenge becomes particularly acute in domains such as hate speech. Social media has amplified the spread of harmful rhetoric (Silva et al., 2021), (Waseem and Hovy, 2016), prompting responses beyond content removal, which can reinforce censorship narratives. As a more constructive alternative, structured counterspeech protocols focus on timely, thoughtful responses that dismantle harmful arguments, avoid fostering further conversations, and align with broader goals. The CONAN dataset (COunter NArratives through Nichesourcing) (Chung et al., 2019), "the first large-scale, multilingual, expert-based dataset of hate speech/counter-narrative pairs", remains the primary resource. To illustrate the nature of this data, we provide a representative example below.

Example 1 (An HS–CN pair from CONAN (Chung et al., 2019))**.** ***HS (hate speech)*** *"I hate Muslims. They should not exist."*
CN (counter-narrative). *"Muslims are human too. People can choose their own religion."*

Despite its widespread use, CONAN has notable limitations. **Crucially**, it does *not* annotate all HS–CN combinations, leaving many appropriate pairs unlabeled. This limits its usefulness, specifically for evaluating CN recommendation systems, as performance metrics reflect only a lower bound. In a pilot study using one of these systems, we found that while only 2 of 10 suggested CNs were labeled as appropriate, manual review judged 8 to be valid—highlighting the risk of underestimating system accuracy.

The lack of full HS–CN pair annotations also limits generation tasks by reducing training data for fine-tuning LLMs. Comprehensive HS–CN annotations would further enable methods such as contrastive learning. Ultimately, unannotated pairs leave valuable latent information unused, decreasing the dataset's utility for downstream applications.

Another characteristic of the dataset is that the guidelines are rather open-ended. This approach stems from the fact that the original annotators had already been trained to

*Corresponding author: jjunqueras@dc.uba.ar

[1]The dataset is publicly available at https://github.com/jnqeras/FC-CONAN-dataset

[2]This work was partially completed while the first author was at the National Institute of Informatics (NII), Tokyo, Japan.

follow NGO guidelines for crafting effective CNs. These guidelines are notably consistent across both languages and organizations, and closely mirror those established in the Get the Trolls Out project[3]. Annotators were encouraged to rely on their intuition, avoid overthinking, and produce reasonable responses (Chung et al., 2019, §3.2). It's important to note that the high level of subjectivity is a characteristic of this field.

Due to resource constraints, we re-annotated a representative *subset* of all possible HS–CN pairs. This paper details that effort, originally motivated by the need to evaluate a counterspeech recommender more accurately. While CONAN covers three languages, our work focuses solely on English. Extending the annotation to other languages is future work.

2 Related work

In the hate speech domain, CONAN (Chung et al., 2019) is among the best-known multilingual resources. Expert-curated and focused on Islamophobia, it features hate speech (HS) and counter-narrative (CN) pairs in English, Italian, and French. Initially, included 4,078 pairs (1,288 in English) based on 136 unique HS messages, each matched with an average of 9.5 CNs. Through translation and paraphrasing, the English portion was expanded to 6,654 pairs, with 408 unique HS messages and 1,270 CNs. The dataset also includes metadata such as expert demographics, CN type, and HS sub-topic.

Several other datasets focus on hate speech and counter-narratives, such as DIALOCONAN (Bonaldi et al., 2022), which features multi-turn dialogues between a hater and an NGO operator, though it is not organized in HS/CN pairs. Another example is Multitarget CONAN (Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco, 2021), a dataset with HS/CN pairs addressing multiple targets of hate. However, these datasets do not consider every possible combination of HS and CN pairs. From (Furman et al., 2023), one finding is most relevant to our work: a small LLM fine-tuned on a few hundred high-quality HS–CN pairs can outperform larger models. Our work complements theirs by focusing not on argumentative cues, but on exhaustively pairing HS–CN examples for robust evaluation and training.

A number of datasets have been developed for the task of hate speech detection, such as the Twitter corpus introduced by (Waseem and Hovy, 2016). Although our dataset could potentially be utilized for this purpose, it is primarily designed with a different focus.

3 Dataset Creation

We present FC-CONAN, a dataset of HS–CN pairs derived from a subset of the CONAN corpus. During annotation, we exhaustively considered all possible combinations of selected HS and CN items, labeling each pair as *appropriate* or *non-appropriate*, with some removed based on quality control criteria.

[3]https://getthetrollsout.org/stoppinghate

From the English partition of the CONAN dataset (the only language common to all annotators), we randomly selected 45 HS messages and collected the 375 CNs originally paired with them. For each CN, we used the SBERT model `all-MiniLM-L6-v2`[4] to retrieve its two most similar CNs from the full CONAN dataset. We then discarded the 375 original CNs while retaining the HS messages and their newly retrieved CNs, ensuring that only novel HS–CN combinations were kept, so that annotators worked with CNs similar to the originals but not identical, preventing the task from being too easy. This process resulted in 133 unique CNs, from which we randomly selected 129 to ensure an even distribution of HS–CN pairs across annotators, producing 5,805 HS–CN pairs. Nine annotators—academically trained volunteers—labeled these pairs following adapted CONAN guidelines: fact-based information and maintaining a non-offensive tone to avoid escalating the conversation. Labels were assigned independently, with overlap enabling inter-annotator agreement checks, and adjudication resolved conflicts. Pairs marked *not sure* or irreconcilably disputed were discarded to avoid label bias, leaving 5,032 labeled pairs (4,143 as *'the CN is not appropriate for the given HS'* and 889 as *'the CN is appropriate for the given HS'*).

To further ensure reliability, 4,000 adjudicated pairs underwent a validation round by four independent reviewers (not involved in initial annotation), each re-assessing 1,000 pairs. Validators applied the same guidelines, could skip distressing items (three pairs were skipped), and judged whether labels conformed. This process ensured every retained pair has both annotator and validator input. The validated pairs were retained, regardless of whether they were deemed valid or not. Pairs outside this set were discarded to prioritize label reliability over dataset size. Although some excluded pairs may still be appropriate, all possible combinations were reviewed during annotation.

The final resource balances reliability and coverage by defining four quality-graded subsets (*Diamond*, *Gold*, *Silver*, *Bronze*), allowing users to trade size for label confidence. Ethical safeguards included warnings, optional skipping, and on-demand debriefing breaks to support annotators. The dataset may pose dual-use risks if inverted to generate hateful replies; we therefore stress the need for responsible downstream use.

4 Results / Analysis.

We begin by describing the dataset itself before transitioning to system-level evaluation. Section 4.1 introduces the four quality-graded partitions generated through our annotation and validation pipeline. Section 4.2 then explores their internal structure. Finally, Section 4.3 presents a retrieval experiment that quantifies the impact of these partitions on downstream performance.

Partition	Total	Appr.	Non-Appr.
Diamond	551	35	516
Gold	663	54	609
Silver	3580	431	3149
Bronze	3997	702	3295

Table 1: Number of hate speech – counter-narrative (HS–CN) pairs in each dataset partition, categorized by appropriateness.

4.1 Dataset Partitions.

Following the annotation and validation processes (§3), we obtained HS–CN pairs annotated by one to three annotators and assessed for validity by one validator. Based on whether annotator labels aligned and the results of the validation process, we defined four distinct dataset partitions, each differing in annotation quality and size. Table 1 summarizes the size of each partition and the distribution of appropriate vs. non-appropriate HS–CN pairs. The characteristics of each partition are detailed below:

- **Diamond Standard Dataset:** This partition includes only HS-CN pairs annotated by two or more annotators who reached unanimous agreement –whether the counter-narrative was deemed appropriate or non-appropriate. Additionally, a validator has confirmed the accuracy of these annotations.
- **Gold Standard Dataset:** This partition extends the *Diamond Standard Dataset* by incorporating additional HS-CN pairs annotated by two or more annotators, regardless of whether the annotators unanimously agreed. In cases where annotators initially disagreed, these disagreements were resolved through the adjudication process (mentioned in § 3). Each resulting annotation was further reviewed and confirmed as accurate by a validator.
- **Silver Standard Dataset:** The *Silver Standard Dataset* includes all HS-CN pairs from the *Gold Standard Dataset*, along with pairs annotated by only one annotator and subsequently confirmed by the validators. Thus, annotations in this partition come from 1 to 3 annotators, with initial disagreements resolved via the aforementioned adjudication process. All annotations in this partition were approved during the validation stage.
- **Bronze Standard Dataset:** This dataset comprises all entries from the *Silver Standard Dataset*, supplemented by all the HS-CN pairs that were not approved during the validation stage. As before, annotations originate from 1 to 3 annotators, with disagreements resolved through the adjudication phase. However, unlike the Silver Standard, this partition also retains pairs that were not approved during the validation stage.

As expected, these partitions differ inversely in size and quality. Higher-quality datasets (Diamond and Gold) require greater annotation agreement and validation, resulting in smaller dataset sizes. Conversely, lower-quality datasets (Silver and Bronze) are larger but contain annotations with potentially reduced reliability. Thus, the datasets, arranged from smallest to largest (highest to lowest quality), are: Diamond, Gold, Silver, and Bronze Standard datasets.

4.2 Qualitative Analysis of the Dataset

Across the six overlapping subsets described in Section 3, we obtain a mean Cohen's $\kappa = 0.34$ ($\sigma \approx 0.20$), computed *before* any additional checks. As noted by Klie et al., "*although it is often treated as such, agreement is no panacea; high agreement does not automatically guarantee high-quality labels.*" (Klie, Eckart de Castilho, and Gurevych, 2024), so we applied the validation procedure outlined in Section 3. This section presents a qualitative analysis of the four dataset partitions: Diamond, Gold, Silver, and Bronze. Understanding these partitions helps users select the most suitable subset—prioritizing annotation reliability (Diamond/Gold) or volume (Silver/Bronze).

Table 1 confirms that many valid HS–CN pairs were unannotated in the original CONAN dataset. By evaluating new combinations, we recovered hundreds of appropriate and inappropriate pairs across all partitions.

Table 2 shows a trade-off between quality and quantity: Diamond and Gold are smaller but fully valid, Silver is larger and still fully validated, while Bronze is the largest yet includes some non-valid pairs—allowing users to choose between size and reliability for downstream tasks.

Table 3 illustrates the distribution of annotated pairs based on the number of annotators involved. Within the Bronze partition, of the 199 pairs annotated by three annotators, only 4 pairs (≈2.01%) were deemed invalid by validators. For the 502 pairs annotated by two annotators, 34 pairs (≈6.77%) were marked invalid. Lastly, among the 3,296 pairs annotated by a single annotator, 379 pairs (≈11.49%) were classified as invalid. These observations indicate a clear trend: pairs annotated by multiple annotators tend to have proportionally fewer invalid instances, underscoring how reliability significantly improves with increased annotator agreement.

In sum, the four-tier partitioning balances label reliability. Diamond and Gold deliver perfect validation, making them ideal for benchmarking model performance under minimal label noise. Silver adds scale without compromising valid pairs, while Bronze boosts volume, introducing the only subset of non-valid pairs. Altogether, the new annotations help fill clear gaps in the original CONAN dataset.

4.3 Experimental Evaluation: Evaluation of Recommendation Systems.

As discussed in §1, incomplete CONAN labels masked appropriate CNs. In what follows, we re-evaluate recommendation systems on our exhaustively annotated FC-CONAN partitions[5].

We compare twelve recommenders trained on an English-only dataset (*conan_not_in_bronze_train*) created by exclud-

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

[5]The datasets used in this experiment are available at https://github.com/jnqeras/FC-CONAN-dataset/tree/main/recommender_experiment_data

Partition	Valid	Non-Valid
Diamond	551	0
Gold	663	0
Silver	3580	0
Bronze	3580	417

Table 2: Number of hate speech – counter-narrative (HS–CN) pairs in each dataset partition, categorized by validity.

Partition	3 annot.	2 annot.	1 annot.
Diamond	195	356	0
Gold	195	468	0
Silver	195	468	2917
Bronze	199	502	3296

Table 3: Number of HS–CN pairs annotated by 3, 2 and 1 annotators.

ing any HS or CN found in the Bronze partition, ensuring no overlap between training and evaluation. Specifically, we evaluate:

- **TF–IDF** – cosine similarity on TF–IDF vectors (sparse baseline); (Salton and Buckley, 1988).
- **BM25** – Okapi BM25 lexical ranker (k_1=1.2, b=0.75); (Robertson and Zaragoza, 2009).
- **random** – uniform sampling of ten candidate CNs.
- **sbert(MiniLM)** – Sentence-BERT *all-MiniLM-L6-v2* (384 d) + cosine; (Reimers and Gurevych, 2019), (Wang et al., 2020).
- **sbert(MPNET)** – Sentence-BERT *all-mpnet-base-v2* (768 d) + cosine; (Song et al., 2020).
- **sbert(RoBERTa-L)** – Sentence-BERT *all-roberta-large-v1* (1024 d) + cosine; (Zhuang et al., 2021).
- **textEmb3L** – OpenAI dense encoder *text-embedding-3-large* (3 072 d); (OpenAI, 2024b).
- **GPT-4o** – zero-shot prompting of *gpt-4o-2024-08-06* to return ten CNs; (OpenAI, 2024a).
- **RR25 system 1 / RR50 system 1** – top-25/50 CNs from textEmb3L, then GPT-4o re-ranks.
- **RR25 system 2 / RR50 system 2** – top-25/50 CNs from sbert(MPNET), then GPT-4o re-ranks.

To ensure a fair comparison, each system ranks a fixed candidate pool: the 100 unique CNs from the Bronze partition. This simulates a deployment where all models share the same, unchanging reply catalogue. To maintain consistency, we limit our query set to the hate speech instances in the Bronze partition (*bronze_HS*). For each HS in *bronze_HS*, each recommender produces its top-10 counter-narratives, yielding one ranking per model.

We then evaluate each ranking against all four partitions (Diamond, Gold, Silver, and Bronze). A recommendation is counted as correct if —and only if— the $\langle \text{HS}; \text{CN} \rangle$ pair is annotated as appropriate in the partition under test. As an example, we show the TF-IDF ranking for the following HS (capped at 2 CNs to keep the paper concise):

- "When Islam hasn't arrived yet in England, things was good. Now we are full of muslim rapists and it is evident that sexual violence is jihad in England".

The ranking is marked ✔ or ✖ for each partition to illustrate how appropriateness varies depending on which annotation standard we use.

1. Muslims aren't rapists or murderers, in fact many Muslims are doing good things to fight the extremists falsely appropriating Islam: https://5pillarsuk.com/2019/01/03/muslim-woman-praised-for-aiding-victims-of-new-years-eve-terror-attack-in-manchester/. [D✖, G✖, S✖, B✖]
2. Rapists are, in most cases, friends, family, or partners of the victim. The proportion of Pakistani people who are rapists is no more than that of White British population. [D✖, G✔, S✔, B✔]

In the example ranking, the top-ranked counter-narrative (CN) is never judged appropriate in any partition. By contrast, the second-ranked CN is judged appropriate in the Gold, Silver, and Bronze partitions, but not in Diamond. This demonstrates that the evaluation metrics computed on a generated ranking can change substantially depending on which partition is used to define "appropriate" pairs.

Figures 1 and 2 show that recommender systems performance improves with partitions containing a greater number of annotated "appropriate" pairs: Diamond scores lowest, followed by Gold, then Silver, with Bronze achieving the highest values across both metrics. Overall, metric values scale roughly in proportion to the number of appropriate pairs in each partition. We observe the same trend for metrics such as NDCG@10, MAP@10, Precision@10, Accuracy@10, and F1@10, although their plots are omitted due to space limitations.

None of the pairs formed from the candidate pool (Bronze CNs) and *bronze_HS* are labeled as appropriate in the CONAN dataset; thus, using CONAN as the sole gold standard for this set yields zero scores across all metrics. As progressively more appropriate pairs are included—from CONAN's subset to Diamond, Gold, Silver, and Bronze—metric scores increase. This suggests that when appropriate pairs remain unannotated, evaluation metrics serve only as lower bounds and fail to reflect true retrieval system performance.

Table 4 shows a clear performance hierarchy. **Embedding-based rankers** (all SBERT variants plus OpenAI's `textEmb3L`) obtain the highest average score ($\mu \approx 0.32$) and the lowest coefficient of variation ($CV\% \approx 46$), indicating that dense vector representations are both *effective* and *robust* to missing annotations. **Hybrid rerankers** (RR50 system 1/2 and RR25 system 2) come next in terms of performance ($\mu \approx 0.23$) yet remain substantially less stable ($CV\% \approx 69$), presumably because the GPT-4o reranking step amplifies noise whenever the embedding pre-filter retrieves weak candidates. Among the **lexical baselines**, TF–IDF matches hybrid effectiveness ($\mu \approx 0.22$) while BM25 lags behind ($\mu \approx 0.18$); both exhibit high variability ($CV\% > 63$), confirming their sensitivity to annotation sparsity. The **LLM zero-shot** strategy (GPT-4o alone) clusters with the Hybrid reranker RR25 system 1

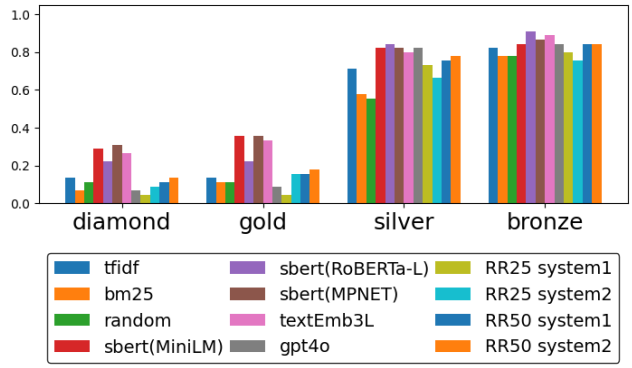

Figure 1: HIT RATIO@10 across the twelve systems (glossary in Section 4.3).

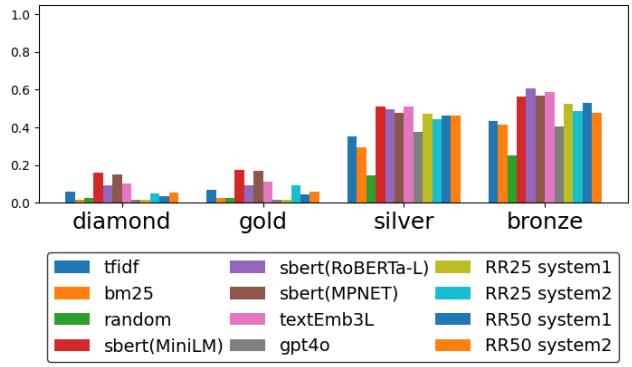

Figure 2: MRR@10 of the twelve evaluated systems (see glossary in Section 4.3).

($\mu \approx 0.21$, $CV\% \approx 89$). As expected, the **random** baseline sits at the bottom ($\mu = 0.15$). Overall, the results support the conclusion that *representation quality—rather than pipeline complexity alone—drives both effectiveness and robustness* in counter-narrative retrieval.

5 Limitations and Future Work

Generative fine-tuning and contrastive learning: Generative models can be fine-tuned on appropriate pairs from each partition to compare performance—we conducted such experiments but omitted them due to space limitations—while contrastive learning could leverage both appropriate and non-appropriate pairs to fully exploit the dataset's structure.

Language scope: We cover only the English portion of CONAN. Extending exhaustive pairing to French and Italian remains future work.

Annotation coverage: Our 45 HS × 129 CN subset produced partitions large enough to reveal evaluation artefacts, yet remains far from a *fully* exhaustive re-annotation of all possible combinations of CONAN. A semi-automatic "LLM-first, human-verify" pipeline could finish that job at lower cost.

Demographic diversity: Our annotator pool is skewed towards young, highly educated English-speaking individuals. Broader demographic sampling would reveal whether cultural background influences appropriateness judgments.

System	Avg.	Min.	Max.	CV%
sbert(MPNET)	0.3283	0.1836	0.4904	38.425
sbert(MiniLM)	0.3270	0.1766	0.4804	39.200
textEmb3L	0.3128	0.1366	0.5098	53.200
sbert(RoBERTa-L)	0.3034	0.1214	0.5138	56.400
RR50 system2	0.2441	0.0658	0.4271	69.175
RR50 system1	0.2416	0.0526	0.4508	75.350
RR25 system2	0.2278	0.0514	0.3992	63.175
TF-IDF	0.2203	0.0703	0.4024	63.275
RR25 system1	0.2165	0.0184	0.4339	91.950
GPT-4o	0.2154	0.0267	0.4118	87.100
BM25	0.1819	0.0296	0.3810	81.650
random	0.1500	0.0414	0.3115	78.100

Table 4: Macro-level robustness of the 12 systems. For each system we average, over four metrics (HIT RATIO@10, MRR@10, NDCG@10, MAP@10), the metric–wise *mean*, *minimum*, *maximum* and coefficient of variation (CV% = (standard deviation/mean) × 100). Higher "Avg." indicates better overall effectiveness, while lower CV% indicates greater stability across the Diamond–Bronze partitions.

6 Conclusions

We introduced **FC-CONAN**, to the best of our knowledge, the *first* hate speech / counter-narrative dataset where *every* possible pairing between two finite sets—45 HS messages and 129 CNs—is explicitly judged.

Consequently, if a pair is labeled *appropriate* in a given partition, it means that—according to the requirements of that partition—it was indeed deemed suitable. Conversely, if it is labeled *non-appropriate*, it reflects that it did not meet those same partition-specific criteria.

Four quality-controlled partitions—DIAMOND, GOLD, SILVER, and BRONZE—let practitioners trade annotation reliability for corpus size.

None of the HS–CN pairs we annotated occurs in the original CONAN corpus; our partitions therefore add hundreds of previously missing *appropriate* CNs, revealing how many unlabeled appropriate pairs the sparse labels in CONAN contained.

Importantly, using the partitions introduced, we confirm that not considering all possible pairs during dataset creation leaves many appropriate pairs unannotated. This results in artificially lower scores for most evaluation metrics in recommendation systems. For metrics that are not negatively affected by a greater number of positives—such as MRR@10 or Hit Ratio@10—this implies that scores obtained when evaluating on partially annotated datasets should be considered *lower bounds*.

We also showed that embedding-based rankers outperform others in both effectiveness and robustness to missing annotations.

In summary, researchers evaluating counter-narrative *retrieval* systems should rely on densely annotated datasets such as ours to avoid underestimating system performance.

Acknowledgements

This work was supported by the "R&D Hub Aimed at Ensuring Transparency and Reliability of Generative AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology and JSPS KAKENHI Grant Numbers, JP22H00543. We also sincerely thank the annotators and validators who generously volunteered their time to contribute to this project. Junqueras was partially supported by the UBA BIICC Fellowship Program, the Fundar FunDatos Fellowship Program, and the NII International Internship Program.

References

Abelson, H.; Sussman, G. J.; and Sussman, J. 1985. *Structure and Interpretation of Computer Programs*. Cambridge, Massachusetts: MIT Press.

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 May, J.; Shutova, E.; Herbelot, A.; Zhu, X.; Apidianaki, M.; and Mohammad, S. M., eds., *Proceedings of the 13th International Workshop on Semantic Evaluation*, 54–63. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Baumgartner, R.; Gottlob, G.; and Flesca, S. 2001. Visual information extraction with Lixto. In *Proceedings of the 27th International Conference on Very Large Databases*, 119–128. Rome, Italy: Morgan Kaufmann.

Bonaldi, H.; Dellantonio, S.; Tekiroglu, S. S.; and Guerini, M. 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*, 8031–8049. Association for Computational Linguistics.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Brachman, R. J., and Schmolze, J. G. 1985. An overview of the KL-ONE knowledge representation system. *Cognitive Science* 9(2):171–216.

Chung, Y.-L.; Kuzmenko, E.; Tekiroglu, 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*, 2819–2829. Florence, Italy: Association for Computational Linguistics.

Chung, Y.-L.; Abercrombie, G.; Enock, F.; Bright, J.; and Rieser, V. 2024. Understanding counterspeech for online harm mitigation. *Northern European Journal of Language Technology* 10:30–49.

Chung, Y.-L.; Tekiroğlu, S. S.; and Guerini, M. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics.

DataCanary. 2024. Quora question pairs.

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.

Dolan, B., and Brockett, C. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*. Asia Federation of Natural Language Processing.

Fanton, Margherita and Bonaldi, Helena and Tekiroğlu, Serra Sinem and Guerini, Marco. 2021. Human-in-the-Loop for Data Collection: a Multi-Target Counter Narrative Dataset to Fight Online Hate Speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Furman, D.; Torres, P.; Rodriguez, J.; Martinez, L.; Alonso Alemany, L.; Letzen, D.; and Martinez, M. V. 2022. Parsimonious argument annotations for hate speech counter-narratives.

Furman, D.; Torres, P.; Rodríguez, J.; Letzen, D.; Martinez, M.; and Alemany, L. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2942–2956. Singapore: Association for Computational Linguistics.

Gao, L., and Huang, R. 2017. Detecting online hate speech using context aware models. In Mitkov, R., and Angelova, G., eds., *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 260–266. Varna, Bulgaria: INCOMA Ltd.

Gillespie, T. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press.

Goffredo, P.; Basile, V.; Cepollaro, B.; and Patti, V. 2022. Counter-TWIT: An Italian corpus for online counterspeech in ecological contexts. In Narang, K.; Mostafazadeh Davani, A.; Mathias, L.; Vidgen, B.; and Talat, Z., eds., *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, 57–66. Seattle, Washington (Hybrid): Association for Computational Linguistics.

Gottlob, G.; Leone, N.; and Scarcello, F. 2002. Hypertree decompositions and tractable queries. *Journal of Computer and System Sciences* 64(3):579–627.

Gottlob, G. 1992. Complexity results for nonmonotonic logics. *Journal of Logic and Computation* 2(3):397–425.

Klie, J.-C.; Eckart de Castilho, R.; and Gurevych, I. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics* 50(3):817–866.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, 79–86.

Levesque, H. J. 1984a. Foundations of a functional approach to knowledge representation. *Artificial Intelligence* 23(2):155–212.

Levesque, H. J. 1984b. A logic of implicit and explicit belief. In *Proceedings of the Fourth National Conference on Artificial Intelligence*, 198–202. Austin, Texas: American Association for Artificial Intelligence.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In Kondrak, G., and Watanabe, T., eds., *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Nebel, B. 2000. On the compilability and expressive power of propositional planning formalisms. *Journal of Artificial Intelligence Research* 12:271–315.

OpenAI. 2024a. Gpt-4o technical report. https://openai.com/research/gpt-4o.

OpenAI. 2024b. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/. Accessed: 2025-08-02.

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

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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)*, 3982–3992.

Robertson, S., and Zaragoza, H. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval* 3:333–389.

Salles, I.; Vargas, F.; and Benevenuto, F. 2025. HateBRXplain: A benchmark dataset with human-annotated rationales for explainable hate speech detection in Brazilian Portuguese. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 6659–6669. Abu Dhabi, UAE: Association for Computational Linguistics.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5):513–523.

Silva, L.; Mondal, M.; Correa, D.; Benevenuto, F.; and Weber, I. 2021. Analyzing the targets of hate in online social media. *Proceedings of the International AAAI Conference on Web and Social Media* 10(1):687–690.

Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc.

Vallecillo Rodríguez, M. E.; Cantero Romero, M. V.; Cabrera De Castro, I.; Montejo Ráez, A.; and Martín Valdivia, M. T. 2024. CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 3677–3688. Torino, Italia: ELRA and ICCL.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20. Red Hook, NY, USA: Curran Associates Inc.

Waseem, Z., and Hovy, D. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In Andreas, J.; Choi, E.; and Lazaridou, A., eds., *Proceedings of the NAACL Student Research Workshop*, 88–93. San Diego, California: Association for Computational Linguistics.

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 Burstein, J.; Doran, C.; and Solorio, T., eds., *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)*, 1415–1420. Minneapolis, Minnesota: Association for Computational Linguistics.

Zhuang, L.; Wayne, L.; Ya, S.; and Jun, Z. 2021. A robustly optimized BERT pre-training approach with post-training. In Li, S.; Sun, M.; Liu, Y.; Wu, H.; Liu, K.; Che, W.; He, S.; and Rao, G., eds., *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, 1218–1227. Huhhot, China: Chinese Information Processing Society of China.