Gendered Language in Job Postings: A Contextual Embedding Analysis using BERT

Katherine Ann Sick
Dep. of Computer Science and Linguistics
University of Illinois at Urbana-Champaign
kasick2@illinois.edu

Abstract

This study investigates the presence and impact of gendered language in job advertisements using contextual word embeddings derived from BERT. Drawing on a dataset of over 3 million LinkedIn job postings, we quantify the gender bias in job descriptions by comparing extracted keywords to prototypical male and female embedding centroids. The centroids in question are based on already-established gender coded vocabulary from previous experiments that have studied similar effects. By calculating a bias score for each posting, we can evaluate how strongly its language aligns with stereotypically masculine or feminine traits. We then analyze how these bias scores correlate with job-level features such as salary, experience requirements, and work type. Our results reveal a significant skew toward masculine coded language across the corpus, suggesting that gender bias remains prevalent in modern job postings. This work contributes to ongoing efforts in fair hiring by using an embedding-based method for detecting and analyzing gender coded language in employment contexts.

1 Introduction

Gendered language in job advertisements can subtly shape applicant behavior, and possibly cause harmful stereotypes to be enforced when it comes to the hiring process. There have been multiple different works done that could suggest that word choice in job postings can reinforce those occupational stereotypes, and contribute to gender disparities in recruitment for jobs. Despite increased attention to inclusive hiring practices as of recent, the presence and consequences of gender-coded language can still be difficult to detect at a large scale.

In this experiment, we conduct a corpus-based analysis of job advertisements to investigate the relationship between gendered language and certain job attributes, such as salary, experience level, and work type. Using a database of over 3 million LinkedIn job postings, we quantify gender bias using contextual word embeddings derived from BERT.

Using the gendered vocabulary decided by previous studies of this topic (Gaucher et al. (2011)), we constructed prototypical "male" and "female" embedding centroids. We then use these centroids to measure the relative alignment of each job's language to those stereotypically masculine or feminine traits. From here, we analyze how these bias scores correlate with other features of the job postings that were provided in our dataset.

By using contextual embeddings from a large transformer model like BERT, we are able to go beyond surface-level word frequency counts and assess how words are used in context, which provides a more flexible measure of gender encoding in text. This methodological shift allows us to revisit existing theories of bias with more data and new methods, potentially uncovering patterns that older techniques might have missed.

The central question that we want to answer through this study is: Is there still gendered language bias in job postings, and if so, to what extent can this language be predictive of job characteristics? It is true that earlier studies provide much evidence of these gendered words and biases occurring, however I want to see if this still prevalent in the current job market today.

Our approach here allows us to evaluate gender bias not only from a linguistic perspective, but also in relation to the actual attributes we see in job postings. We will be able to use this data to answer our question succinctly, and make further predictions based on our results.

2 Problem Definition

Language used in job advertisements can subtly reflect societal biases, including gender stereotypes, which may influence how individuals perceive and respond to employment opportunities. It can also cause harmful or negative effects on those who are represented by these stereotypes. Previous research has shown that even seemingly neutral word choices can carry gendered associations, potentially affecting applicant behavior and contributing to occupational segregation.

Understanding how gendered language manifests in job postings and whether such language correlates with factors such as salary, experience requirements, or job type is crucial for improving fairness and inclusivity in hiring practices.

While prior work has established that gendered wording can affect job appeal and applicant selection, much of this research could possibly be outdated or done with smaller scale studies. Its could be the case that these methods fail to capture the contextual nuance of how words are used in modern job postings, and they may overlook evolving language patterns across industries or job levels. Addressing these issues is essential for deepening our understanding of these gendered biases.

This study investigates whether job advertisements on LinkedIn exhibit these systematic gender biases in their word choice and whether this bias correlates with other job attributes.

For our purposes, we represent job posting keywords as embeddings using a BERT model. We then compare these embeddings against gendered centroids derived from word lists that are associated with stereotypically male or female terms. By computing the relative similarity of job language to each centroid, we can obtain a measure of gender bias for each posting. This approach allows us to study the relationship between biased language here, and actual characteristics of the job on a large scale.

3 Related Works

The presence of gendered language in professional contexts has been studied multiple times in different subjects. A study by Gaucher et al. (2011) showed that job advertisements for male-dominated fields usually contain a disproportionate amount of words associated with agency such as "competitive" or "dominant", while postings in female-dominated domains more often include supportive descriptors such as "supportive" or "interpersonal". Their research found that such language not only reflects occupational gender stereotypes, but also discourages women from applying to roles framed with masculine-coded terms.

Other studies have further explored these concepts, investigating the psychological and behavioral effects of gendered words on applicant pools, resume evaluations, and hiring outcomes. Some hiring platforms have even begun incorporating gender de-biasing recommendations in job ad templates, but it seems that there is still biases present in many of these platforms.

More recent studies have begun exploring the use of word embeddings and machine learning models to detect this gender bias in text, offering a more in-depth understanding of how gendered language functions beyond predefined word lists. For example, researchers have used models like Word2Vec or GloVe to identify biased associations in professional language. However, these approaches often lack contextual sensitivity. Our work contributes to this area by using BERT-based contextual embeddings, which allow us to account for the meaning of words in specific contexts. This approach enables us to explore how this bias aligns with certain job attributes based on context as well.

But in general, our study in this paper keeps the ideas from these previous works in mind and analyzes how gendered language correlates with the characteristics of employment opportunities. While these previous studies do a good job of exploring this topic, many of them are not as recent, and through this paper we will get to see if those findings are still present.

4 The Dataset and Corpus

For the data we are using in the experiment, we chose the LinkedIn Job Postings dataset, a large corpus of over 3.3 million job advertisements scraped from LinkedIn and made publicly available through Kaggle. This data was published by user Arsh Koneru, and spans a wide range of industries, geographic locations, and experience levels, and includes both textual and characteristic information about each job.

Each job posting contains multiple fields, including a unique job id (to identify each separate posting), job title, full job description, a skills summary, salary information, work type (full-time, part-time, etc.), and experience level. There are more data fields provided such as company name, posting time, location, post views, etc. however they are not important for our purposes.

Focusing on specific fields of this data though, we have a perfect basis for our experiment of studying bias in hiring posts. It provides access to the language used by employers across a diverse set of job categories, and gives us many characteristics of these posts to work with. Specifically some of the most helpful fields include salary, experience level, and job type which enables us analyze the relationship between the linguistic features we find and these other aspects of employment.

This dataset contains all of the data we would need to compare job characteristics after we discover gendered word embeddings in the text. Due to this, we won't need any other datasets to aid in our experiment. However, there is still some preprocessing that we need to do for this data to be usable. We will cover our steps for this in the next section.

5 Processing the Data

Before we can actually start training our models, we have to clean and process our corpus, and make sure we have a good dataset to work with. So, to prepare our LinkedIn job postings dataset for analysis, we first isolate the fields that are most relevant to our experiment.

The original corpus that we downloaded had over 30 different columns of data. This is of course much more data than we actually needed, and could cause our later processes to be much slower if we keep them in. So in our "cleaned" data file, we only retain the columns with the job ID, title, description, salary, experience level, work type, location, skills summary, and remote status. This amount of features, while heavily cutting down on the original amount that we had, allows us access to all of the key information we need to capture both the linguistic content and key job attributes in our analysis.

Then, after we isolated only the columns that we needed, we filter the data to remove rows with missing or extremely short descriptions. This is because these entries are unlikely to contain meaningful linguistic information, or at the very least will not be informative as other longer descriptions. Specifically, we exclude postings with missing values in the description field and drop any descriptions shorter than 100 characters.

On top of removing rows with unimportant data, this also allows us to reduce the size of our dataset to make for a shorter and more efficient runtime later in our experiment. In total, removing these rows of data allowed us to reduce the size of our dataset from over 3 million job postings down to less than 1 million postings.

From here, one of the final things we have to do is clean the description text for each of the postings. Seeing as these descriptions are what we will be training BERT with, we need them to be as uniform as possible to get the best results that we can. To do this, we applied a couple different strategies.

First, we converted all characters to lowercase to ensure that the same words with different capitalization's will not be counted as separate words and possibly skew the data. Then, we stripped all punctuation and removed any extra whitespace. For the purposes of our experiment, punctuation is not helpful in analyzing gender bias in these texts, so it is important to remove these to improve our results. This ensures consistency across all of the entries and reduces noise that may otherwise later interfere with our keyword extraction and embeddings.

Finally, after we finished processing the corpus, we saved the cleaned dataset to a new file for us to use in later stages of the analysis. The final corpus is reduced to a smaller and more concise version of our original dataset, while still containing sufficient textual content for extracting our representative keywords and computing gender bias scores.

6 Methodology and Experiment

For our actual experiment, to quantify and analyze gendered language in these job advertisements, we developed a process combining keyword extraction, contextual embedding generation, and bias scoring. Our methods build on prior literature (specifically (Gaucher et al., 2011)).

6.1 Gender-Coded Word Lists and Embeddings

For our experiment to work properly, we first need to create our gendered centroids that we want to base each of our "bias scores" off of. To do this, we used the publicly available lists created in the paper by Gaucher et al. (2011). In this paper, Gaucher provides two lists of adjectives that are commonly associated with male and female stereotypes. In the "male" list, we see words such as "dominant", "analytical", "aggressive", etc. Whereas in our "female" centroid, we see words such as "affectionate", "nurturing", and "pleasant". Since these words have already been pre-labeled into these two categories, we can use these to come up with our gendered centroids.

To do this, we embed these words using bert-base-uncased, which is a transformer model from the Hugging Face Transformers library. For each word, we generate an embedding by placing it in a neutral sentence ("This is word.") and then extracting the hidden state of the CLS token. We then compute a male centroid and female centroid by averaging embeddings of each gendered word list.

6.2 Extracting Keywords from Postings

In order to compute a "bias score" for each of the entries in our corpus, we first have to identify the most informative words from each job description. For this we used scikit-learn's TF-IDF vectorizer. Specifically, we created this vectorizer with a maximum vocabulary size of 10,000 and English stopword filtering.

Then, for each data entry, we extract the top 10 keywords that had the highest TF-IDF weights to capture the terms that are most representative of the content in the posting.

6.3 Computing Bias Scores

Now that we have our gendered centroids, and the top informative words from each job description, we can finally begin to compute the actual bias scores for each entry.

For each keyword extracted from a job posting, we calculate its gender bias score as the difference in cosine similarity to the female and male centroids. A positive score indicates greater similarity to feminine-coded language, while a negative score suggests a greater similarity to masculine-coded language. The bias score for an entire job description is computed as the average bias score across its top keywords.

When initially calculating and storing this data, we ran into a problem with the run time being much too large and taking too long for us to reasonably run it. This is because initially we were weren't caching our scores, and calculating it for every single word. However, to ensure scalability and efficiency, we changed this to pre-compute and cache embeddings for all unique keywords so that we could reuse them during scoring. This reduces redundant computations and significantly accelerates processing time.

By the end of this process, we store the bias information for each post in our corpus so that we can use it to visually compare with the other attributes we have in the dataset.

7 Results

Now with all of our collected data fully stored, we can see what the distribution is and use our results to explore our leading question.

7.1 Bias Score Distribution

Based on the bias scores that we calculated, it actually appears that there is a large skew of job postings that are masculine encoded. The full distribution of our corpus's scores are shown below in figure one.

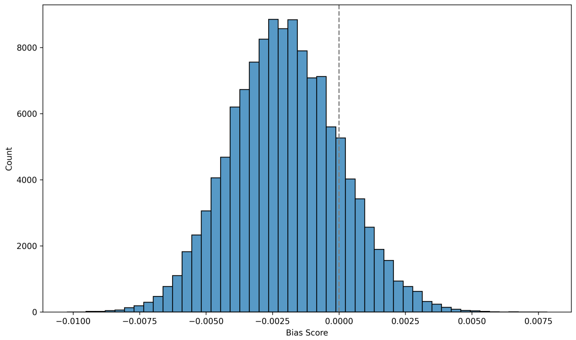

Figure 1: Distribution of Gendered Bias Scores in Job Postings

As discussed in earlier sections, a positive score indicates a female leaning gender encoding, whereas a negative score indicates a male leaning gender encoding. In figure one, the dotted vertical line indicates a score of 0, where a job posting is neither female nor male leaning. In this case, the line splitting the posts between these two encodings shows that there is an extremely heavy skew towards male encodings.

The data as a whole generally follows a bell curve with its scoring, however the maxima of this curve occurs when there is still a decent bias towards masculine language. The cutoff for feminine encoding postings doesn't occur until almost the 75 percent point of the curve, which leaves much less job postings in this category.

7.2 Top Keywords per Category

The next thing we observed is the top words in each category that contributed the most to the calculated bias scores.

In table one below, we display the most strongly masculine and feminine associated terms that were found in our dataset contributing to the bias scores. These are specifically the top ten words for each category which are representative of our data in each centroid as a whole.

Masculine	Feminine
obtain	warmth
optimally	caring
independently	comfort
generate	nice
assume	kind
holds	warm
satisfy	touch
solves	caregiving
determines	kindness
achieved	compassionate

Table 1: Top Masculine and Feminine Coded Words by Bias Score Contribution

From this table, we can see that these top contributing words definitely lean into their stereotypes. The top feminine words are extremely similar to the original biased words that we used to make our centroid, with some of them being exactly the same.

The male words are not as similar to their original centroid, however they still tend to lean towards "dominant" or agency words that are more associated with masculine traits. They are certainly a bit more vague than before, but this is likely due to the sheer volume of postings that fall into this category. In order to be representative of so many texts, the words that are in common are more general. On the other hand, for the feminine coded descriptions, since there are less total entries, the words are more specific.

7.3 Comparing with Salary

When comparing our scores with the salary attribute, we first normalized our salary data to make it easier to visualize in our graph. Again in our plot, we use the dotted vertical line as a reference for the split point between male and female bias encodings. We can observe this distribution in figure two below.

Although in normalizing our data, much of the salary is around the same level, we can still see the distribution of the higher salary ranges. There is not as much difference in the vast majority of postings as expected, however we can still see that most of the highest salaries belong to male encoded job postings.

There are only four exceptions of female encoded postings having higher that average salary values, whereas there an many more for male encoded ones. Additionally, the ones with those negative scores are much higher than those with positive scores. This suggests at least a small bias where jobs associated with masculine stereotypes having high salaries and wages.

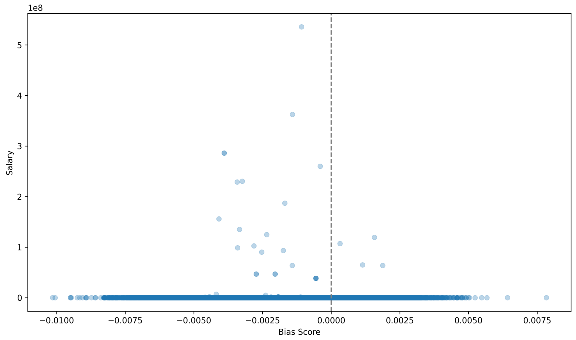

Figure 2: Distribution of Salary vs. Bias Scores in Job Postings

7.4 Comparing with Experience Levels

The next attribute we compare our scores to is the experience level of the job. This allows us to see what experience levels are most commonly associated with which gender based on their job posting descriptions. We can see the distribution in our corpus as shown below in figure 3.

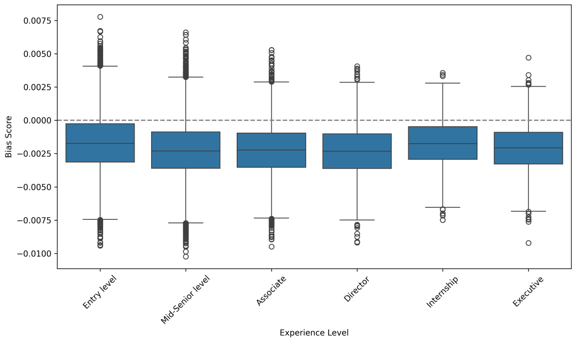

Figure 3: Distribution of Experience Level vs. Bias Scores in Job Postings

In this figure, since we are using box plots to show our data, the dotted line is now horizontal to show the male vs. female bias score split. For this graph, we see a fairly large distribution of data for each type of work, however the middle parts of each plot falls on the side of male bias.

This would indicate that all experience levels of jobs are biased towards male descriptions, with "entry level" being the closest to neutral. On the other hand, the experience level leaning the most towards male bias is "Director".

While there is evidently bias present here, there is also likely a skew from the ratio of postings between male and female descriptions. With there being so many more male coded postings, even if the ratio of experience levels in female encoded texts is higher than the ratio in male encoded texts, we will not see those specific values presented here.

7.5 Comparing with Work Type

Finally, the last attribute we want to compare our bias scores with is work type. This includes categories of full time, part time, internship, contract, temporary, volunteer, and other. We can see these distributions in figure four below where again, the dotted horizontal line shows the male vs. female bias score split.

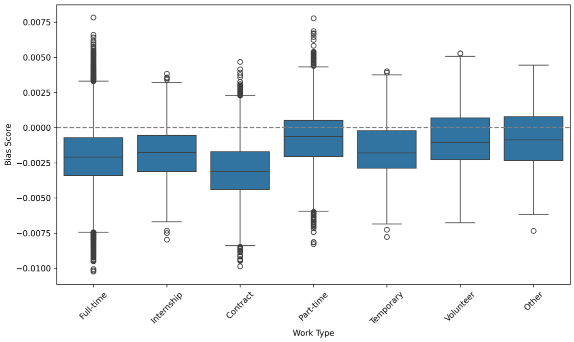

Figure 4: Distribution of Work Type vs. Bias Scores in Job Postings

The work types we see here all tend to lean on the side of male-bias similar to the trend we see when comparing experience levels. There is some variation with our box plot ranges, but in general it seems that we have a general male bias for all of our data here.

We do still want to keep in mind though that this dataset is subject to the same issue we see in figure three. There is likely a skew from the ratio of postings between male and female descriptions.

However, what's interesting here is that even with that skew, we see a much more significant lean towards female-encoded language in three of these categories where the third quartile is actually on the positive side of the scale. These three categories are part-time, volunteer, and other. This would suggest that there is a much larger bias in those three work types considering the ratio of male and female coded postings.

8 Discussion

Our findings indicate a pronounced skew toward masculine-coded language in job postings, even in recent data from LinkedIn. This aligns with earlier studies such as Gaucher et al. (2011), suggesting that despite growing awareness of inclusive hiring practices, subtle linguistic biases still persist in these environments.

One possible explanation is that masculine-coded language continues to dominate in traditionally male-dominated industries, such as technology, finance, and engineering, which may be disproportionately represented in our dataset. Additionally, companies may unconsciously adopt language that signals competitiveness or assertiveness which is generally thought to be masculine coded language.

It seems though, that a lot of our data contains enough masculine encoded language that it causes our actual results to be skewed. What we see here suggests that there is heavy male bias present, however there could be many explanations for this. There is also the possibility that if we changed our parameters to narrow down the words that are accepted for our male centroid, we would see different results.

These patterns raise concerns about the unintended effects on potential applicants, especially women, who may be less likely to apply to roles framed in a more masculine context. Our analysis further highlights how these issues can be present in these professional spaces.

Another important consideration is the role that embedding-based methods play in detecting bias. While contextual models like BERT allow us to capture more nuanced representations of language, they also inherit and potentially intensify biases present in the data they were trained on. This raises the possible question of whether our measurements are reflecting real-world patterns or biases of the model itself. We attempted to circumvent this by using centroids and bias scores, but it could still possibly have effected our results.

Future work might compare multiple embedding models on our corpora to better isolate these variables. In either case, our findings do suggest that embedding-based approaches can be effective, but we still need to keep in mind possible limitations such as this. We discuss this more thoroughly in the next section.

9 Limitations

While we were able to thoroughly explore our leading question through this study, there are several limitations that should be considered.

First, our method for measuring gender bias relies on pre-defined word lists from prior literature, which may not capture the full nuance of current gender-coded language. The study that contributed these word lists is from 2011, which has the possibility of being outdated. The cultural connotations of words evolve, and terms once coded as masculine or feminine may no longer hold the same associations.

Second, BERT embeddings are context-dependent, but our analysis embeds each keyword in a generic sentence. This limits the model's ability to fully capture the context specific meaning of a word within a job description. However, for long lists of words, it is difficult to generate sentences with good contextual support for each entry. To get the best results with this, we would have to manually create these sentences, which would take too much time on a larger scale. For our smaller experiment, this would be too inefficient for our purposes. With more time and resources, this may be a good way to improve the quality of our results.

Additionally, while our dataset is large, it only contains data from LinkedIn, and may not generalize to other job platforms or regions. Other websites such as Handshake or Indeed may have different types of descriptions even for the same jobs, but this experiment does not consider that data. Job descriptions from small businesses or international job boards may also differ significantly in tone and content.

Overall, there are multiple limitations that we faced in this experiment, but were were still able to produce quality results to analyze. While addressing these limitations could certainly improve quality, it would also take much more time and could become less efficient in the process.

10 Future Work

There are several possible avenues for future exploration on this subject. If we were to investigate this topic further, we could gain even more valuable insights using other methods.

First, it would be interesting to investigate if these same biases exist in other similar job postings websites. As we touched on lightly in the previous section, this experiment only contains data gathered from LinkedIn, and not any other sites that have the same or similar purpose. It is possible that if we used data from these other websites instead, or in tandem with the data from LinkedIn, we would find drastically different results. This would allow us to consider possible biases between social media sites on top of biases that are present in the job postings themselves.

It would also be valuable to investigate how gender bias varies across industries, regions, and job levels. Pairing our bias scores with applicant data such as who applies to which jobs could reveal whether or not gendered language truly affects behavior at scale.

Another possible avenue we could explore is whether the bias we observe has changed over time. The dataset that we used in this paper is fairly recent (2023-2024), but it could provide valuable insight to see if running the same experiment on corpora from different years would have varying results. The study that we used for creating our centroids was a very similar experiment, however it was conducted back in 2011, so it is not very recent. There are likely other studies as well that explore same or similar topics that we could use as a basis for tracking bias overtime.

Lastly, it could be interesting to apply this bias study to other types of bias that could appear in job descriptions. This could include biases in race, disabilities, age, and other categories that have stereotypical language associated with them. For the study of gender biases however, there are some fairly clear examples of language that is associated with these stereotypes. This may not be the case so explicitly for other categories of bias.

11 Conclusion

This study has demonstrated that gendered language remains a prevalent feature in job advertisements and that its presence can be effectively analyzed using contextual word embeddings such as BERT. By constructing gendered centroids and comparing job-related terms to these embeddings, we were able to quantify possible linguistic biases associated with gender.

Our results show that words commonly associated with male coded language tend to appear more frequently in higher paying and full time positions, whereas female coded language often correlates with lower compensation and part-time or caregiving related roles. These findings not only highlight the persistence of gender stereotypes in professional recruitment but also show how methods such as this can identify these potential biases.

This work allows us to contribute to ongoing conversations about fairness and inclusivity in hiring. The insights gained from this analysis could be helpful in preventing biased phrasing. Ideally, organizations would be inclined to revise their job postings and promote more fair descriptions of employment opportunities.

Additionally, while our focus was on gender, the methodology could be extended to study other dimensions of bias such as race, age, or disability. There are also plenty more opportunities to explore other veins of our gender bias question with the data we have here.

Overall, this research emphasizes the critical role of language in shaping workplace diversity and calls for continued efforts to address bias at every stage of the recruitment process.

References

Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. https://ideas.wharton.upenn.edu/wp-content/uploads/2018/07/Gaucher-Friesen-Kay-2011.pdf.

Arsh Koneru. 2024. Linkedin job postings (2023 - 2024).