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. *Bolukbasi et al. (2016) define*

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

where they define $\nu = \mu - \mu_B$. This vector is a unit vector, i.e. $||\vec{w}||_2 = 1$.

Proof.

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

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

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. Following Bolukbasi et al., we define $\vec{e}$ and $\vec{w}$ as follows:

$$\begin{aligned}
\vec{e} &:= \frac{\vec{e} - \vec{e}_B}{||\vec{e} - \vec{e}_B||} = \frac{\vec{e}_{\perp B}}{||\vec{e}_{\perp B}||} \\
\vec{w} &:= \nu + \sqrt{1 - ||\nu||^2} \frac{\vec{w}_B - \vec{\mu}_B}{||\vec{w}_B - \vec{\mu}_B||}
\end{aligned}$$

Now, we have the result that

$$\vec{e} \cdot \vec{w} = \vec{e} \cdot \nu \tag{3}$$

which is the same for any $\vec{e}$. Now, we may compute the distance between $\vec{w}$ and any vector $\vec{e}$ as

$$\begin{aligned}
||\vec{e} - \vec{w}||^2 &= (\vec{e} - \vec{w}) \cdot (\vec{e} - \vec{w}) \\
&= \vec{e} \cdot \vec{e} - 2\vec{e} \cdot \vec{w} + \vec{w} \cdot \vec{w} \\
&= 2 - 2\vec{e} \cdot \nu
\end{aligned}$$

□

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.

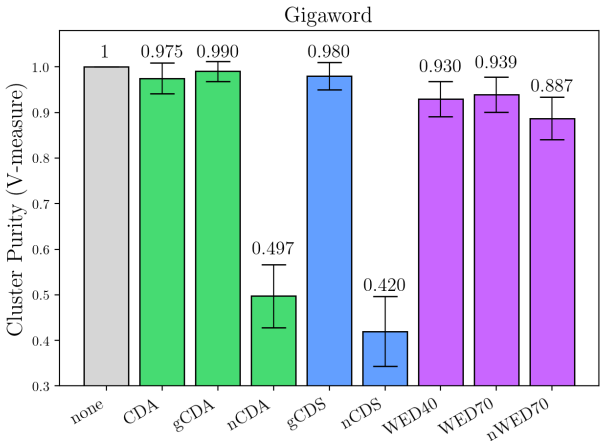

Figure 10: Most biased cluster purity results

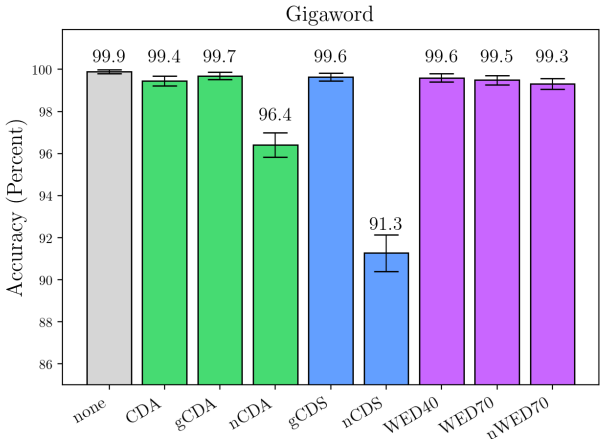

Figure 11: Reclassification of most biased words results

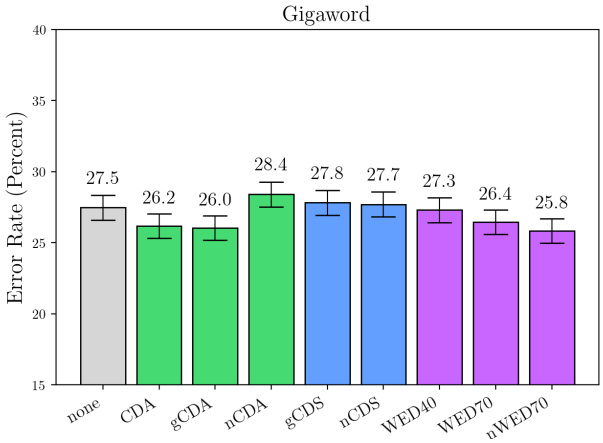

Figure 12: Sentiment classification results

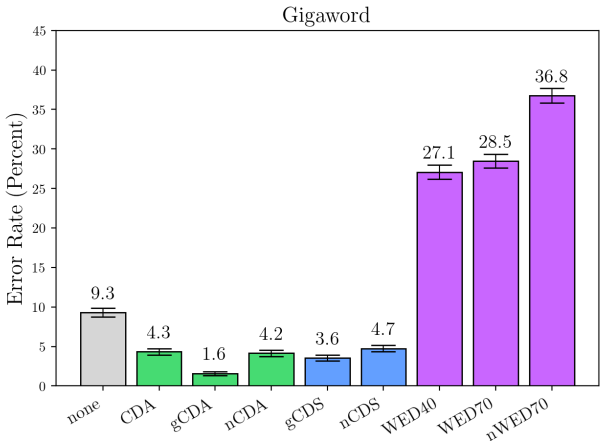

Figure 13: Non-biased gender analogy results