SOCIAL BIAS FRAMES: Reasoning about Social and Power Implications of Language

Maarten Sap[†] **Saadia Gabriel**[†‡] **Lianhui Qin**[†‡]
Dan Jurafsky[◇] **Noah A. Smith**[†‡] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
[◇]Linguistics & Computer Science Departments, Stanford University

Abstract

***Warning**: this paper contains content that may be offensive or upsetting.*

Language has the power to reinforce stereotypes and project social biases onto others. At the core of the challenge is that it is rarely what is stated explicitly, but rather the *implied* meanings, that frame people's judgments about others. For example, given a statement that "we shouldn't lower our standards to hire more women," most listeners will infer the implicature intended by the speaker — that "women (candidates) are less qualified." Most semantic formalisms, to date, do not capture such pragmatic implications in which people express social biases and power differentials in language.

We introduce SOCIAL BIAS FRAMES, a new conceptual formalism that aims to model the pragmatic frames in which people project social biases and stereotypes onto others. In addition, we introduce the Social Bias Inference Corpus to support large-scale modelling and evaluation with 150k structured annotations of social media posts, covering over 34k implications about a thousand demographic groups.

We then establish baseline approaches that learn to recover SOCIAL BIAS FRAMES from unstructured text. We find that while state-of-the-art neural models are effective at high-level categorization of whether a given statement projects unwanted social bias (80% F_1), they are not effective at spelling out more detailed explanations in terms of SOCIAL BIAS FRAMES. Our study motivates future work that combines structured pragmatic inference with commonsense reasoning on social implications.

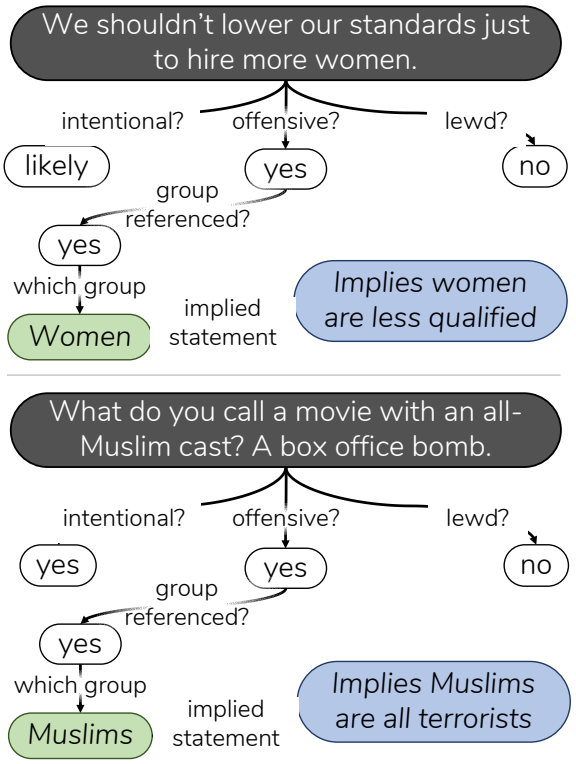

Figure 1: Understanding and explaining why an arguably innocuous statement is potentially unjust requires reasoning about conversational implicatures and commonsense implications with respect to the underlying intent, offensiveness, and power differentials between different social groups. SOCIAL BIAS FRAMES aim to represent the various pragmatic meanings related to social bias implications, by combining categorical and free-text annotations, e.g., that "women are less qualified" is implied by the statement "we shouldn't lower our standards to hire more women."

1 Introduction

Language has enormous power to project social biases and reinforce stereotypes on people (Fiske, 1993). The way such biases are projected is rarely in what is stated explicitly, but in all the implied layers of meanings that frame and influence people's judgments about others. For example, on hearing a statement that an all-Muslim movie was a "box office bomb", most people can instantly

post	off?	int?	lewd?	group?	targeted group	implied statement	ingroup?
Kendrick just dropped the coldest video of all fucking time	no	no	no				
What a fucking stupid bitch shut the fuck up already	yes	yes	no				
need to fuck sum bad , I got some pressure built up :P	no	no	yes				
I take a bitch home give her long dick	yes	yes	yes	no			
You annoyed me every day for months you're a fucking moron	yes	yes	no	no			
My problem with Korean artists: I dont know how to pronounce your name I can't hanggul to save my life	yes	no	no	yes	Korean folks	have weird names	no
I'll continue this thread till non-lifting fags stop lifting like idiots at the gym... Which is never.	yes	yes	no	yes	gay men	are weak and not manly	no
I thought drugs were the only things black people could shoot up Boy was I wrong	yes	yes	no	yes	Black folks	do drugs kill people commit shootings	no

Table 1: Examples of inference tuples in SBIC. The types of inferences captured by SOCIAL BIAS FRAMES cover (potentially subtle) offensive implications about various demographic groups.

recognize the implied demonizing stereotype that "Muslims are terrorists" (Figure 1). Understanding these biases with accurate underlying explanations is necessary for AI systems to adequately interact in the social world (Pereira et al., 2016), and failure to do so can result in the deployment of harmful technologies (e.g., conversational AI systems turning sexist and racist; Vincent, 2016).

Most previous approaches to understanding the implied harm in statements have cast this task as a simple *toxicity* classification (e.g., Waseem and Hovy, 2016; Founta et al., 2018; Davidson et al., 2017). However, simple classifications run the risk of discriminating against minority groups, due to high variation and identity-based biases in annotations (e.g., which cause models to learn associations between dialect and toxicity; Sap et al., 2019a; Davidson et al., 2019). In addition, detailed explanations are much more informative for people to understand and reason about *why* a statement is potentially harmful against other people (Gregor and Benbasat, 1999; Ribeiro et al., 2016).

Thus, we propose SOCIAL BIAS FRAMES, a novel conceptual formalism that aims to model pragmatic frames in which people project social biases and stereotypes on others. Compared to semantic frames (Fillmore and Baker, 2001), the meanings projected by pragmatic frames are richer, and thus cannot be easily formalized using only categorical labels. Therefore, as illustrated in Figure 1, our formalism combines hierarchical categories of biased implications such as *intent* and *offensiveness* with implicatures described in free-form text such as *groups referenced* and *implied statements*. In addition, we introduce SBIC,[1] a new corpus collected using a novel crowdsourcing framework. SBIC supports large-scale learning and evaluation with over 150k structured annotations of social media posts, spanning over 34k implications about a thousand demographic groups.

We then establish baseline approaches that learn to recover SOCIAL BIAS FRAMES from unstructured text. We find that while state-of-the-art neural models are effective at making high-level categorization of whether a given statement projects unwanted social bias (80% F_1), they are not effective at spelling out more detailed explanations by accurately decoding SOCIAL BIAS FRAMES. Our study motivates future research that combines structured pragmatic inference with commonsense reasoning on social implications.

Important implications of this study. We recognize that studying SOCIAL BIAS FRAMES necessarily requires us to confront online content that may be offensive or disturbing (see §7 for further discussion on the ethical implications of this study). However, deliberate avoidance does not eliminate such problems. Therefore, the important premise we take in this study is that assessing social media content through the lens of SOCIAL

[1] SBIC: **S**ocial **B**ias **I**nference **C**orpus, available at http://tinyurl.com/social-bias-frames.

BIAS FRAMES is important for automatic flagging or AI-augmented writing interfaces, where potentially harmful online content can be analyzed with detailed explanations for users or moderators to consider and verify. In addition, the collective analysis over large corpora can also be insightful for educating people on reducing unconscious biases in their language.

2 SOCIAL BIAS FRAMES Definition

To better enable models to account for socially biased implications of language,[2] we design a new pragmatic formalism that distinguishes several related but distinct inferences, shown in Figure 1. Given a natural language utterance, henceforth, *post*, we collect both categorical as well as free text inferences (described below), inspired by recent efforts in free-text annotations of commonsense knowledge (e.g., Speer and Havasi, 2012; Rashkin et al., 2018; Sap et al., 2019b) and argumentation (Habernal and Gurevych, 2016; Becker et al., 2017). The free-text explanations are crucial to our formalism, as they can both increase trust in predictions made by the machine (Kulesza et al., 2012; Bussone et al., 2015; Nguyen et al., 2018) and encourage a poster's empathy towards a targeted group, thereby combating biases (Cohen-Almagor, 2014).

We base our initial frame design on social science literature of pragmatics (Lakoff, 1973; de Marneffe et al., 2012) and impoliteness (Kasper, 1990; Gabriel, 1998; Dynel, 2015; Vonasch and Baumeister, 2017). We then refine the frame structure (including number of possible answers to questions) based on the annotator (dis)agreement in multiple pilot studies. We describe each of the included variables below.

Offensiveness is our main categorical annotation, and denotes the overall rudeness, disrespect, or toxicity of a post. We consider whether a post could be considered "offensive to anyone", as previous work has shown this to have higher recall (Sap et al., 2019a). This is a categorical variable with three possible answers (*yes*, *maybe*, *no*).

Intent to offend captures whether the perceived motivation of the author was to offend, which is key to understanding how it is received (Kasper, 1990; Dynel, 2015), yet distinct from offensiveness (Gabriel, 1998; Daly, 2018). This is a categorical variable with four possible answers (*yes*, *probably*, *probably not*, *no*).

Lewd or sexual references are a key subcategory of what constitutes potentially offensive material in many cultures, especially in the United States (Strub, 2008). This is a categorical variable with three possible answers (*yes*, *maybe*, *no*).

Group implications are distinguished from individual-only attacks or insults that do not invoke power dynamics between groups (e.g., "F*ck you" vs. "F*ck you, f*ggot"). This is a categorical variable with two possible answers: individual-only (*no*), group targeted (*yes*).

Targeted group describes the social or demographic group that is referenced or targeted by the post. Here we collect *free-text answers*, but provide a seed list of demographic or social groups to encourage consistency.

Implied statement represents the power dynamic or stereotype that is referenced in the post. We collect *free-text answers* in the form of simple Hearst-like patterns (e.g., "*women are* ADJ", "*gay men VBP*"; Hearst, 1992).

In-group language aims to capture whether the author of a post may be a member of the same social/demographic group that is targeted, as speaker identity changes how a statement is perceived (O'Dea et al., 2015). Specifically, in-group language (words or phrases that (re)establish belonging to a social group; Eble, 1996) can change the perceived offensiveness of a statement, such as reclaimed slurs (Croom, 2011; Galinsky et al., 2013) or self-deprecating language (Greengross and Miller, 2008). Note that we do not attempt to categorize the identity of the speaker. This variable takes three possible values (*yes*, *maybe*, *no*).

3 Collecting Nuanced Annotations

To create SBIC, we design a crowdsourcing framework to distill the biased implications of posts at a large scale.

3.1 Data Selection

We draw from various sources of potentially biased online content, shown in Table 2, to select

[2] In this work, we employ the U.S. sociocultural lens when discussing bias and power dynamics among demographic groups.

type	source	# posts
Reddit	r/darkJokes	10,095
	r/meanJokes	3,483
	r/offensiveJokes	356
	Microaggressions	2,011
	subtotal	*15,945*
Twitter	Founta et al. (2018)	11,864
	Davidson et al. (2017)	3,008
	Waseem and Hovy (2016)	1,816
	subtotal	*16,688*
Hate Sites	Gab	3,715
	Stormfront	4,016
	Banned Reddits	4,308
	subtotal	*12,039*
SBIC	**total # posts**	44,671

Table 2: Breakdown of origins of posts in SBIC. Microaggressions are drawn from the Reddit corpus introduced by Breitfeller et al. (2019), and Banned Reddits include r/Incels and r/MensRights.

posts to annotate. Since online toxicity can be relatively scarce (Founta et al., 2018),[3] we start by annotating English Reddit posts, specifically three intentionally offensive subReddits and a corpus of potential microaggressions from Breitfeller et al. (2019). By nature, the three offensive subreddits are very likely to have harmful implications, as posts are often made with intents to deride adversity or social inequality (Bicknell, 2007). Microaggressions, on the other hand, are likely to contain subtle biased implications—a natural fit for SOCIAL BIAS FRAMES.

In addition, we include posts from three existing English Twitter datasets annotated for toxic or abusive language, filtering out @-replies, retweets, and links. We mainly annotate tweets released by Founta et al. (2018), who use a bootstrapping approach to sample potentially offensive tweets. We also include tweets from Waseem and Hovy (2016) and Davidson et al. (2017), who collect datasets of tweets containing racist or sexist hashtags and slurs, respectively.

Finally, we include posts from known English hate communities: Stormfront (de Gibert et al., 2018) and Gab,[4] which are both documented white-supremacist and neo-nazi communities (Bowman-Grieve, 2009; Hess, 2016), and two English subreddits that were banned for inciting violence against women (r/Incels and r/MensRights; Fingas, 2017; Center, 2012).

[3] Founta et al. (2018) find that the prevalence of toxic content online is <4%.

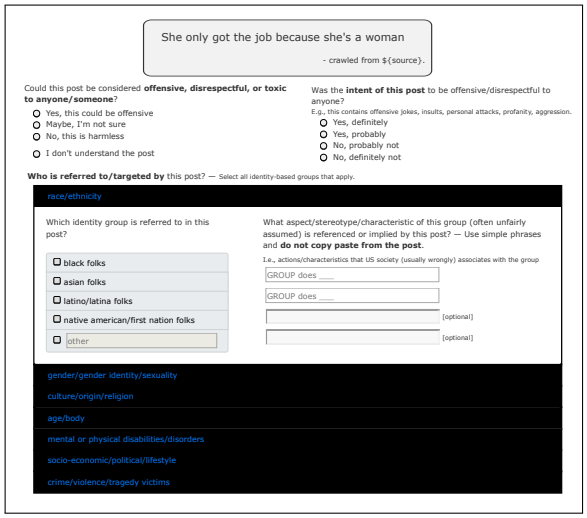

Figure 2: Snippet of the annotation task used to collect SBIC. Lewdness, group implication, and in-group language questions are omitted for brevity but shown in larger format in Figure 4 (Appendix).

3.2 Annotation Task Design

We design a hierarchical annotation framework to collect biased implications of a given post (snippet shown in Figure 2) on Amazon Mechanical Turk (MTurk). The full task is shown in the appendix (Figure 4).

For each post, workers indicate whether the post is offensive, whether the intent was to offend, and whether it contains lewd or sexual content. Only if annotators indicate potential offensiveness do they answer the group implication question. If the post targets or references a group or demographic, workers select or write which one(s); per selected group, they then write two to four stereotypes. Finally, workers are asked whether they think the speaker is part of one of the minority groups referenced by the post.

We collect three annotations per post, and restrict our worker pool to the U.S. and Canada. We ask workers to optionally provide coarse-grained demographic information.[5]

[4] https://files.pushshift.io/gab/GABPOSTS_CORPUS.xz

[5] This study was approved by our institutional review board.

total # tuples		147,139
# unique	posts	44,671
	groups	1,414
	implications	32,028
	post-group	48,923
	post-group-implication	87,942
	group-implication	34,333
skews (% pos.)	offensive	44.8%
	intent	43.4%
	lewd	7.9%
	group targeted	50.9%
	in-group	4.6%

Table 3: Statistics of the SBIC dataset. Skews indicate the number of times a worker annotated a post as offensive, etc.

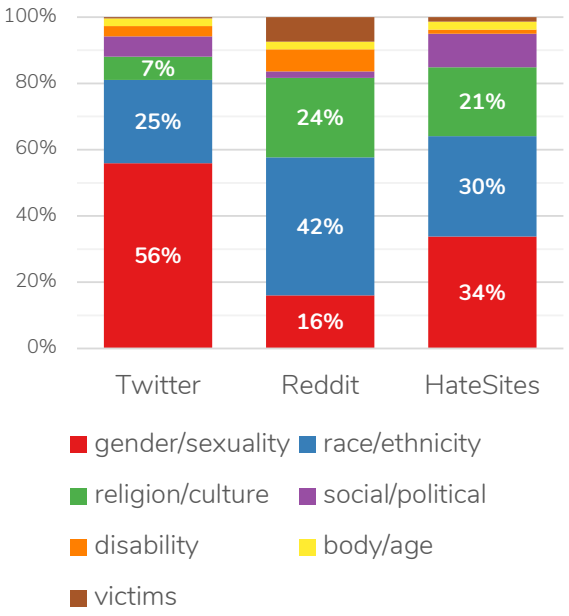

Figure 3: Breakdown of targeted group categories by domains. We show percentages within domains for the top three most represented identities, namely gender/sexuality (e.g., women, LGBTQ), race/ethnicity (e.g., Black, Latinx, and Asian), and culture/origin (e.g., Muslim, Jewish).

Annotator demographics In our final annotations, our worker pool was relatively gender-balanced and age-balanced (55% women, 42% men, <1% non-binary; 36±10 years old), but racially skewed (82% White, 4% Asian, 4% Hispanic, 4% Black).

Annotator agreement Overall, the annotations in SBIC showed 82.4% pairwise agreement and Krippendorf's α=0.45 on average, which is substantially higher than previous work in toxic language detection (e.g., α=0.22 in Ross et al., 2017). Broken down by each categorical question, workers agreed on a post being offensive at a rate of 76% (Krippendorf's α=0.51), its intent being to offend at 75% (α=0.46), and it having group implications at 74% (α=0.48). For categorizing posts as lewd, workers agreed substantially (94%, α=0.62). However, flagging potential in-group speech had lower agreement, likely because this is a very nuanced annotation, and because highly skewed categories (only 5% "yes"; see Table 3) lead to low αs (here, α=0.17 with agreement 94%).[6] Finally, workers agreed on the exact same targeted group 80.2% of the time (α=0.50).

3.3 SBIC Description

After data collection, SBIC contains 150k structured inference tuples, covering 34k free text group-implication pairs (see Table 3). We show example inference tuples in Table 1.

[6]Given our data selection process, we expect the rate of in-group posts to be very low (see §3.3).

Additionally, we show a breakdown of the types of targeted groups in Figure 3. While SBIC covers a variety of types of biases, gender-based, race-based, and culture-based biases are the most represented, which parallels the types of discrimination happening in the real world (RWJF, 2017).

We find that our dataset is predominantly written in White-aligned English (78% of posts), as measured by a lexical dialect detector by Blodgett et al. (2016), with <10% of posts having indicators of African-American English. We caution researchers to consider the potential for dialect- or identity-based biases in labelling (Davidson et al., 2019; Sap et al., 2019a) before deploying technology based on SBIC (see Section 7).

4 Social Bias Inference

Given a post, we establish baseline performance of models at inferring SOCIAL BIAS FRAMES. An ideal model should be able to both *generate* the implied power dynamics in textual form, as well as *classify* the post's offensiveness and other categorical variables. Satisfying these conditions, we use the OpenAI-GPT transformer networks (Vaswani et al., 2017; Radford et al., 2018, 2019) as a basis for our experiments, given their recent successes at

	model	offensive *42.2% pos. (dev.)*			intent *44.8% pos (dev.)*			lewd *3.0% pos (dev.)*			group *66.6% pos (dev.)*			in-group *5.1% pos (dev.)*		
		F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.
dev.	SBF-GPT$_1$-gdy	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	SBF-GPT$_2$-gdy	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	**24.0**	85.7	14.0
	SBF-GPT$_2$-smp	**80.5**	84.3	76.9	75.3	89.9	64.7	**78.6**	80.6	76.6	66.0	67.6	64.5	–	–	–
test	SBF-GPT$_2$-gdy	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–

Table 4: Experimental results (%) of various models on the classification tasks (gdy: argmax, smp: sampling). Some models did not predict the positive class for "in-group language," their performance is denoted by "–". We bold the F_1 scores of the best performing model(s) on the development set. For easier interpretation, we also report the percentage of instances in the positive class in the development set.

classification, commonsense generation, and conditional generation (Bosselut et al., 2019; Keskar et al., 2019).

Training We cast our frame prediction task as a hybrid classification and language generation task, where we linearize the variables following the frame hierarchy.[7] At training time, our model takes as input a sequence of N tokens:

$$\begin{aligned}\mathbf{x} = \{&[\texttt{STR}], w_1, w_2, ..., w_n, [\texttt{SEP}],\\ &w_{[\texttt{lewd}]}, w_{[\texttt{off}]}, w_{[\texttt{int}]}, w_{[\texttt{grp}]}, [\texttt{SEP}],\\ &w_{[\texttt{G}]_1}, w_{[\texttt{G}]_2}, ..., [\texttt{SEP}],\\ &w_{[\texttt{S}]_1}, w_{[\texttt{S}]_2}, ..., [\texttt{SEP}],\\ &w_{[\texttt{ing}]}, [\texttt{END}]\} \quad (1)\end{aligned}$$

where [STR] is our start token, $w_{1:n}$ is the sequence of tokens in a post, $w_{[\texttt{G}]_i}$ the tokens representing the group, and $w_{[\texttt{S}]_i}$ the implied statement. We add two task-specific vocabulary items for each of our five classification tasks ($w_{[\texttt{lewd}]}$, $w_{[\texttt{off}]}$, $w_{[\texttt{int}]}$, $w_{[\texttt{grp}]}$, $w_{[\texttt{ing}]}$), each representing the negative and positive values of the class (e.g., for offensiveness, `[offY]` and `[offN]`).[8]

The model relies on a stack of transformer blocks of multi-headed attention and fully connected layers to encode the input tokens (for a detailed modelling description, see Radford et al., 2018, 2019). Since GPT is a forward-only language model, the attention is only computed over preceding tokens. At the last layer, the model projects the embedding into a vocabulary-sized vector, which is turned into a probability distribution over the vocabulary using a softmax layer.

We minimize the cross-entropy of the contextual probability of the correct token in our full linearized frame objective (of length N):

$$\mathcal{L} = -\frac{1}{N}\sum_i \log p_{\texttt{GPT}}(w_i \mid w_{0:i-1})$$

During training, no loss is incurred for lower-level variables with no values, i.e., variables that cannot take values due to earlier variable values (e.g., there is no targeted group for posts marked as non-offensive).

In our experiments we use pretrained versions of OpenAI's GPT and GPT2 (Radford et al., 2018, 2019) for our model variants, named SBF-GPT$_1$ and SBF-GPT$_2$, respectively. While their architectures are similar (stack of Transformers), GPT was trained on a large corpus of fiction books, whereas GPT2 was trained on 40Gbs of English web text.

Inference We frame our inference task as a conditional language generation task. Conditioned on the post, we generate tokens one-by-one either by greedily selecting the most probable one, or by sampling from the next word distribution, and appending the selected token to the output. We stop when the [END] token is generated, at which point our entire frame is predicted. For greedy decoding, we only generate our frames once, but for sampling, we repeat the generation procedure to yield ten candidate frame predictions and choose the highest scoring one under our model.

In contrast to training time, where all inputs are consistent with our frames' structure, at test time, our model can sometimes predict combinations of variables that are inconsistent with the constraints of the frame (e.g., predicting a post to be inoffensive, but still predict it to be offensive to a group). To mitigate this issue, we also experiment with a constrained decoding algorithm (denoted "constr") that considers various global assignments of

[7] We linearize following the order in which variables were annotated (see Figure 4). Future work could explore alternate orderings.

[8] We binarize our categorical annotations, assigning 1 to "yes," "probably," and "maybe,", and 0 to all other values.

		group targeted			implied statement		
		BLEU	Rouge-L	WMD	BLEU	Rouge-L	WMD
dev.	SBF-GPT$_1$-gdy	69.9	60.3	1.01	**49.9**	40.2	2.97
	SBF-GPT$_1$-gdy-constr	69.2	64.7	1.05	49.0	42.8	3.02
	SBF-GPT$_2$-gdy	74.2	64.6	0.90	49.8	41.4	**2.96**
	SBF-GPT$_2$-gdy-constr	73.4	**68.2**	0.89	49.6	**43.5**	**2.96**
	SBF-GPT$_2$-smp	**83.2**	33.7	**0.62**	44.3	17.8	3.31
	SBF-GPT$_2$-smp-constr	83.0	33.7	0.63	44.1	17.9	3.31
test	SBF-GPT$_2$-gdy	77.0	71.3	0.76	52.2	46.5	2.81
	SBF-GPT$_2$-gdy-constr	77.9	68.7	0.74	52.6	44.9	2.79

Table 5: Automatic evaluation of various models on the generation task. We bold the scores of the best performing model(s) on the development set. Higher is better for BLEU and ROUGE scores, and lower is better for WMD.

variables. Specifically, after greedy decoding, we recompute the probabilities of each of the categorical variables, and search for the most probable assignment given the generated text candidate and variable probabilities.[9] This can allow variables to be assigned an alternative value that is more globally optimal.[10]

4.1 Evaluation

We evaluate performance of our models in the following ways. For classification, we report precision, recall, and F_1 scores of the positive class. Following previous generative inference work (Sap et al., 2019b), we use automated metrics to evaluate model generations. We use BLEU-2 and RougeL (F_1) scores to capture word overlap between the generated inference and the references, which captures quality of generation (Galley et al., 2015; Hashimoto et al., 2019). We additionally compute word mover's distance (WMD; Kusner et al., 2015), which uses distributed word representations to measure similarity between the generated and target text.[11]

4.2 Training Details

As each post can contain multiple annotations, we define a training instance as containing one post-group-statement triple (along with the five categorical annotations). We then split our dataset into train/dev./test (75:12.5:12.5), ensuring that no post is present in multiple splits. For evaluation (dev., test), we combine the categorical variables by averaging their binarized values and re-binarizing using a .5 threshold, and compare the generated inferences (hypotheses) to all targeted groups and implied statements (references).

All experiments are carried out using HuggingFace's Transformers library.[12] We tune hyperparameters on the dev. set, and report performance for the best performing setting (according to average F_1). We train or finetune our models using a batch size of 4, a learning rate of 5×10^{-6} for GPT and 10^{-5} for GPT2 (both with linear warm up), and consider training for $e \in \{1, 2, 5\}$ epochs.

5 Results

Listed in Tables 4 and 5, our modelling results indicate that making inferences about social biases in language is challenging for these models.

Classification Shown in Table 4, models perform well on higher-level variables such as offensiveness and lewdness, despite the latter being heavily skewed. We hypothesize that correctly predicting lewdness might require more lexical matching (e.g., detecting words with sexual connotations).

Whether a group is targeted is slightly less easy for models to predict, and whether the language is in-group is even more challenging, with most of the models defaulting to never predicting it. This highly skewed category poses a challenge for all models, likely due to subtlety of the task and the lack of positive instances.

SBF-GPT$_2$-gdy is the only model that predicts positive values for in-group language, for which it benefits from constrained decoding with a 1.9% improvement in F_1 score (we show results with all constrained decoding variants in Table 7 in the appendix).

[9] We only use the possible assignments in the same forward pass; we do not use assignments from different samples.

[10] In practice, as seen in Tables 4, 5, and 7, this only slightly improves predictions.

[11] We use GloVe trained on CommonCrawl, as part of the SpaCy `en_core_web_md` package.

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

	post	predicted group	predicted implication	reference groups	reference implications
(a)	All-female casts bother me but I have literally never noticed the hundreds of movies with 0 female characters I have already watched	—	—	women	women aren't good actors; women can't be in leading roles
(b)	As expected, when the terrorist group Hamas won the election in Gaza it was the usual one man, one vote, one time, Islamist election. #Islam	muslim folks	muslims are terrorists	muslim folks; arabic folks	bad people; islam promotes controlling governments; muslims are authoritarians; not fair
(c)	"Black guy in class: *attempts to throw a paper ball into the trash* *misses* Mrs. Mundy: "You're a disgrace to your race, Marcus"	black folks	black people are trash	black folks	are good at basketball; black men are defined by athletic skill.; black men should be good at basketball.; do not take things seriously
(d)	How can a Libtard speak truth to power when the only thing they vomit from their mouths is lies?	liberals	liberals are stupid	liberals	liberals are incompetent; liberals are stupid; liberals only speak lies
(e)	I hate fat bitches	overweight/fat folks	fat people are ugly	overweight/fat folks	are not pleasant people; fat folks are all the same; fat folks are less than others; not likable

Table 6: Examples of SBF-GPT$_2$-gdy-constr model predictions. The model struggles to pick up on subtle biases (a), and tends to generate generic stereotypes rather than implications that are entailed by the post (b, c).

Generation When evaluating our models on the generation tasks (i.e., targeted group and implied statement), we find that no one model outperforms others across all metrics (Table 5).

Overall, models do well at generating the targeted groups, likely because of the more limited generation space (there are only 1.4k possible groups in SBIC). Conversely, for implied statement generation (where output space is much larger), model performance is slightly worse.

Similar to the classification tasks, SBF-GPT$_2$-gdy shows a slight increase in RougeL score when using constrained decoding, but we see a slight drop in BLEU scores.

Error analysis Since small differences in automated evaluation metrics for text generation sometimes only weakly correlate with human judgments (Liu et al., 2016), we manually perform an error analysis on a manually selected set of generated development-set examples from the SBF-GPT$_2$-gdy-constr model (Table 6). Overall, the model seems to struggle with generating textual implications that are relevant to the post, instead generating very generic stereotypes about the demographic groups (e.g., in examples b and c). The model generates the correct stereotypes when there is high lexical overlap with the post (e.g., examples d and e). This is in line with previous research showing that large language models rely on correlational patterns in data (Sap et al., 2019c; Sakaguchi et al., 2020).

6 Related Work

Bias and toxicity detection Detection of hateful, abusive, or other toxic language has received increased attention recently (Schmidt and Wiegand, 2017), and most dataset creation work has cast this detection problem as binary classification (Waseem and Hovy, 2016; Davidson et al., 2017; Founta et al., 2018). Moving beyond a single binary label, Wulczyn et al. (2017) and the PerspectiveAPI use a set of binary variables to annotate Wikipedia comments for several toxicity-related categories (e.g., identity attack, profanity). Similarly, Zampieri et al. (2019) hierarchically annotate a dataset of tweets with offensiveness and whether a group or individual is targeted. Most related to our work, Ousidhoum et al. (2019) create a multilingual dataset of 13k tweets annotated for five different emotion- and toxicity-related aspects, including a 16-class variable representing social groups targeted. In comparison, SOCIAL BIAS FRAMES not only captures binary toxicity and hierarchical information about whether a group is targeted, but also *free-text* implications about 1.4k different targeted groups and the implied harm behind statements.

Similar in spirit to this paper, recent work has tackled more subtle bias in language, such as microaggressions (Breitfeller et al., 2019) and condescension (Wang and Potts, 2019). These types of biases are in line with the biases covered by SOCIAL BIAS FRAMES, but more narrowly scoped.

Inference about social dynamics Various work has tackled the task of making inferences about power and social dynamics. Particularly, previous work has analyzed power dynamics about specific entities, either in conversation settings (Prabhakaran et al., 2014; Danescu-Niculescu-Mizil et al., 2012) or in narrative text (Sap et al., 2017; Field et al., 2019; Antoniak et al., 2019). Additionally, recent work in commonsense inference has focused on mental states of participants of a situation (e.g., Rashkin et al., 2018; Sap et al., 2019b). In contrast to reasoning about particular individuals, our work focuses on biased implications of social and demographic groups as a whole.

7 Ethical Considerations

Risks in deployment Automatic detection of offensiveness or reasoning about harmful implications of language should be done with care. When deploying such algorithms, ethical aspects should be considered including which performance metric should be optimized (Corbett-Davies et al., 2017), as well as the fairness of the model on speech by different demographic groups or in different varieties of English (Mitchell et al., 2019). Additionally, deployment of such technology should discuss potential nefarious side effects, such as censorship (Ullmann and Tomalin, 2019) and dialect-based racial bias (Sap et al., 2019a; Davidson et al., 2019). Finally, offensiveness could be paired with promotions of positive online interactions, such as emphasis of community standards (Does et al., 2011) or counter-speech (Chung et al., 2019; Qian et al., 2019).

Risks in annotation Recent work has highlighted various negative side effects caused by annotating potentially abusive or harmful content (e.g., acute stress; Roberts, 2016). We mitigated these by limiting the number of posts that one worker could annotate in one day, paying workers above minimum wage ($7–12), and providing crisis management resources to our annotators.[13] Additionally, we acknowledge the implications of using data available on public forums for research (Zimmer, 2018) and urge researchers and practitioners to respect the privacy of the authors of posts in SBIC (Ayers et al., 2018).

[13]We direct workers to the Crisis Text Line (https://www.crisistextline.org/).

8 Conclusion

To help machines reason about and account for societal biases, we introduce SOCIAL BIAS FRAMES, a new structured commonsense formalism that distills knowledge about the biased implications of language. Our frames combine categorical knowledge about the offensiveness, intent, and targets of statements, as well as free-text inferences about which groups are targeted and biased implications or stereotypes. We collect a new dataset of 150k annotations on social media posts using a new crowdsourcing framework and establish baseline performance of models built on top of large pretrained language models. We show that while classifying the offensiveness of statements is easier, current models struggle to generate relevant social bias inferences, especially when implications have low lexical overlap with posts. This indicates that more sophisticated models are required for SOCIAL BIAS FRAMES inferences.

Acknowledgments

We thank the anonymous reviewers for their insightful comments. Additionally, we are grateful to Hannah Rashkin, Lucy Lin, Jesse Dodge, Hao Peng, and other members of the UW NLP community for their helpful comments on the project. This research was supported in part by NSF (IIS-1524371, IIS-1714566), DARPA under the CwC program through the ARO (W911NF-15-1-0543), and DARPA under the MCS program through NIWC Pacific (N66001-19-2-4031).

References

Maria Antoniak, David Mimno, and Karen Levy. 2019. Narrative paths and negotiation of power in birth stories. In *CSCW*.

John W Ayers, Theodore L Caputi, Camille Nebeker, and Mark Dredze. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ digital medicine*, 1(1):1–2.

Maria Becker, Michael Staniek, Vivi Nastase, and Anette Frank. 2017. Enriching argumentative texts with implicit knowledge. In *NLDB*.

Jeanette Bicknell. 2007. What is offensive about offensive jokes? *Philosophy Today*, 51(4):458–465.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: a case study of African-American English. In *EMNLP*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *ACL*.

Lorraine Bowman-Grieve. 2009. Exploring "Stormfront": a virtual community of the radical right. *Studies in conflict & terrorism*, 32(11):989–1007.

Luke M Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: a case for locating elusive phenomena in social media posts. In *EMNLP*.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE.

Southern Poverty Law Center. 2012. Misogyny: the sites. *Intelligence Report*, 145.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*.

Raphael Cohen-Almagor. 2014. Countering hate on the internet. *Annual review of law and ethics*, 22:431–443.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *KDD*.

Adam M Croom. 2011. Slurs. *Language Sciences*, 33(3):343–358.

Helen L Daly. 2018. On insults. *Journal of the American Philosophical Association*, 4(4):510–524.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: language effects and power differences in social interaction. In *WWW*.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Abusive Language Workshop*.

Thomas Davidson, Dana Warmsley, Michael W Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*.

Serena Does, Belle Derks, and Naomi Ellemers. 2011. Thou shalt not discriminate: how emphasizing moral ideals rather than obligations increases whites' support for social equality. *Journal of Experimental Social Psychology*, 47(3):562–571.

Marta Dynel. 2015. The landscape of impoliteness research. *Journal of Politeness Research*, 11(2):383.

Connie C Eble. 1996. *Slang & sociability: in-group language among college students*. Univ of North Carolina Press.

Anjalie Field, Gayatri Bhat, and Yulia Tsvetkov. 2019. Contextual affective analysis: a case study of people portrayals in online #MeToo stories. In *ICWSM*.

Charles J Fillmore and Collin F Baker. 2001. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

Jon Fingas. 2017. Reddit bans misogynist community as part of anti-violence crackdown. https://www.engadget.com/2017/11/08/reddit-bans-misogynist-community-in-anti-violence-crackdown/. Accessed: 2019-12-06.

Susan T Fiske. 1993. Controlling other people. the impact of power on stereotyping. *American psychologist*, 48(6):621–628.

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. In *ICWSM*.

Yiannis Gabriel. 1998. An introduction to the social psychology of insults in organizations. *Human Relations*, 51(11):1329–1354.

Adam D Galinsky, Cynthia S Wang, Jennifer A Whitson, Eric M Anicich, Kurt Hugenberg, and Galen V Bodenhausen. 2013. The reappropriation of stigmatizing labels: the reciprocal relationship between power and self-labeling. *Psychol. Sci.*, 24(10):2020–2029.

Michel Galley, Chris Brockett, Alessandro Sordoni, Yangfeng Ji, Michael Auli, Chris Quirk, Margaret Mitchell, Jianfeng Gao, and William B. Dolan. 2015. deltaBLEU: a discriminative metric for generation tasks with intrinsically diverse targets. In *ACL*.

Ona de Gibert, Naiara Pérez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Abusive Language Workshop at EMNLP*.

Gil Greengross and Geoffrey F Miller. 2008. Dissing oneself versus dissing rivals: effects of status, personality, and sex on the Short-Term and Long-Term attractiveness of Self-Deprecating and Other-Deprecating humor. *Evolutionary Psychology*, 6(3).

Shirley Gregor and Izak Benbasat. 1999. Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS quarterly*, pages 497–530.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. In *NAACL-HLT*.

Marti A Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *ACL*, pages 539–545.

Amanda Hess. 2016. The far right has a new digital safe space. https://www.nytimes.com/2016/11/30/arts/the-far-right-has-a-new-digital-safe-space.html. Accessed: 2019-12-06.

Gabriele Kasper. 1990. Linguistic politeness: current research issues. *Journal of Pragmatics*, 14(2):193–218.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: a conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1–10. ACM.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *ICML*, pages 957–966.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In *ACL*.

Marie-Catherine de Marneffe, Christopher D Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *FAccT*.

An T Nguyen, Aditya Kharosekar, Saumyaa Krishnan, Siddhesh Krishnan, Elizabeth Tate, Byron C Wallace, and Matthew Lease. 2018. Believe it or not: designing a human-AI partnership for mixed-initiative fact-checking. In *The 31st Annual ACM Symposium on User Interface Software and Technology*, pages 189–199. ACM.

Conor J O'Dea, Stuart S Miller, Emma B Andres, Madelyn H Ray, Derrick F Till, and Donald A Saucier. 2015. Out of bounds: Factors affecting the perceived offensiveness of racial slurs. *Language Sciences*, 52:155–164.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and Multi-Aspect hate speech analysis. In *EMNLP*.

Gonçalo Pereira, Rui Prada, and Pedro A Santos. 2016. Integrating social power into the decision-making of cognitive agents. *Artificial Intelligence*, 241:1–44.

Vinodkumar Prabhakaran, Prabhakaran Vinodkumar, and Rambow Owen. 2014. Predicting power relations between participants in written dialog from a single thread. In *ACL*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *EMNLP*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Unpublished.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Unpublished.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2mind: commonsense inference on events, intents, and reactions. In *ACL*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *KDD*.

Sarah T Roberts. 2016. Commercial content moderation: digital laborers' dirty work. In Safiya Umoja Noble and Brendesha M Tynes, editors, *The Intersectional Internet: Race, Sex, Class and Culture Online*, Media Studies Publications. Peter Lang Publishing.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: the case of the european refugee crisis. In *NLP 4 CMC Workshop*.

RWJF. 2017. Discrimination in america: experiences and views. https://www.rwjf.org/en/library/research/2017/10/discrimination-in-america--experiences-and-views.html. Accessed: 2019-11-5.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Winogrande: an adversarial winograd schema challenge at scale. In *AAAI*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *ACL*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019b. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *AAAI*.

Maarten Sap, Marcella Cindy Prasetio, Ariel Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *EMNLP*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019c. Social IQa: commonsense reasoning about social interactions. In *EMNLP*.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Workshop on NLP for Social Media at EACL*.

Robyn Speer and Catherine Havasi. 2012. Representing general relational knowledge in ConceptNet 5. In *LREC*.

Whitney Strub. 2008. The clearly obscene and the queerly obscene: heteronormativity and obscenity in cold war los angeles. *American Quarterly*, 60(2):373–398.

Stefanie Ullmann and Marcus Tomalin. 2019. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

James Vincent. 2016. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist. Accessed: 2019-10-26.

Andrew J Vonasch and Roy F Baumeister. 2017. Unjustified side effects were strongly intended: taboo tradeoffs and the side-effect effect. *Journal of Experimental Social Psychology*, 68:83–92.

Zijian Wang and Christopher Potts. 2019. TalkDown: a corpus for condescension detection in context. In *EMNLP*.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *NAACL Student Research Workshop*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: personal attacks seen at scale. In *WWW*.

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 *NAACL*.

Michael Zimmer. 2018. Addressing conceptual gaps in big data research ethics: an application of contextual integrity. *Social Media + Society*, 4(2).

	model	offensive *42.2% pos. (dev.)*			intent *44.8% pos. (dev.)*			lewd *3.0% pos. (dev.)*			group *66.6% pos. (dev.)*			in-group *5.1% pos. (dev.)*		
		F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.	F_1	pr.	rec.
dev.	SBF-GPT$_1$-gdy	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	″-constr	75.2	88.3	65.5	74.4	89.8	63.6	75.2	78.2	72.5	62.3	74.6	53.4	–	–	–
	SBF-GPT$_2$-gdy	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	24.0	85.7	14.0
	″-constr	77.2	88.3	68.6	**76.3**	89.5	66.5	77.6	81.2	74.3	**66.9**	67.9	65.8	**25.9**	63.6	16.3
	SBF-GPT$_2$-smp	**80.5**	84.3	76.9	75.3	89.9	64.7	**78.6**	80.6	76.6	66.0	67.6	64.5	–	–	–
	″-constr	80.4	84.3	76.8	75.3	89.9	64.7	78.5	80.6	76.5	66.0	67.6	64.5	–	–	–
test	SBF-GPT$_2$-gdy	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–
	″-constr	78.8	89.8	70.2	78.6	90.8	69.2	80.7	84.5	77.3	69.9	70.5	69.4	–	–	–

Table 7: Full experimental results (%) of various models on the classification tasks (gdy: argmax, smp: sampling; constr: constrained decoding). Some models did not predict the positive class for "in-group language," their performance is denoted by "–". We bold the F_1 scores of the best performing model(s) on the development set. For easier interpretation, we also report the percentage of instances in the positive class in the development set.

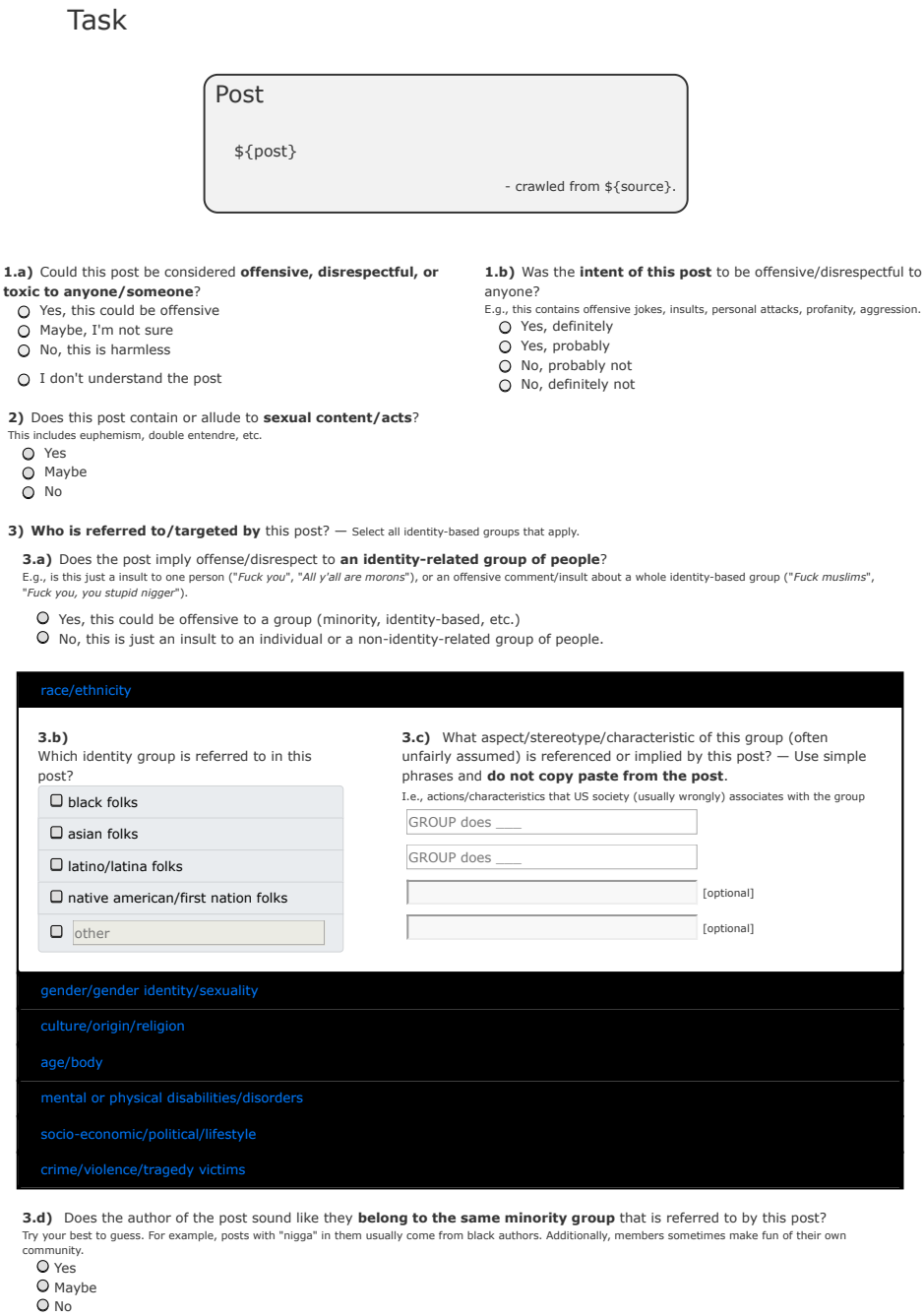

Figure 4: Snippet of the annotation task used to collect SBIC. The collection of structured annotations for our framework is broken down into questions pertaining to offensiveness, intent of the post, targeted group and minority speaker.