ARE BRANDS GENDERED? LEVERAGING GENDER BIAS FOR APPEAL AND ENGAGEMENT

ANN Developer Intern, Utkarshini Edutech

CONTENTS

PROBLEM STATEMENT

To analyze if brands leverage gender bias for appeal and engagement.

OBJECTIVES

1. Define metrics to measure gender bias in a sentence.
2. Extract YouTube transcripts of brands targeted at men and women along with other data and select brands with most bias.
3. Perform statistical analysis on these brands' ads and formulate meaningful conclusions.

CONCEPTS

1. GenBit Library

Responsible-AI's GenBit is a python library that aims to measure gender bias in NLP text corpora. The main goal of GenBit is to analyze your corpora and compute metrics that give insights into the gender bias present in a corpus. The computations in this tool are based primarily on ideas from Shikha Bordia and Samuel R. Bowman, "Identifying and reducing gender bias in word-level language models" in the NAACL 2019 Student Research Workshop.

Identifying and reducing gender bias in word-level language models:

We quantify and reduce gender bias in word level language models by defining a gender subspace and penalizing the projection of the word embeddings onto that gender subspace. We device a metric to measure gender bias in the training and the generated corpus.

In a text corpus, we can express the probability of a word occurring in context with gendered words as follows:

P(w|g) = (c(w, g)/Σic(wi, g)) / (c(g)/Σic(wi))

where c(w, g) is a context window and g is a set of gendered words that belongs to either of the two categories: male or female.

The bias score of a specific word w is then defined as:

biastrain(w) = log(P(w|f)/P(w|m))

This bias score is measured for each word in the text sampled from the training corpus and the text corpus generated by the language model. A positive bias score implies that a word cooccurs more often with female words than male words.

We take a fixed context window size and measure the bias scores. We generated bias scores for several context window sizes in the range (5, 15). For a context size k, there are k words before and k words after the target word w for which the bias score is being measured.

GenBit: Measure and Mitigate Gender Bias in language datasets

We create a word cooccurrence matrix across the tokenized input data and, from these counts, calculate conditional probabilities via the maximum likelihood method (MLE) denoted as,

$P(w|g)$ = $count(w, g)$ /$\sum count(wi, g)$/ $count(g)$/$\sum count(wi)$

For each word in the corpus, w, the above formula calculates the probability of it cooccurring with any male-gendered or female gendered word, g, from the gender definition lexicon. Co-occurrence counts between w and g are collected if w and g occur with a pre-defined context window of length c. To add greater importance to words that appear in proximity to gender definition words, we apply back-off weighting so that each co-occurrence count is multiplied by the discount value 0.95*$distance(w, g)$. To avoid non-zero counts, add 1/N smoothing, with N being the number of unique tokens in the dataset. Probabilities are returned as log values to prevent overflow. To quantify the bias measurement scores, we choose two key critical metrics for bias assessment. The method leverages co-occurrence frequency counts and conditional probability as

described above, and iterative benchmarking was performed to validate the final metrics used in the framework.

• Average absolute bias score:

$avg(abs(count(w|gm)/count(w|gf)))$

• Average absolute bias conditional score:

$avg(abs(P(w|gm)/P(w|gf)))$

The algorithmic implementation accepts a list of strings; the length of each element of the list is not a constraint. However, each component of the list may represent text from an entire file, a single paragraph, or a single sentence. The nonempty list constraint is applied i.e., a list should contain at least one element, and that the members of the list are python string types.

Metrics:

1. GenBit Score: P(w|g)
2. Average Bias Ratio:
3. % Female definition words
4. % Male definition words:
5. Frequency of Female definition words
6. Frequency of Male definition words:

2. Independent T-Test

The independent t-test, also called the two-sample t-test, is an inferential statistical test that determines whether there is a statistically significant difference between the means in two unrelated groups. It compares means for two groups of cases.

Null and alternative hypotheses for the independent t-test:

The null hypothesis for the independent t-test is that the population means from the two unrelated groups are equal:

H0: u1 = u2

In most cases, we are looking to see if we can show that we can reject the null hypothesis and accept the alternative hypothesis, which is that the population means are not equal:

HA: $u1 \neq u2$

To do this, we need to set a significance level (also called alpha) that allows us to either reject or accept the alternative hypothesis. Most commonly, this value is set at 0.05.

Unrelated groups, also called unpaired groups or independent groups, are groups in which the cases (e.g., participants) in each group are different.

The independent t-test assumes the variances of the two groups you are measuring are equal in the population. If your variances are unequal, this can affect the Type I error rate. The assumption of homogeneity of variance can be tested using Levene's Test of Equality of Variances, which is produced in SPSS Statistics when running the independent t-test procedure.

This test for homogeneity of variance provides an F-statistic and a significance value (p-value). We are primarily concerned with the significance value – if it is greater than 0.05 (i.e., $p > 0.05$), our group variances can be treated as equal. However, if $p < 0.05$, we have unequal variances and we have violated the assumption of homogeneity of variances.

3. Regression Analysis

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. It helps us to understand how the value of the dependent variable changes corresponding to an independent variable when other independent variables are held fixed.

Regression is used for two primary purposes: 1. To study the magnitude and structure of the relationship between variables. 2. To forecast a variable based on its relationship with another variable.

By performing the regression, we can confidently determine the most important factor, the least important factor, and how each factor is affecting the other factors.

If two or more variables are correlated, their directional movements are related. If two variables are positively correlated, it means that as one goes up or down, so does the other. Alternatively, if two variables are negatively correlated, one goes up while the other goes down.

Causation means that one variable caused the other to occur. Proving a causal relationship between variables requires a true experiment with a control group (which doesn't receive the independent variable) and an experimental group (which receives the independent variable).

While regression analysis provides insights into relationships between variables, it doesn't prove causation. It can be tempting to assume that one variable caused the other—especially if you want it to be true—which is why you need to keep this in mind any time you run regressions or analyze relationships between variables.

In regression analysis, R-squared and adjusted R-squared are statistical measures used to assess the goodness-of-fit of a regression model. They provide insights into how well the model fits the observed data.

1. R-squared (Coefficient of Determination):

R-squared, also known as the coefficient of determination, represents the proportion of the variance in the dependent variable that is explained by the independent variables in the regression model. It is a value between 0 and 1.

R-squared is calculated as:

R-squared = Explained variation / Total variation

The explained variation is the sum of squares of the differences between the predicted values and the mean of the dependent variable, while the total variation is the sum of squares of the differences between the actual values and the mean of the dependent variable.

A higher R-squared value indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables in the model. However, R-squared has a limitation in that it tends to increase with the addition of more predictors, even if they do not contribute significantly to the model's predictive power.

The R-squared statistic suffers from a major flaw. Its value never decreases no matter the number of variables we add to our regression model. That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables. This clearly does not make sense because some of the independent variables might not be useful in determining the target variable. Adjusted R-squared deals with this issue.

2. Adjusted R-squared:

Adjusted R-squared is an adjusted version of R-squared that considers the number of predictors (k) and the sample size (n) in the regression model. It penalizes the addition of unnecessary predictors that do not contribute significantly to the model's explanatory power. In doing so, we can determine whether adding new variables to the model increases the model fit.

Adjusted R-squared is calculated using the formula:

Adjusted R-squared = 1 - [(1 - R-squared) * (n - 1) / (n - k - 1)]

The adjusted R-squared value ranges from negative infinity to 1. Like R-squared, a higher adjusted R-squared value indicates a better fit of the model. However, adjusted R-squared accounts for model complexity and provides a more conservative estimate of the model's explanatory power by adjusting for the degrees of freedom.

Adjusted R-squared is often preferred over R-squared when comparing models with different numbers of predictors, as it balances goodness-of-fit with model simplicity.

Both R-squared and adjusted R-squared are useful measures to evaluate the quality and predictive power of a regression model. However, it is important to interpret them in conjunction with other considerations such as the context of the analysis, the nature of the variables, and the overall significance of the model.

METHODOLOGY

1. Collected transcripts of 10 most viewed Ads of 10 brands targeted at men and 10 brands targeted at women.

2. Calculated gender bias metrics of transcripts of each brand and based on this, selected 3 brands targeted at men and 3 brands targeted at women.
3. Collected the transcripts, dates, views, likes, and comments of all the ads from these brands over the past two years.
4. Calculated gender bias metrics on the transcripts collected.
5. Performed Independent T-Test and Regression Analysis.
6. Formulated inferences and derived conclusion from the obtained results.

INFERENCES

1. From average gender bias metrics

I gave a ranking system where highest value is given 3 score, second highest is given 2 score and third highest is given 1 score. Based on the sum of these scores, 6 brands were selected.

Masculine Brands: Old Spice, The Man Company and Gillette
Feminine Brands: Chanel, Shein and Sephora

2. From Independent T-Test

- It was observed that p-value of number of views, number of comments, GenBit score, average bias ratio, % female words and % male words is less than 0.05.
- Hence, it can be concluded that there is statistically significant difference between the means of the two groups in terms of the above-mentioned metrics.
- That is, there is not much similarity between the Ads targeted at male and female in terms of popularity and gender bias metrics.
- Although, this cannot be said about the number of likes.

3. From Regression Analysis

Independent Variables: GenBit score, Avg. Bias ratio, Frequency of Male definition words, Frequency of Female definition words and all of these multiplied by binary

variable (Binary Variable: 1-Male, 2-Female), Title GenBit score, Title avg bias ratio, Length and Days old.

1. Dependent Variable: Views

- Title GenBit score has –3.3e6 regression co-efficient which says that titles with more gender bias are less likely to get views.
- Title avg bias ratio has 1.4e6 regression co-efficient which says that use of female gender bias words is more likely to generate more views.
- % Female words has 1e8 regression co-efficient whereas % Male words has –3.5e6 regression co-efficient which says that use of female definition words is more likely to generate more views and use of male definition words is more likely to generate less views.
- GenBit score has –3.67e7 regression co-efficient whereas GenBit score*Binary Variable has 1.7e10 regression co-efficient.
- Average bias ratio has 2.9e7 regression co-efficient whereas Average bias ratio* Binary Variable has –1.5e7 regression co-efficient.

2. Dependent Variable: Likes

- % Female words has –4.2e-2 regression co-efficient whereas % Male words has –3.9e-2 regression co-efficient which says that use of female and male definition words is more likely to generate less likes.
- GenBit score has –2.4e-2 regression co-efficient whereas GenBit score*Binary Variable has 4.55e-3 regression co-efficient.
- Average bias ratio has 3.4e-2 regression co-efficient whereas Average bias ratio* Binary Variable has –1.6e-2 regression co-efficient.

CONCLUSION

- There is a significant difference between gender bias in ads and virality of ads of brands targeted at men and women. However, the same cannot be said about the appeal of the ads.
- While Ads with titles containing gender bias generally generate less views, ads with titles exhibiting bias towards women tend to attract more views.

- That is, titles exhibiting bias towards women generally contribute to the virality and controversy surrounding the ads.
- Ads that feature female definition words are more likely to generate higher viewership compared to those with male definition words.
- Ads that feature male and female definition words are less likely to a generate higher number of likes.
- Similarly, ads that display gender bias are generally less likely to generate a higher number of likes.