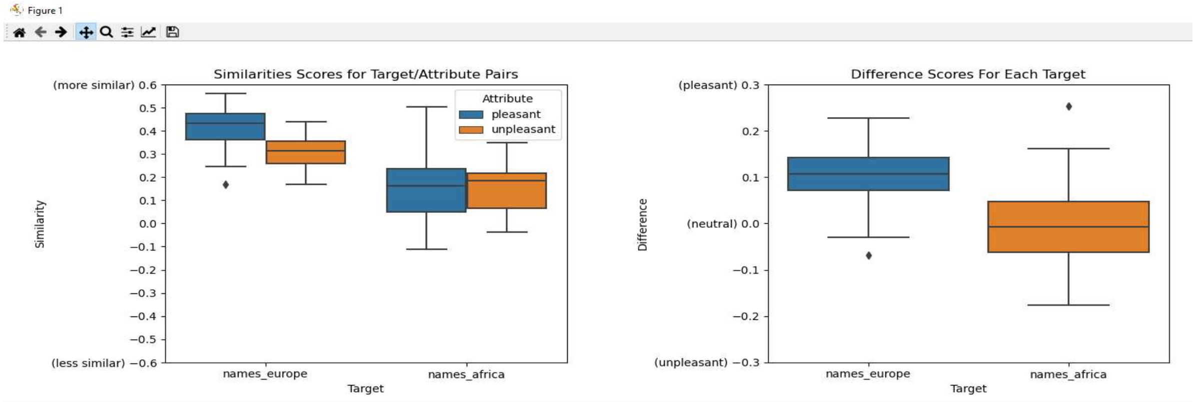
Willy Esquivel-Lopez (G01127937)

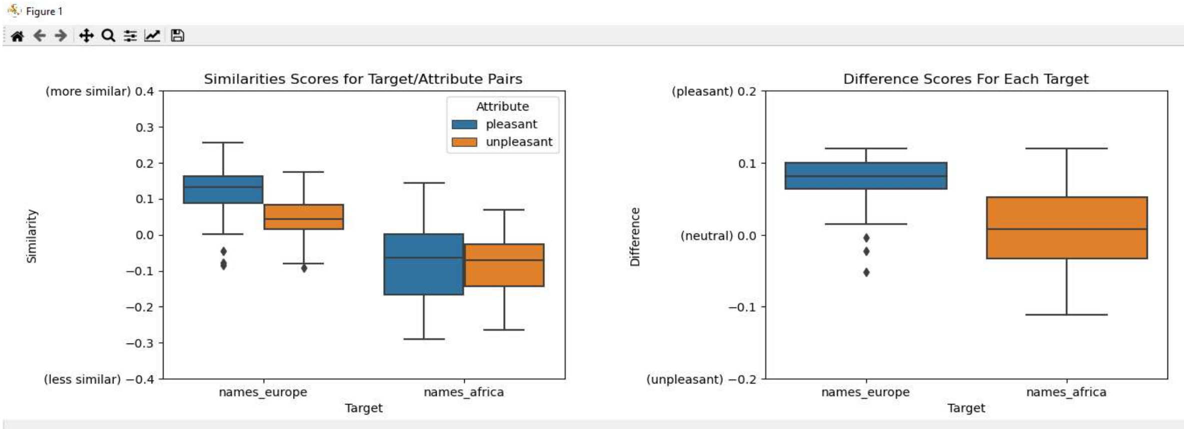
HW3 Report

Bias in Word Embeddings

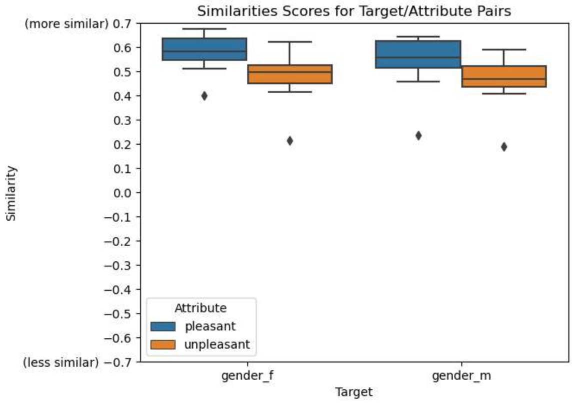
Upon re-running the word pairing, I noticed several implications when running comparisons of European and African names with unpleasant and pleasant words. There were noticeable differences when running the data on either the twitter word vectors and the Wikipedia. Mostly the choices in related word attributes were surprising. When running **weatTest.py twitter names_europe names_africa pleasant unpleasant,** it returns the top 5 most similar attribute words to names_europe are: happy, lucky, family, love, and friend while the top 5 most similar attribute words to names_africa are: diamond, filth, divorce, pollute, and caress. Below is the provided the data from twitter word vectors:

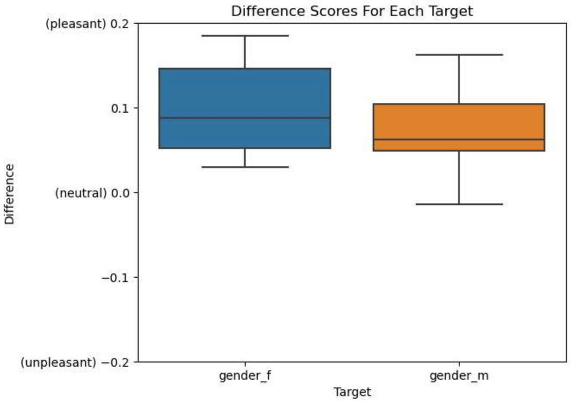


weatTest.py wikipedia names_europe names_africa pleasant unpleasant, returns the top 5 most similar attribute words to names_europe were: Lucky, happy, family, love, and friend. While the top 5 most similar attribute words to names_africa: divorce, murder, diamond, caress, and friend.



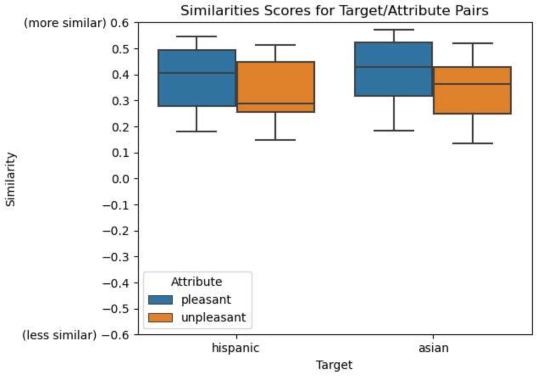
when comparing genders to pleasant and unpleasant words there was still some biasness, but definitely a lot less noticeable:

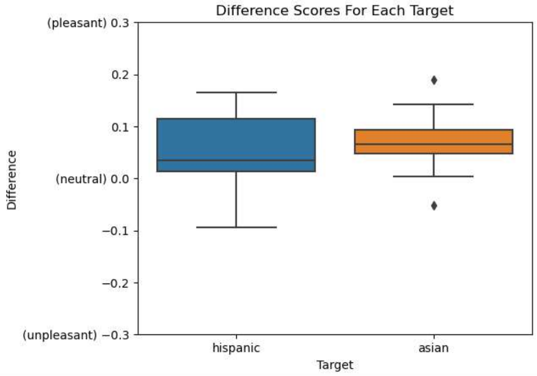


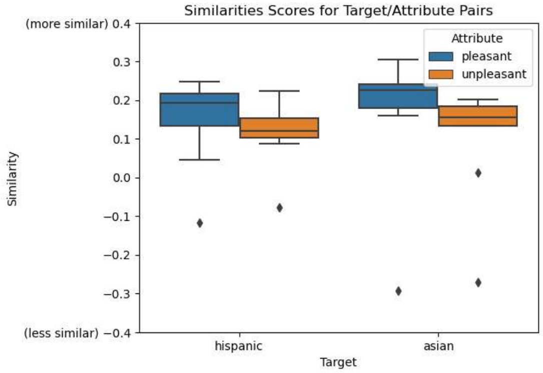


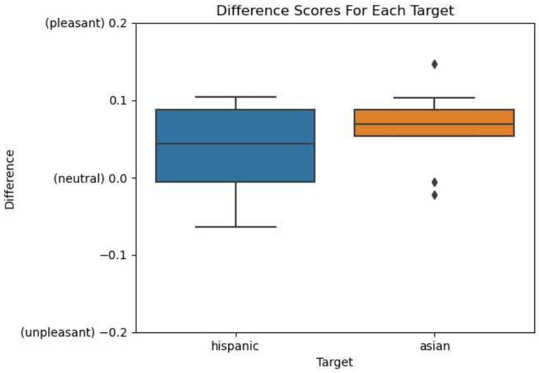
Although when comparing Top 5 most similar attribute words they were interestingly quite similar with the exception of two words. For Males it was: Kill, Love, Lucky, family, friend. For females it was: Love, Ugly, Lucky, Family, Friend.

Proposing my word list depicting nationalities of both Hispanic and Asian descendants/nationalities, I applied them to the twitter and Wikipedia vectors to compare them to pleasant and unpleasant words, both were very similar in this realm and showed nearly identical outcomes:

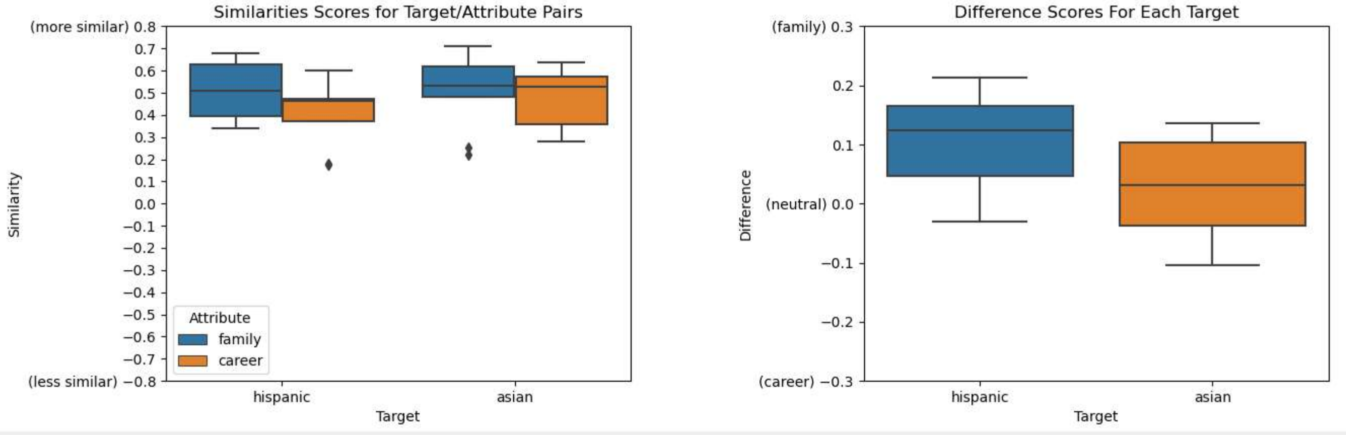








The real bias was discovered upon comparing Hispanic and Asian nationalities to family and career-oriented words. Where the bias depicted Hispanics as very family oriented and not career oriented and Asian nationals as very career oriented and somewhat family.



With the tops 5 similar attribute words for Hispanic nationals being: relatives, cousins, marriage, children, and family. For Asian nationals: business, office, wedding, family, and children. Though this bias is not harmful or makes for any implications it could lead to financial decisions made in circumstances less favorable towards Hispanics.

The list of words used for Hispanic nationals: Hispanic, Latino, Spanish, Mexico, Mexican, Guatemalan, Salvadorian, Brazilian, and Chile.

The list of words used for Asian nationals: Asian, Chinese, Korean, Vietnamese, china, Cantonese, Japanese, and Thailand.

Ethical case study

To safely use machine learning in life threatening decision making, the models must be nearly 100% certain. But even then, there would still be concerns regarding ethics. The main concerns being privacy/confidentiality, the responsibility of Machine learning in making time-critical life-threatening decisions, and grouping/discrimination. The first concern regarding privacy and confidentiality is one already monitored by organizations like HIPAA. Ensuring that health information is kept confidential, but to create models the data must be analyzed by developer. There must be some sort of agreement between the patients and the hospitals to allow this information to be analyzed deeply, otherwise that would be considered a breach of trust, not to mention, not HIPAA compliant. The second concern being time critical imminent danger, as making this queue could pose harm due to false negative or high amount of false positives. By either prioritizing non important emergency rooms being filled or not prioritizing a real emergency, causing possible deaths in both case scenarios. This would be difficult to put any responsibility on any one person as the hospital and developers were both at fault in this situation. without a model that predicts in 100% certainty there is always a possibility of these scenarios to occur. Using a natural language processor for this task would also be bad practice, as it could lead to low risk and high risk grouping, and further lead to forms of discrimination for those types of groupings, whether it be race or gender biasness if the issue is ignored.