Bias in Language Models

G#: G01348097 **Name:** Ramaswamy Iyappan

Objective:

Explore the bias in natural language word embeddings by testing associations with pairs of words and varying the underlying training corpus used to learn the word embeddings.

Technical Findings:

Performing WEAT using different Target and attribute pairs by varying the training corpus (Wikipedia/Twitter) used to learn the word embeddings:

Targets	**Attributes**	**Training corpus**	**Effect size**	**Tar => atrribute1**
Names_male, names_female	Pleasant, unpleasant	Wiki.50d	**0.38**	Very little effect
		Twitter.25d	**-1.53**	Names_female
Names_male, names_female	Family, career	Wiki.50d	**-1.76**	Names_female
		Twitter.25d	**-1.32**	Names_female

- As you can see the effect sizes, pleasant attribute is more associated with females and unpleasant attribute is more associated with males, when the word embedding model is learnt using the twitter corpus, whereas the effect size is close to 0 when learnt using the wiki corpus, indicating a weak association, which means unbiased. This shows that there are many unpleasant associations with males than over females in twitter, and hence it shows a **very strong -ve effect** size when the training corpus is twitter, implying this bias in language against males with pleasant attribute.
- With same gender as target pairs, Family & career as attribute pairs, we notice that it is biased towards females implying biased against males with family attribute in both corpuses because in real-world and in natural language, family co-occur more frequently with females than it does with males, also career co-occur more frequently with males than it does with females. So, this indicates the bias in natural language in favor of females with family attribute, as well as in favor of males with career attribute.

Names_europe, names_africa	Pleasant, unpleasant	Wiki.50d	**1.14**	Names_europe
		Twitter.25d	**1.16**	Names_europe
Flowers, Insects	Positive-words, Negative-words	Wiki.50d	**0.83**	Very little effect
		Twitter.100d	**0.99**	Flowers
Art, Science	gender_m, gender_f	Wiki.50d	**-1.56**	Science
		Twitter.50d	**0.68**	Very little effect

- When considering race names (European names, African names) as target pairs, and (pleasant, unpleasant) as attributes, the word embeddings are biased against African names implying biased towards European names with respect to the pleasant attribute when trained using both wiki and twitter corpuses. This indicates that there are more unpleasant co-occurrences with African names than with European names and more pleasant co-occurrences with European names than with African names in both Wikipedia and twitter. Hence this is a bias in the natural language in favor of European names and against African names with the pleasant attribute.
- Similarly, it is biased towards flowers implying bias against insects in natural language with respect to the positive-words only by a marginal score, since negative-words co-occur more frequently with insects than with flowers and positive-words co-occur more frequently with flowers than with insects in both Wikipedia as well as twitter.
- When the Word Embedding Association Test is performed using target pairs (Art, Science) and gender as attribute, males are more associated with science than with art when learnt using the wiki corpus, whereas males are more associated with art than with science but only by a little effect (0.68) when the model is learnt using the twitter corpus. This shows that, since Wikipedia is a collection of general facts and accepted theories, it seems to relate more men with science and more women with art. But twitter includes various kinds of people opinions and works of all budding artists and innovations. So, when the model is trained using twitter, it cancels out the bias in the previous step showing that twitter involves almost equal art & science associations with males as well as females, thus being fair enough.

Hypothesis1: performing WEAT with new set of target pairs and existing attributes.

Pet_animals.txt => dog, cat, mice, parrot, rabbit, cow, fish, horse.

Wild_animals.txt => snake, lion, bear, tiger, leopard, wolf, crocodile, monkey.

These lists of words in hypothesis1 are created based on the significant differences between pet animals and wild animals which we don't want to be biased and are paired as targets to find relationships with the existing attributes to check for additional biases.

Pet_animals, Wild_animals	Pleasant, Unpleasant	Wiki.100d	**-0.25**	very little effect
		Twitter.50d	**-0.45**	Very little effect

- It is found that the effect size is very close to 0 by using both the training corpus, implying a weak association, which means it is not biased towards or against any target. Also, pet/wild animals in associations with pleasant/unpleasant may not be present in the given literatures and hence there is no bias here in both Wikipedia and twitter corpus. This was my intuition while I formed the target pairs, and the output also confirms the same. So, this is an example of the right kind of model that can be taken and applied to larger datasets in order to arrive at a fair proposition eliminating the implicit bias.

Hypothesis2: performing WEAT with target pairs created above and new set of attribute pairs.

Weak.txt => fragile, sick, faulty, lacking, imperfect, faint, dull, mild.

Strong.txt => powerful, vigorous, tough, robust, muscular, fierce, solid, resistant.

These lists of words in hypothesis2 were created to describe the targets in hypothesis1 based on how strong/weak they are.

Note: these weak & strong wordlists can also be used as target pairs with attributes such as pleasant/ unpleasant, male/female etc.

Pet_animals, Wild_animals	Weak, strong	Wiki.100d	**1.10**	Pet_animals
		Twitter.25d	**0.85**	Pet_animals

- With respect to weak and strong attribute, my intuition was pet animals will be more associated with weak and wild animals will be more associated with being strong, and also that the score would be high while using Wikipedia compared to that of twitter. So, after working it out, I was surprised to find all those intuitions to be captured exactly by the model. The +ve effect scores in both the cases conveys the same intuition quantitatively as well.

Therefore, from analyzing all the above quantitative results by comparing the effect sizes between different sets of WEAT pairings and by varying the training corpus used to learn the embeddings, the findings are:

- The bias in natural language using the given WEAT model mostly seems to depend on the training corpus that is used to learn the word embeddings.
- We could find the pre-existing bias in natural language for many hypotheses that seems to be mostly against or in favor of a particular target in both Wikipedia and twitter corpus which reflects the real-world associations.
  Example: It is biased against names_africa and in favor of names_europe with the pleasant attribute regardless of the training corpus used to learn the word associations.
- In some cases, the effect size turns out to be very close to 0, which indicates that there is no bias towards or against any target lists or there is not much of co-occurrences between the target and attribute pairs found in the training corpuses.
- There are cases when it's biased against a target when trained using Wikipedia, but it's in favor of the same target when the training corpus is twitter. This is because Wikipedia is a universal representation of widely accepted propositions and theories, whereas twitter is a public space where people express all their interests and opinions about almost everything. Hence, this should be the reason why the bias shifts for some cases.

Conclusion:

Therefore, the model was implemented successfully by using the given tools to explore the bias in natural language based on Wikipedia & twitter corpuses, thereby discussing all associations and technical findings as above.