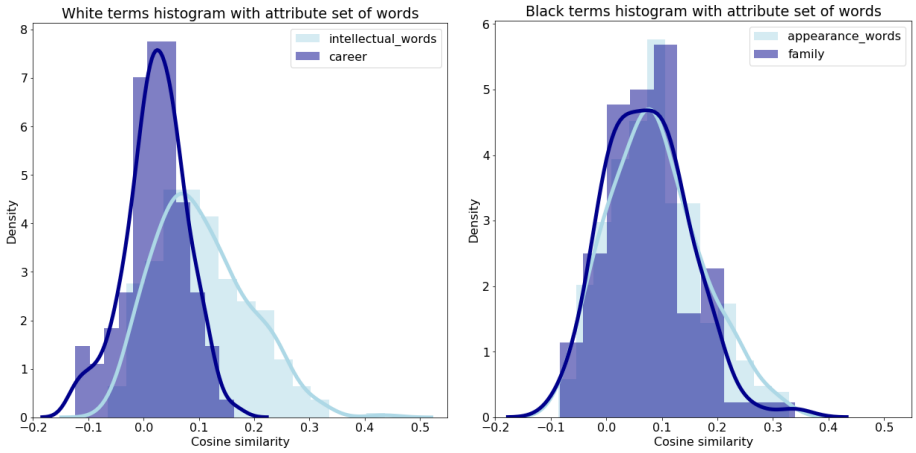
A Prior Analysis



(a) No irregularities here: similar angle outputs for both attribute sets.

(b) Removing career set of words from white sets may not be necessary.

Figure 1: Kernel density estimate of cosine similarities between black, white word sets and their corresponding attribute word sets ready for debiasing (Word2Vec)

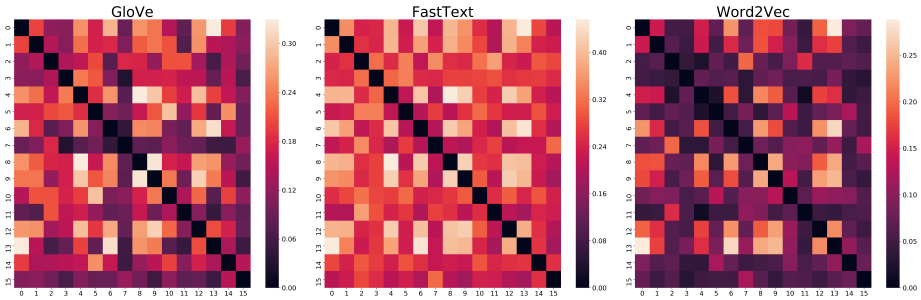


Figure 2: Median cosine similarity value between each pair of attribute sets (Not measured between words within the same set due to easier visibility). Index-set name pairs are defined as 0: aggressive, 1: appearance, 2: art, 3: career, 4: competent, 5: family, 6: incompetent, 7: instruments, 8: intellectual, 9: likable, 10: pleasant, 11: science, 12: shy, 13: unlikable, 14: unpleasant, 15: weapons.

B Bias results

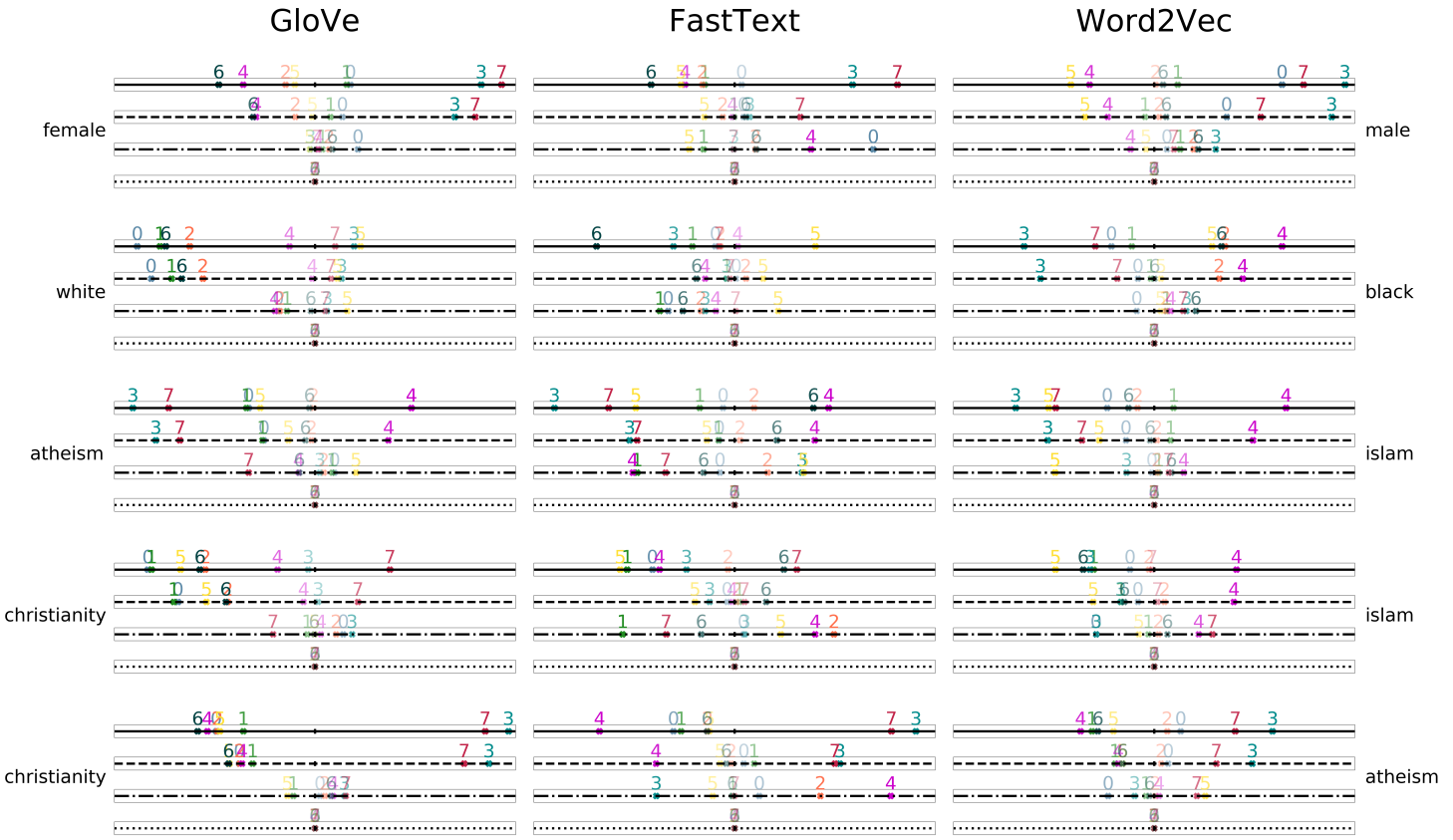


Figure 1: **WEAT results before Regular Embedding and after debiasing (SoftWEAT $\lambda = 0.2$, SoftWEAT $\lambda = 1$ and HardWEAT).** Target words pair X is on the right, Y on the left, whereas attribute pairs A, B are numbers, indexed as: 0: likable vs unlikeable, 1: competent vs incompetent, 2: shy vs aggressive, 3: intellectual vs appearance, 4: family vs career, 5: instruments vs weapons, 6: pleasant vs unpleasant, 7: science vs artIf the number is on the left side of the scale, it suggest negative d value (between -1.55 and 0) and $Y - A, X - B$ relationship. In case of positive value (between 0 and 1.55) $X - A, Y - B$.

C Quality of Embeddings

	Regular	HardWEAT	SoftWEAT $\lambda = 0.2$	SoftWEAT $\lambda = 1$
RG65	**76.031***	63.422	75.677	71.425
WS	73.804	69.731	**73.997***	70.336
RW	**46.147***	46.093	46.146	45.939
MEN	80.131	77.515	**80.339***	78.738
MTurk	**71.513***	69.785	71.251	68.635
SimLex	40.881	40.437	41.464	**42.185***
SimVerb	28.735	28.744	28.897	**29.776***
Mikolov	0.652	0.635	**0.652***	0.639

Table 1: **Quality tasks on GloVe:** In 5 out of 8 tests, our methods maintain/increase quality of embeddings. HardWEAT makes a significant drop with RG65.

	Regular	HardWEAT	SoftWEAT $\lambda = 0.2$	SoftWEAT $\lambda = 1$
RG65	**86.713***	72.653	86.136	76.897
WS	**72.204***	68.279	72.175	64.411
RW	57.114	56.84	**57.43***	56.283
MEN	**81.35***	79.343	81.245	76.848
MTurk	75.101	73.757	**75.137***	70.238
SimLex	47.153	46.219	**47.517***	45.209
SimVerb	38.216	**38.238***	38.2	38.1
Mikolov	**0.853***	0.833	0.851	0.842

Table 2: **Quality tasks on FastText:** In half of the tests, quality is maintained/improved. Significant drop seen with RG65 and WS with HardWEAT and SoftWEAT with $\lambda = 1$.

	Regular	HardWEAT	SoftWEAT $\lambda = 0.2$	SoftWEAT $\lambda = 1$
RG65	**74.943***	64.025	74.86	69.69
WS	69.998	65.943	**70.128***	69.325
RW	**53.394***	53.119	53.356	52.954
MEN	**77.068***	75.307	77.025	76.304
MTurk	**67.139***	66.326	66.999	66.184
SimLex	44.255	43.954	44.301	**44.302***
SimVerb	36.749	**36.758***	36.745	36.673
Mikolov	0.74	0.722	**0.74***	0.737

Table 3: **Quality tasks on Word2vec:** Similar output as with FastText, on half tests were results maintained/improved, with RG65 achieving lower quality on HardWEAT and SoftWEAT with $\lambda = 1$.

D Sensitivity Analysis

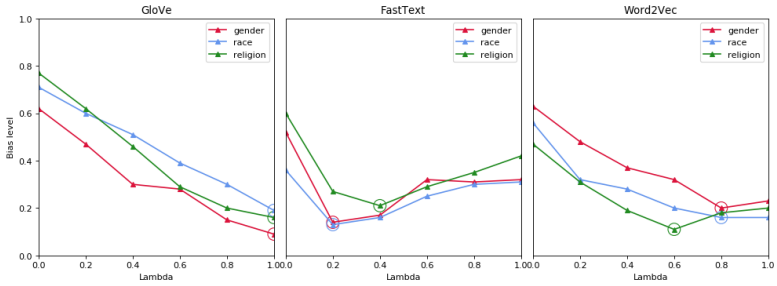


Figure 1: **WEAT levels SoftWEAT:** SoftWEAT leads to bias decrease, however not always in complete correlation with λ, like in GloVe.

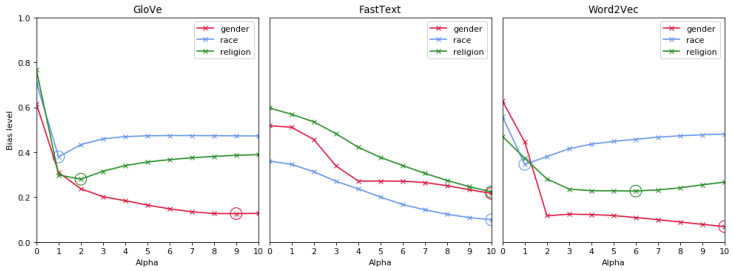


Figure 2: **WEAT levels Conceptor:** Similarly as with SoftWEAT, higher method's parameter does not always α lead to less of bias levels.

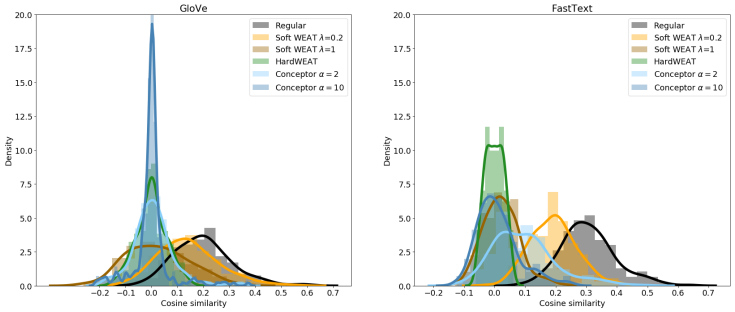


Figure 3: **Kernel Density Estimate of cosine similarity values between words of pairs of sets: atheism - science, christianity - art**: While Conceptor does reduce overall angles between target and attribute sets of words, it also decreases angles between words of actual target sets. We performed summation of sums (double sum) of cosine similarity values between each word pair within each target set. Results are following, before and after Conceptor Debiasing ($\alpha = 2$): GloVe ($1745 \rightarrow 260$), FastText ($1558 \rightarrow 490$), Word2Vec ($438 \rightarrow 85$). These values remain the same when applying SoftWEAT.

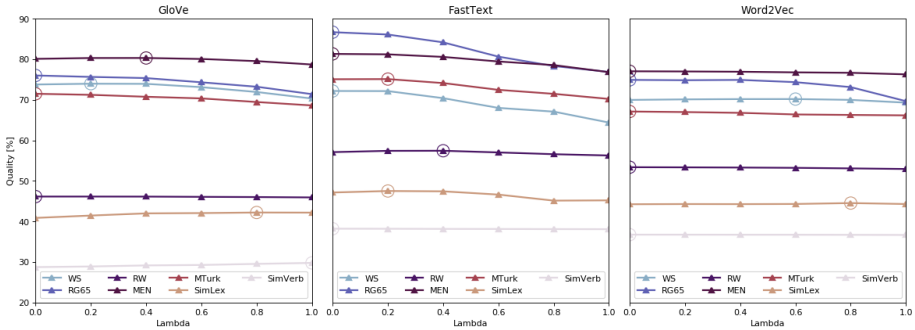


Figure 4: **Quality explained via Similarity Task - SoftWEAT**

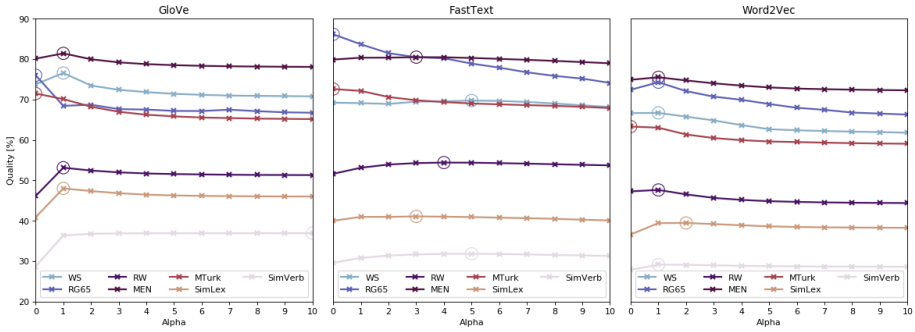


Figure 5: **Quality explained via Similarity Task - Conceptor**

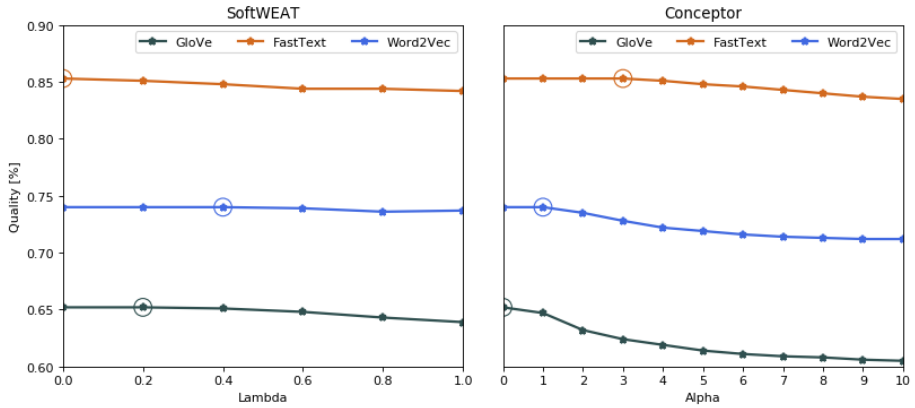


Figure 6: **Comparison of Mikolov analogy task based on different SoftWEAT λ and Conceptor's α values**

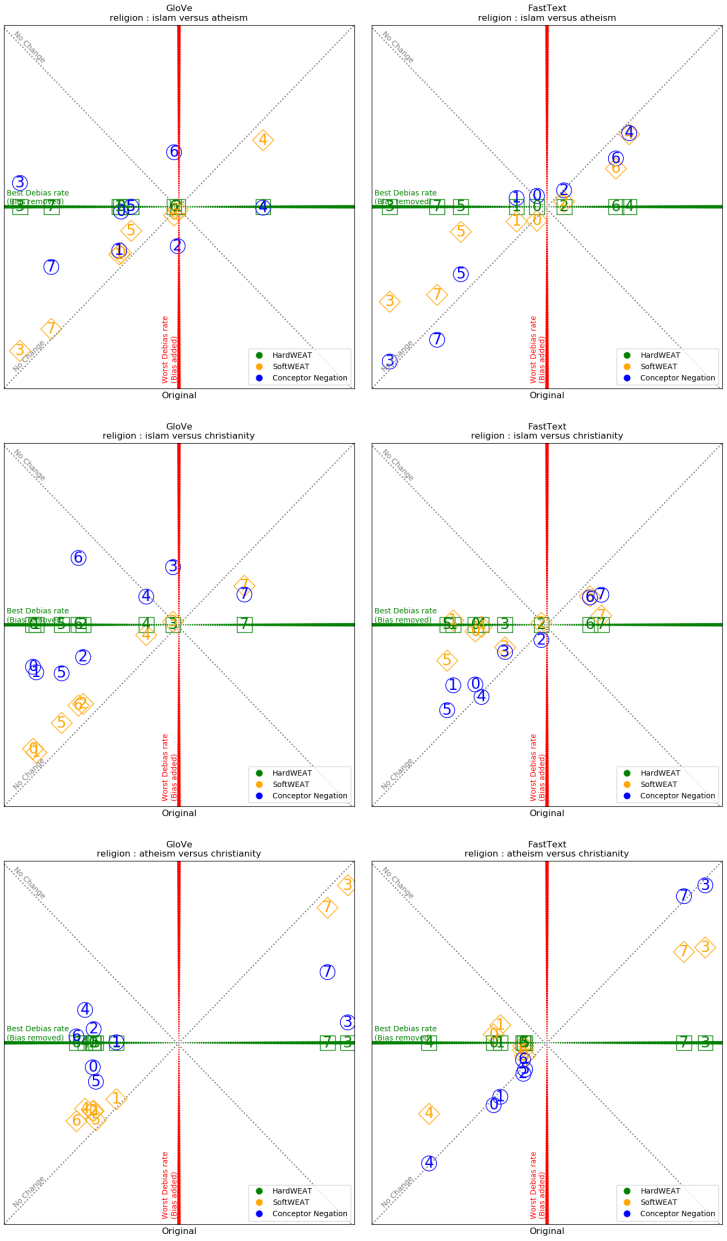


Figure 7: **Before and after debiasing for HardWEAT, SoftWEAT** ($\lambda = 0.2$ **and Conceptor** $\alpha = 2$)

E Sentiment Analysis

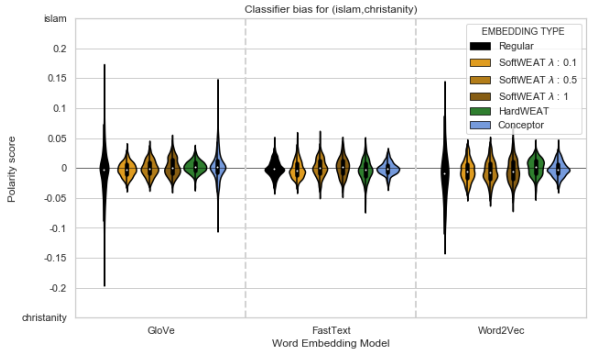


Figure 1: **Classifier Bias on (islam, christianity) words**

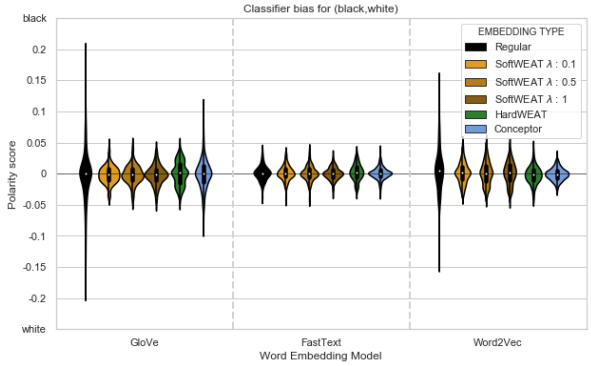


Figure 2: **Classifier Bias on (black, white) words:**

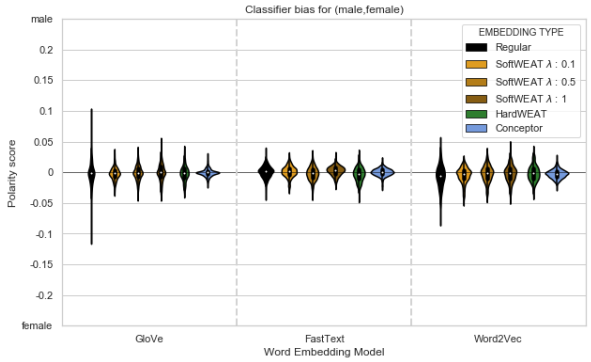


Figure 3: **Classifier Bias on (male, female) set of words**

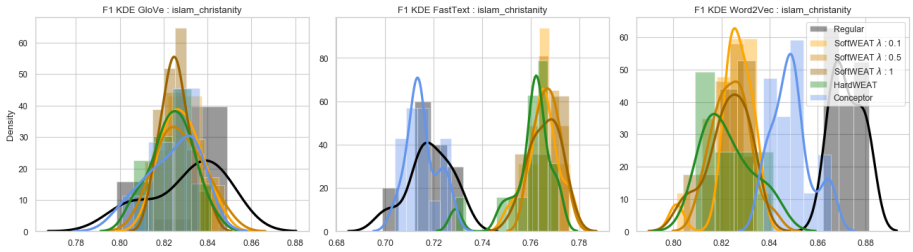


Figure 4: Kernel density estimate on F1 scores for (islam, christianity) modified embeddings.

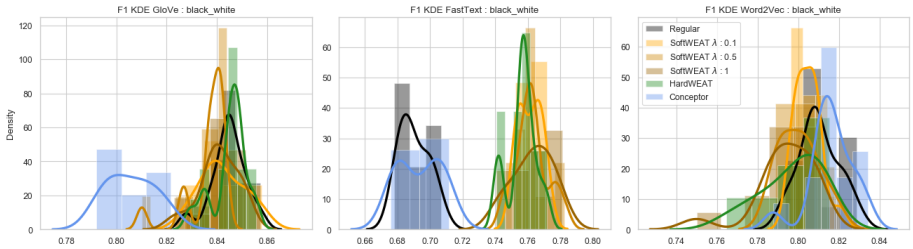


Figure 5: Kernel density estimate on F1 scores for (black, white) modified embeddings.

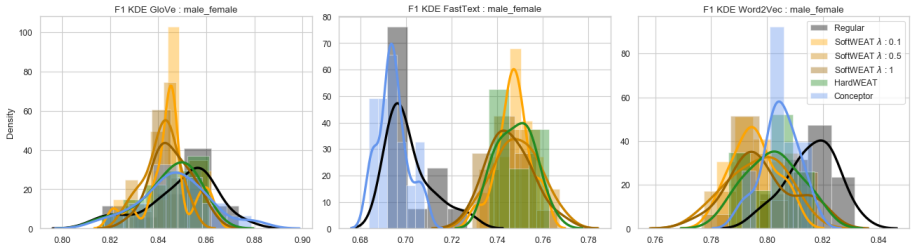


Figure 6: Kernel density estimate on F1 scores for (male, female) modified embeddings.

F Target attribute sets for debiasing within SoftWEAT.

GloVe : Male terms [science, intellectual, unpleasant]
Female terms [pleasant, appearance, art]
Black names [aggressive, unlikable, unpleasant, incompetent]
White names [competent, likable, pleasant, shy]
Islam terms [unlikable, art, weapons, family, incompetent, appearance, aggressive, unpleasant]
Atheism terms [unlikable, science, intellectual, weapons, career, aggressive, unpleasant]

Christianity terms [competent, pleasant, art, shy, family, instruments, likable, appearance]
Word2Vec : Male terms [science, intellectual, weapons, likable]
Female terms [instruments, unlikable, appearance, art]
Black names [family, appearance]
White names [intellectual, career]
Islam terms [weapons, family, appearance, art]
Atheism terms [science, career, instruments, intellectual]
Christianity terms [career, instruments, appearance, art]
Fasttext : Male terms [science, intellectual, unpleasant]
Female terms [pleasant, appearance, art]
Black names [instruments, unpleasant]
White names [weapons, pleasant]
Islam terms [unlikable, pleasant, art, weapons, family, incompetent, appearance]
Atheism terms [science, intellectual, instruments, career, unpleasant]
Christianity terms [competent, art, family, instruments, likable, appearance]

Number of changed words within SoftWEAT

Embedding \ Subclass	Male terms	Female terms	Black names	White names	Islam words	Atheism words	Christianity words
GloVe	50	47	156	244	80	146	102
Word2Vec	31	34	108	92	47	52	39
Fasttext	57	54	470	254	126	157	120

Table 4: Number of changed target set words within SoftWEAT.

G Sets

Target sets

male[1] = {'male', 'man', 'boy', 'brother', 'he', 'him', 'his', 'son', 'father', 'uncle', 'grandfather'}
female[1]= {'female', 'woman', 'girl', 'sister', 'she', 'her', 'hers', 'daughter', 'mother', 'aunt', 'grandmother'}
white[1] = {adam, chip, harry, josh, roger, alan, frank, ian, justin, ryan, andrew, fred, jack, matthew, stephen, brad, greg, jed, paul, todd, brandon, hank, jonathan, peter, wilbur, amanda, courtney, heather, melanie, sara, amber, crystal, katie, meredith, shannon, betsy, donna, kristin, nancy, stephanie, bobbiesue, ellen, lauren, peggy, sue-ellen, colleen, emily, megan, rachel, wendy, brendan, geoffrey, brett, jay, neil, anne, carrie, jill, laurie, kristen, sarah}
black[1] = {alonzo, jamel, lerone, percell, theo, alphonse, jerome, leroy, rasaan, torrance, darnell, lamar, lionel, rashaun, tyree, deion, lamont, malik, terrence, tyrone, everol, lavon, marcellus, terryl, wardell, aiesha, lashelle, nichelle, shereen, temeka, ebony, latisha, shaniqua, tameisha, teretha, jasmine, latonya, shanise, tanisha, tia, lakisha, latoya, sharise, tashika, yolanda, lashandra, malika, shavonn, tawanda, yvette, hakim, jermaine, kareem, jamal, rasheed, aisha, keisha, kenya, tamika}
christianity[2] = {baptism, messiah, catholicism, resurrection, christianity, salvation, protestant, gospel, trinity, jesus, christ, christian, cross, catholic, church}
islam[2] = {allah, ramadan, turban, emir, salaam, sunni, koran, imam, sultan, prophet, veil, ayatollah, shiite, mosque, islam, sheik, muslim, muhammad}
atheism[3] = {atheism, atheist, atheistic, heliocentric, evolution, darwin, galilei, agnostic, agnosticism, pagan, science, disbelief, scepticism, philosophy, university, kopernikus}

Attribute sets

science[1] = {math, algebra, geometry, calculus, equations, computation, numbers, addition, science, technology, physics, chemistry, einstein, nasa, experiment, astronomy}
art[1] = {poetry, art, dance, literature, novel, symphony, drama, sculpture, shakespeare}
intellectual[3] = {precocious, luminous, brilliant, stupid, logical, idiotic, trivial, analytical, thoughtful, reflective, astute, inquiring, venerable, dull, resourceful, ingenious, smart, discerning, sagacious, clever, wise, inventive, unintelligent, adaptable, sage, judicious, imaginative, intuitive, unwise, naive, dumb, apt, clumsy, foolish, shrewd, inquisitive, genius, intelligent}
appearance[2] = {alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong}
career[1] = {executive, management, professional, corporation, salary, office, business, career}

family[1] = {home, parents, children, family, cousins, marriage, wedding, relatives}
pleasant[1] = caress, freedom, health, love, peace, cheer, friend, heaven, loyal, pleasure, diamond, gentle, honest, lucky, rainbow, diploma, gift, honor, miracle, sunrise, family, happy, laughter, paradise, vacation, joy, wonderful}
unpleasant[1] = abuse, crash, filth, murder, sickness, accident, death, grief, poison, stink, assault, disaster, hatred, pollute, tragedy, divorce, jail, poverty, ugly, cancer, kill, rotten, vomit, agony, prison, terrible, horrible, nasty, evil, war, awful, failure}
shy[5] = {soft, quiet, compromising, rational, calm, kind, agreeable, servile, pleasant, cautious, friendly, supportive, nice, mild, demure, passive, indifferent, submissive}
aggressive[5] = {shrill, loud, argumentative, irrational, angry, abusive, obnoxious, controlling, nagging, brash, hostile, emasculating, mean, harsh, sassy, aggressive, opinionated, domineering}
shy[5] = {soft, quiet, compromising, rational, calm, kind, agreeable, servile, pleasant, cautious, friendly, supportive, nice, mild, demure, passive, indifferent, submissive}
aggressive[5] = {shrill, loud, argumentative, irrational, angry, abusive, obnoxious, controlling, nagging, brash, hostile, emasculating, mean, harsh, sassy, aggressive, opinionated, domineering}
competent[5] = {competent, productive, effective, ambitious, active, decisive, strong, tough, bold, assertive]}
incompetent[5] = {incompetent, unproductive, ineffective, unambitious, passive, indecisive, weak, gentle, timid, unassertive}
likable[5] = {agreeable, fair, honest, trustworthy, selfless, accommodating, likable, liked}
unlikable[5] = {abrasive, conniving, manipulative, dishonest, selfish, pushy, unlikable, unliked}
flowers[5] = {aster, clover, hyacinth, marigold, poppy, azalea, crocus, iris, orchid, rose, bluebell, daffodil, lilac, pansy, tulip, buttercup, daisy, lily, peony, violet, carnation, gladiola, magnolia, petunia, zinnia}
insects[5] = {ant, caterpillar, flea, locust, spider, bedbug, centipede, fly, maggot, tarantula, bee, cockroach, gnat, mosquito, termite, beetle, cricket, hornet, moth, wasp, blackfly, dragonfly, horsefly, roach, weevil}

Extended non-neutral words list for HardWEAT
male[7] = {countryman, fraternal, wizards, manservant, fathers, divo, actor, bachelor, papa, dukes, barman, countrymen, brideprice, hosts, airmen, andropause, penis, prince, governors, abbot, men, widower, gentlemen, sorcerers, sir, bridegrooms, baron, househusbands, gods, nephew, widowers, lord, brother, grooms, priest, adultors, andrology, bellboys, his, marquis, princes, emperors, stallion, chairman, monastery, priests, boyhood, fellas, king, dudes, daddies, manservant, semen, spokesman, tailor, cowboys, dude, bachelors, barbershop, emperor, daddy, masculism, guys, enchanter, guy, fatherhood, androgen, cameramen, godfather, strongman, god, patriarch, uncle, chairmen, sir, brotherhood,

host, testosterone, husband, dad, steward, males, cialis, spokesmen, pa, beau, stud, bachelor, wizard, sir, nephews, fathered, bull, beaus, councilmen, landlords, grandson, fiances, stepfathers, horsemen, grandfathers, adultor, schoolboy, rooster, grandsons, bachelor, cameraman, dads, him, master, lad, policeman, monk, actors, salesmen, boyfriend, councilman, fella, statesman, paternal, chap, landlord, brethren, lords, blokes, fraternity, bellboy, duke, ballet_dancer, dudes, fiance, colts, husbands, suitor, paternity, he, businessman, masseurs, hero, deer, busboys, boyfriends, kings, brothers, masters, stepfather, grooms, son, studs, cowboy, mentleman, sons, baritone, salesman, paramour, male_host, monks, menservants, mr., headmasters, lads, congressman, airman, househusband, priest, barmen, barons, abbots, handyman, beard, fraternities, stewards, colt, czar, stepsons, himself, boys, lions, gentleman, penis, his, masseur, bulls, uncles, bloke, beards, hubby, lion, sorcerer, macho, father, gays, male, waiters, sperm, prostate, stepson, prostatic_utricle, businessmen, heir, waiter, headmaster, man, governor, god, bridegroom, grandpa, groom, dude, gay, gents, boy, grandfather, gelding, paternity, roosters, prostatic_utricle, priests, manservants, stailor, busboy, heros}

female[7] = {countrywoman, sororal, witches, maidservant, mothers, diva, actress, spinster, mama, duchesses, barwoman, countrywomen, dowry, hostesses, airwomen, menopause, clitoris, princess, governesses, abbess, women, widow, ladies, sorceresses, madam, brides, baroness, housewives, godesses, niece, widows, lady, sister, brides, nun, adultresses, obstetrics, bellgirls, her, marchioness, princesses, empresses, mare, chairwoman, convent, priestesses, girlhood, ladies, queen, gals, mommies, maid, female_ejaculation, spokeswoman, seamstress, cowgirls, chick, spinsters, hair_salon, empress, mommy, feminism, gals, enchantress, gal, motherhood, estrogen, camerawomen, godmother, strongwoman, goddess, matriarch, aunt, chairwomen, ma'am, sisterhood, hostess, estradiol, wife, mom, stewardess, females, viagra, spokeswomen, ma, belle, minx, maiden, witch, miss, nieces, mothered, cow, belles, councilwomen, landladies, granddaughter, fiancees, stepmothers, horsewomen, grandmothers, adultress, schoolgirl, hen, granddaughters, bachelorette, camerawoman, moms, her, mistress, lass, policewoman, nun, actresses, saleswomen, girlfriend, councilwoman, lady, stateswoman, maternal, lass, landlady, sistren, ladies, wenches, sorority, bellgirl, duchess, ballerina, chicks, fiancee, fillies, wives, suitress, maternity, she, businesswoman, masseuses, heroine, doe, busgirls, girlfriends, queens, sisters, mistresses, stepmother, brides, daughter, minxes, cowgirl, lady, daughters, mezzo, saleswoman, mistress, hostess, nuns, maids, mrs.,
headmistresses, lasses, congresswoman, airwoman, housewife, priestess, barwomen, barnoesses, abbesses, handywoman, toque, sororities, stewardesses, filly, czarina, stepdaughters, herself, girls, lionesses, lady, vagina, hers, masseuse, cows, aunts, wench, toques, wife, lioness, sorceress, effeminate, mother, lesbians, female, waitresses, ovum, skene_gland, stepdaughter, womb, businesswomen, heiress, waitress, headmistress, woman, governess, godess, bride, grandma, bride, gal, lesbian, ladies, girl, grandmother, mare, maternity, hens, uterus, nuns, maidservants, seamstress', busgirl, heroines}

gender[6] = {cowboy, cowgirl, cowboys, cowgirls, camerawomen, cameramen,

cameraman, camerawoman, busboy, busgirl, busboys, busgirls, bellboy, bellgirl, bellboys, bellgirls, barman, barwoman, barmen, barwomen, tailor, seamstress, tailors, seamstress', prince, princess, princes, princesses, governor, governess, governors, governesses, adultor, adultress, adultors, adultresses, god, godess, gods, godesses, host, hostess, hosts, hostesses, abbot, abbess, abbots, abbesses, actor, actress, actors, actresses, bachelor, spinster, bachelors, spinsters, baron, baroness, barons, barnoesses, beau, belle, beaus, belles, bridegroom, bride, bridegrooms, brides, brother, sister, brothers, sisters, duke, duchess, dukes, duchesses, emperor, empress, emperors, empresses, enchanter, enchantress, father, mother, fathers, mothers, fiance, fiancee, fiances, fiancees, priest, nun, priests, nuns, gentleman, lady, gentlemen, ladies, grandfather, grandmother, grandfathers, grandmothers, headmaster, headmistress, headmasters, headmistresses, hero, heroine, heros, heroines, lad, lass, lads, lasses, landlord, landlady, landlords, landladies, male, female, males, females, man, woman, men, women, manservant, maidservant, manservants, maidservants, marquis, marchioness, masseur, masseuse, masseurs, masseuses, master, mistress, masters, mistresses, monk, nun, monks, nuns, nephew, niece, nephews, nieces, priest, priestess, priests, priestesses, sorcerer, sorceress, sorcerers, sorceresses, stepfather, stepmother, stepfathers, stepmothers, stepson, stepdaughter, stepsons, stepdaughters, steward, stewardess, stewards, stewardesses, uncle, aunt, uncles, aunts, waiter, waitress, waiters, waitresses, widower, widow, widowers, widows, wizard, witch, wizards, witches}

religion[4] = {synagogue, synagogues, altar, altars, parish, parishes, biblical, bishop, bishops, jihadist, clergy, bible, bibles, mosque, mosques, mullah, church, churches, sermon, sermons, papacy, imam, pew, chancel, pope, priest, priests, baptism, jihad, confessional, holy_eucharist, evangelical, jesus, burqa, vicar, vicars, judaism, christianity, islam, jew, christian, muslim, torah, quran, rabbi}

References

[1] Caliskan, A., Bryson, J.J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science (2017)

[2] Garg, N., Schiebinger, L., Jurafsky, D., Zou, J.: Word embeddings quantify 100 years of gender and ethnic stereotypes. Proceedings of the National Academy of Sciences (2018)

[3] Knoche, M., Popović, R., Lemmerich, F., Strohmaier, M.: Identifying biases in politically biased wikis through word embeddings. In: Proceedings of the 30th ACM Conference on Hypertext and Social Media. p. 253–257 (2019)

[4] Manzini, T., Lim, Y.C., Tsvetkov, Y., Black, A.W.: Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. CoRR **abs/1904.04047** (2019), `http://arxiv.org/abs/1904.04047`

[5] May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: On measuring social biases in sentence encoders. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 622–628. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), `https://www.aclweb.org/anthology/N19-1063`

[6] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.: Gender bias in coreference resolution: Evaluation and debiasing methods. CoRR **abs/1804.06876** (2018), `http://arxiv.org/abs/1804.06876`

[7] Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.: Learning gender-neutral word embeddings. CoRR **abs/1809.01496** (2018), `http://arxiv.org/abs/1809.01496`