**Capstone Research Project**

This project studies how training language models on a variety of languages affects racial and gender bias in in these models. Specifically, I ask how the choice of languages used in training different versions of a model impacts the racial and gender biases presented within these NLP models? Does the inclusion of other languages, and therefore the inclusion of the societal norms and customs that come from the cultures who speak and write in these languages have any impact on the potential biases these models may show to a certain ethnic group or gender? To answer this I examine the shifts in racial and gender bias across BERT models trained for a variety of languages.

While language models, such as BERT, have revolutionized NLP tasks, there are concerns that these models are trained to exhibit bias unintentionally. In a study done by Caliskan et al. (2017), the authors evaluated the cosine similarity between a subset of words pertaining to a variety of categories. Their tests discovered the model’s racial and gender biases. Through this study, the authors showed not only that the GloVe model exhibits these commonly held biases, but also that these biases could lead to disproportionate representation in automated decision-making processes.

I propose a similar design to Caliskan et al. in which I evaluate the cosine similarity of a subset of target words and concepts to determine if the selected models exhibit bias. All the models evaluated will be trained in both English and in at least one other language. This ensures that when running the tests for cosine similarity, no translation will be needed. The models being evaluated are available from Hugging Face. The data I will evaluate these models on will change depending on the kind of bias I want to look at.

To measure bias in these models, I test how similar a set of words is to another. For example, in the test for gender bias, I evaluate the cosine similarity between two sets of the most popular names for both men and women respectively and two sets of job labels, one consisting of STEM-related jobs and one of non-STEM-related fields. In doing so, I compare how closely related these sets of words are to each other, and whether one set of names is more closely linked to one of the two sets of job labels.

To test for biases, I will create a set of both target words and a set of concepts in my evaluations of cosine similarity. The set of target words used in each test will be consistent throughout to keep results consistent. These two sets will consist of words which are considered Pleasant (e.g. happy, kind, enjoyable) and Unpleasant (e.g. sad, angry, disgusting) I will first test for race using subsets of names that are commonly classified as European/English-American, African-American, Chinese-American, and Latin-American, and evaluating the cosine similarity of these subsets with both the group of Pleasant and Unpleasant words. I will then test for gender bias following a similar format, in which I will gather two sets of the most common names in the US for both men and women respectively, and evaluate the cosine similarity between each set and the set of target words. With regards to gender bias, I will also evaluate the cosine similarity between these sets of names and two sets of job labels. These two sets of labels will consist of labels pertaining to STEM careers, such as engineer or data scientist, and non-STEM fields, such as journalist or teacher.

The models I will be testing are outlined in the List\_of\_Models.csv file. This csv file contains a list of 9 models, 3 of which are trained only in English and 6 multilingual models. The three English models are as follows: bert-base-cased, roberta-base, and distilbert-base-cased. These three models will serve as the control group in this experiment, as these are some of the most used variations of the BERT model, and the majority of the models I will be testing are derived from these base models. The other 6 models are: bert-base-cased-multilingual, bert-base-uncased-multilingual, xlm-roberta-base, ernie-2.0-base-en, distilbert-base-multilingual-cased, and stsb-xlm-r-multilingual.

To organize these models, I’ve constructed a dataset of a variety of one-hot encoded variables outlining the languages they’ve been trained in, training data, and the uses of these models, as well as if the models have been finetuned. The variables outlining the language the model was trained in are: English, Chinese, Spanish, German, French, and Multi. The Multi variable in this case represents that a model was trained in more than one language, whether that be 2 or more of the five previously listed languages, or that the model was trained in more languages than just the five listed ones. I’ve also included another variable, Number\_of\_Languages, which as the name suggests, contains the total number of languages the model was trained in through training and finetuning.

The next group of variables represents the training data that was used to train the model, or the underlying model each variation was built from. The variables are as follows: Wikipedia, BookCorpus, CommonCrawl, CC100, Ted2020, and Other. The Wikipedia variable represents the Wikipedia datasets, a collection of cleaned Wikipedia articles written in every available language. The BookCorpus variable represents the BookCorpus dataset, a collection of text from over 11,000 unpublished books scraped from the internet. The CommonCrawl dataset represents the CommonCrawl dataset, a collection of 3.15 billion webpages scraped from the internet, 46% of which are in English, with the rest being in a variety of other languages. The CC100 variable represents the CC100 corpus which was used to train XLM-R. This corpus contains data for 116 languages from the 2018 CommonCrawl snapshot. The Ted2020 variable represents the Ted2020 corpus, a collection of transcripts from nearly 4000 TED talks in 2020. Finally, the OtherData variable represents if were any other datasets used to train the model.

The intended uses of each model are also outlined in the List\_of\_Models dataset. The MLM variable represents if one of the primary uses of the model is for Masked Language Modeling, where the model is able to predict missing words in a sentence by looking at the context of the rest of the words in the sentence. The NSP variable represents the model’s ability to perform Next Sentence Prediction, where the model is given two sentences and is tasked with determining if the two sentences were following one another in whatever text they were taken from. The Sequence\_Classification variable represents if the model is capable of sequence classification, where the model is fed a sequence of data and is tasked with predicting a category for the sequence of data. The Token\_Classification variable represents if the model is capable of token classification, where the model is given a string of text and is tasked with assigning each token, in this case each word, to a category. The QnA variable represents if the model can provide a response to a question posed by a human. Lastly, the NLI variable represents if the model is capable of Natural Language Inference, in that when presented with a hypothesis, the model can determine if the hypothesis is true, false, or undetermined. If there are any other primary uses for the model which do not require finetuning, the Other\_Uses variable is used.

The next set of variables focuses on if a model was finetuned for some task(s), and the datasets that were used in the process of doing so. In the case of many of these models, they were finetuned to complete certain tasks or to be able to understand and communicate in more than one language. The finetuning variables used in this dataset represent the names of different finetuning datasets that a model uses. The variables in this section are as follows: XNLI, XQuAD, GLUE, SuperGLUE, MNLI, and OtherFT. All of these variables, apart from OtherFT, represent a specific well-known and trusted dataset used in finetuning language models. The OtherFT variable represents if the model was finetuned on any other dataset not listed previously.

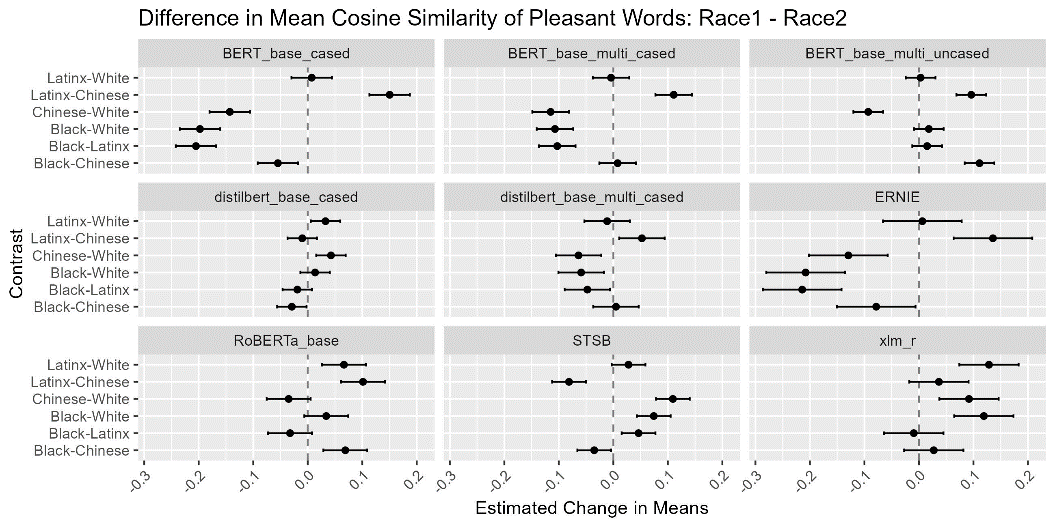
Finally, the last section of variables focuses on the structure of each model. The first two columns indicate if the model is either an encoder or a decoder model, or both. For a model to be an encoder, it must take in an input, in this case, strings of text, and maps the data into a more compact representation that focuses on the specific patterns and features of this input data. On the other hand, a decoder model takes this compact representation of the data and turns it back into the original form of the data. In the case of these 12 models, only 1 is capable of being considered both an encoder and a decoder model, that being the mDeBERTA v3 model that was finetuned on the XNLI and MNLI datasets. The rest of these models are all considered to be encoder models. The final two variables, named Attention\_Heads and Layers, describe how many attention heads each neural network has, as well as the number of layers each network has. The number of attention heads reflects ow many different mechanisms that capture different aspects of the input data in the model, and discern the patterns present in the data. As for the layers in a model, each layer focuses on a specific operation performed on the data, and produces an output that is passed down to the next layer.

Moving forward, the focus of this project will be on testing each model for cosine similarity. As discussed, there will be 3 separate tests performed on each model. The first of these tests will focus on testing for cosine similarity between 20 of the most popular baby names, according to Names.org, of both African American, Latin American, Asian American and European/English American origin against two lists of words similar to Pleasant and Unpleasant. These cosine similarity values will be used to determine the amount of underlying racial biases each of these models may hold. The list of African American baby names will consist of the 10 most popular male names (Reginald, Kameron, Kendrick, Javon, Tyrell, Jamar, Camron, Tyree, Jamari, and Reggie) and the 10 most popular female names (Jada, Latoya, Jayla, Tamika, Latoyna, Journey, Tameka, Journee, Lawanda, and Janiya). Similarly, each list will follow a similar structure as the African American list. The list of European/English American names is as follows: James, John, Robert, Michael, William, David, Joseph, Richard, Charles, and Thomas for the males, and Mary, Elizabeth, Patricia, Jennifer, Linda, Barbara, Margaret, Susan, Sarah, and Jessica for the females. For the list of Latin American names, the 10 most popular male names are Paul, Vincent, Victor, Adrian, Marcus, Leo, Miles, Roman, Sergio, and Felix, with the 10 most popular female names being Patricia, Laura, Amanda, Victoria, Julia, Gloria, Diana, Clara, Paula, and Norma. Finally, the list of the most popular Chinese American names is as follows: Lian, Shan, Lew, Long, Quan, Jun, Tou, Jin, Cai, and Chan for the males, and Lue, China, Lu, Maylee, Tennie, Maylin, Chynna, Jia, Mei, and Tylee for the females. For each of these lists, the names are in order of their popularity, with the first name being the most popular and the last beign the 10th most popular name. As for the lists of words similar to Pleasant and Unpleasant these names will be compared too, the lists were chosen on a more subjective approach. I’ve chosen 10 words for each category that I believe will give both a broad range of results as well as covering a variety of different characteristics. It is also important to note these lists will also be used in the second round of testing I will complete for each of these models involving gender biases. The list of “pleasant” words I’ve chosen are Happy, Agreeable, Polite, Civil, Charming, Gracious, Gentle, Approachable, Love, and Cool. As for the “unpleasant” words, I’ve chosen to include the words Rude, Lazy, Disagreeable, Lousy, Sad, Hate, Violent, Bitter, Harsh, and Angry.

When moving to the second and third rounds of testing, the focus of these tests will be on the potential gender biases within these models. The list of male and female names used in these two rounds of testing will be the top 20 male and female baby names in the US according to Names.org. To note, for both lists of names, as well as the previously discussed lists of names, these baby names are the most popular names in the US from 1880 to the present. The list of male names I will use in these tests will be: James, John, Robert, Michael, William, David, Joseph, Richard, Charles, Thomas, Christopher, Daniel, Matthew, George, Anthony, Donald, Paul, Mark, Andrew, and Edward. As for the list of female names I will use, they are: Mary, Elizabeth, Patricia, Jennifer, Linda, Barbara, Margaret, Susan, Dorothy, Sarah, Jessica, Helen, Nancy, Betty, Karen, Lisa, Anna, Sandra, Emily, and Ashley. The first of these two tests for cosine similarity will follow almost the exact same format as the test for racial bias, in that this group of commonly male and female names will be compared with the same list of “pleasant” and “unpleasant” words that I’ve outlined previously.

However, the second test for gender bias will follow a slightly different design, in that rather than comparing the cosine similarity between the two groups of names and the groups of pleasant and unpleasant words, this test will compare each group of names to two sets of professions, namely those that are STEM-related careers and those that are not. The reason for this specific test is to not only determine the potential biases held by these models of the characteristics of men and women, but also to test the biases these models may possess regarding what roles men and women have in the workforce. There has been a strong emphasis in recent years on improving gender relations in the STEM fields and reform to help these fields become more inclusive, however certain training data that these models are trained on are full of older material from times where it was uncommon for women to work in these professions. It is also important to note that the training and/or finetuning data for the multilingual models possess texts from a variety of different cultures and languages, where it may be more commonplace that women are not to be working in specific fields (let alone at all), as these cultures and their teachings put the focus on the men to earn for the family and for women to take care of the home. To test for these potential biases, we need to first identify the STEM and non-STEM careers we will use to compare with the lists of male and female names. For STEM careers, the list will consist of 10 of the best jobs in the field according to US News’s best jobs in STEM in 2023: Software Developer, Nurse Practitioner, Health Services Manager, Physicians Assistant, Security Analyst, IT Manager, Web Developer, Dentist, Orthodontist, and Computer Systems Analyst. For the list of non-STEM careers, the list of 10 professions is selected from Indeed’s list of best paying non-STEM careers: Artist, Marketing Manager, Social Worker, Attorney, Journalist, Musician, Teacher, Media Manager, Graphic Designer, and Judge. The reason that the list of non-STEM professions was selected through a more subjective process is that there is an extremely wide range of professions that are not in the STEM fields, and as such are much more difficult to narrow down into a list of the 10 most popular jobs. As such, I chose what I believe to be the most diverse and appropriate professions to include in the list.

Test 1 consisted of testing the names commonly associated with the 4 previously outlined ethnic groups (European/English Americans, African Americans, Chinese Americans, and Latinx Americans) against both lists of Pleasant and Unpleasant words. After calculating the cosine similarity values between each of the 80 names and 20 target words, I was left with a dataset consisting of 1600 observations for each of the 9 models, each with the name, the word, and the cosine similarity value of those two words. I have also included the name of the model as another column to be used in the cross-model analysis portion of my results, as well as two columns with the ethnic group the name in question is commonly associated with and whether the word it was being compared to was a Pleasant or Unpleasant word.

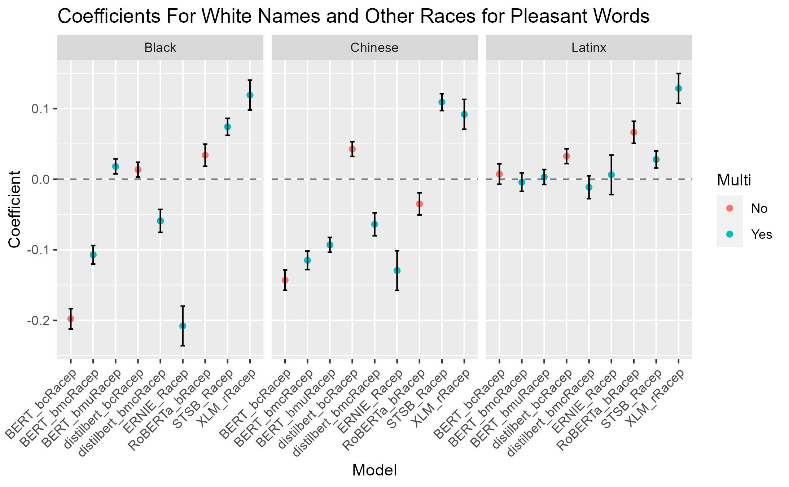
After gathering this data from all models, I first began with a simple analysis of variance test on the cosine similarity values for each of the 4 ethnic groups, as well as a complementary post-hoc Tukey test. In performing these tests, we can decern graph the differences in mean cosine similarities of two ethnic groups to help visualize the potential biases these models may hold against certain ethnic groups. Shown to the right is a graph displaying the change in means of the cosine similarities for a pair of ethnicities and the group of Pleasant words for each model, accompanied by an error bar showing a confidence interval with which the true difference in means is most likely to be found. Looking at this graph, we can determine how the mean cosine similarity value of the first ethnicity listed in the pair changes when moving from the second ethnicity to the first. For example, we can that for the BERT-base-cased model, the mean cosine similarity value is greater for Latinx names than it is for White names when compared with the list of Pleasant words, as the value of the difference of means is greater than 0. This set-up allows us to study what races specifically are more closely related to the set of target words relative to each of the other three races in this round of testing.

For starters, looking at BERT-base-cased model, there is a clear bias for Latinx names compared to the groups of White, Black, and Chinese names. We can also see for this model that Black names are shown to be less favorable than all three of the other races, with the difference in means for Black names and all three other races being negative. When looking at the other models in this graph, the ERNIE model shows the exact same relationship between the differences in the mean cosine similarity values, suggesting that both models show a strong bias against names commonly associated with African Americans, and a slight bias for names commonly associated with Latinx Americans. The rest of the models, however, are not as simple regarding the specific biases that possess. The BERT-base-multilingual-cased model possesses a strong bias against Black names when comparing the difference in mean cosine similarity values of White names and Latinx names, but shows a slight positive difference when compared to Chinese names. And while Chinese names have a smaller mean cosine similarity to Pleasant words than White names do, they are favored over the group of Latinx names. When looking at the disparity in the difference in means, we can see that White names are primarily favored over all the other races for this model, albeit only slightly when compared to Latinx names, followed by Latinx names, Black names, and Chinese names in order of how closely related these groups of names are to Pleasant words.

For BERT-base-multilingual-uncased, the model seems to show bias for Black names more than the other three races, followed by Latinx names, Chinese names, and shockingly White names being last in order of favorability. I believe that for this model specifically, this may be due to the model’s ability to take in an interpret the context of a word regardless of the case of any and/or all of the letters in this word. In this case, the words “Happy”, “happy”, and “hApPy” would all be read as the same base uncased version, happy. Since the list of names all start with a capital letter, and all of the target words are written entirely in lowercase letters, the models that do take into account the case of the letters, like BERT-base-cased and multilingual-cased, read a name or target word as a separate word from any other variation of the same word, and the cosine similarity values will overall be lower relative to the cosine similarity values found from the uncased versions of BERT.

Moving to the distilbert models, the base-cased model possesses a slight favoritism towards Chinese names over the other racial groups, followed by Latinx names, Black names, and again lastly White names. This to me was the most shocking discovery regarding which groups of names these models would favor, as I had originally hypothesized that models which were only trained on a variety of English texts would produced the most skewed results in favor of White names, due to the history of racism and prejudice that is engrained in many English-speaking nations’ cultures. Surprisingly, it was the base-multilingual-cased model of distilbert which possessed a strong bias for White names, in that relative to all the other racial groups, the mean cosine similarity between the sets of White names and Pleasant words was greater than all the other groups. The most plausible cause for these unexpected biases might be the way in which these distilbert models are trained. As the name suggests, the distilbert models are models distilled from the original BERT base model and utilize a much smaller training dataset. These smaller datasets have a smaller variation in what the text data has information about, and in this case, may have much less text data regarding the names listed in the racial groups these models show the most bias against.

Finally, moving to the three variations of RoBERTa models, we again see this unexpected bias against the group of White names relative to the other lists of names. The RoBERTa-base model is again and English-only model, but shows a bias for Black and Latinx names relative to White names. However, unlike with the distilbert-base-cased model, White names are still on average closer related to the group of Pleasant words than that of Chinese names. This bias for White names over Chinese names however is again lost when looking at both the Sentence Transformers and base variations of the XLM RoBERTa model. Both of these models are multilingual models, and show a sizeable bias against White names when comparing the mean cosine similarity value to that of the mean value for the three other groups of names. Specifically, the Sentence Transformers model possesses biases primarily for Chinese names, whereas the XLM-r model possesses biases primarily for Latinx names. While each of the models in this study possess their own individual biases for certain groups of names, I next want to focus in on how the biases they possess compare from model-to-model, and which models look to be the least biased models of the bunch, in that they don’t particularly favor one group of names over the other.

Following these analysis of variance tests and the associated post-hoc Tukey Honest Significant Difference tests for each model, the challenge of comparing the potential biases shown in these models must be tackled. To start, I began by normalizing the cosine similarity values for each model using min-max normalization. To perform min-max normalization, the lowest cosine similarity value from all three rounds of testing is subtracted from the value being normalized, which then get divided by the largest possible cosine similarity value minus the smallest possible cosine similarity value. By using this min-max normalization technique, I could run a linear regression model for each of the 9 models that would determine the change in cosine similarity when moving from one ethnic group to another. Before creating the model however, the Race attribute, which states what ethnic group/race the name in the observation is commonly associated with is, needed to have a reference factor set. I chose to set the reference factor to “White”, as one of my original hypotheses was that these NLP models would show biases for European/English names. In doing so, the linear model that will be produced will display the coefficients for the change in cosine similarity when moving from the group of White names to any of the other three ethnic groups. After running these models for both the dataset for Pleasant and Unpleasant words, I was able to derive these two graphs:

A graph of different types of numbers

Description automatically generatedThese graphs show the coefficients of cosine similarity when moving from the group of White names to Black, Chinese, and Latinx names for each other 9 models. A negative coefficient suggests that the regression model predicts the cosine similarity value for a Pleasant/Unpleasant word to decrease when moving from comparing the word to a commonly associated White (European/English) name to an African/Chinese/Latinx American name, whereas a positive coefficient suggests the model predicts the cosine similarity value for a Pleasant/Unpleasant word to increase when moving from comparing the word to a White name to comparing the same word to an African/Chinese/Latinx American name. In the case of a negative coefficient, we can interpret that as the NLP model in question is more likely to associate the word in question with a White name rather than a name associated with a different ethnic group, and a positive coefficient shows that the NLP model is more likely to associate the word in question with a name from one of the other ethnic groups rather than a White name.

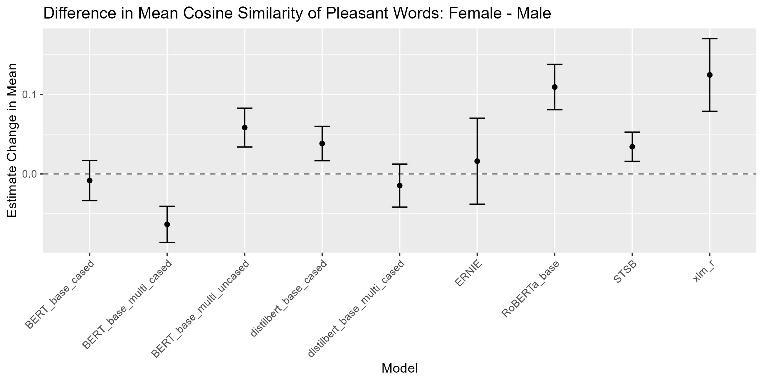
The first graph looks at the relationship between White names and the other three ethnic groups and how their cosine similarity values with Pleasant words changes. We can see that there are a multitude of models which show a clear bias for White names compared to Black and Chinese names, while there surprisingly seems to be a bias against White names when compared to Latinx names. The BERT-base-cased, BERT-base-multilingual-cased, distilbert-base-multilingual-cased, and ERNIE models all have negative coefficients for Black names, indicating the cosine similarity value decreases when moving from the group of White names to the group of Black names. This negative coefficient shows there is a clear bias in these four models against names which are more commonly associated with African Americans when comparing them to Pleasant words. We can also see that all three base BERT models (base-cased, base-multilingual-cased, and base-multilingual-uncased), distilbert-base-multilingual-cased, ERNIE, and RoBERTa-base all have negative coefficients for Chinese names, suggesting that there is a bias towards White names when comparing their cosine similarities to Pleasant words to Chinese names’ cosine similarities. However, when looking at the coefficients for Latinx names, not a single one of the 9 models has a negative coefficient. This is to say that when looking at the cosine similarity values for Pleasant words and White names and Pleasant words and Latinx names, each of these models is more likely to associate a Pleasant word with a Latinx name than a White name.

The second of these two graphs is set up in a similar fashion, as this graph shows the coefficients of cosine similarities when comparing a White name and an Unpleasant word and a name associated with one of the other three ethnic groups in this study and an Unpleasant word. We can see in this graph a similar pattern of bias against Black and Chinese names and a bias for Latinx names. Only three models, BERT-base-cased, BERT-base-multilingual-cased, and ERNIE, have a negative coefficient for Black names, suggesting that the other 6 models are likely to associate and Unpleasant word with a Black name rather than a White name. We can also see that the same three models, BERT-base-cased, BERT-base-multilingual-cased, and ERNIE, as well as RoBERTa-base, all have negative coefficients for Chinese names, suggesting that the other 5 models are more likely to associate an Unpleasant word with a Chinese name than a White name. Finally, we can see that BERT-base-multilingual-cased and uncased models, distilbert-base-multilingual-cased, and stsb-xlm-r-multilingual all have very slightly negative coefficients for Latinx names, indicating these 4 models are more likely to associate White names with an Unpleasant word than Latinx names.

There are a few strange and concerning patterns we can see in these graphs however. First, looking at Black names specifically, all three models which possessed negative coefficients of cosine similarity for Unpleasant words (BERT-base-cased, BERT-base-multilingual-cased, and ERNIE) also had negative coefficients for Pleasant words. We can also see this same pattern within the both graph for our Chinese name coefficients, as these same three models have negative coefficients for both Pleasant and Unpleasant words. This doesn’t necessarily rule out that these models might not possess any biases one way or another towards Black and Chinese names, but rather these models may not have enough training data including uses of the names in our Black names or Chinese names lists relative to our White names list. From the results of our regression models, there are only 5 instances where a coefficient appears to not be statistically significant to a p-value < 0.05 for Black names and Chinese names; BERT-base-multilingual-uncased’s coefficient for Black names and Unpleasant words (p-value of 0.087) and coefficient for Chinese names and Unpleasant words (0.547), both of distilbert-base-multilingual-cased’s coefficients for Black names (p-values of 0.197 and 0.255 for Pleasant and Unpleasant words respectively) , and RoBERTa-base’s coefficient for Chinese names and Unpleasant words. In the case of BERT-base-multilingual-cased specifically, the combination of having only negative coefficients for both Black names and Chinese names and having not statistically significant coefficients for both sets of names for Unpleasant words suggests that there may be discernable pattern of bias within this model, and that there may not have been enough training data included in this model to account for the specific names and words chosen to compare. This same conclusion is also the most likely suitable conclusion for the other two models in question (BERT-base-cased and ERNIE), as even though the coefficients may be statistically significant, the appearance of a negative coefficients for both Pleasant and Unpleasant words for both Black names and Chinese names is likely due to a smaller number of appearances of these names in their training data, as BERT-base-cased is an English only model, and ERNIE is only trained in English and Chinese.

After running through these results, I believe that of all the models tested, the least biased model regarding the race the name being compared to both sets of Pleasant and Unpleasant words looks to be the BERT-base-multilingual-uncased model. The coefficients for this specific model, apart from the coefficient for Chinese names and Pleasant words, are all within a range of -0.05 to 0.05. This model was the only model which showed a good variation of both positive and negative coefficients for each group of names (3 positive, 3 negative) that were as a group relatively close to 0. Even though there are coefficients for this model that is not statistically significant to a p-value of 0.05, being the coefficient for the group of Black names and Unpleasant words’ p-value is only just slightly above the 0.05 threshold at 0.087, it is very much still plausible this coefficient is not equal to 0.

Moving to Test 2, this test consisted of again testing for the cosine similarity values for both Pleasant and Unpleasant words, however this time with just two sets of names commonly associated with Males and Females. This test, and the following Test 3, follows a similar structure to Test 1 for Racial Biases, in that I have performed an analysis of variance test and a post-hoc Tukey’s Honestly Significant Difference test for the cosine similarities between both sets of names and target words to determine if the group of Male names or Female names is favored by the model. I also performed a linear regression on the cosine similarity values for each of the 9 models with the group of Male names set to be the reference factor to determine which of the models is the least biased towards one group of names or the other.

The first two graphs covered in this round of testing displays the results of the TukeyHSD tests for each model, showing the difference in mean cosine similarity values between the group of Male names and Female names. The first of these two graphs displays the difference in mean cosine similarity values for Pleasant words and the group of Female names and Males names. Right off the bat, it’s very clear that both BERT-base-cased and BERT-base-multilingual-cased, as well as distilbert-base-multilingual-cased, show a bias towards males, as the difference in mean cosine similarities between the two sets of names is negative, with the BERT-base-multilingual-cased model by far having the largest negative difference of the three. On the other hand, the remaining 6 models show varying levels of positive bias towards the group of female names, as the difference in mean cosine similarities for both groups of names is positive. Specifically, the base XLM-r model shows the largest positive difference in means, with the standard RoBERTa-base model just slightly behind it.

A graph of a number of men and women

Description automatically generatedThe second of these two graphs showing the difference in mean cosine similarity values for both Male and Female names is for the set of Unpleasant words. Similar to the results of the TukeyHSD tests for each model for Pleasant words, the same 3 models, BERT-base-cased, BERT-base-multilingual-cased, and distilbert-base-multilingual-cased, all show bias for Male names, while the remaining 6 show bias for Female names. These relatively similar findings for both the set of Pleasant and Unpleasant words is extremely surprising, as this test suggests that regardless of the meaning of the word, either “good” or “bad”, each of these models predicts either the set of Male names or the set of Female names to be more closely related to both sets of words. This suggests that for every model tested in this round, it’s very likely that rather than there being some kind of bias towards either Males or Females based on the sentiment of the word being compared to these names, the overall model may in fact have more text data used in its training process which contained more instances of Male names or Female names being used. In this case, rather than determining what model shows the least amount of bias towards a specific gender, it may be effective to determine which models have consistent coefficients for cosine similarity, regardless of whether the group of words the sets of Male and Female names are being compared to is Pleasant or Unpleasant. By focusing on this relationship, we may discover which of these models are less likely to possess biases based on sentiment, but rather possess bias on the basis that they may be lacking a sufficient amount of training data for a certain gender.

A graph showing the number of women and men

Description automatically generatedA graph showing the number of women and men

Description automatically generatedLooking at both graphs used in my cross-model analysis, we can see a similar pattern that we found from the results of the TukeyHSD tests. As expected, the same models which held biases for the group of Male names for both Pleasant and Unpleasant words possessed negative coefficients for Female names for both Pleasant words and Unpleasant words. On the other hand, the same 6 models which possessed biases for Female names regardless of the sentiment of the word it is being compared to all possessed positive coefficients for Female names for both Pleasant and Unpleasant words. In finding these same relationships for both sets of target words, we must shift the focus from finding which model has the most consistent coefficients, as these models are most likely the models which possess the least amount of gender biases based on the sentiment of the word. The models that stand out the most in this regard are distilbert-base-cased and XLM-r. These three models all have coefficients for Female names for both sets of target words within less than four thousandths (0.004) of each other. The coefficients for distilbert-base-case are 0.0382 for Pleasant words, and 0.0413 for Unpleasant words, and the coefficients for XLM-r are 0.1213 for Pleasant words and 0.1246 for Unpleasant words. All four of these coefficient estimates are also statistically significant to a p-value < 0.001, suggesting that the true coefficient values are almost certainly not equal to 0. So, when looking at these two models, we can hypothesize that they both more than likely were trained on data that contained the use of the Female names I’ve used in this testing. Of these two models however, I do believe that through further training these two models for gender-specific tasks, the better of the two models to continue training with text data that uses the names listed in my set of Male names would be distilbert-base-cased. While it is true this model has less room to work with involving what training data it can interpret, as it is a model only trained in English, the coefficients for Female names are nearly a tenth (0.1) closer to 0 than that of the coefficients for the XLM-r model.

For the third and final test, Test 3 focuses not only on each model’s biases for a certain gender, but specifically how closely each gender is related to a set of careers. In this case, I’d like to see which model’s are more likely to relate Males or Females to different STEM and Non-STEM professions. Keeping in mind that there may also be a possibility that the results of this round of testing are similar to that of the test for overall gender biases in Test 2, if similar relationships are discovered, I will again shift my analysis to which of the models shows the least bias for both sets of professions in an attempt to establish which model(s) would be most effective for future training.

A graph of a number of individuals

Description automatically generated with medium confidenceA graph of stem careers

Description automatically generatedThe first two graphs are again the results of TukeyHSD tests, specifically for the difference in mean cosine similarity between Female names and Male names for both the set of STEM careers and Non-STEM careers. As I feared, the same exact relationships for each of the 9 models that were present in Test 2 were again present in Test3. BERT-base-cased, BERT-base-multilingual-cased, and distilbert-base-multilingual-cased all possessed negative differences in cosine similarity means for the set of Female names and Males names for both sets of professions, as well as the rest of the model having positive differences in means. While there seems to be a more noticeable difference between the majority of models’ differences in mean cosine similarities for STEM words and Non-STEM words, these same biases towards one of two lists of names for Males and Females regardless of the type of word they are having the cosine similarity calculated for still remains. As such, it’s imperative we move straight into how these models compare to one another to determine which models would be most effective in limiting any potential gender biases given additional training.

A graph of different career paths

Description automatically generated with medium confidenceA graph showing the number of careers

Description automatically generatedAgain, from these two graphs of the coefficients for Female names’ cosine similarities, we can see that both distilbert-base-cased and XLM-r are the two most likely candidates to receive additional training in the event they are trained for gender-based tasks. It is also important to note that while there is very little difference between the coefficients for the ERNIE model, the coefficients themselves are not statistically significant to any considerable p-value, as the p-values for both STEM and Non-STEM professions is greater than 0.1. The coefficients for distilbert-base-cased are 0.0268 for the STEM careers cosine similarities and 0.0393 for the Non-STEM careers cosine similarities. The coefficients for XLM-r are 0.1339 and 0.1520 for STEM and Non-STEM careers respectively, indicating that this model again potentially has received less training than the distilbert-base-cased model on text that includes the names I’ve included in my list of Male names. All of these coefficients are indeed statistically significant to a p-value < 0.05, suggesting that we are able to reject the null hypothesis that the coefficients are equal to 0. As was the case in Test 2, I would conclude that the most likely candidate to receive additional training in an attempt to train on more text data using the Male names I’ve used in this study while maintaining a focus on removing any potential semantic biases for a certain gender would be the distilbert-base-cased model.

While these results, specifically for Tests 2 and 3 regarding how these models show or do not show biases towards a certain gender, are not what I had expected, I do believe that these finding are vital for the development and improvement of Natural Language Processing, and the Machine Learning world as a whole. It is extremely important for us to ensure that our AI/ML models are not only accurate in the predictions they make, but all mitigate as much potential for bias as possible. If we feed our models during the training process data that contains biases towards a certain group of observations/individuals/etc., these biases will only be further compounded when the model is put into use, and can drastically affect the results we wish to reach through their utilization.

As a result of my findings, I not only wish to dive further into exploring the potential biases that may be found across a wide range of NLP models, butt would also look to begin training my own NLP models and/or re-training some of the models involved in this study. I am quite shocked to find that of all the models used in this study, distilbert-base-cased, an English-only model, was found to be the most likely candidate for future re-training. While this assumption would need to be further explored through more analysis, I do believe it is plausible that given more data using the Males names I have included in my list in this study, as well as more gender-related text data, would further improve the overall model’s performance. I also believe that a model such as this could eventually lead to an extremely effective model for certain NLP tasks, such as being utilized by companies to scan through candidates’ resumes and applications in the hiring process. With the big push for AI and Machine Learning expansion across the corporate world, and the world in general, ensuring that our models are as effective as possible at completing the tasks we give them is imperative for the growth of this technology.

Sources

Aarsen, Tom, et al. “Sentence-Transformers/STSB-XLM-R-Multilingual · Hugging Face.” *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks*, Association for Computational Linguistics, Nov. 2019, huggingface.co/sentence-transformers/stsb-xlm-r-multilingual.

“Best Stem Jobs | Best Jobs Rankings | US News Careers.” *US News World Report*, 2023, money.usnews.com/careers/best-jobs/rankings/best-stem-jobs.

“Best Non-STEM Majors with High-Paying Salaries | Indeed.Com.” *Indeed*, 27 Oct. 2022, www.indeed.com/career-advice/finding-a-job/highest-paying-non-stem-majors.

Caliskan, Aylin, et al. “Semantics Derived Automatically from Language Corpora Contain Human-like Biases.” *Papers With Code*, 25 Aug. 2016, paperswithcode.com/paper/semantics-derived-automatically-from-language.

Conneau, Alexis, et al. “Unsupervised Cross-lingual Representation Learning at Scale.” *Hugging Face*, 11 Nov. 2019, huggingface.co/xlm-roberta-base.

Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.” *Hugging Face*, 30 Oct. 2018, huggingface.co/bert-base-uncased.

Liu, Yinhan, et al. “RoBERTa: A Robustly Optimized BERT Pretraining Approach.” *Hugging Face*, 1 Aug. 2019, huggingface.co/roberta-base.

“Most Popular African American Baby Names - the Meaning of Names.” *Names.Org*, www.names.org/lists/by-origin/african-american/. Accessed 17 July 2023.

“Most Popular Baby Names in the United States.” *Names.Org*, www.names.org/lists/most-popular/all-time/. Accessed 17 July 2023.

“Most Popular Chinese Baby Names - the Meaning of Names.” *Names.Org*, www.names.org/lists/by-origin/chinese/. Accessed 17 July 2023.

“Most Popular English Baby Names - the Meaning of Names.” *Names.Org*, www.names.org/lists/by-origin/english/. Accessed 17 July 2023.

“Most Popular Latin Baby Names - the Meaning of Names.” *Names.Org*, www.names.org/lists/by-origin/latin/. Accessed 17 July 2023.

Sanh, Victor, et al. “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” *Hugging Face*, 2019, huggingface.co/distilbert-base-cased.

Wang, Haifeng, et al. “ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding.” *Hugging Face*, 2019, huggingface.co/nghuyong/ernie-2.0-base-en.

.