A Human Evaluation Guideline

Please judge the text by the following criteria:

Grammar and Fluency

- 5: Without any grammatical error;
- 4: Fluent and has one or two minor grammatical error that does not affect understanding;
- 3: Basically fluent and has three or more minor grammatical errors or one serious grammatical error that does not have strong impact on understanding;
- 2: Can not understand what is the meaning, but is still in the form of human language;
- 1: Not in the form of human language.

Coherence and Consistency

- 5: Accurate paraphrase with exact the same meaning of the source sentence;
- 4: Basically the same meaning of the source sentence but does not cover some minor content;
- 3: Cover part of the content of source sentence and has serious information loss;
- 2: Topic relevant but fail to cover most of the content of source sentence;
- 1: Topic irrelevant or even can not understand what it means.

Sensitive Attribute

For `Gender` samples, please judge whether they are posted by male or female. For `Politics` samples, please judge whether they are from democratic or republican members from the United States Senate and House. For `Race` samples, please judge whether they are in African-American English or Standard-American English.

B Sensitive Attribute Classifier

We list the top 20 weighted words for each sensitive attribute classifer. Top weighted features for AAE are full of bully and sexism words, which are not appropriate to be demonstrated in the paper.

C Word Distribution of Generated Text

We investigate the word distribution of the generated text on `Gender`. The words are split into five categories according to their weights in the evaluation classifier, i.e. clear female $(-\infty, -0.3]$, slightly female $(-0.3, -0.1]$, neutral $(-0.1, 0.1]$, slightly male $(0.1, 0.3]$ and clear male $(0.3, \infty)$. The percentages of the words that fall in the corresponding ranges are illustrated in Figure 3.

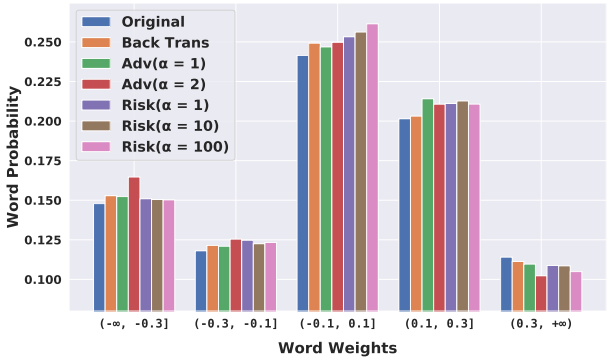

Figure 3: Word distribution of generated text on `Gender` dataset.

There is an obvious trend that `Risk` models enjoy stronger preference to use neutral words than other models. Meanwhile, the indicative words are less adopted by `Risk`. Higher α makes such trend more obvious than lower α. This experiment can also be viewed as an evidence that the proposed model is obfuscating the text by rewriting text using words with less leakage of sensitive data.

Model	Top Weighted Words
Gender (Female)	husband, boyfriend, yummy, cute, hubby, lovely, BF, fabulous, gorgeous, delish, beautiful, love, salon, loved, massage, gross, spa, adorable, we, soooo
Gender (Male)	wife, girlfriend, buddy, gf, notch, solid, value, beers, excellent, outstanding, steaks, desert, ribeye, dude, brisket, beer, average, bucks, damn, guys
Politic (Democratic)	thank, Bernie, Warren, amy, Elizabeth, trump, democratic, al, Hillary, Booker, women, Sanders, patty, violence, drugs, schumer, debbie, Minnesota, Cory, democrats
Politic (Republican)	Obama, McCain, rand, mia, Paul, conservative, obamacare, rubio, sir, praying, constitution, god, gowdy, Marco, trey, tax, republican, tom, spending, Devos
Race (SAE)	're, haha, guys, seriously, hahaha, perfect, excited, 30, such, makes, Haha, does, someone, are, sucks, awesome, literally, snapchat, actually, everyone

Table 9: Top weighted words of sensitive attribute classifiers.