Bias Misperceived: The Role of Partisanship and Misinformation in YouTube Comment Moderation (Supplementary Materials)

Shan Jiang, Ronald E. Robertson, Christo Wilson
Northeastern University, USA
{sjiang, rer, cbw}@ccs.neu.edu,

1 Lexicon

We selected eight categories from ComLex. These categories are determined by a preliminary linear regression model that showed significant ($p < 0.001$) effect between the given work category and moderation likelihood for comments. All words in these categories are shown in Table 1.

2 Partisanship Score Distribution

The distribution of partisanship scores for comments in our dataset is shown in Figure 1. We observe that the distribution of partisanship scores is unbalanced. Therefore, setting a minimum threshold for left/right labels filters out different portions of comments from any subsequent analysis.

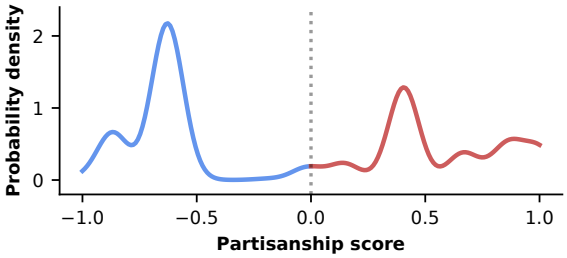

Figure 1: **Distribution of partisanship score.** The right/left distribution is unbalanced.

3 Label Distribution

The label distributions for each treatment and control are shown in Figure 2-16. Each figure shows the distribution of a label split by moderation outcome. All selected labels are distributed distinctly between moderated and unmoderated comments.

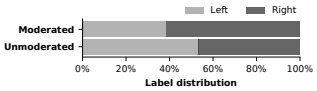

Figure 2: **Right/Left.**

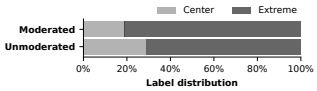

Figure 3: **Extreme/Center.**

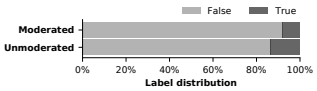

Figure 4: **True/False.**

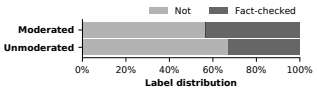

Figure 5: **Fact-check/Not.**

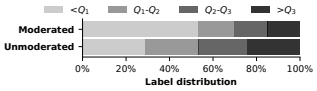

Figure 6: **Views.**

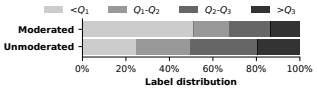

Figure 7: **Likes.**

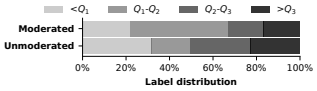

Figure 8: **Dislikes.**

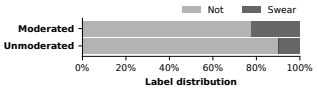

Figure 9: **Swear.**

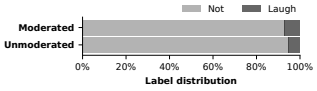

Figure 10: **Laugh.**

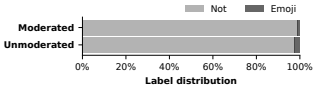

Figure 11: **Emoji.**

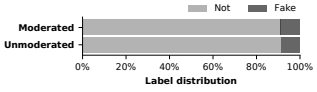

Figure 12: **Fake.**

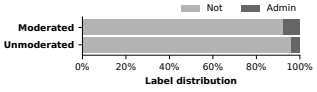

Figure 13: **Administration.**

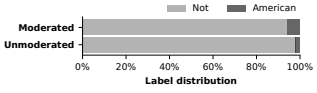

Figure 14: **American.**

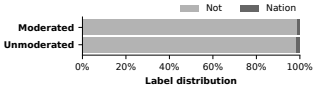

Figure 15: **Nation.**

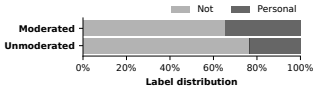

Figure 16: **Personal.**

Table 1: **Lexicon.** Words in all selected categories from ComLex, which we use to derive linguistic controls.

Category	Words
Swear	moron, fool, loser, clown, cunt, dumbass, jerk, douche, prick, jackass, dummy, turd, lunatic, twat, douchebag, looser, buffoon, dipshit, dickhead, psycho, twit, fucktard, bozo, goof, wanker, dolt, tard, nitwit, dork, dimwit, schmuck, simpleton, putz, dumbfuck, blowhard, quack, dunce, nutter, doofus, dingbat, fuckhead, bitch, dude, nigger, nigga, fucker, negro, monkey, faggot, redneck, cracker, homo, fag, motherfucker, sob, queer, ghetto, mf, juggalos, hick, coon, gangster, hillbilly, inbred, weirdo, slob, kettle, chimp, whitey, gorilla, cocksucker, gangsta, baboon, mofo, bich, spook, wetback, spic, kike, *, f, #, , sh, fu, ck, ing, bullsh, cking, pu, fuc, tch, shi, fuck, fuckin, dam, fuk, fucken, goddamn, fk, fkn, fricken, damm, fn, f'n, f'ing, fck, fricking, frigging, fukin, fing, gd, fking, fuking, fuckn, fing, fcking, friggen, bastard, scum, sheep, parasite, scumbags, degenerate, ungrateful, goon, maggot, cockroach, celebs, fucktards, swine, junky, asshats, peasant, cretin, shithead, douchebags, complainer, vermin, neanderthal, naysayer, slug, dipshits, dweller, a-holes, politicans, subhuman, knucklehead
Laugh	lol, omg, wtf, lmao, haha, xd, hahaha, lmfao, hahahaha, wth, hahahahaha, omfg, hahah, hahahahahaha, jk, lmaooo, lolol, lololol, lmbo, bahahaha, bahaha, lmfaooo, lolololol, yooo, lmaoo, lmmfao
Emoji	:), :d, <3, ;), :-), :p, :/, ;-), =), (:, :-(, :'(, xx, xxx, [other graphic emojis]
Fake	hoax, failure, genius, tool, conspiracy, usual, prank, scam, myth, distraction, scheme, stunt, maverick, disappointment, slogan, coincidence, pawn, travesty, ploy, non-issue, bunk, farce, setup, metaphor, shocker, scapegoat, diversion, disservice, ponzi, prat, gimmick, grandstand, fake, false, bias, twist, spin, mislead, incorrect, debunk, photoshopped, stag, bogus, contradict, distort, untrue, blatantly, fabricate, exaggerate, inaccurate, parrot, skew, disingenuous, deceptive, baseless, slant, misrepresent, rehearse, regurgitate, malicious, unfounded, repetitive, falsify, misinterpret, contrive, trickery, lie, propaganda, corruption, rhetoric, tactic, spew, narrative, blatant, deception, rumor, outright, misinformation, manipulation, deceit, speculation, distortion, falsehood, hysteria, fabrication, innuendo, disinformation, hyperbole, misrepresentation, propoganda, untruth, fear-mongering, joke, disgrace, trick
Administration	state, former, county, lawyer, chief, sheriff, speaker, attorney, secretary, district, mayor, minister, holder, council, session, clerk, assistant, assembly, chairman, ag, deputy, lt, state's, trooper, continental, administrator, oversee, parliament, connecticut, col, colonial, commissioner, govenor, supervisor, governor's, univ, provincial, delaware, sheriff's
American	detroit, houston, tx, sc, nyc, ga, az, wv, valley, rural, atlanta, dakota, cleveland, kansa, seattle, montana, brooklyn, toronto, phoenix, tn, denver, mn, philly, wyoming, delta, columbus, killeen, midwest, tampa, idaho, pittsburgh, portland, nashville, nebraska, memphis, southeast, vancouver, fla, southwest, northeast, sw, nm, texas, california, florida, nc, carolina, ny, alabama, michigan, arizona, ohio, georgia, alaska, wisconsin, arkansas, colorado, hawaii, virginia, massachusetts, missouri, louisiana, oklahoma, utah, mississippi, illinois, indiana, fl, iowa, cali, miami, minnesota, nj, wi, pennsylvania, tennessee, oregon, kentucky, oakland, nevada, maine, greensboro, maryland, vermont, calif
Nation	canada, mexico, europe, france, uk, germany, india, england, australia, britain, kenya, brazil, greet, italy, sweden, ireland, indonesia, spain, norway, venezuela, greece, abroad, nigeria, poland, scotland, philippine, switzerland, denmark, netherlands, bangladesh, iceland, belgium, portugal, ontario, alberta, sri, finland, quebec, lanka
Personal	your, my, their, her, ur, people's, whose, someone's, everyone's, man's, one's, thier, woman's, person's, anyone's, other's, somebody's, everybody's, everyones, carrie's, anyones, another's