Bias in Word Embeddings

Ankit Pant 2018201035
Tarun Mohandas 2018201008
Team: *The Lost Linguists*

Outline

Introduction

Introduction

- Word Embeddings (WE) - extensively used in NLP
- WE inherently contain various types of biases
- NLP models tend to amplify the biases
- Hence removing bias from WE increasingly crucial
- Project attempts to use existing debiasing techniques
- Shortcomings are highlighted and alternative proposed

Literature Review

Word Embeddings

- Word Embeddings
  - Most popular representation of document vocabulary
  - Captures context of words
  - Various models include *Word2Vec* and *GLoVe*
  - Both use neural network to form word representations

Types of Biases in Datasets

- Historical Bias:
  - Unwanted biases that were present in society years ago
- Representational Bias:
  - Certain parts of the input space are under-represented
- Measurement Bias:
  - Imperfect measuring of the data
  - assuming that the measured data is proxy of some other desired feature
- Aggregation Bias:
  - Same model is used for groups with different conditional distributions
- Evaluation Bias:
  - Evaluation and benchmark data for model does not represent actual target population

Social Bias in Datasets

- Gender Bias:
  - Biased due to associating stereotype to gender
  - Man is to computer programmer as woman is to homemaker
- Racial Bias:
  - Biased due to associating stereotype to race
  - Modern is to American as medieval is to Indian
- Religious Bias:
  - Biased due to associating stereotype to religion
  - Smart is to Christian as cheapskate is to Jew

Existing Debiasing Methods

- Post-processing debiasing (Bolukbasi et al.):
  - Make change to word vector to reduce encoded gender bias
  - Done by zeroing the gender projection of each word on a predefined gender direction
  - $\vec{w} = \frac{(\vec{w} - \vec{w_b})}{\left\|(\vec{w} - \vec{w_b})\right\|}$
  - Ensure all neutral words are equally close to the two words
- Train word embeddings from scratch (Zhao et al.)
  - Alter the loss of GloVe model
  - Concentrate most gender information to last coordinate of each vector
  - Use word representation excluding the gender coordinate
  - Representation of gender neutral words orthogonal to gender direction

Experimental Setup

Experimental Setup

- Various kinds of biases were identified in word embeddings
- Graphically projected to check clustering of biased words
- Applying debaising techniques
- Check result of debiasing quantitatively and graphically

Experiments and Results

Identifying Bias

- **Gender Bias**
  - manager is to man as mastercard is to woman
  - programmer is to man as nutritionist is to woman
- **Racial Bias**
  - filthy is to black as elitists is to white
  - manager is to American as barber is to Indian
- **Religious Bias**
  - rich is to Christian as homeless is to Muslim
  - smart is to Christian as cheapskate is to Jew

Most biased words - Gender i

Bias in Professions- descending order (top 10) (man-woman)

('magician', 0.11574505)
('carpenter', 0.046477903)
('butcher', 0.035740647)
('gamer', 0.027397368)
('soldier', 0.018920997)
('servant', 0.00930707)
('barber', 0.007335724)
('engineer', -0.0022402927)
('player', -0.01447518)
('programmer', -0.016492786)

Most biased words - Gender ii

Bias in Misc. words - descending order (top 10) (man-woman)

('swear', 0.21706516)
('filthy', 0.20151067)
('sweet', 0.19195326)
('roar', 0.18197575)
('weep', 0.18006238)
('pretty', 0.17052722)
('beautiful', 0.15033276)
('think', 0.14811862)
('brave', 0.14424863)
('savage', 0.14332578)

Most biased words - Race i

Bias in Professions- descending order (top 10) (Indian-American)

('butcher', 0.30157873)
('barber', 0.20129806)
('florist', 0.18413721)
('nurse', 0.15209064)
('vet', 0.14899719)
('cashier', 0.1279751)
('tutor', 0.12523493)
('waiter', 0.11155586)
('chemist', 0.108406395)
('shopkeeper', 0.105821185)

Most biased words - Race ii

Bias in Misc. words - descending order (top 10) (Indian-American)

('lovely', 0.2813881)
('backpack', 0.25974354)
('bag', 0.20628145)
('cute', 0.19745344)
('purse', 0.19460957)
('sweet', 0.18077426)
('beautiful', 0.1678942)
('grumpy', 0.14973086)
('cool', 0.14963488)
('sassy', 0.14638355)

Most biased words - Religion i

Bias in Professions- descending order (top 10) (Muslim-Christian)

('vet', 0.12283135)
('cop', 0.11075719)
('soldier', 0.076730676)
('shopkeeper', 0.060545065)
('worker', 0.018872034)
('cashier', 0.016069816)
('butcher', -0.005377178)
('landlord', -0.00816943)
('nurse', -0.010036133)
('nanny', -0.011244484)

Most biased words - Religion ii

Bias in Misc. words - descending order (top 10) (Muslim-Christian)

('terrorist', 0.2741907)
('car', 0.10418006)
('fear', 0.10198632)
('bag', 0.07560906)
('creep', 0.06417281)
('brute', 0.057166893)
('yell', 0.041089058)
('poor', 0.04030589)
('roar', 0.039248332)
('mad', 0.038921878)

Clustering on Profession

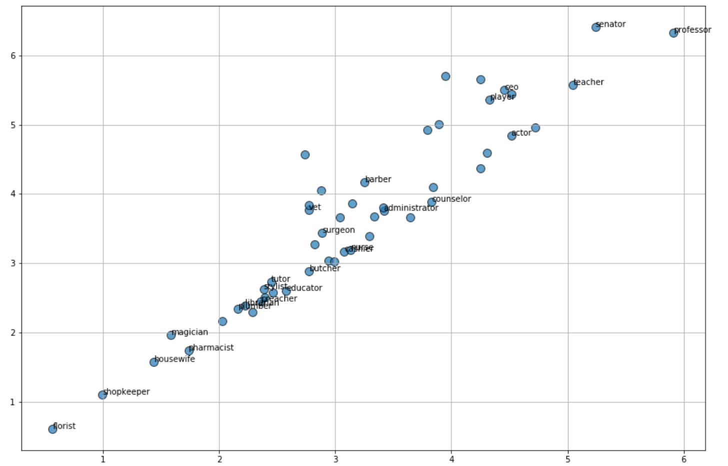

Figure 1: Result of clustering on male and female biased words

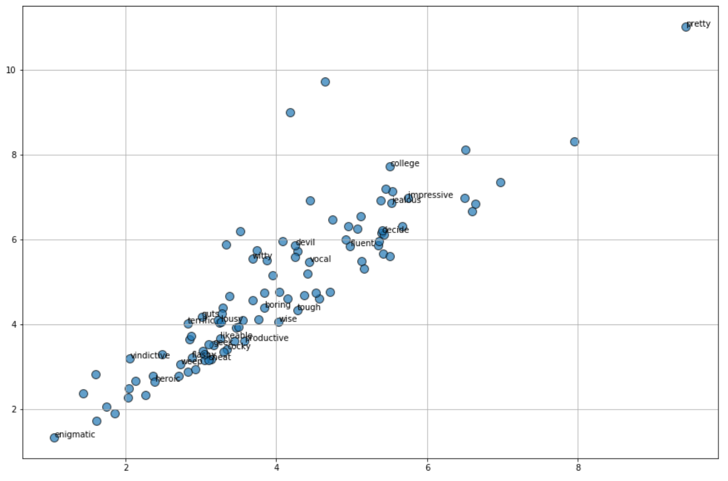

Figure 2: Result of clustering on male and female biased words

Debiasing Word Embeddings

Debiasing using neutralisation of hard-debiasing

Debiasing Professions based on Gender – descending order

('magician', 0.11574504) ⇒ ('magician', 0.025625029836058758)

('carpenter', 0.046477906) ⇒ ('carpenter', 0.0213426987806078)

('gamer', 0.027397364) ⇒ ('gamer', 0.0047532092934476355)

('soldier', 0.018920997) ⇒ ('soldier', -0.0024024025014908745)

('servant', 0.009307076) ⇒ ('servant', -0.0011262382124894582)

('barber', 0.0073357197) ⇒ ('barber', 0.00432715694605746)

('engineer', -0.0022402948) ⇒ ('engineer', -0.010558215937537)

('programmer', -0.016492) ⇒ ('programmer', -0.0184568722608)

Clustering on Profession after Debiasing

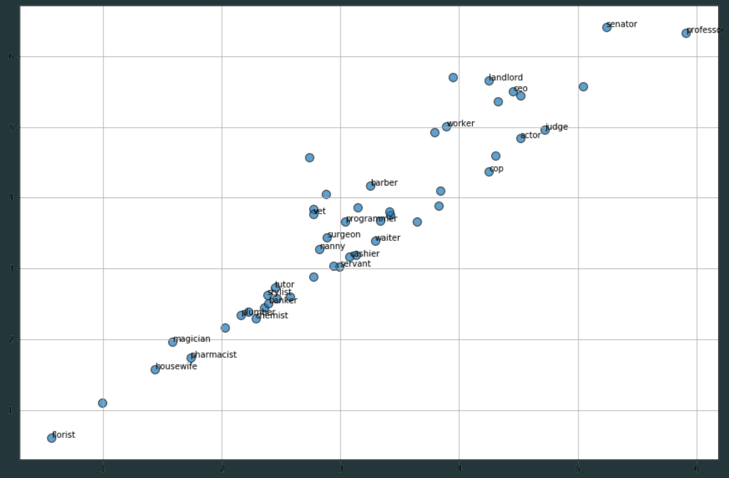

Figure 3: Result of clustering on male and female biased words

Debiasing Professions based on Race – descending order

('servant', -0.014473432) $\Rightarrow$ ('servant', -0.01782729110568683)

('lawyer', -0.014795305) $\Rightarrow$ ('lawyer', -0.019276932830898832)

('banker', -0.024144737) $\Rightarrow$ ('banker', -0.03891313915416421)

('engineer', -0.039301794) $\Rightarrow$ ('engineer', -0.06321906512113795)

('programmer', -0.06562963) $\Rightarrow$ ('programmer', -0.117997700482)

('senator', -0.06934556) $\Rightarrow$ ('senator', -0.12828888982156872)

('business', -0.08860369) $\Rightarrow$ ('business', -0.15978598382964143)

('scientist', -0.09481263) $\Rightarrow$ ('scientist', -0.16940044153193437)

('worker', -0.11409591) $\Rightarrow$ ('worker', -0.19914736635361133)

('ceo', -0.11503146) $\Rightarrow$ ('ceo', -0.1961109605431922)

Clustering on Profession after Debiasing

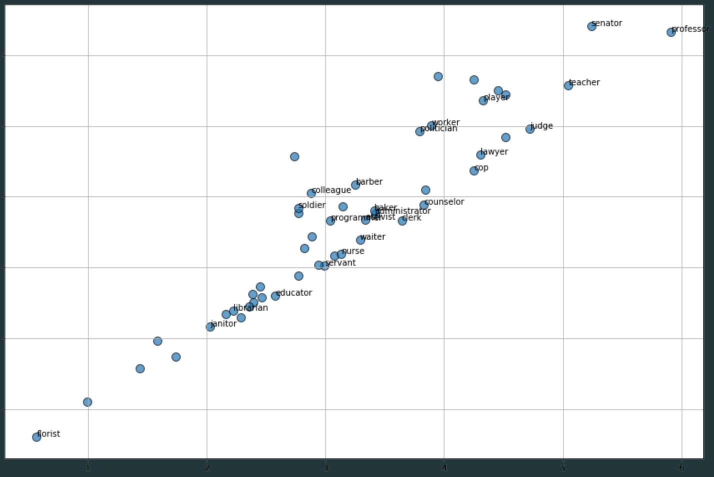

Figure 4: Result of clustering on race biased words

Debiasing Professions based on Religion – descending order

('butcher', -0.0053771744) ⇒ ('butcher', -0.005928319604247454)

('landlord', -0.008169426) ⇒ ('landlord', -0.013176119539724127)

('nurse', -0.010036137) ⇒ ('nurse', -0.014683282999863613)

('nanny', -0.011244472) ⇒ ('nanny', -0.0167685528948944)

('programmer', -0.06562963) ⇒ ('programmer', -0.117997700482)

('pharmacist', -0.015825728) ⇒ ('pharmacist', -0.0209037176323)

('clerk', -0.018447457) ⇒ ('clerk', -0.0534647674024996)

('scientist', -0.09481263) ⇒ ('scientist', -0.16940044153193437)

('business', -0.027096914) ⇒ ('business', -0.0641848280060503 3)

('player', -0.034285568) ⇒ ('player', -0.0754654672619229)

Clustering on Profession after Debiasing

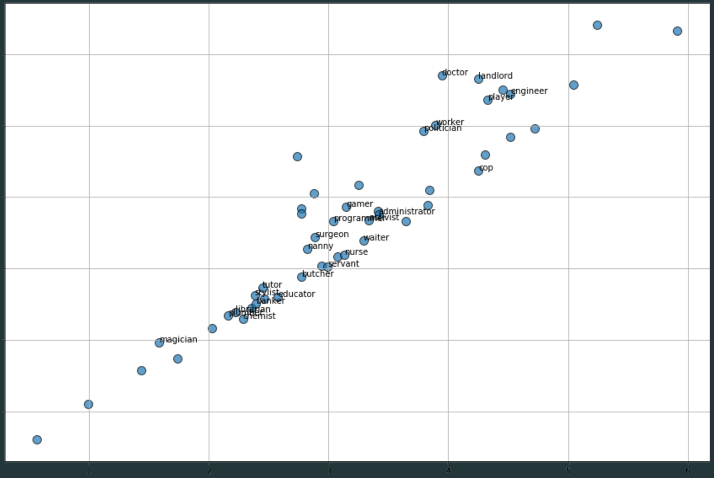

Figure 5: Result of clustering on religion biased words

Debiasing Misc. Words based on Gender – descending order

('swear', 0.21706519) $\Rightarrow$ ('swear', 0.04715369477049303)

('filthy', 0.20151068) $\Rightarrow$ ('filthy', 0.026021153148268236)

Debiasing Misc. Words based on Race – descending order

('comical', -0.0012468412) $\Rightarrow$ ('comical', -0.007484548413770923)

('maths', -0.015644163) $\Rightarrow$ ('maths', -0.027991168738123316)

Debiasing Misc. Words based on Religion – descending order

('guts', 0.0026711777) $\Rightarrow$ ('guts', 0.0017492761220381202)

('backpack', -0.0016981252) $\Rightarrow$ ('backpack', -0.01208588318522)

Clustering on Profession after Debiasing

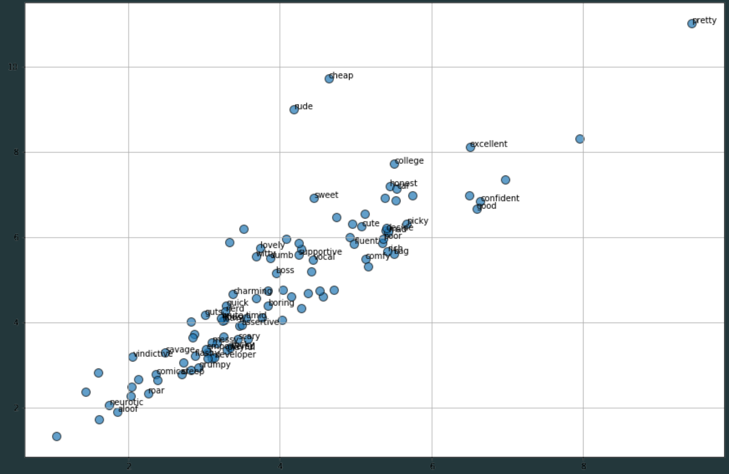

Figure 6: Result of clustering on gender biased words

Clustering on Profession after Debiasing

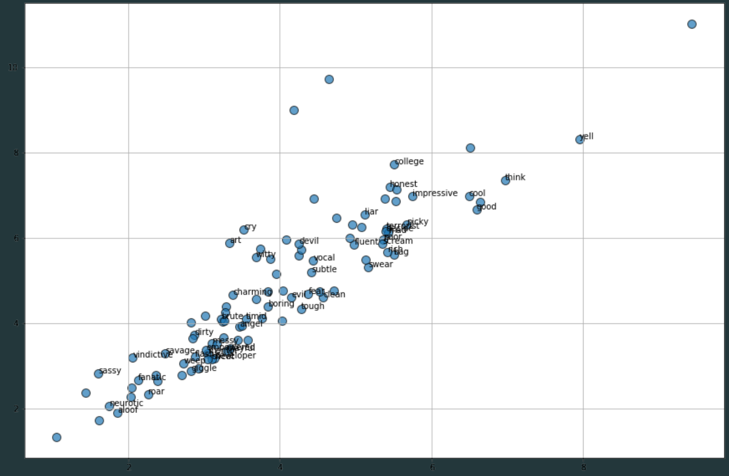

Figure 7: Result of clustering on race biased words

Clustering on Profession after Debiasing

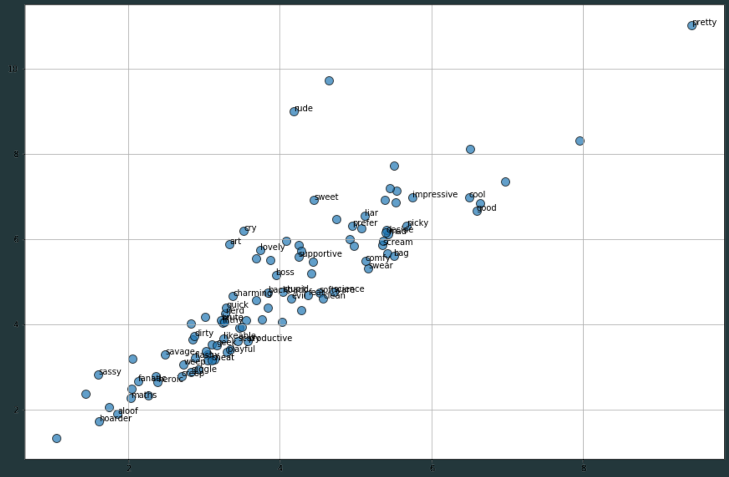

Figure 8: Result of clustering on religion biased words

Proposed Algorithm

Proposed algorithm

- Pre-processing dataset –¿ add sentences by inverting the gender, races, religion, etc. –¿ appending them to create a new dataset.
- Debiasing during pre-processing as with (Zhao et al.)
- Analyse ‘biased’ words –¿ projecting them to biased sub-space and debias (Bolukbasi et al.)
- After identifying these words –¿ correlated with each other –¿ distances between the words clustering together should also be equalised.

Conclusion

Conclusion

- Project involved exploring various biases in word embeddings
- Word embeddings trained on the Reddit dataset were explored
- Attempt to debias using the traditional hard-debiasing method was done
- Debiasing happened on a superficial level
- Approach proposed that may address the issue

Future Scope

Future Scope

- Implement proposed algorithm
- Empirically test performance of proposed model
- Explore how biases are encoded in word embeddings

References

References i

- Hila Gonen, Yoav Goldberg, *Lipstick on a Pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them*, https://arxiv.org/pdf/1903.03862.pdf
- Tolga Bolukbasi, et al., *Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings*, https://arxiv.org/pdf/1607.06520.pdf
- Jieyu Zhao, et al., *Learning Gender-Neutral Word Embeddings*, https://arxiv.org/pdf/1809.01496.pdf
- Sevtap Duman, et al., *(Visualization of) gender bias in word embeddings*, http://wordbias.umiacs.umd.edu/

- Aylin Caliskan, et al., *Semantics derived automatically from language corpora contain human-like biases*, https://arxiv.org/pdf/1608.07187.pdf
- Nikhil Garg, et al., *Word embeddings quantify 100 years of gender and ethnic stereotypes*, https://arxiv.org/pdf/1711.08412.pdf
- Thomas Manzini, et al., *Black is to Criminal as Caucasian is to Police:Detecting and Removing Multiclass Bias in Word Embeddings*, https://arxiv.org/pdf/1904.04047v1.pdf

- Harini Suresh, John V. Guttag, *A Framework for Understanding Unintended Consequences of Machine Learning*, https://arxiv.org/pdf/1901.10002.pdf
- Alex Fefegha, *Racial Bias and Gender Bias Examples in AI systems*, https://peopleofcolorintech.com/articles/racial-bias-and-gender-bias-examples-in-ai-systems/
- Marianne Bertrand, Sendhil Mullainathan, *Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on LaborMarket Discrimination*, https://www2.econ.iastate.edu/classes/econ321/Orazem/bertrand_emily.pdf

References iv

- Thomas Manzini, et al., *Debiasing Multiclass Word Embeddings*, https://github.com/TManzini/DebiasMulticlassWordEmbedding
- Ella Rabinovich, Shuly Wintner, *The Reddit-L2 corpus*, http://cl.haifa.ac.il/projects/L2/