A Supplemental Material

In this section, we list the identity terms we used as fairness priors and toxic terms we used as toxic priors.

A.1 Identity Terms

lesbian	gay	bisexual
transgender	trans	cis
queer	lgbt	homosexual
straight	heterosexual	male
female	african	black
white	european	hispanic
latino	latina	mexican
canadian	american	asian
indian	chinese	japanese
christian	muslim	jewish
buddhist	catholic	protestant
sikh	old	older
young	younger	teenage
elderly	blind	deaf

Table 1: Identity terms for fairness experiments.

A.2 Toxic Terms

bullshit	fuck	fucking
crap	jerk	shit
moron	hell	asshole
loser	dick	anal
bitch	penis	faggot
fucked	fucker	wtf
cock	cunt	pussy
morons	bastard	piss
damn	dickhead	dumbass
retarded	retard	arse
bastards	motherfucker	jackass
coward	douchebag	pricks
prick	fu	bitches
fuckhead	cocksucker	fcuk
fuckers	fuckin	arsehole
penises	whore	assholes
fuckwit	scumbag	

Table 2: Toxic terms for scarce data experiments.