Interpretable Word Embeddings via Informative Priors

Miriam Hurtado Bodell* Linköping University miriam.hurtado.bodell@liu.se

Martin Arvidsson* Linköping University martin.arvidsson@liu.se

Måns Magnusson Aalto University mans.magnusson@aalto.fi

A Pre-processing

We pre-process each corpus with the following steps:

1. Transform the text to lowercase, remove all punctuation, replace numbers with *X* and apply stemming.
2. Set vocabulary as the V most frequent word types in the corpora, and remove other word types from the text.
3. To speed up estimation, we follow (Mikolov et al., 2013; Rudolph et al., 2016) and remove each token with probability $1 - \sqrt{\frac{10^{-5}}{f_i}}$ where f_i is the frequency of token i's word type.

B Prior anchor word types

In this section, we list the word types that are used to specify the priors. Bold font marks the words used in the "few"-setting, while all words are used in the "many"-setting. Note that all words are pre-processed by stemming (see Sec. A).

B.1 Gender

Positive: ***man***, *men, male, father, he, him, son, boy, himself, brother, uncle, nephew*

Negative: ***woman***, *women, female, mother, she, her, daughter, girl, herself, sister, aunt, niece*

B.2 Sentiment

Positive: ***posit, good***, *accomplish, admir, advantag, adventur, amus, approv, ardent, attract, bargain, bliss, celebr, cherish, clean, comfort, courag, dare, defend, delight, desir, eager, ecstat, enchant, energet, enlighten, enterpris, entertain, ethic, excit, fearless, festiv, fond, freedom, gain, gallant, glori, gracious, guarante, hardi, help, hero, heroic, honest, honor, hope, humor, import, impress, improv, influenti, inspir, intellig, interest, kudo, luck, merci, merri, miracl, nobl, passion, perfect, picturesqu, play, pleas, power, prais, progress, promis, protect, reassur, recommend, rejoic, safe, satisfi, smile, solid, stabl, support, sweet, tender, thank, triumph, triumphant, unbias, visionari, willing, winner, worthi*

Negative: ***negat, bad***, *abus, accident, ach, afraid, aggrav, alien, anger, anguish, animos, annoy, antagonist, anxieti, anxious, appal, arrog, attack, aw, bastard, bias, bitch, bitter, bizarr, bomb, bore, broken, cancer, casualti, catastroph, chao, childish, clash, complain, condemn, confus, contagi, contempt, controversi, coward, cramp, crash, crime, crisi, critic, cruel, cri, damag, deadlock, death, deceiv, defect, despair, destruct, devast, die, dirt, dirti, disast, disastr, discord, dishonest, disorgan, disparag, disrupt, distract, distress, dizzi, doom, doubt, dubious, dumb, embarrass, enemi, enslav, erron, error, exagger, excus, exhaust, falsifi, farc, fear, fiasco, foolish, fraudul, frenzi, furious, haunt, helpless, hindranc, horribl, hostil, humili, hurt, hypocrit, hysteria, hyster, idiot, illeg, ill, impati, inact, inadequ, incompet, indecis, indiffer, indign, inferior, insignific, insult, irrat, lack, lag, loath, loss, lost, lurk, mad, manipul, mediocr, melancholi, menac, mischief, miseri, mistak, mistaken, mourn, murder, nasti, needi, nervous, noisi, obliter, obscen, pain, panic, passiv, pathet, pollut, powerless, prick, problem, prosecut, punish, rape, rash, reject, remors, reveng, risk, scare, scream, shaki, shit, shock, shortag, sick, sin, spite, strike, suck, suicid, suspect, suspici, terribl, terror, threat, tortur, traumat, treason, unaccept, unbeliev, uncertain, undecid, undermin, uneasi, unequ, unhappi, unjust, unsettl, unsupport, upset, urgent, weird, whore, worsen, worthless, wreck, wrong*

* Equal contribution

B.3 Neutral

Neutral: *the, it, a, an, and, as, of, at, by*

C Hold-out test words

In this section, we list all the word types used as hold-out words to test the accuracy of the different prior specifications.

C.1 Gender

Positive: *aaron, adam, alan, albert, alexand, andrew, anthoni, arthur, benjamin, bobbi, brandon, brian, carl, charl, christoph, daniel, david, denni, dougla, edward, ethan, eugen, gabriel, georg, gerald, gregori, harold, henri, jack, jacob, jame, jason, jeremi, jerri, jess, joe, john, johnni, jonathan, jordan, jose, joseph, joshua, juan, justin, keith, kenneth, kevin, larri, lawrenc, loui, matthew, michael, nathan, nichola, noah, patrick, paul, peter, philip, randi, raymond, richard, robert, roger, russel, samuel, scott, stephen, steven, terri, thoma, timothi, vincent, walter, willi, william, zachari, king, mr, sir, princ, gentleman, gentlemen, knight, lad, mankind, monk, pope, grandfath, papa, baron, clergyman, workmen, waiter, workman, brotherhood, schoolboy, masculin, brotherinlaw, grandson, fatherinlaw, boyhood, superman, grandpapa, godfath, dad, stepfath, grandpa, greatgrandfath, cowboy, daddi, fatherhood, grandnephew, granddad, businessman, businessmen, bradley, bruce, bryan, donald, dylan, eric, gari, jeffrey, kyle, logan, ralph, ronald, roy, ryan, sean, tyler, wayn, boyfriend, batman, fanboy, boyz, playboy, stepdad, homeboy, frat, exboyfriend, boyband, babyboy, penis, granda, congressman, vicechairman*

Negative: *abigail, amber, ami, andrea, angela, ann, anna, barbara, betti, brittani, carol, catherin, christin, cynthia, deborah, dian, diana, donna, dori, dorothi, elizabeth, emma, evelyn, gloria, hannah, heather, helen, jacquelin, jane, janet, jessica, joan, joyc, judith, juli, julia, katherin, kathleen, kelli, laura, lori, margaret, mari, maria, martha, nanci, natali, olivia, pamela, rachel, rebecca, ruth, sara, sarah, sharon, shirley, sophia, susan, teresa, theresa, victoria, mrs, ladi, queen, princess, breast, mistress, duchess, goddess, grandmoth, hostess, nun, landladi, feminin, gentlewoman, sisterinlaw, mama, stepmoth, womanhood, actress, granddaught, motherinlaw, frenchwoman, godmoth, nunneri, schoolgirl, princesss, grandmama, womankind, sisterhood, grandmamma, waitress, grandma, motherhood, greatgrandmoth, alexi, amanda, ashley, bever, brenda, carolyn, cheryl, christina, daniell, debra, denis, janic, jennif, judi, karen, kathryn, kayla, kimber, lauren, linda, lisa, madison, marilyn, megan, melissa, michell, nicol, patricia, samantha, sandra, stephani, mommi, girlfriend, momma, girli, lesbian, fangirl, babygirl, homegirl, stepmom, cowgirl, girlz, uterus, superwoman, breastfeed, feminist, grandmom, femin, congresswoman, chairwoman, servicewomen, churchwomen, businesswomen, businesswoman, vagin, femalehead, spokeswoman*

C.2 Sentiment

Positive: *masterpiec, heaven, tranquil, heartfelt, clever, commend, pardon, earnest, remark, bless, treasur, pretti, faith, privileg, benefit, fortun, pleasant, encourag, loyalti, loyal, effect, ador, creativ, hug, friend, raptur, glee, joy, fame, thought, affect, fair, peac, optim, happi, grate, superior, sparkl, swift, award, eas, excel, lucki, vigil, reviv, favorit, wonder, robust, rest, innov, cheer, calm, outstand, eleg, generous, glad, hail, virtuous, confid, fascin, fun, agreeabl, beauti, hilari, comprehens, brilliant, steadfast, grace, advanc, joyous, superb, reward, respons, cute, compassion, worth, enjoy, sincer, marvel, prosper, charm, healthi, proud, top, cool, splendid*

Negative: *defiant, retard, bankrupt, danger, propaganda, contenti, greedi, dull, insan, moodi, hardship, disregard, curs, greed, undesir, bulli, steal, misunderstand, deni, guilt, delay, vagu, broke, decept, interrupt, inhibit, lazi, liar, useless, ruin, naiv, unhealthi, outcri, shame, cruelti, weak, dire, limit, horrifi, mock, grief, choke, dread, asham, jealous, punit, angri, dark, mislead, dismal, oppress, accus, threaten, fool, fall, guilti, stab, mess, prison, bloodi, resign, distrust, stolen, worri, lonesom, fraud, horrif, scorn, dump, collaps, scandal, frantic, sabotag, skeptic, disdain, regret, stupid, hopeless, frustrat, chaotic, question, pretend, peril, disgrac, reckless, betray, blame, worn, apathi, disrespect, fatigu, miser, lunat, injustic, blind, stereotyp, thwart, unstabl, ugli, disput, ignor, weari, tire, penalti, disgust, intimid, ridicul, inabl, havoc, resent, recess, injuri, insecur, irrit, boycott, rage, persecut, cheat, disturb, sad, outrag, fright, neglect,*

denounc, sorrow, poverti, hell, banish, tragic, infring, silli, struggl, tragedi, selfish, nonsens, sore, restrict, uncomfort, numb, crimin, subvers, aggress, chagrin, refus, bother, wast, desper, isol, alarm, hoax, evil, poison, wick, weep, obstacl, wors, kill, lose, suffer, exclus, gross, harm, scold, screw, lone, leak, distort, depress, apprehens, meaningless, emerg, violent, fake, crush, damn, offend, disappoint, displeas, conflict, ineffect, crazi, debt, degrad, deceit, vicious, disord, timid, jeopardi, expel

References

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. 2016. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486.