Toximatics: Towards Understanding Toxicity in Real-Life Social Situations

Mayukh Das and Wolf-Tilo Balke Institute for Information Systems, TU Braunschweig Mühlenpfordtstraße 23, 38106 Braunschweig, Germany {mayukh,balke}@ifis.cs.tu-bs.de

Abstract

The rise of social media has amplified the visibility and impact of hate speech, prompting the development of NLP solutions to identify both explicit and implicit forms of hate speech. These approaches assess toxicity in isolation, neglecting context and limiting models to sentence-level understanding. Therefore we study, how contextual factors influence perceived toxicity, thereby anchoring assessments in a more nuanced semantic framework. We introduce a novel synthetic data generation pipeline designed to create contextutterance pairs at scale with controlled polarity. This pipeline can enhance existing hate speech datasets by adding contextual information to utterances, either preserving or altering their polarity, and also generate completely new pairs from seed statements. We utilised both features to create Toximatics, a dataset that includes context-dependent utterances and it's toxicity score. To address biases in state-of-the-art hate datasets, which often skew towards specific sensitive topics such as politics, race, and gender, we propose a method to generate neutral utterances typical of various social settings. These are then contextualized to show how neutrality can shift to toxicity or benignity depending on the surrounding context. Toximatics' approach to hate speech detection extends beyond the sentence level, rendering it suitable for discourse analysis and also revealing that current models underperform on this dataset.

1 Introduction

Toxicity classifiers are normally fine-tuned with hate speech datasets that contain explicit or overtly abusive lexicons (Davidson et al., 2017; Founta et al., 2018) or implicit, coded, indirect framing of offensiveness (ElSherief et al., 2021; Hartvigsen et al., 2022). Explicit hate datasets suffer from topic bias like over-reliance on sensitive attributes (race, gender, religion, nationality, etc) (Basile et al., 2019) which can inflate model performance on phrases containing indirect offense. Implicit hate speech introduces diverse hate classes based on coded language such as irony, sarcasm, euphemism, metaphor, circumlocution, etc (Talat and Hovy, 2016; Magu and Luo, 2018; Gao and Huang, 2017; Warner and Hirschberg, 2012; Qian et al., 2019). This line of work approaches the problem of detecting toxicity/hate speech as: Given a sequence of n tokens $u \in \{x_1, x_2...x_{target}...x_n\}$ with x_{target} being a sensitive attribute or target group type, can a machine flag whether u is *toxic*? (Sap et al., 2020) defines the task differently, like given a sequence u, can a machine determine if a possible toxicity is intentional, offensive, lewd towards any identity group. The sequence u could be interpreted as an utterance by a person in online or real-life social situations. In NLP datasets, these sequences are usually mined via social media or machine-generated with specialized prompts (Hartvigsen et al., 2022).

However, does toxicity depend solely on the semantics of the utterance u? Our hypothesis is that this current method is simplistic as it bases the concept of toxicity solely on the utterance. To accurately determine toxicity, it is necessary to take into account the broader context. This paper illustrates how the context of spoken dialogue can alter the human toxicity score of the same utterances (see Figure 1). Consequently, we also highlight other flawed assumptions made by the previous research. Previous research on hate speech has highlighted the importance of considering characteristics such as abusive language, speaker intention, sensitive attributes, target group, and framing when determining what constitutes hate speech. However, we speculate human perceived toxicity can occur for any social category if grounded in context, e.g. {Irony towards an introvert in a social gathering. Recent implicit datasets (Sap et al., 2020; Hartvigsen et al., 2022; Zhou et al.,



Figure 1: Toximatics dataset with it's parent seed utterances. Direct augmentation do and multistage augmentation do are generation methods introduced in section 4. Labels are the chat-gpt-4 legacy, the hate-bert, the roberta-toxigen, lama-2-chat-7b, with an annotation. Toximatics makes the polarity of the utterance context-dependent and fools the models.

2023) overemphasize dimensions like intention, power dynamics, or target groups as hate determiners. On the other hand, we highlight that toxic perlocution can occur regardless of power dynamic, identity group. Furthermore, we observed distinct examples for such cases are clearly missing in stateof-the-art papers and datasets.

To address all this gap we introduce *Toximatics*. A dataset to understand pragmatic toxic utterance which encompasses deeper level semantics than implicit datasets. We ground the notion of toxicity to the context of the utterance, rather than grounding it solely to the utterance as done in previous work. Unlike previous work, our samples consist of an utterance-context pair. The crowdworkers were presented with the context while annotating the utterance to make sure the validity of the toxicity score becomes depended on the context. We also introduce a generation pipeline that utilizes state-of-the-art language models and expertly crafted prompts and methods. This enables to generate similar examples at scale without the need for further quality checks, ensuring the high level of accuracy. Our generation pipeline uses utterances from implicit datasets as seeds to create context with controlled polarity and also generates entirely new utterance-context pairs from these seeds. Additionally, we produce seemingly neutral utterances, atypical for certain social situations, and generate contexts for them with specific polarity control. Following the later step we explore whether phrases like { You're so lucky to be *able to work from home and have more flexibility*} could be perceived as toxic without any allusion to power dynamics, intent, or identity groups (see Figure 1). We release a dataset of 19,800 utterancecontext pairs and their toxicity labels. Toximatics was evaluated with sota toxicity classification models, foundational models and chat models, which showed poor efficacy on zero-shot classification task. This dataset is the first of its scale to pivot hate speech detection research towards a contextdependent framework. The dataset and generation method codes are available via the provided link 1 .

2 Related Work

Early papers on hate datasets emphasized explicit abusive language and profane use of slurs racial identifiers, minority mentions, hateful keywords, etc (Basile et al., 2019; Davidson et al., 2017; Warner and Hirschberg, 2012; Silva et al., 2016;

¹https://github.com/Mayukhga83/Toximatics

Burnap and Williams, 2014; de Gibert et al., 2018). These examples are collected using keyword-based, bootstrap scrapping, or adversarial data collection (Davidson et al., 2017; Zampieri et al., 2019; Founta et al., 2018; Waseem, 2016; Dinan et al., 2019; Vidgen et al., 2021). These datasets have an over-reliance on lexical cues and specific topics. In response to this, researchers have tried to curate a newer corpus that labels hate considering the rhetorical framing grounded in sociology and psychology (Kennedy et al., 2018; Sap et al., 2019). ElSherief et al. (2021) introduced a taxonomy of implicit hate speech and a benchmark corpus mined from online hate groups. Hartvigsen et al. (2022) uses GPT 3, demonstration-based prompting and constrained decoding to generate large-scale implicit hate corpus. Pavlopoulos et al. (2020) investigated the potential effect of context on human judgment of toxicity scores through an analysis of Wikipedia discussions. Xenos et al. (2021) created a toxicity dataset where the annotators had access to one previous comment. Zhou et al. (2023) developed a formalism to explain the intentions, reactions, and harms of offensive or biased statements based on their social context.

Unlike most previous works which focuses on a single statement, we have a situational context in which the statement was uttered. While the previous work examines the extent to which the framing of an utterance determines its level of toxicity, our work investigates the extent to which the context determines the toxicity level of an utterance. Previous studies (Pavlopoulos et al., 2020; Xenos et al., 2021) narrowly examined context, focusing solely on preceding comments and discussion headings. However, this limited approach may fail to capture the circumstances of the utterance. Thus, we propose contextualizing the situation with a situation descriptor (see Figure 1). Zhou et al. (2023) heavily relies on the identity group of both the speaker and listener. Additionally, the context description is limited in scope. In contrast, we solely use detailed situational descriptor as context as an explaination of the entire scenario. We also have curated examples to show how toxicity can be perceived without any allusion to identity group. Zhou et al. (2023) also does not generate large scale polarity controlled context, they only have 928 counterfactual context. In contrast, our work solely deals with generating polarity-controlled context. None of the previous works have focused on generating completely new implicit hate utterances in a given context, nor have they attempted to uncover the toxic nature of arbitrary social statements in a contextualized manner, unlike us.

3 Pragmatics, Meaning, and Toxicity

To adapt toxicity detection (Founta et al., 2018) to dialogs, we formalise toxicity as something that can potentially affect the climate of discourse in a negative way. In technical terms, if we have a hypothetical value function $V(\mathcal{D}/C)$ that can estimate the state-value of discourse \mathcal{D} at a specific time given context C, an utterance u_t at time t is a potential contestant for hate speech if

$$V(\mathcal{D}_{t+i}/C_{< t+i}) \ll V(\mathcal{D}_{t-1}/C_{< t-1}) : i \ge 1$$

Empirically $V(\cdot)$ is impossible to estimate due to the subjective perception of language among humans and lack of consensus on what to include in $C_{<t}$. Defining hate speech in this way highlights the limitations of basing toxicity levels solely on snippets of utterances. In this paper, we consider free text situational descriptor as $C_{<t}$.

We hypothesise that toxicity is performative. In linguistics, performatives are speech acts that not only convey information but also perform an action and have a perlocutionary effect on the listener's mind (Austin, 1962). For instance, "I would like some Kimchi!" at a dinner table implies "pass me the Kimchi". Perlocutionary effects include persuading, convincing, enlightening, and commanding. We propose that conveying hate or offense is a valid perlocution, potentially affecting perceived toxicity scores when annotators have full context.². We aim to investigate how perceived toxicity changes across different contexts and nuanced situations.

4 Generation Pipeline

In this section we formalise a general overview of the pipeline, a straightforward summary of which is presented in Figure 2. The pipeline utilizes supervised finetuned language (SFT) models (Ouyang et al., 2022), contrastive search decoder (Li et al., 2022) and carefully curated prompts as the base elements. After conducting several preliminary experiments, we propose the prompt should have a template designed to achieve prespecified goal as

²Please note that hate or offense was never tied to perlocution by Austin (1962), this is one contribution of our dataset

done in prompt engineering (Sahoo et al., 2024) and it should also contain few in-context examples. Following our findings (appendix A), we propose using contrastive search (over top-p or temperature) becasue it along with in-context prompts reduces hallucination and improve the quality of generations while maintaining relevance to the instruction. These claims are supported by findings from other sources (O'Brien and Lewis, 2023). The pipeline supports three types of context augmentation, depending on the number of iterations and the dynamic addition of statements. This is controlled by the target polarity and other hyperparameters.

4.1 Direct Augment

Let $\mathcal{L}_{\theta}^{(\alpha,\kappa,H)}$ be a pretrained language model parameterised by θ coupled with contrastive search decoder parameterised by α and κ and set $H \in (h_1, h_2, ..)$ containing hyperparameters that modifies the output logits. *H* includes properties like *repition-pentalty, max-token, repeat-ngram*, etc. α and κ controls the trade-off between model confidence and degeneration penalty. Formally given the input prompt $x_{< t}$ the selection of output x_t will follow:

$$x_t = \underset{v \in \mathcal{V}^{(\kappa)}}{\arg \max} \{ (1 - \alpha) p_{\theta}(v | x_{< t}) - \alpha(\max\{s(v, x_j)\}) \}$$

Where $\mathcal{V}^{(\kappa)}$ is the *top-k* preditiction from the LMs probability distribution $p_{\theta}(./x_{< t})$. Model confidence, is the probability of candidate v predicted by the LMs $p_{\theta}(v|x_{< t})$. Degeneration penalty $max\{s(v, x_j) : 1 \leq j \leq t - 1\}$, measures the maximum cosine similarity between the candidate v and the tokens in the input prompt. In case of direct augment if u be any predefined utterance and t_p be the target polarity of the utterance then the context generated by direct augment is given by:

$$C = \mathcal{L}_{\theta}^{(\alpha,\kappa,H)}(u, P_{cont}(n, t_p))$$

Where $P_{cont}(n, t_p)$ is the taylored prompt having n in-context examples and instruction to generate context given utterance u.

4.2 Multistage Augment

This method generates completely new utterancecontext pair by passing the input through LMs at multiple steps with distinct polarity objectives. Using three chains of target polarity adds dynamic to the connotation of the utterance-context pair and its framing. For example, a seemingly neutral context u could first be made toxic along with a generated context C. Then a new utterance u_{new} could be constructed which along with the previous context sounds benign. Then again a new context C_{new} could be constructed which along with u_{new} sounds toxic. If $P_{utt}(n, t_p)$ is the taylored prompt having n in-context examples and instruction to generate utterance given context C then the process can be written as:

$$C = \mathcal{L}_{\theta_1}^{(\alpha_1, \kappa_1, H_1)}(u, P_{cont}(n, t_{p_1}))$$
$$u_{new} = \mathcal{L}_{\theta_2}^{(\alpha_2, \kappa_2, H_2)}(C, P_{utt}(n, t_{p_2}))$$
$$C_{new} = \mathcal{L}_{\theta_3}^{(\alpha_3, \kappa_3, H_3)}(u_{new}, P_{cont}(n, t_{p_3}))$$

Where t_{p_i} is the target polarity at *i*th step. The dynamic nature of this method improves the quality of counterfactual examples greatly.

4.3 N-iter Multistage Augment

This methods further extends multistage augment with new utterance at N intermediate steps (typically N = 2, 3, 4, ...). This further adds dynamic to the utterance and context quality and helps even improve the counterfactual examples. The steps could be written as follows

$$(u_1, C_1) = M_{\Theta}(u, P)$$

$$\forall i \in (2, 3, 4...N - 1)$$

$$u_i = \mathcal{L}_{\theta_2}^{(\alpha_2, \kappa_2, H_2)}(C_{i-1}, P_{utt}(n, t_{p_2}))$$

$$C_i = \mathcal{L}_{\theta_3}^{(\alpha_3, \kappa_3, H_3)}(u_i, P_{cont}(n, t_{p_3}))$$

Where M_{Θ} is the multistage augment step with Θ containing all the hyperparameters associated with that step. u_i and C_i being the generated utterance and context at *i*th step.

5 Dataset Generation

All augmentation methods were utilised in the pipeline for creation of Toximatics.

5.1 Models

We utilized the largest available open-source model 70 billion parameter LLama 2 chat model (Touvron et al., 2023) supervised finetuned with Orca dataset (Mitra et al., 2023).We conducted a side experiment to compare different SFT versions of the model for our task. We generated 5 generations using the direct augmentation method and crowd-validated

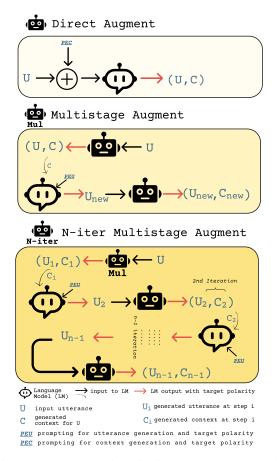


Figure 2: The generation pipeline supports 3 methods: Direct Augment adds context to the seed utterance, while Multistage and N-iter Multistage can generate novel utterance-context pairs given the seed.

the quality as the relevance of the generation to the prompt. Results reveal that the Oraca finetuned version outperformed the others with an agreement of 0.71. To streamline the process, we substituted the same model at all stages of the multistage augmentation.

quality	llama2 Orca	llama2 Oasst	llama2 chat	llama2 base
relevance	0.79	0.72	0.58	0.51

Table 1: Generation quality for various SFT versions of LLama-2, Oasst is OpenAssistance sft version while chat is meta's llama2 sft version

5.2 Seed Utterance

We generate Toximatics from both state of the art implicit hate datasets and socially grounded neutral statements which were also curated with generative models.

5.2.1 Implicit Hate Dataset

We primarily used ToxiGen (Hartvigsen et al., 2022) which is a large-scale machine-generated

dataset containing human annotated toxicity score, framing and perceived intent. We first divide the dataset (train set) into three parts by thresholding over the human toxicity score. We taxonomize them as *benign-batch* $(0.5 \le h_{tox} \le 1.5)$ with 3230 samples, *neutral-batch* $(1.5 \le h_{tox} \le 3.5)$ with 3230 samples and *toxic-batch* $(3.5 \le h_{tox} \le$ 4.5) with 1145 samples, where h_{tox} is the human toxicity score in a scale of (1, 5). Extreme toxic statements were left out as they contain lexical cues of overt negative words. This distinction based on a threshold was established to enable the creation of experiments with precise goals, such as modifying the toxicity of samples to benign, toxic to neutral, neutral to benign and so on (see section 5.4).

5.2.2 Socially Grounded Neutral Statement

The primary objective of these seed utterances is to challenge preconceived notions of toxicity linked to power dynamics, identity groups, race, politics, and gender. Instead, we aim to ground the analysis in more generic contexts, such as whether an utterance in a restaurant, a birthday celebration, or a friendly environment can be perceived as toxic. This approach allows us to analyze the polarity of utterances within valid social contexts, termed *"base-context"*, as opposed to online comments. We mined the base-context as detailed below (see Figure 3).

Conversational Topic Extraction: First we apply a topic model algorithm based on BertTopic (Grootendorst, 2022) on two conversation data sets Daily-Dialog (Li et al., 2017) and Blended-Skill-Talk (Smith et al., 2020). Firstly, the dialogues were converted to embeddings using Sentence Transformer (Reimers and Gurevych, 2020), and then reduced in dimensionality using UMAP (McInnes et al., 2018) with key-parameters like nearest neighbour size as 15 and min-dist as 0.25 (the minimum distance between points in low-dimensional space). Setting both parameters to low helps to emphasize the local structure of conversational data. HDBSCAN (McInnes and Healy, 2017) was employed as the clustering algorithm, with Euclidean as the distant metric and minimum cluster size of 200 so that we don't end up having too many clusters. The topic theme was generalized from the topic cluster keywords using chat-gpt-4, and it was then taken as the conversational topic. In this way, the two datasets yielded 413 conversation themes.

Social Location Extraction: We define a social location as any place that has a social environment

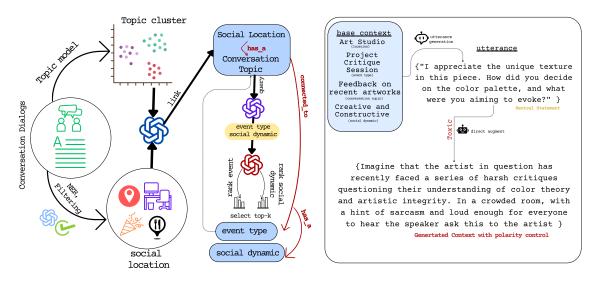


Figure 3: We first mine utterances atypical to a social topic and then augment it with polarity controlled context. We utilised LLMs to assists the mining process. Like chat-gpt-4 during S linking phase, S filter and augment social location, S to generate candidate event types and social dynamics S base gpt4 to rerank the generated list

and can stimulate civil conversation, such as restaurants, parks, bars. We employed Named Entity Recognition (NER) to the same dialog datasets with spacy (Honnibal and Montani, 2017) to extract *FAC* (facilities) location data; a total of 685 samples. Consequently, duplicates were removed and chat-gpt-4 was instructed to generalise several categories of social location from the remaining set e.g., *Entertainment and Recreation* \approx {*Disneyland*, *Saikei Ski Resort, Zoo, Kangaroo Club, etc*}. Taking these categories and it's associated location as example we instruct chat-gpt-4 to mine a set of 150 social locations (appendix B).

Linking: This stage links the conversational topic to the social location via chat-gpt-4. Set S containing social location and set C_t having conversational topics has many-to-many relationship with overlapping association (e.g, almost any conversation can happen in a caffe but only some in hospital). The set of links L is therefore the subset of Cartesian product of S and C_t .

$$L \subseteq S \times C_t$$

We present each element c_i in C_t and entire set S to chat-gpt-4 and instruct to link c_i to elements of S with one in-context example (appendix C).

Event and Social Dynamic Prediction: As it is non-trivial to mine events and social dynamics from conversation datasets, we use LLMs as the a retrieval system. For each linked location and conversation $l_i \in L$, chat-gpt-4 was employed to generate a preliminary list of potential event types. This list was re-ranked with base GPT4 (appendix D & E). The decision was influenced by (Sun et al., 2023), who demonstrated the efficacy of LLMs in retrieval tasks and identified that base GPT4 outperforms all other models in ranking tasks. We then applied a top-k threshold to select k entities from the list as a measure of most relevance. In this paper we used k = 3 to account for high relevance and brevity of our dataset. After appending the event type to the base context, we repeat the same step for social dynamic.

The aforementioned procedure yielded 1554 base-context units, with approximately 2% of these removed by three crowd validators (appendix G) with an agreement of 0.88, resulting in 1523 units (examples in F). We generated 1523 seed utterance associated with the units with our generation pipeline and name this batch as *social-neutral-batch*.

5.3 Prompt Engineering

For the context generation task at hand, the unpredictable nature and absence of validation data made it challenging to create a prompt using a Chain of Thought (CoT) (Wei et al., 2022) or other CoT-based approach. Furthermore, the utilisation of recursive prompting techniques similar to Self-Refine (Madaan et al., 2023; Saunders et al., 2022; Yang et al., 2022), represents a potential bottleneck within our pipeline, particularly when utilising multistage augmentation techniques. This is due to the fact that these methods already have iterations, which could even worsen the time complexity. We structure our prompt inspired by (Rajagopal et al., 2021) which curate prompt as $concept \xrightarrow{\text{qualifier}} concept$ where concept slot contains abstract category of concepts. For our task, the concepts become the *context* and *utterance* while the qualifier becomes target polarity like benign, mildly – toxic, toxic. As a consequence, it reduces to $context \xrightarrow{qualifier} utterance$. For each objective in section 5.4, we first generate a few examples of (utterance, context) pair with the instruction prompt "Add <context> to the <utterance> such that the statement becomes <qualifier>". Then we manually correct and refine the generated context to construct our in-context examples. Then we used the same prompts and in-context examples to create context for the rest of the batches in few-shot mode. In preliminary experiments, increasing the number of examples beyond six did not improve generation quality but impacted generation time. Therefore, we used six example in the few-shot setting for the rest of the generations (appendix H).

5.4 Batches and Polarity Control

We sample 2000 utterances from *benign-batch* and generated 8000 counterfactual-toxic samples by augmenting using final polarity toxic with direct-augment, multistage augment, 2-iter and 4-iter multistage augment. Subsequently, we sample 1000 utterances from *toxic-batch* and generated 4000 counterfactual-benign samples with final polarity benign and using the same methods. 1500 samples from *social-neutral-batch* was used with direct augment to generate 1500 toxic and benign samples each. 2000 sampled units from *neutral-batch* was used with direct augment to generate 1500 toxic and benign samples and 3000 benign samples (to balance the dataset). The dataset finally contains approx. 56% toxic samples and 44% benign samples.

6 Human Toxicity Annotation

The samples emanating from section 5 were passed to crowd workers. The workers were provided both the utterance and context. They were tasked to respond in 5 point Likert scale if they agree that the *utterance* sounded toxic if it was actually uttered in real life contextual scenario provided in the *context*. We interpret the 5-point Likert scale in the range (1, 5) with 1 being completely benign and 5 very toxic. 10 responses per example were considered and the mean score was accepted as the final toxicity score. As Mturk workers often cheats (Marshall et al., 2023), the work was divided into batches of 30 examples with 3 attention check questions appearing quarterly like age, date of birth and age group. We rejected workers who failed the attention checks. Also, we restricted the participation from only people residing in the USA and have a previous HIT approval rate greater than 95% and had at least 50 HIT approved. The application of filters to the annotations allows for the improvement of the quality of the annotations themselves. The kappa agreement score was 0.57. We hypothesised that the level of agreement will be low due to the subjective nature of the task. As the process of labelling toxicity is prone to individual bias, such as that derived from a person's social background, culture, age, and so forth, it is likely that there will be a lack of consensus. However, the agreement score inspite of being low is empirically consistent with kappa scores recorded by similar generation task (Amidei et al., 2018, 2019; Celikyilmaz et al., 2020).

7 Evaluation of Model Performance

The performance of our dataset is evaluated in comparison to state-of-the-art toxicity classifiers and text generation models, including both foundational and chat models. For the classification task, the problem is formulated as a binary toxicity classification. This is achieved by concatenating the context and the utterance. With regard to the text generation model, the problem is framed as a zero-shot classification task. For the purposes of evaluation, 1,100 examples of toxic content and 900 examples of benign content were randomly sampled from the dataset. For the classifier, we considered base transformer models like Bert (Devlin et al., 2018), HateBert (Caselli et al., 2020), Roberta (Liu et al., 2019), DistilRoberta, finetuned with explicit or implicit hate datasets like Toxigen, Jigsaw³, None⁴, RAL-E, social-bias-dataset⁵. For text generation models, we evaluated T5 (Raffel et al., 2019), Flang-T5 (Chung et al., 2022), OPT (Zhang et al., 2022), OPT-iml (Iver et al., 2022), Llama-2 (Touvron et al., 2023), Llama-2chat, Chat-Gpt. Where Flang-T5, OPT-iml and Llama-2-chat are the supervised finetuned versions of the base model.

³https://www.kaggle.com/c/

jigsaw-toxic-comment-classification-challenge
 ⁴https://www.kaggle.com/datasets/subhajournal/

normal-hate-and-offensive-speeches
⁵https://github.com/rpryzant/

neutralizing-bias

Model	Fintune-Data	Accuracy (%)	Recall (%)	Precision (%)	F1
Bert-base	Jigsaw 2020	43.3	3.90	50.0	0.07
HateBert	RAL-E, None	47.7	21.0	61.1	0.31
HateBert	ToxiGen	44.0	7.80	57.0	0.13
RoBERTa	Jigsaw 2018, 19, 20	43.3	3.92	50.0	0.07
DistilRoBERTa base	wikirev-bias	52.2	27.2	70.0	0.39
RoBERTa	ToxiGen	46.6	5.80	98.0	0.11
T5-xl	-	54.8	72.2	54.4	0.62
T5-xxl	-	50.4	75.8	53.6	0.62
Flang-T5-xl	-	58.4	51.5	81.8	0.63
Flang-T5-xxl	-	69.9	54.0	84.0	0.66
OPT-13b	-	62.8	83.3	53.5	0.65
OPT-30b	-	48.7	97.0	55.0	0.70
OPT-13b-iml	-	61.9	72.1	63.7	0.67
OPT-30b-iml	-	51.3	70.6	58.5	0.64
Llama-2-7b	-	43.3	3.30	28.5	0.06
Llama-2-13b	-	55.7	31.1	70.3	0.43
Llama-2-7b-chat	-	70.8	47.8	68.0	0.56
Llama-2-13b-chat	-	71.7	70.5	75.4	0.73
Chat-Gpt-3.5-turbo	-	68.1	61.8	70.8	0.66
Chat-Gpt-4	-	72.0	54.3	86.4	0.67

Table 2: State-of-the-art fine-tuned toxicity classifiers, foundation and chat model's performance on Toximatic samples, here accuracy, recall, and precision is in percentage

7.1 Findings

The finding for this experiment is depicted in Table 2. From the table, we can see our dataset successfully fools the state-of-the-art classifier model. The models failed to detect many valid samples; hence we see an extremely low recall. This is because, in Toximatics, we intended to alter the toxic polarity with context. The classifier was not trained on such an objective. Moreover, we see that finetuning with implicit datasets will rarely improve performance on pragmatic understanding. Sometimes high precision was achieved as the models flagged an extremely small number of actual positive examples, as positive. For example, ToxiGen roberta scored a high precision with only guessing 35 TP (true positives). For the zero-shot classification problem, taking the F1-score as the main measure of performance, we see the instruction finetuned / chat models perform better than their base counterparts (with the exception of OPT-30b-iml). We also observed within all the chat models (instruction finetuned), the newer chat models like Chat-Gpt, Llama-2 are more accurate than older ones. Even for the same model type, scaling up improves both accuracy and F1 score (excluding OPT models). The best-performing model was Llama-2-13b-chat

with an F1-score 0.73 and balanced recall and precision. Chat-Gpt-4 had the highest accuracy but with less recall indicating a higher number of false negatives. We also observe that Chat-Gpt-4 does not significantly outperform Chat-Gpt-3.5-turbo with our dataset. This experiment illustrates the power of such a dataset and why it will raise the bar in natural language understanding.

8 Conclusion

In this paper, we introduce Toximatics, a dataset of toxic and benign statements (19.8k) where toxicity is context-dependent. This dataset offers a novel approach to hate speech detection, examining how contextual scenarios can shift the polarity of an utterance. Toximatics addresses the topical biases of previous datasets, such as those focused on race, identity, gender, and power, by presenting neutral social statements and contextualizing them to render them toxic. Our findings show that generative models and state-of-the-art toxicity classifiers are often misled by this dataset, demonstrating the increased difficulty of this task compared to sentence-level toxicity detection. We also present a mined base-context for grounding social utterances, providing a foundation for further research. Additionally, we curate a novel scalable data generation pipeline. We propose that a research direction focusing on pragmatic hate speech understanding, which considers holistic contextual information, should be pursued. This would facilitate the development of more suitable toxicity detection techniques for long dialogues and discourse.

Ethical Considerations

In this section, we will briefly highlight some of the ethical concerns and limitations of this work. We would like to bring to your attention that the dataset contains political references and opinions that may be subjectively provocative. For simplicity, we are only checking raw toxicity scores but not fine-grain categories like framing, abuse, vulgar, obscene, etc. Context can go far beyond situational descriptor and base-contexts mentioned in this paper. But we leave it open for future works. The subjective nature of interpreting toxicity still reamins a challenging task. To mitigate this, future studies could develop more robust automated techniques to improve reliability.

References

- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2018. Rethinking the agreement in human evaluation tasks. In Proceedings of the 27th International Conference on Computational Linguistics, pages 3318–3329, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jacopo Amidei, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings* of the 12th International Conference on Natural Language Generation, pages 344–354, Tokyo, Japan. Association for Computational Linguistics.
- John Langshaw Austin. 1962. *How to Do Things with Words*. Clarendon Press, Oxford.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *International Workshop on Semantic Evaluation*.
- Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: interpretation and communication for policy decision making.
- Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. 2020. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472.

- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *CoRR*, abs/2006.14799.
- Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.
- Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.
- Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *ArXiv*, abs/1809.04444.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Conference on Empirical Methods in Natural Language Processing*.
- Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *ArXiv*, abs/1802.00393.
- Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Recent Advances in Natural Language Processing*.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv* preprint arXiv:2203.05794.
- Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Srinivas Iyer, Xiaojuan Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. 2022. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *ArXiv*, abs/2212.12017.
- Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Y. Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Aadila Hussain, A. Lara, olmos g, Asmaa Al Omary, C. G. Park, C. C. Wang, X Wang, Y. Zhang, and Morteza Dehghani. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech.
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.
- Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Workshop on Abusive Language Online*.
- Catherine C. Marshall, Partha S.R. Goguladinne, Mudit Maheshwari, Apoorva Sathe, and Frank M. Shipman. 2023. Who broke amazon mechanical turk? an analysis of crowdsourcing data quality over time. In Proceedings of the 15th ACM Web Science Conference 2023, WebSci '23, page 335–345, New York, NY, USA. Association for Computing Machinery.

- L. McInnes, J. Healy, and J. Melville. 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv e-prints*.
- Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *Data Mining Workshops (ICDMW), 2017 IEEE International Conference on*, pages 33–42. IEEE.
- Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.
- Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *arXiv preprint arXiv:2309.09117*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155.
- 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.
- Jing Qian, Mai Elsherief, Elizabeth M. Belding-Royer, and William Yang Wang. 2019. Learning to decipher hate symbols. In *North American Chapter of the Association for Computational Linguistics*.
- Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21:140:1–140:67.
- Dheeraj Rajagopal, Vivek Khetan, Bogdan Sacaleanu, Anatole Gershman, Andy E. Fano, and Eduard H. Hovy. 2021. Template filling for controllable commonsense reasoning.
- Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the* 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.
- Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

- Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Annual Meeting of the Association for Computational Linguistics*.
- Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.
- William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. arXiv preprint arXiv:2206.05802.
- Leandro Araújo Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the targets of hate in online social media. In *International Conference on Web and Social Media*.
- Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.
- Weiwei Sun, Lingyong Yan, Xinyu Ma, Pengjie Ren, Dawei Yin, and Zhaochun Ren. 2023. Is chatgpt good at search? investigating large language models as re-ranking agent. *arXiv preprint arXiv:2304.09542*.
- Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *North American Chapter of the Association for Computational Linguistics*.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. ArXiv, abs/2307.09288.

- Bertie Vidgen, Tristan Thrush, Zeerak Talat, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. *ArXiv*, abs/2012.15761.
- William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web.
- Zeerak Waseem. 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.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.
- Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. Re3: Generating longer stories with recursive reprompting and revision. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *North American Chapter of the Association for Computational Linguistics*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. ArXiv, abs/2205.01068.
- Xuhui Zhou, Haojie Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. *ArXiv*, abs/2306.01985.

A Generation quality vs decoding technique

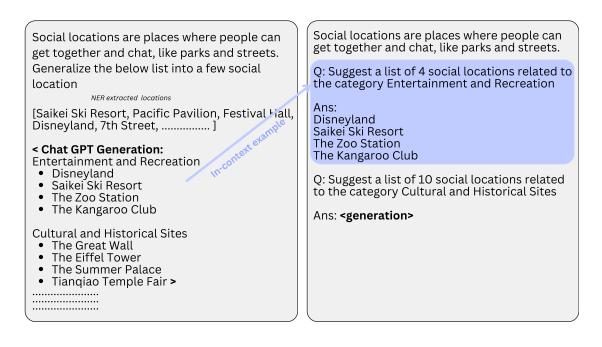
In this sub-experiment, we used our generation pipeline with a range of decoding techniques, as detailed in Table 3. We kept the task description and all other parameters like in-context prompts exactly similar and used 5 generations each with direct augment method. We used top-p sampling with temperature

top-p	quality	top-k	quality	penalty-alpha	quality
0.9	0.43	60	0.31	0.9	0.52
0.7	0.51	40	0.55	0.7	0.63
0.6	0.48	30	0.42	0.6	0.71
0.5	0.49	20	0.45	0.5	0.57
0.4	0.45	10	0.59	0.4	0.58
0.3	0.36	5	0.51	0.3	0.33

Table 3: Generation quality for various decoding paradigm with our incontext prompts

0.7, top-k sampling with temperature 0.7 and contrastive decoding with top- κ 8. We provided all the generations to 3 crowd workers to label the quality within a scale of 0 to 1 and then mean the score for 5 samples. The kappa aggrement score was 0.683. The results clearly shows contrasting decoding performs better for majority of the case in our task of context generation.

B Prompts to generalise and augment social locations in Section 5.2.2



C Prompts to link conversation topic and social location in Section 5.2.2

Q: The conversation topic "*Networking opportunities and meeting new people*" can occur in which of the given locations:

Ans:

Workplace Parties University Onlines Art Venue Community Center

Q: The conversation topic "*<conversational topic>*" can occur in which of the given locations *<set containing all social location>*

Ans: **<generation>**

D Prompts to generate event types from chat-gpt in Section 5.2.2

Q: Suggest 10 list of candidate event type for the following setting Location: Botanical garden, Conversation: Sharing tips on capturing natural beauty

Ans:

- 1. Photography Workshop
- 2. Plein Air Painting Session
- 3. Nature Sketching Class
- 4. Botanical Illustration Workshop
- 5. Floral Arrangement Demonstrations
- 6. Eco-Friendly Gardening Seminar
- 7. Nature Writing Retreat
- 8. Landscape Design Workshop
- 9. Wildlife Photography Talks

10. Guided Nature Walks with Expert Commentary

Q: Suggest 10 list of candidate event type for the following setting Location: *<social location>*, Conversation: *<conversation topic>*

Ans: <generation>

E Prompts to rank event types from chat-gpt in Section 5.2.2

Q:

- 1. Workshop on Presentation Skills
- 2. Research Methodology Seminar
- 3. Poster Design Session
- 4. Mock Conference Presentations
- 5. Literature Review Techniques Class
- 6.Q&A Preparation Session
- 7. Technology in Research Showcase
- 8. Ethics in Research Discussion panel
- 9. Grant writting Bootcamp
- 10. Data Analysis Workshop

Rerank the above list of candidate event type that is highly relevant to the following catergories Location: Library, Conversation: Preparing for an upcoming academic conference

Ans:

- 1. Mock Conference Presentations
- 2.Q&A Preparation Session
- 3. Poster Design Session
- 4. Ethics in Research Discussion panel
- 5. Workshop on Presentation Skills
- 6. Data Analysis Workshop
- 7. Literature Review Techniques Class
- 8. Research Methodology Seminar
- 9. Technology in Research Showcase
- 10. Grant writting Bootcamp

Q:

<EVENT TYPES>

Rerank the above list of candidate event type that is highly relevant to the following catergories Location: *<social location>*, Conversation: *<conversation topic>*

Ans: <generation>

F Examples of Generated Base Context in Section 5.2.2

Location	Event-Type	Conversation-Topic	Social-Dynamic
Workplace	Team lunch,	Discussing team roles and responsibilities	Supportive and collaborative.
University	Department seminar	Planning a new collaborative research project	Eager and intellectually stimulating.
Conference room	Strategy meeting	Setting long-term goals for the department	Visionary and goal-oriented
Local library	Book club meeting	Discussing the next book selection	Inquisitive and engaging
Public park	Volunteer clean-up day	Organizing teams and areas for cleanup	Community-minded and cooperative
cafe	Poetry reading	Sharing and discussing personal works	Expressive and encouraging
Conference	Panel discussion	Debating ethical implications of research methods	Engaged and respectful
Workplace	Safety training	Learning about new safety protocols in the lab	Safety-conscious and attentive
University	Guest lecture	Engaging with an expert visiting from another institution	Enthusiastic and receptive
Workshop	Professional development workshop	Learning new skills for professional growth	Eager to learn and apply new knowledge
Office	Planning meeting	Coordinating logistics for an upcoming international conference	Organized and detail-oriented
Art museum	Guided tour	Learning about different art periods	Curious and appreciative
Health clinic	Health workshop	Discussing nutrition and wellness strategies	Proactive and health-conscious
Gym	Fitness class	Setting personal fitness goals	Motivational and supportive
Cooking school	Cooking class	Deciding on recipes for the session	Collaborative and fun
Sports club	Team practice	Strategies for the next game	Competitive and team-spirited
Music studio	Band rehearsal	Arranging a new song	Creative and harmonious
Film studio	Film screening	Discussing the thematic elements of the film	Analytical and insightful
Theater	Rehearsal	Perfecting scenes and lines	Artistic and detailed
Botanical garden	Photography walk	Sharing tips on capturing natural beauty	Artistic and sharing
Planetarium	Astronomy night	Discussing constellations and celestial events	Enthusiastic and awe-inspired
Historical society	Lecture series	Discussing local history and significant events	Engaged and respectful of heritage
Dance studio	Dance workshop	Learning new dance moves and routines	Energetic and rhythmic
Local brewery	Craft beer tasting	Learning about brewing processes and flavors	Sociable and relaxed
Local cafe	Business brunch	Discussing a new marketing strategy	Collaborative and innovative
High school classroom	Teacher meeting	Planning semester curriculum adjustments	Supportive and consensus-seeking
startup office	Weekly tech sync	Reviewing product development timelines	Energetic and fast-paced
Art studio	Project critique session	Feedback on recent artworks	Creative and constructive
Corporate boardroom	Quarterly financial review	Analyzing profit and loss statements	Analytical and strategic
Nonprofit organization office	Grant writing workshop	Brainstorming for funding opportunities	Motivated
Public library	Book club meeting	Discussing this month's book selection	Informal and friendly
City hall	Urban planning session	Outlining new public transportation options	Formal and regulatory
Sports room	Pre-game strategy talk	Assigning player roles and strategies	Energetic and motivational
University laboratory	Scientific research collaboration	Sharing experimental results	Curious and data-driven
Architecture firm	Design review	Discussing a new building project	Visionary and detail-conscious
International conference center	Diplomatic summit	Negotiating international treaties	Formal and cautious
Community center	Volunteer coordination meeting	Organizing a local food drive	Enthusiastic and community-oriented
Airline operations center	Flight scheduling session	Managing flight crew assignments	Precision-oriented and time-sensitive
Fashion design studio	Trend forecasting meeting	Selecting fabrics for next season	Trend-sensitive and collaborative
Music production studio	Album planning meeting	Deciding on track listings and collaborations	Creative and expressive
Government agency office	Policy development workshop	Crafting new regulations for public safety	Analytical and policy-oriented
Luxury cruise ship	Staff training session	Reviewing safety protocols and guest relations	Professional and guest-centric
Local government office	Community feedback forum	Addressing citizen concerns about local policies	Responsive and community-focused

Table 4: Few examples of base context generated with methods discussed in section 5.2.2

G Crowd Task for validating base-context

A social-context is a combination of location, event-type, conversation theme and social dynamic that bears resemblance to real life social scenario

An example of a valid social context in given below

Location: City parkEvent Type: Community clean-up coordinationConversation: Assigning areas and tasks for a neighborhood clean-upTeam Dynamics: Community-minded and cooperative

Do you think the given combination of social-context is also valid?

Location: Auto repair shop Event Type: Technician training session Conversation: Introducing new diagnostic tools and techniques Team Dynamics: Hands-on and knowledge-sharing

yes □ no □

H Example of prompts used in Generation Pipeline

System: You are an assistance which adds contextual scenario to any given utterance as per the given instrution.

User: Add contextual scenario to the utterance \'there is no such thing as the "______ nation, there is no such thing as the ______ nation, there is no such thing as\', such that the statement becomes toxic **Assistant:** In a heated online debate about national identity and politics between two users named John and Sarah, they argue over whether certain nations exist or not. John says, He believes he does not recognize XYZ as legitimate countries.<eos>

User: Add contextual scenario to the utterance \'this is not to suggest that there is no such thing as bisexuality or that bisexuality does not exist, but rather that the concept of bisexuality\', such that the statement becomes toxic

Assistant: A popular social media influencer with millions of followers making this claim during Pride Month, dismissing the experiences and struggles faced by members of the LGBTQ+ community who identify.<eos>

User: Add contextual scenario to the utterance \'you have to be careful when you go to a restaurant because you might get food poisoning \', such that the statement becomes toxic

Assistant: Someone is trying to discourage people from going out and supporting local businesses during these difficult times due to the pandemic. They want to create fear among potential customers by exaggerating the risks of getting sick.<eos>

User: Add contextual scenario to the utterance \'they are just like any other race \', such that the statement becomes toxic

Assistant: In a discussion about racial equality, one person tries to dismiss concerns by saying the above comment and implying that there is no need for special attention or consideration towards marginalized groups because they should be treated exactly the same as everyone else without acknowledging their unique experiences and challenges.<eos>

User: Add contextual scenario to the utterance \'if youre so smart, how come youre not rich like me \', such that the statement becomes toxic

Assistant: In a competitive business environment where success is measured by wealth and status, two colleagues engage in a heated argument over their accomplishments. One of them, who has recently become quite successful financially, taunts his less fortunate coworker with this phrase as they argue about intelligence versus financial achievements.<eos>

User: Add contextual scenario to the utterance \'disability is not the state of being incomplete\', such that the statement becomes toxic

Assistant: In a society where people with disabilities face constant discrimination and marginalization, someone uses this phrase as part of their argument against providing equal opportunities or resources for disabled individuals.<eos>

User: Add contextual scenario to the utterance <seed utterance>, such that the statement becomes toxic Assistant: <generation>