Bias in Language Models

ANDRES REYES
Team Member, George Mason University (CS 484), areyes24@gmu.edu

Siddharthkumar Patel

Team Member, George Mason University (CS 484), spatel53@gmu.edu

Prashant Shrestha

Team Member, George Mason University (CS 484), pshrest@gmu.edu

As technology advances each year, the use and strive of AI (Artificial Intelligence) to ingrain in our lives to make it more meaningful has increased to the extent that we can claim we experience AI every day. Tech companies worldwide have adopted AI in most of their products used by most of the world's population. We see AI in our phones, computers, and even services such as banking. Although AI is a great way to increase revenue, it also brings severe problems. One of many implementations of AI and Machine Learning can be seen in language processing, a field-specific to interpreting our natural language and making sense by the machine to give our language meanings and make it computational ready. One of the most embraced methods to help machines perceive and interpret our natural language is by taking note of the correlation and co-occurrence of words in a sentence. In this report, we visit the main issue with using such natural datasets in training machine learning models. Training models with natural language also propagate any existing similarities with other words. These models are then used in our everyday lives, making biased decisions.

1. INTRODUCTION

Throughout this project, our team will analyze the effects of bias on language models in scenarios such as gender and race to name a few. It is essential to mention the background and motivation for this project. The idea is based on word embeddings, mappings from words to number vectors. The types of methods aim to represent words that incorporate the context in which words get used. The most critical part of that description is the context. Without context, we would not understand the definition of a word we do not know. For example, the words "man" and "human" are distinct, but we understand that they might have the same meaning due to the context. We need to vectorize how words are used to successfully train our models and have this same technique to learn novel words. By placing these words close to each other, the word embedding model would be determined as similar. Algorithms such as GloVe and word2vec use large corpora examples from the human language to train the model with these embeddings. This is where the problem presents itself and why we are motivated to explain the problem and what we believe is the solution.

Using a natural language set is that the model will incorporate natural, cultural biases with the natural dataset. For this project, we are using large corpora from Twitter and Wikipedia to conduct experiments and show how bias can be incorporated into the model, with or without implicit intent.

Before moving forward, it is essential to present the questions produced, which will further pave our understanding of this project. Our first question was whether social and informational platforms such as Twitter and Wikipedia have gender and racial bias? Many people refer to Wikipedia as their source of knowledge since it is the first search result for any word or sentence most of the time. On the other hand, Twitter is a source of entertainment that many people use. Whether information or entertainment, we might constantly be fed bias through these large platforms. Our second question was whether increasing the size or dimension of the corpora that we were going to use to test the GloVe algorithm increases the model's association confidence of the model? Since an increase in the corpora size would also mean a wider variety of views, perspectives, and facts to associate in the model, such a scenario may increase confidence.

2. ANALYSIS/IMPLEMENTATION

To demonstrate the existence and implications of models trained with biased datasets, we used vectorized datasets collected, parsed, and generated from the contents of the two most popular websites - i.e., Twitter and Wikipedia. The datasets can be found on the GloVe website [1]. The exists options of datasets to choose from as they are available to download as a bundle. We needed to select the dimension we wished to test the model with the available dimensions. The questions raised above are yet to be answered. For example, we chose 100 dimensions vectors for Twitter datasets and 50 for the Wikipedia datasets as our initial test. Specific dataset with .txt extension needed to be extracted from the downloaded bundle and converted to a compatible NumPy file, the only format accepted by the model. In addition to the GloVe datasets, we were provided with various word lists. The wordlist's purpose is to check the distinction between the vectorized datasets and the words in the wordlists.

The word embedding model provided to us consisted of various Python files but had only two main entry points, i.e., `findSimilarWords.py` and `weatTest.py.` The `weatTest.py` takes five input arguments, NumPy file, target file #1, target file #2, attribute file #1, attribute file #2. The `weatTest.py` performs the Word Embedding Association Test to detect bias. The resulting output is called effect size, and the output value is bounded by `[-2, 2].`

3. RESULT

Our model yielded for most of the attributes was, as expected, a bias for sensitive attributes such as race, religion, and gender. Therefore, we cross-tested all the related wordlists paired accordingly.

3.1 BIAS IN WORD EMBEDDINGS

3.1.1 COMPARE AND CONTRAST *(Effect size replication)*

We conducted the first run regarding race, namely European Names vs. African Names, against the unpleasant/pleasant wordlist. We did not expect the names to contribute to the bias. However, we were caught off-guard by the effect size. We hoped to get close to 0 although not 0, which means that the attribute does not contribute to the bias. The resulting bias for African and European Names against pleasant vs. Unpleasant wordlist resulted in –1.11 for Twitter datasets and –1.14 for the Wikipedia datasets, as seen in the table below.

Source	Attributes	Targets	Effect size
Twitter, Wikipedia	names_africa, names_european	pleasant, unpleasant	-1.11, -1.14

Table 2. African names vs. European names against pleasant/unpleasant.

The effect size is cross-evaluable, meaning, in the effect size for the source, attributes, and targets from the table, European names are strongly associated with pleasant words, and vice versa with positive (firm) effect size. The effect size computed by our WEAT model, although not as high and the same as the data observed by Caliskan et al., showed prominent similarity. The paper consulted by Caliskan et al. to compare their observation, Greenwald et al., seems consistent with the data we observed. Caliskan et al. used multiple statistically significant techniques such as p-value to strengthen the implications of their findings; they observed highly similar values, and so did we. This means that when African names were checked against the words like pleasant words in our wordlist against the mapped vectors in Twitter and Wikipedia datasets, it returned borderline low associativity to pleasant words.

Source	**Attributes**	**Targets**	**Effect size**
Twitter, Wikipedia	names_africa, names_european	insects, flowers	-0.35, -0.33
Twitter, Wikipedia	names_africa, names_european	pleasant, unpleasant	-1.11, -1.14
Twitter, Wikipedia	names_africa, names_european	positive, negative-words	-1.25, -1.40
Twitter, Wikipedia	names_africa, names_european	art, science	-0.66, -0.08
Twitter, Wikipedia	names_africa, names_european	tech_and_egr, medical	-1.25, -1.41

Table 3. Effect size table for all the target wordlists.

Similar to the observations made by Caliskan et al. and Greenwald et al., the consistent and robust effect size can be noticed when gender or name for male individuals are used to determine associativity towards career vs. family, as mentioned in the table below.

Source	**Attributes**	**Targets**	**Effect size**
Twitter, Wikipedia	names_male, names_female	career, family	1.26, 1.77

Table 3. Associativity strength for male names vs. female names against career and family

The effect sizes our WEAT model yielded seem consistent with the effect size reported by Caliskan et al. and Greenwald et al. for the male name attributes towards career and female names towards family. We created unique wordlists to test our hypothesis besides African and European names computed against pleasantness and unpleasantness of their names and various other attributes and targets permutations. For example, we found a compiled CSV dataset [3] consisting of various majors and fields of study. We later parsed and filtered it into two fields: technologies/engineering and medical. Upon checking the effect size for sensitive attributes such as names and gender, Twitter and Wikipedia were inconsistent to each other.

Source	**Attributes**	**Targets**	**Effect size**
Twitter, Wikipedia	names_male, names_female	tech_and_egr, medical	1.52, 1.85
Twitter, Wikipedia	names_male, names_female	art, science	-0.64, -0.97
Twitter, Wikipedia	names_male, names_female	insects, flowers	1.80, 1.09
Twitter, Wikipedia	names_male, names_female	pleasant, unpleasant	-1.17, 0.06
Twitter, Wikipedia	names_male, names_female	positive-words, negative-words	-0.58, 0.32
Twitter, Wikipedia	names_male, names_female	career, family	1.26, 1.77
Twitter, Wikipedia	gender_m, gender_f	tech_and_egr, medical	0.53, -0.97
Twitter, Wikipedia	gender_m, gender_f	art, science	0.94, -0.97

Twitter, Wikipedia	gender_m, gender_f	insects, flowers	0.77, 1.09
Twitter, Wikipedia	gender_m, gender_f	pleasant, unpleasant	-0.09, 0.06
Twitter, Wikipedia	gender_m, gender_f	positive-words, negative-words	-0.13, 0.32
Twitter, Wikipedia	gender_m, gender_f	career, family	0.53, 0.83

Table 4. Effect size for targets related to gender.

There does seem to be a noticeable change in the effect size, simply observing the names and gender attributes cross-checked against all available target attributes. However, there is a less strong implication of association when gender (pronouns) is involved. Gender attribute, although with more weight to the importance of association compared to the names, seems to have lessened the effect size, confidence in associativity between the attributes and the target.

3.1.2 PROPOSAL, DESIGNS, & IMPLICATIONS

We compiled and tested our existing target attributes with gender, race, or religion but with "region/nation," and the result surprised us with how biased yet factual the results were. Finally, we compiled two separate wordlists of all the countries based on Africa and Europe, and the table displays notable effect size results below.

Source	Attributes	Targets	Effect size
Twitter, Wikipedia	countries_africa, countries_europe	pleasant, unpleasant	-1.50, -0.07
Twitter, Wikipedia	countries_africa, countries_europe	positive-words, negative-words	-1.34, -0.69
Twitter, Wikipedia	countries_africa, countries_europe	computers_and_maths, medical	-0.79, -1.45
Twitter, Wikipedia	countries_africa, countries_europe	insects, flowers	0.98, 1.00

Table 5. Custom wordlist with same targets (high bias from Twitter).

The bias against African countries also exists, given the datasets, even more from Twitter than Wikipedia. Wikipedia is a factual (informational) website, whereas Twitter is a social platform where opinions are widely welcomed, ignoring the factual data, introducing bias but not guaranteeing falseness.

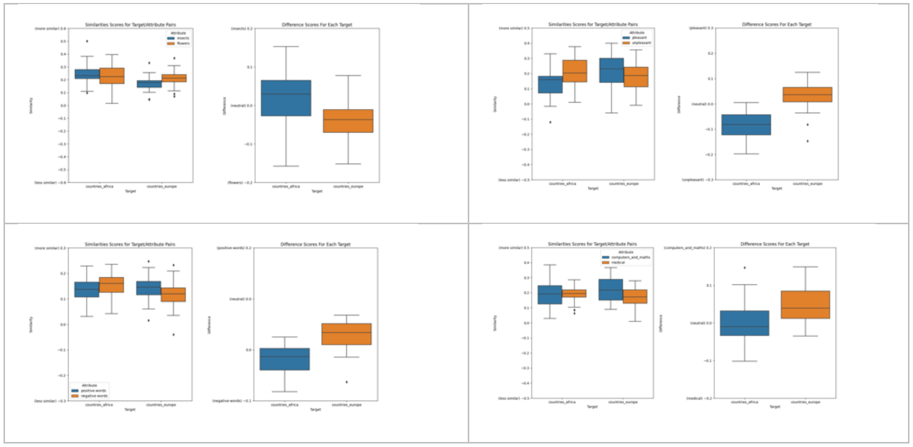

Figure 1. Similarity and Difference plots for countries in Africa and Europe against the target attributes

What we can infer from our observations and findings above regarding the sensitive attributes such as names, gender, religion, and race, is that any datasets with the involvement to/about the sensitive attributes can undoubtedly introduce bias to the model resulting in strong skew on the output generated by the model. A complex question is whether we should ignore the sensitive attributes and feed our model a cleaned and filtered version of the dataset? We believe it truly depends on the context. For example, suppose a context requires the language model to map associations in terms of sensitive attributes. In that case, we cannot avoid making biased decisions or sway our model to make unbiased decisions. We also believe that trifling with the sensitive attributes related to specific context could deteriorate and nullify the facts implied by the context itself, resulting in the void of language models' fundamentals such as co-occurrence. A simple example would be the usage of the word "not" in "not healthy" and "healthy", has two different meanings, the inclusion of the word "not" changes the meaning of the words. It is not so much the correctness of the decision made by the model, but the unexpected noise included in the decision-making process.

According to our questions listed above, we initially produced a hypothesis that said there would be bias in our datasets. Twitter's data is heavily influenced by multiple factors that might deem specific attributes comparison biased. For example, suppose the data is collected from Twitter around a political campaign. In that case, with high possibility, data have a big chance to be heavily biased towards political context, when fed onto a model, the model would learn and replicate the bias from the dataset, and decisions made by the model will be biased towards political terms. This turns out to be, in a way, "supervised" learning rather than unsupervised since we know that bias exists in our dataset. We are indirectly "guiding" the language models (whichever it may be, for whatever tasks/goals it may have) to learn the bias, yet we act surprised when we see the biased decisions made by our model. Based on the results listed above, we can conclude that there is a fair bit of bias throughout these data. However, there can be numerous implications of such bias. For instance, career path programs that use machine learning-based models to decide the suitable field/major for students will pair up masculine names with Technical and Engineering fields more often than feminine names. On the other hand, feminine names would be strongly paired up with medical fields. This example is also backed up by the data results shown above, where male names are strongly paired with Engineering fields. Finally, based on the thorough analysis discussed above, these datasets have strong biases against gender and race, not to mention naming conventions associated with gender.

As per the instructions to include the list of words, the words are, as mentioned above, for re-iteration and to save valuable space in this paper, first 5 African country names – Algeria, Angola, Benin, and Botswana. Similarly, first 5 European countries – Albania, Andorra, Armenia, Austria, and Azerbaijan, to name the first five. We included many words for another custom wordlist (tech_and_egr, medical). However, the first five words from both the wordlists are computer, mathematics, programming, networking, and technologies—administrative assisting astronomy, biochemical, and biological for the medical wordlist.

4. ETHICAL CASE STUDY

We are given the task of reviewing a proposal by a regional hospital. According to the proposal, the regional hospital is looking to improve the system to improve every patient's experience that comes to the emergency room. It seems to be the case that nurses' jobs are too focused on paperwork rather than detecting the severity of each patient. Using natural language processing to suggest severity represents a good intention from the regional hospital. However, they ignore that it could bring other significant problems to the prioritization system they want to implement. For example, using AI to suggest priority for a patient, problems such as biased suggestions or erroneous suggestions can negatively affect the hospital. We discuss this in further detail, including how these problems can violate the ethics of the medical field.

Historically, the nurse position has been dominated by the female gender. The concern here is that any biased decision taken by most female nurses will be propagated into the model. However, the outcome would not change if men dominated the nurse field. When such data is universally valid in human history, it is bound to propagate into the training model that will suggest priority. Such universal data can be seen when comparing flowers with pleasant words and insects with unpleasant words [2]. Therefore, it is vital to bring ethical violations in discussing these concerns. Having a model that learns from potential biased decisions f from nurses violates the ethics of a nurse's responsibility towards their patients. The nurse's role carries many ethical obligations since it deals with another human's life. With this implementation, we rely on an algorithm to follow all nurses' ethical obligations in each decision.

The second concern to consider is human nature; making mistakes is naturally part of our being. The data that AI will be learning comes from nurses who are also human beings. If these nurses made errors while prioritizing patients, this could also be propagated into the model. As shown, along with biased decisions, errors can also be propagated into models. Ethically, allowing a machine learning model to make suggestions on priority would violate the ethical theory of Deontology. For reference, Deontology is derived from the word duty and science of. Deontology judges the morality of choices by criteria different from the states of affairs those choices bring about [5]. Therefore, each decision taken by any individual will have consequences. If we follow the deontology theory, it is morally wrong to allow AI to make decisions hoping that it will not violate the nurse's duties. Also, we are hoping that previous nurses did not violate their duties that would reflect on the corpora used to train the models.

If the regional hospital does not address these concerns, there exists a possibility that this implementation could result in havoc for the hospital. If patients were to find out that an AI is deciding their priority in the emergency room, there would most likely exist an individual who thinks of this as ethically wrong. Deontology stood out because patients have confidence in the nurses to make ethical decisions as it is their duty. Patients could grow anger with the result of this implementation that could cause them to sue the hospital. A patient's well-being is a huge responsibility and should not be taken lightly. Therefore, implementing AI to improve the process for the patient's sound might not be the best implementation.

5. CONCLUSION

The bias perpetuated by the dataset can be seen in the model's resulting output for the effect size. Using WEAT, we measured effect size, but we can observe that if we incorporate the same dataset into language models. However, as a subset of NLP (Natural Language Processing), we can guarantee the inclusion of bias in the model's decision-making process and the decision itself. As humans, we rely on technology more often than we used to a decade ago, especially for knowledge, intellect, and information. If the information accessed through web sources is biased, such information will impact societal cultures, such as how we see different races, cultures, genders, or religions. Such bias in tech will blur our perception of how people depict the world clearly, leading to negative impacts such as gender discrimination and racism.

REFERENCES

[1] J. Pennington, “GloVe: Global Vectors for Word Representation,” *Glove: Global vectors for word representation*. [Online]. Available: https://nlp.stanford.edu/projects/glove/. [Accessed: 13-Dec-2021].

[2] G. M. Sullivan and R. Feinn, “Using effect size-or why the P value is not enough,” *Journal of graduate medical education*, Sep-2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3444174/. [Accessed: 13-Dec-2021].

[3] fivethirtyeight, “Data/college-majors at master · fivethirtyeight/data,” *GitHub*. [Online]. Available: https://github.com/fivethirtyeight/data/blob/master/college-majors. [Accessed: 13-Dec-2021].

[4] R. Goodman, “Why Amazon's Automated Hiring Tool discriminated against women,” *American Civil Liberties Union*, 15-Oct-2018. [Online]. Available: https://www.aclu.org/blog/womens-rights/womens-rights-workplace/why-amazons-automated-hiring-tool-discriminated-against. [Accessed: 13-Dec-2021].

[5] L. Alexander and M. Moore, “Deontological ethics,” *Stanford Encyclopedia of Philosophy*, 30-Oct-2020. [Online]. Available: https://plato.stanford.edu/entries/ethics-deontological/. [Accessed: 13-Dec-2021].

Our video presentation link: https://youtu.be/VJSSCvlRVwI