Report on Fairness

|  |  |  |  |
| --- | --- | --- | --- |
| **Corpus** | **Target Words** | **Attributes** | **Effect Size** |
| Twitter -25d |  |  | 1.16 |
| Twitter -50d | names\_europe vs names\_africa |  | 1.11 |
| Twitter -100d |  |  | 1.18 |
| Twitter -200d |  |  | 1.17 |
| Twitter -25d |  |  | 1.53 |
| Twitter -50d | names\_female vs names\_male |  | 1.17 |
| Twitter -100d |  |  | 1.26 |
| Twitter -200d |  | Pleasant vs unpleasant | 1.05 |
| Twitter -25d |  |  | 1.25 |
| Twitter -50d | Flowers vs insects |  | 1.23 |
| Twitter -100d |  |  | 1.12 |
| Twitter -200d |  |  | 1.10 |
| Twitter -25d |  |  | -1.22 |
| Twitter -50d | Science vs art |  | -1.14 |
| Twitter -100d |  |  | -0.84 |
| Twitter -200d |  |  | -0.30 |
| Twitter -25d |  |  | 1.32 |
| Twitter-50d | names\_female vs names\_male |  | 1.26 |
| Twitter-100d |  |  | 1.28 |
| Twitter-200d |  |  | 1.21 |
| Twitter -25d |  |  | 0.70 |
| Twitter -50d | names\_europe vs names\_africa |  | 0.64 |
| Twitter -100d |  | Family vs career | 0.70 |
| Twitter -200d |  |  | 0.80 |
| Twitter-25d |  |  | 0.53 |
| Twitter-50d | gender\_f vs gender\_m |  | 0.53 |
| Twitter-100d |  |  | 0.47 |
| Twitter-200d |  |  | 0.40 |
| Twitter-25d |  |  | -0.25 |
| Twitter-50d | science vs arts | gender\_f vs gender\_m | 0.68 |
| Twitter-100d |  |  | 0.05 |
| Twitter-200d |  |  | -0.44 |
| Wikipedia-50d |  |  | 1.14 |
| Wikipedia-100d | names\_europe vs names\_africa |  | 1.14 |
| Wikipedia-200d |  |  | 1.22 |
| Wikipedia-300d |  |  | 1.21 |
| Wikipedia-50d |  |  | -0.38 |
| Wikipedia-100d | names\_female vs names\_male |  | 0.48 |
| Wikipedia-200d |  |  | -0.24 |
| Wikipedia-300d |  |  | -0.38 |
| Wikipedia-50d |  | Pleasant vs unpleasant | 1.08 |
| Wikipedia-100d | Flowers vs insects |  | 1.18 |
| Wikipedia-200d |  |  | 1.15 |
| Wikipedia-300d |  |  | 1.41 |
| Wikipedia-50d |  |  | -1.36 |
| Wikipedia-100d | Science vs art |  | -1.17 |
| Wikipedia-200d |  |  | -1.13 |
| Wikipedia-300d |  |  | -1.05 |
| Wikipedia-50d |  |  | 0.15 |
| Wikipedia-100d | names\_europe vs names\_africa |  | -0.08 |
| Wikipedia-200d |  |  | 0.11 |
| Wikipedia-300d |  |  | 0.05 |
| Wikipedia-50d |  |  | 1.76 |
| Wikipedia-100d | names\_female vs names\_male |  | 1.77 |
| Wikipedia-200d |  | Family vs career | 1.80 |
| Wikipedia-300d |  |  | 1.75 |
| Wikipedia-50d |  |  | 0.98 |
| Wikipedia-100d | gender\_f vs gender\_m |  | 0.83 |
| Wikipedia-200d |  |  | 0.79 |
| Wikipedia-300d |  |  | 0.71 |
| Wikipedia-50d |  |  | -1.56 |
| Wikipedia-100d | Science vs art | gender\_f vs gender\_m | -1.10 |
| Wikipedia-200d |  |  | -1.34 |
| Wikipedia-300d |  |  | -1.36 |
| Wikipedia-100d |  |  | -1.17 |
| Twitter-100d | gender\_f vs gender\_m | sports vs fashion | -0.93 |
| Wikipedia-100d |  |  | 1.32 |
| Twitter-100d | musical-instruments vs weapons | pleasant vs unpleasant | 1.04 |

**Part-1:** From the above table it is evident that both twitter and Wikipedia have significant biases in the trained word embedding models. Incase of female and male names vs pleasant and unpleasant attributes twitter model showed greater bias (1.53) compared to Wikipedia (-0.38). Also, for European and African names vs family and career twitter model has more bias (0.8) compared to Wikipedia ( -0.08). For the rest of the mappings the bias value is almost similar/close for twitter and Wikipedia. For twitter model, with attributes pleasant and unpleasant, female and male names had the highest bias (1.53) and for family vs career attributes, European vs African names has the highest bias (0.8). For the Wikipedia model, with attributes family vs career, female and male names has the highest bias (1.8). Wikipedia model also shows highest bias based on gender for science vs art target pair (-1.56).

**Part-2:**

Additional list of words used:

1. Sports: cricket, basketball, soccer, tennis, football, baseball, volleyball, swimming, racing, chess, archery, table tennis, badminton, gymnastics.
2. Fashion: footwear, makeup, clothing, chains, bracelets, rings, spectacles, belts.
3. Musical-instruments: piano, guitar, saxophone, ukulele, violin, guitar, trumpet, cello, drum, clarinet, accordion, xylophone, keyboard, flute, oboe, harmonica, trombone, harp, mandolin, bassoon, tuba, French horn, double bass, bagpipes, banjo, tambourine, lute, marimba, theremin, lyre, cajon, harpsichord, organ, oud, didgeridoo, ocarina, bell, gong, melodica, sitar.
4. Weapons: Gun, Cannon, Mortar, bola, bow and arrow, crossbow, longbow, grapeshot, flamethrower, sling, spear, bayonet, club, dagger, halberd, lance, pike, sword, tomahawk, grenade, shrapnel, TNT, missile, musket, pistol, revolver, machine gun, atomic bomb, neutron bomb, catapult, culverin.

When running the pairings female vs male with sports vs fashion attributes both Wikipedia and twitter models showed effect sizes -1.17 and 0.93 which is significant enough to be considered as bias based on gender. Also, for musical-instruments vs weapons as target pair and pleasant vs unpleasant attributes the models showed corresponding effect sizes of 1.32 and 1.04 respectively which is a potential bias.