A Proofs for method from Bolukbasi et al. (2016)

We found the equations suggested in Bolukbasi et al. (2016) on the opaque side of things. So we provide here proofs missing from the original work ourselves.

Proposition 1. *The neutralise step of Bolukbasi et al. (2016) yields a unit vector. Specifically, Bolukbasi et al. (2016) define*

$$\vec{w} = \nu + \sqrt{1 + ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||^2} \qquad (2)$$

Proof. We want to prove that $||\vec{w}||_2 = 1$

$$
\begin{aligned}
||\vec{w}||_2^2 &= \vec{w}^\top \vec{w} \\
&= \left(\nu + \sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right)^\top \\
&\quad \times \left(\nu + \sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right) \\
&= ||\nu||_2^2 + 2\nu^\top \left(\sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right) \\
&\quad + \left(\sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right)^\top \\
&\quad \times \left(\sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right) \\
&= ||\nu||_2^2 + 2\nu^\top \left(\sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right) \\
&\quad + 1 - ||\nu||_2^2 \\
&= 2\nu^\top \left(\sqrt{1 - ||\nu||_2^2} \frac{\vec{w}_B - \mu_B}{||\vec{w}_B - \mu_B||_2} \right) + 1 \\
&= 1
\end{aligned}
$$

where we note that $\nu = \mu - \mu_{\perp B}$ so it is orthogonal to both $\vec{w}_B$ and $\vec{\mu}_B$. □

Proposition 2. *The equalise step of Bolukbasi et al. (2016) ensures that gendered pairs, e.g. man–woman, are equidistant to all gender-neutral words.*

Proof. The normalized vectors for gendered words are orthogonal to those gender-neutral words by construction. Thus, the distance in both cases is simply ν. □

B WEAT word sets

Below are listed the word sets we used for the WEAT to test direct bias, as defined by Nosek et al. (2002). Note that for the careers–family test, the target and attribute words have been reversed; that is, gender is captured by the target words, rather than the attribute words. Whilst this distinction is important in the source psychological literature (Greenwald et al., 1998), mathematically the target sets and attribute sets are indistinguishable and fully commutative.

Art–Maths Target_X: *math, algebra, geometry, calculus, equations, computation, numbers, addition*; Target_Y: *poetry, art, dance, literature, novel, symphony, drama, sculpture*; Attribute_A: *male, man, boy, brother, he, him, his, son*; Attribute_B: *female, woman, girl, sister, she, her, hers, daughter*

Arts–Sciences Target_X: *science, technology, physics, chemistry, Einstein, NASA, experiment, astronomy*; Target_Y: *poetry, art, Shakespeare, dance, literature, novel, symphony, drama*; Attribute_A: *brother, father, uncle, grandfather, son, he, his, him*; Attribute_B: *sister, mother, aunt, grandmother, daughter, she, hers, her*

Careers–Family Target_X: *John, Paul, Mike, Kevin, Steve, Greg, Jeff, Bill*; Target_Y: *Amy, Joan, Lisa, Sarah, Diana, Kate, Ann, Donna*; Attribute_A: *executive, management, professional, corporation, salary, office, business, career*; Attribute_B: *home, parents, children, family, cousins, marriage, wedding, relatives*

C Additional Gigaword results

Additional results for the Annotated English Gigaword are given here.
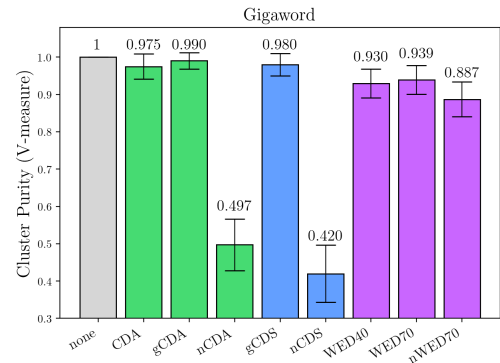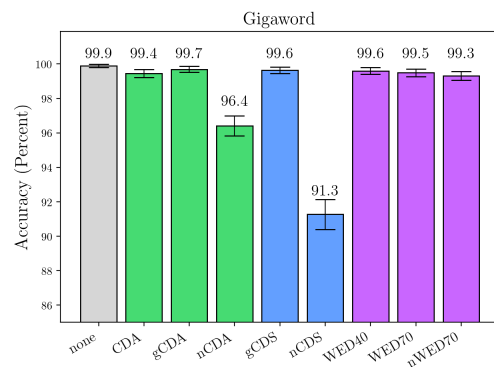

Figure 10: Most biased cluster purity results

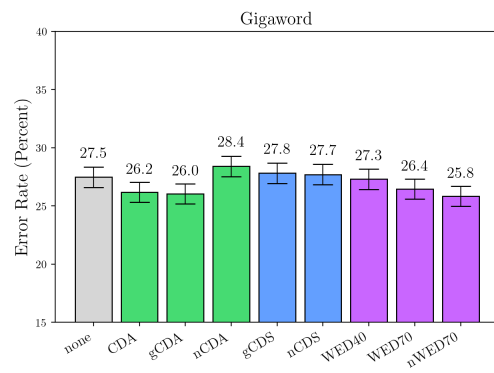Figure 11: Reclassification of most biased words results
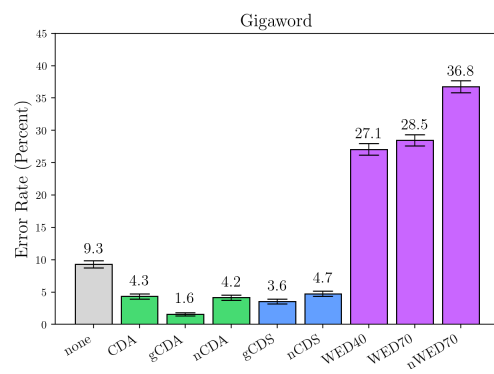

Figure 12: Sentiment classification results

Figure 13: Non-biased gender analogy results