Using Skip-Gram and CBOW Embeddings to Evaluate Twitter and Reddit for Gender Bias Against a Wikipedia Baseline

John Speaks
jspeaks2@illinois.edu

Abstract

In the digital age, online platforms serve as crucial platforms for communication, shaping societal and communal interactions. Written language, as the primary form of expression in these spaces, must thus reflect societal biases, including those related to gender. This paper investigates gender biases across prominent online platforms (Reddit, Twitter, and Wikipedia) using skip-gram and CBOW embeddings, with a focus on sentiment association. Through systematic analysis and custom evaluation metrics using a corpus of positive and negatively aligned terms, surprising patterns emerge. While skip-gram models reveal balanced alignments between gendered terms and sentiment, CBOW models unveil unexpected biases. Reddit subtly favors positive language linked to men, Twitter exhibits a pronounced preference for positive sentiment alongside bias towards male terms, and Wikipedia depicts a significant skew towards negative sentiment in discussing both genders, with a slight emphasis on male terms. The results are contingent on the comparison of small similarity values, but still make clear the complexity of gender dynamics in online communication and highlight the need for further research to understand and address these underlying biases in order to develop more equitable digital environments.

1 Introduction

In today's world, online platforms and social media are vital for everyday communication, shaping discourse and influencing societal perceptions, and forming specialized communities. Within these digital spaces, language plays a central role, serving as the primary way for people to express their ideas, beliefs, and biases. In recent years, considerable focus has been dedicated to the examination of gender biases occurring in online communication, with many researchers trying to identify causes.

The question explored in this paper is how gender changes the language used in online communities. This research paper endeavors to contribute to this growing field by exploring gender biases prevalent across three prominent online platforms: Reddit, Twitter, and Wikipedia. Reddit and Twitter were both chosen for their roles in active communication and their strong communities. Wikipedia was chosen as a neutral baseline, representing a large corpus of informational digital data created in the online age. Using embedding spaces and similarity measures, the paper explores how the gender of terms used changes the sentiment of the overall langauge used.

Reddit and Twitter are both platforms that are typically thought to be hostile environments, with communities that vary from extreme misogynists to strong feminist communities. It is important to evaluate these communities as a place accessible and to people of all genders in order to ensure the internet is an equitable place. Language use is a good way to perform these evaluations as natural language processing allows for salable analysis, but directly looks at the discourse without abstraction.

When creating the embeddings it was thought that the Reddit and Twitter data would be biased in favor of male terms, but that did not end up being entirely true. When the embedding models were evaluated using a process of comparing the similarity of gendered terms and positive/negative terms, it was not found that the two social media platforms had the strongest bias, and instead Wikipedia had the strongest lean across its metrics.

All of the code and processes used in the paper are detailed in the GitHub repository for the

project[1]. All code was executed through Jupyter Notebooks and interacted with files that existed in the root repository. The datasets are not included in the repository, but are linked to when possible.

This paper will outline a detailed approach to the analysis through expansions of the corpora being used, the methodology used to perform the evaluations of gender bias (through both embedding creation and evaluation), the results and data gathered through the study, the evaluation of those results, and finally possible avenues for future research in the field based on the discoveries of the paper.

2 Background on the Corpora

To explore gender bias on the online platforms Reddit and Twitter, I used several corpora, one for each platform, and a third corpus of Wikipedia data in order to establish a baseline. Additionally, for evaluating gender biases, I used a corpus of negative and positive terms in combination with a small list of terms in male/female pairs.

All of the datasets were standardized to be in lower case and have punctuation removed, but remained in standard English sentences and were not modified into any other format.

The Reddit data[2] comes from a corpus representing all Reddit comments from May 2015. This dataset is massive, with over 54 million comments in the dataset. This is a small sample of a larger several terabyte dataset released by the company, but is still substantial on its own. As a dataset of comments from across the site, it comprised many languages, not just English. For the model in this paper, all the data was used to train the embedding space in order to create the strongest model possible with the given data. The data is made up of comments that are variable in length. Reddit limits the length of comments to under 10,000 characters, which is very large, but the majority of comments are far below this maximum size. In this sample, each comment was an average of 137.76 characters long.

The Twitter data (provided by Professor Jonathan Dunn) used for this paper was a collection of over one million rows of English tweets tagged with their country of origin. Each row was a concatenation of several Tweets with an average character length of 2165.78 characters. Tweets have a limited size of 280 characters, so each line contained approximately ten full Tweets (of length 216.58 characters) which was also qualitatively observed by manually reading a few rows of the data and estimating where one Tweet ends, and another starts. Each individual Tweet was then a comparable length to the Reddit comment data, making them fit for comparison. All of the data was used in training the models for this paper.

The Wikipedia corpus[3] used was a large corpus of English Wikipedia articles that was created as a snapshot of the site on July 1st, 2023. It contains 6,286,775 articles and contains their titles, raw text, and categories assigned by Wikipedia. The data was separated into parquet files A-Z for the starting letter of each article with two additional files for number and other. The Wikipedia data was significantly longer than either the Reddit or Twitter data because each article is large, containing thousands of characters. Wikipedia was chosen as data despite this, because it is a good sample of modern-day academic language which is thought to be unbiased. This provides a good baseline corpus because it is from the same time as the other data, but in a register expected to be less controversial and full of individual biases. Other similar corpora or models created for generalized language either include older data, leading to a different language use than on the modern internet, or include more data from the internet, making the data less clean and of several registers. All the data in the in the corpus was used in training in order to create the most robust possible model.

For the evaluation of the perception of masculine and feminine terms, a dataset of words in sentiment analysis[4] typically associated with positive and negative feelings was used. There are 4783 negative terms and 2006 positive terms in the corpus. This provided a broad range of terms to collect sentiment from relative to gendered terms. This was used in combination with a small set of 32 gendered word pairs like girl/boy

[1] https://github.com/JTSIV1/gender-bias-embeddings-on-wikipedia-reddit-and-twitter

[2] https://www.kaggle.com/datasets/kaggle/reddit-comments-may-2015?resource=download

[3] https://www.kaggle.com/datasets/jjinho/wikipedia-20230701/code?datasetId=3521629&sortBy=relevance

[4] https://www.kaggle.com/datasets/prajwalkanade/sentiment-analysis-word-lists-dataset

or actor/actress. Together, these data were used to evaluate every embedding model that was created using Reddit, Twitter, and Wikipedia data.

3 Methodology for Gender Evaluation

In order to evaluate the corpora for gender bias the following method was used. First, embedding models were trained on each corpus being compared. Next, a unique evaluation metric created for this paper was used to determine how language was being used with respect to both men and women. Each of these two steps will be detailed in the following subsections. The metrics are based on the cosine similarity measure included in the Word2Vec model. All of the values calculated depend on the similarity of gendered terms and terms that carry strong sentiment in either a positive or negative direction. The rationale behind this is that if one gender's terms are more similar to negative terms than the other's, there would be a bias against them. The opposite could be said for positive terms. This metric also enables calculating the ratio of positive and negative similarity within one gender's terms, to see if negative/positive words are used disproportionately with one gender.

3.1 Creating Embeddings

When creating embeddings for each dataset, the data had to be cleaned in the same manner, then was used to train both a skip-gram and CBOW model with the same parameters and vocabulary in order to be constant across all datasets. The vocabulary used for each model was the top 50,000 most common English words from a dataset on Kaggle[5]. This was done in order to have a broad vocabulary that would encompass all of the words needed, but also to be consistent and fair across every model created.

First, for just the Reddit model, the language of each comment needed to be identified as it was the only multilingual corpora used. The Lingua LanguageDetectorBuilder was used to identify the language of each comment, and any non-English comments were thrown out and not included in the training data. For each model, the data was made into all lowercase, and all punctuation was removed. The data was split into sentences according to row in its respective corpus, and each sentence was split into words based on white space. This processing made using a consistent vocabulary and processing ensured that each model was trained identically and had words formatted consistently regardless of the source. This was important because formal article writing like on Wikipedia varies significantly in form from online forum posting like the Reddit data.

For training embedding spaces, a skip-gram model and a CBOW model were created for each dataset, for a total of six models. Each model was trained using the Gensim Word2Vec model with a vector size of 100, window size of 5, and 1 epoch. The vector size of 100 was chosen because it provided a good general-purpose model as the task did not require any more dimensions than that. A window size of 5 allowed the model to see only the relevant parts of the sentence while still getting some context for each word. An epoch size of 1 was chosen due to time constraints on the computing hardware available. Each dataset was so large that it was not feasible to train all six models with more epochs, so just one was used. This is comparable because the datasets were so large that even one epoch across the data created a robust embedding space comparable to ones trained with more epochs.

3.2 Creating the Evaluation Metric

In order to evaluate the models, the metrics created help to compare the average similarity between male and positive terms, male and negative terms, female and positive terms, and female and negative terms. For each male/female and positive/negative combination, the similarity was calculated from each gendered term to each term in the sentiment vocabulary corpus and the similarity was added to an ongoing mean. Because some of the sentiment words were not included in the vocabulary that the models were trained on, it was necessary to check that both words being compared were in the vocabulary before accessing their similarity and including it in the average. These similarity averages could then be divided to create a ratio for comparing the speech towards male and female individuals. For example, the similarity of male/negative terms over the similarity of female/negative terms expresses the rate that male terms appear with negative terms in terms of that rate for female terms. A ratio of 1 would indicate equality.

[5] https://www.kaggle.com/datasets/wheelercode/english-word-frequency-list

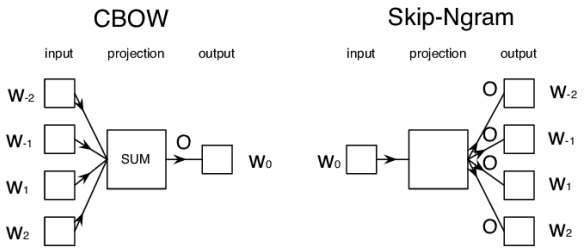

Figure 1: CBOW and Skip-Gram representation (courtesy of Research Gate)

It was important to calculate these metrics for each platform and each of the CBOW and skip-gram models. This is because the CBOW and skip-gram models give us different information about how the terms are interacting. CBOW predicts a word based on surrounding words in the sentence, while skip-gram predicts the surroundings based on the word present. This can be observed above in figure 1. Predicting the word based on surroundings will lead to embeddings that aim to put words that can be used in the same spot in the same embedding space. This is often a more syntactic approach, and similar part of speech words will end up near each other, rather than necessarily being based on meaning. Skip-gram models on the other hand will predict surrounding words. This creates models that will put words used in the same topics together and is a much more semantic approach. This means that CBOW embeddings will have higher similarity scores if the terms can be used interchangeably in part of speech, while the skip-gram embeddings will be more based on relatedness, not necessarily interchangeably.

4 Results

All six models were trained and are available to be seen on the GitHub repository for this paper. They were run through the previously mentioned evaluation metrics and the results from this evaluation are as follows in figure 2.

Model	M_Pos	M_Neg	F_Pos	F_Neg
reddit_c	0.0991	0.0951	0.0828	0.0904
reddit_s	0.7702	0.7573	0.7225	0.7111
twitter_c	0.0376	0.0299	0.0273	0.0210
twitter_s	0.2334	0.2218	0.2192	0.2075
wikipedia_c	0.0318	0.0488	0.0450	0.0646
wikipedia_s	0.2215	0.2159	0.2375	0.2242

Figure 2: Similarity between gender and sentiment terms (c - CBOW and s - Skip-Gram)

We can see that on every platform, the similarities for the skip-gram model are higher than that of the CBOW. This is likely occurring based on the structure of the respective models. Sentiment analysis vocabulary words are likely not to be interchangeable and the same part of speech as gendered terms. Sentiment related words would typically be adjectives while terms referring to individuals of a certain gender are more likely to be nouns. For this reason, the CBOW model will identify them as less similar as a whole as it is much more affected by syntax than the skip-gram model is. Still, the CBOW results are helpful as they provide relative rates of how acceptable it would be to replace a gendered term with a direct positive/negative term which gives valuable information about the language used describing male/female individuals. To make this data easier to analyze, it was further processed to create ratios to analyze the relative similarity between male/female terms, and also positive and negative terms within one gender. They are as follows in figure 3.

Model	M Neg/Pos	F Neg/Pos	Neg M/F	Pos M/F
reddit_c	0.9602	1.0917	1.0527	1.1968
reddit_s	0.9833	0.9843	1.0650	1.0661
twitter_c	0.7937	0.7697	1.4234	1.3803
twitter_s	0.9503	0.9466	1.0688	1.0647
wiki_c	1.5362	1.4349	0.7549	0.7051
wiki_s	0.9748	0.9441	0.9631	0.9328

Figure 3: Processed metric for similarity between gender and sentiment terms (c - CBOW and s - Skip-Gram)

Figure 3 contains values that indicate the relative similarity of different pairs of values from figure 2. M Neg/Pos indicates the ratio of the similarity between male terms and negative terms over the similarity between male and positive terms. F Neg/Pos is the same but for female terms. These two metrics can give us details within one gender about if positive or negative terms occur more similar within that gender. The Neg M/F ratio is the similarity between male and negative terms over the similarity between female and negative terms. Pos M/F is the same but for positive terms. These ratios give us details about the relative similarity between genders to negative and positive terms and are most closely related to the aim of this paper.

Most of the ratio values for the skip-gram models hover around a value of 1. This indi-

cates that in all of the models, terms relating to the male/female genders appear very similarly to each other in terms of closeness to negative and positive terms. Within each gender also, the similarity of negative and positive terms in similar. This indicates that each gender is not talked about more positively than it is negatively or vice versa as modeled with a skip-gram approach.

Using a CBOW approach, the results are very different. The Reddit model hovers closest to 1 in all values, but is still further than any of the skip-gram models. The M and F Neg/Pos columns are very similar to 1 meaning that within each gender the rate of negative and positive terms is likely similar, but for Neg and Pos M/F, male terms are 1.05 times as similar to negative terms and male terms are 1.20 times as similar to positive terms. This indicates that men are evaluated more with words associated with sentiment analysis as a whole, but also even more when it comes to positive terms. In the Twitter model, both M and F Neg/Pos are below 1. This indicates that both genders experience more positive sentiment than negative sentiment. The male value is slightly higher, indicating a slightly higher use of negative terms for men. Looking at the Neg and Pos M/F, the values are significantly above 1. Male terms are 1.42 times as similar to negative terms as female ones and 1.38 times as similar to positive terms. This indicates a strong lean towards male terms when in terms of sentiment analysis words as a whole, with a stronger lean for negative terms. The Wikipedia data was almost opposite to that from Twitter. The M and F Neg/Pos values were significantly above 1, indicating that for both genders, individuals are talked about significantly more negatively. The value for male terms is higher at 1.54 (compared to female terms at 1.43) indicating a stronger negative skew for men. The Neg and Pos M/F terms are both significantly below 1, meaning that the denominator of the ratio is larger, and women are talked about more with sentiment words as a whole. The Pos M/F ratio is slightly lower than the Neg M/F indicating more of a skew towards female terms.

5 Evaluation

The results gathered from this experiment are far different than expected. In the skip-gram model, ratios were very close to 1 while the most change was observed in the CBOW model. This indicates that male and female terms are surrounded by positive and negative terms similarly, but that the rate of interchangeability between them varies.

The ratios for the Reddit data are closest to 1 indicating that the Reddit comment data is most equal to men and women in terms of sentiment. This was the expected behavior of the baseline Wikipedia. Twitter showed a strong preference for positive sentiment as a whole across gender. This was unexpected, because the reputation of social media and expeccially Twitter is that it is a very hostile environment and could even be considered 'toxic.' Additionally, Twitter showed a preference for men for both positive and negative sentiment, with an even higher value for positive. This could indicate a slight bias towards male terms as they are talked about more positively than female terms as a higher rate that their higher negative term similarity. The Wikipedia data had the most drastic biases across the data. Both male and female terms were used far more similarly to negative terms at 1.53x and 1.43x respectively. This has a slight bias with male terms being closer to negative terms. Also, both Neg M/F and Pos M/F had a strong lean toward female terms with male terms being 0.75x and 0.71x as similar respectively. This indicates a slight bias against women with negative terms being used at a higher proportion with female terms than the ratio for positive terms. A qualitative summary of these results can be seen below in figure 4.

Platform	Pos/Neg Bias	Male/Female Bias
Wikipedia	Negative	Female[6]
Reddit	–	Male[7]
Twitter	Positive	Male[8]

Figure 4: Qualitative summary of results 5 - Women were talked about more with both positive and negative terms - even more with positive. 6 - The skew is slight but significant with men having a proportionally higher positive similarity. 7 - Men are talked about more with both negative and positive terms - even more with negative terms

The increased use of negative terms on Wikipedia is likely due to the nature of Wikipedia articles. The articles on the site are much more likely to be about negative topics as a whole, because negative topics and events are more likely to be notable if they are negative. This goes for the many people who have articles about them

on the site. For example, every cruel dictator will have an extensive article detailing their crimes and immoralities, but most good leaders will only have a short article about their life.

Both Twitter and Reddit differ from Wikipedia in that they have a far larger pool of people generating content, and they host countless communities that each talk about and do different things. This can make it difficult to know at face value what is going on. While Reddit and Twitter both have a reputation to the public of being gross or unpleasant places to be on the internet, they have many communities that would be considered positive to the public, but may just not be as vocal. Further exploration into these platforms would be needed to discover the source of the positive bias and male gender bias for twitter and the slight gender bias for Reddit.

A possible issue with this evaluation comes from the original magnitude of the similarity values as seen in figure 2. The values for CBOW similarity averages were all very low (<0.1). This means that even small differences in similarity could lead to large differences in the observed ratios in figure 3. Still, this could be significant. To explore the differences in gender, what matters is not necessarily the absolute level of similarity, but that relative to words relating to the other gender being observed. To understand the way women are talked about, it is necessary to understand how other genders are talked about first to be able to make a comparative judgement. For this reason, the results are valid, but it is unclear the scale on which they are. More testing should be done on this method of evaluation and its derived metrics to understand the true scale and significance of these differences.

6 Future Research

This paper creates many opportunities for future research as, with such a large topic, many smaller yet still significant questions are left unanswered.

The primary topic of research left to be explored, is the origin of these values within the data. More research should be done exploring the causes of the bias that was observed in the data. Communities within each platform should be explored to create a more detailed picture of the situation, not just an overall general view of the site.

Another avenue of research could be into Wikipedia data and exploring the cause of the large bias observed. Given that there is such a bias, this would indicate a large issue with the site. With women being talked about far more in the context of words related to sentiment, this raises a question of why they are not being portrayed with a more objective and non-emotion related terminology. Men already are disproportionately represented in our recorded history (which is mirrored by Wikipedia), so adding less objective writing about women is even more problematic.

A final important topic for research is the validity of these measures. Is a ratio of similarity measures an appropriate way to determine a skew in the data along the lines of gender? If yes, then the methods used in this paper would probably require adjustment in terms of scale. Additional measures of significance are needed to determine the impact of the metrics used. It is unclear how important differences in the ratios really are without a measure of significance. This research could be done using a corpus of data that is labeled with gender biases. With data known to have this skew, the same methods could be used to evaluate the bias and then the ratios generated by that data can be used to scale the output of these methods for use with unknown data.

As a whole this paper provides a good foundation for future more specified research. With the discoveries of these gender biases it is important to explore their implications and sources so that going forward we can work to eliminate them.

7 Conclusions

In summary, this paper aimed to dissect gender biases present within online platforms, focusing particularly on Reddit and Twitter, with a baseline of Wikipedia. Through a systematic methodology using corpus analysis through CBOW and skip-gram embeddings and custom evaluation metrics based on Word2Vec similarity, the study aimed to unravel underlying patterns of language usage and sentiment association concerning gendered terms.

The outcomes were surprising and found interesting patterns in the data that differed from what was originally expected. Notably, the skip-gram models showed a balanced alignment between gendered terms and sentiment, indicative of a more fair and even portrayal of men and

women. By contrast, the CBOW models exposed nuanced biases, with Reddit subtly favoring positive language associated with men, Twitter displaying a pronounced preference for positive sentiment overall alongside a notable bias toward male terms, and Wikipedia depicting a significant skew towards negative sentiment in discussing both genders, with a slight emphasis on male terms.

This paper generated new information regarding the way that language differs based on gender. The similarity between gendered terms and ones associated with sentiment concretely varied based on the gender of the terms being compared. This provided an answer to the question at hand and provided more insight into how language is affected by gender in different online communities.

The ratios of average similarity that were calculated were based on low similarity scores, thus some of the metrics may have been more or less extreme than the bias present in the platforms. With future research these methods should be quantified and made more robust.

This research offers valuable insights into online platforms and highlights the need for further exploration in the field. Future research endeavors should delve deeper into the cause of these biases, delve into community-specific research within platforms, and verify and analyze the validity and significance of the evaluation metrics employed in this paper.

Ultimately, this study addresses the first step of resolving online gender bias: identifying the problem. By studying and evaluating online communities, we can try to foster more equitable and inclusive digital environments.