CS584 HOMEWORK 4
NAME : Preeti Bhattacharya (G01302375)

Introduction:

New programs are being released all the time, putting complex algorithms to work on vast, often updated data sets. The rapidity with which this is happening demonstrates the technology's appeal, but the lack of experience poses genuine dangers. One of the most serious dangers is algorithmic bias, which jeopardizes machine learning's core goal. This often-overlooked flaw can lead to costly mistakes and, if left unaddressed, can lead to projects and organizations going in completely different ways. Effective attempts to address this issue from the start will pay off handsomely, allowing machine learning's true potential to be fulfilled in the most efficient way possible.

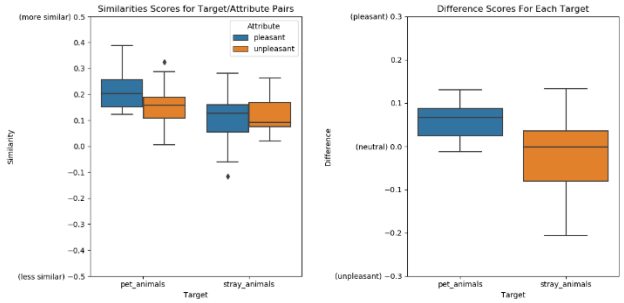
Machine-learning algorithms, which automate business processes, make choices faster and at a lower cost than human decision makers. Due to the alleged absence of human biases, machine learning claims to increase decision quality. Human decision-makers, for example, may be prone to emphasizing their own personal experiences. This is an example of anchoring bias, which is one of many that can influence business decisions. Another is availability bias. When faced with a decision, people use this mental shortcut (heuristic) to make familiar assumptions. The assumptions may have served you well in the past, but they may be unfounded in new circumstances. Loss-aversion bias puts unnecessary conservatism on decision-making processes, whereas confirmation bias tends to pick evidence that supports preconceived assumptions.

Many judgments with commercial ramifications, like as loan approvals in banking, and personal implications, such as diagnostic decisions in hospital emergency rooms, are being made using machine learning. The advantages of eliminating detrimental biases from such judgments are evident and desirable, whether they be financial, medical, or other in nature.

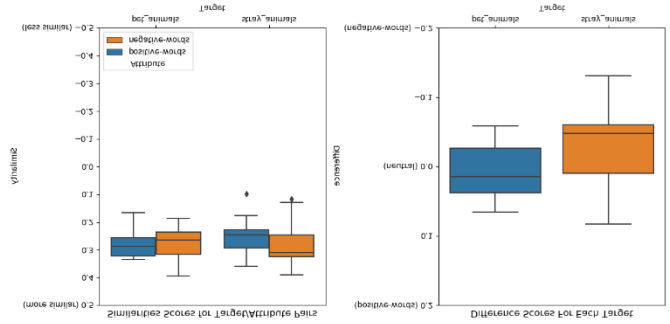
Approach:

1) We have used the tools provided for testing associations with pairs of words that can be encoded in natural language word embeddings.
2) For part 1- I have rerun the pairings of names_males and names_females with family and career and go effect size as 1.32 for twitter and effect size is 1.76 of wiki which shows the biases as female are more inclined towards family and males are more inclined towards career.
3) For part-2 -Then we have created two text files categorized as pet_animals and stray_animals. In pet animals we have **Siberian Husky, Panda Bear, Horse, cat, dog, golden fish, dolphins, kitten, puppy, golden retriever, parrot, butterfly, rabbit, Turtle, Guinea pig, duck, swan, sheep, squirrel, bichon fries, poodle.** And in stray_animals we have **spider, bedbug, bee, cockroach, mosquito, termite, beetle, moth, shark, fox, python, anaconda, bats, crocodile, leopard, owl, octopus, rats, Cane Toads, scorpions.**
4) I have classified these wordlist on the basis of animals which are wanted at homes by humans and which are not.
5) After that we rerun the pairings by varying the underlying training corpus used to learn the word embeddings.
6) I tried each of Wikipedia and Twitter and got the below results.

A) For Wikipedia and Twitter:- With pleasant and unpleasant I got Effect size: 1.05 which shows that there is biased data in wiki and got Effect size: 0.68 which shows that there are biased data in Twitter which classify animals based on how pleasant they are. As they are getting below similar attribute words :- love, lucky, poison, paradise, rainbow, vomit and stink.



B) For Wikipedia and Twitter:- With positive words and negative words I got Effect size: 0.94 which shows that there is biased data in wiki and got Effect size: 0.63 which shows that there are biased data in Twitter which classify animals based on how positive and negative impact they give. As they are getting below similar attribute words :- monster, pig, wild, shark and bug.



Conclusion, Interpretation of Result and Future Work:

From the above result we can say that the algorithm is biased on the data provided as we have got an effect size of around one. And few of the animals such as dog, cat is loved more than any other animals when compared with mosquitoes, bedbugs etc.

The researchers discovered that the network is better able to generalize to new images or viewpoints if the dataset is more diverse — that is, if more photographs portray objects from various perspectives. To overcome bias, it is necessary to have a diverse set of data. So, we should select training data that is appropriately representative and large enough to counteract common types of machine learning bias. Future is all about finding this balance, which can only be done through a combination of algorithms and human intelligence.

References:-

https://www.mckinsey.com/business-functions/risk-and-resilience/our-insights/controlling-machine-learning-algorithms-and-their-biases