Pretraining Data Gender Biases

Kanika Selvapandian
Carnegie Mellon University, Heinz College
kselvapa@andrew.cmu.edu

kanikaselva/genderBias

Abstract

Biases in training data could translate to increased risks of perpetuating stereotypes and perceptions of society and, hence, influencing downstream AI models in biased ways. This study examines gender bias in Wikipedia content, beginning with a general comparison of male and female representations. This analysis showed the presence of strongly gendered words and differences in domain concentration. We then focus on occupation-gender patterns, particularly the arts, where a similar trend was observed.

1 Introduction

Large language models (LLMs) and other foundation models are trained on vast pretraining datasets drawn from web text, books, and other large-scale data sources. Therefore, biases present in these datasets get reflected during the deployment of these models. One such common bias is gender bias. In fact, research shows that LLMs are 3 to 6 times more likely to place users in an occupation based on gender stereotypes, and hence amplify this bias that is different from the ground truth, but rather grounded in perception [Kotek et al., 2023]. One such large-scale data source is Wikipedia biographies that are widely used in Natural Language Processing (NLP). This analysis aims to quantify gender bias using normalized pointwise mutual information (nPMI).

2 Methodology

2.1 Pre-processing

For the dataset, a processed dataset that included randomly sampled 100,000 biographies that was stratified by gender was used [wik, 2026]. The dataset had the columns *text*, *person*, *gender*, and *split*. For this analysis, only the columns *text* and *gender* were used. For preprocessing, the text was first tokenized using `text.lower().split()`. English stopwords were removed using NLTK package, and all punctuation defined in `string.punctuation` was filtered out. Tokens with a total corpus frequency of less than 10 were ignored to reduce noise.

2.2 nPMI definition

Normalized pointwise mutual information (nPMI) is a metric used to measure the strength of association between two events or tokens to identify words that co-occur more than expected [Watford et al., 2018]. It is normalized to a range between –1 and 1, where -1 means a complete negative association and vice versa. Let:

- N be the total number of tokens across all biographies.
- $c(w)$ be the total token count of word w across all documents.
- For a gender label g, let D_g be the set of documents with that gender.
- $c(g)$ be the number of tokens in documents with gender g.
- $c(w, g)$ be the total token count of word w inside documents of gender g.

The pointwise mutual information between a word w and gender g is computed as:

$$\text{PMI}(w, g) = \log \frac{c(w, g) \cdot N}{c(w) \cdot c(g)}$$

Normalized PMI, which accounts for the relative frequency of words, is defined as:

$$\text{NPMI}(w, g) = \frac{\text{PMI}(w, g)}{-\log \frac{c(w,g)}{N}}$$

2.3 Frequency Threshold

For this analysis, the top 30 tokens with the highest nPMI were selected and compared between the two gender groups.

3 Results

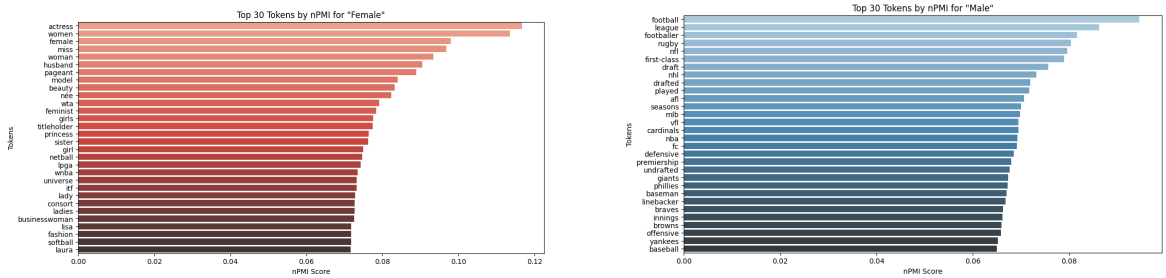

Figure 1: Comparison of representations for women (left) and men (right).

4 Discussion/bias statement

4.1 Patterns Noticed

Firstly, in terms of magnitude, it is observed that the nPMI scores for the top tokens in females are higher than those of males, which could indicate more consistent co-occurrence patterns for female-linked tokens. For instance the top two tokens for females pass the 0.1 mark while none pass it for males. The nPMI tokens for males tend to be more domain-concentrated in sports compared to women. Female-associated terms are more spread across multiple lifestyle domains (media, fashion, relationships, health, etc), indicating a broader but less sharply defined cluster.

When analyzing the tokens themselves, it is also observed that the language is strongly gendered in each group. Tokens associated with "female" are dominated by words related to appearance and the body (such as makeup and beauty), along with social or relational concepts like relationships and lifestyle. In contrast, tokens associated with "male" skew toward sports and competition (football, league). This contrast suggests that the model has learned highly stereotypical gender associations from the training data rather than neutral or evenly distributed language patterns. Particularly, it is interesting to notice that husband was the sixth most common word for females, while there are no terms referencing family or relationship for the male tokens. Extrapolating from this point, the general theme for women is personal and individual, while that for male cluster around public arenas like sport and politics. For instance, the corpus contains biographies of both Draginja Obrenović and Aleksandar Obrenović, who were the queen and king of Serbia from 1889 to 1903. However, Draginja's biography only talks about being the queen and wife of Aleksandar and that she was a former lady-in-waiting. Meanwhile, Aleksandar's biography highlights that he was the king, followed by details of his assassination and his wife. This example demonstrates how gendered biases in the corpus translate into divergent narrative framing.

4.2 Potential Harms

Given the widespread usage of LLMs these days, gendered language biases present in the datasets they are trained on and translate from can produce tangible harms to society. One such use case would be in healthcare, where it was observed that GPT-4 produced clinical vignettes that perpetuate demographic stereotypes [Zack et al., 2024]. This could potentially influence how practitioners interpret symptoms, probably leading to misdiagnosis, delayed treatment, or inappropriate care for certain groups, which is dangerous.

On a psychological level, gendered or demographic biases in LLM outputs can normalize and legitimize existing stereotypes, making them appear objective or data-driven rather than socially constructed. As seen in research, when prompted, LLMs generate factually inaccurate explanations and are likely to hide how they actually got their answers. As a user, when reading the reasoning, one might be compelled to adapt to that reasoning, leading to a change in mindset that is leaning more biased [Kotek et al., 2023].

4.3 Limitations

While nPMI is useful for identifying associations between words and gender labels (as seen above), it has some limitations. Firstly, nPMI is highly sensitive to low-frequency events because it is computed using joint and marginal probabilities, and this could lead to some extreme values for some word–gender co-occurrences even when they may not be statistically meaningful. Secondly, the measure highlights correlation but not causation, so we cannot fully conclude that there is a causal relationship between gender and word usage. Lastly, from a methodological standpoint, since this was a randomly sampled corpus, while there is equal gender composition (50000 each), there could still be differences in document length and topical coverage that can affect association strength. An example is

that Raoul Servais's biography text contains only 19 words while D.H. Peligro's biography was the longest with 566 words. This extreme example shows the difference in document length, and this could result in not all biographies being assessed equally since the token is the metric.

5 Advanced Analysis

To further this analysis, occupation was introduced as another subgroup. The occupation groups identified are STEM, Politics, Art, Sports, and Others. Mapping was used in the creation of this new column, where common occupations that fall into these categories were added to a dictionary, and based on the first sentence of the biography, the group was assigned. After this processing, we had 31299 in Arts, 29959 in Sports, 22283 in Other, 14320 in Politics, and 2139 in STEM. We recognize there is an imbalance in this representation, which could limit the analysis. Similar analysis with nPMI scores for the top 30 tokens was done with the occupational groups. We further performed subgroup analyses by both occupation and gender to identify differences within occupational categories. For this analysis, only the arts group will be discussed since this group had the largest set of biographies.

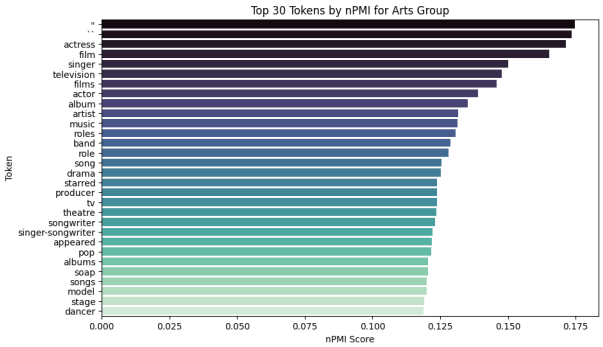

Figure 2: Top 30 tokens in Art subgroup

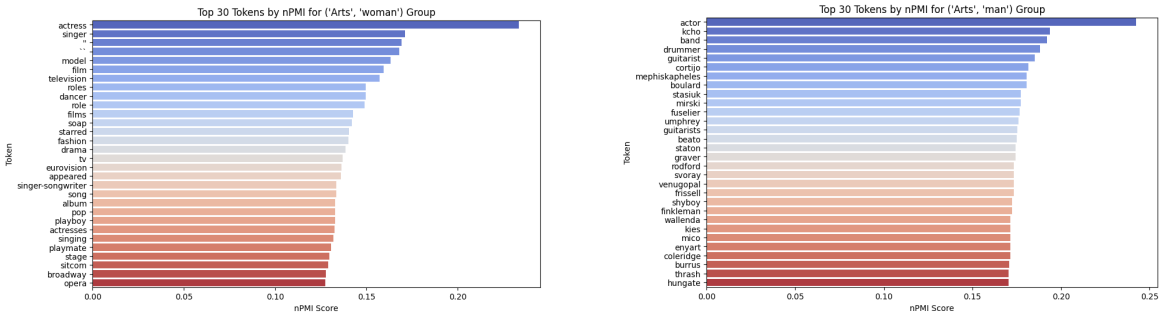

Figure 3: Comparison of representations for women (left) and men (right) in the Art subgroup

Referencing Figure 2, it appears that the tokens usually refer to words synonymous with this field, such as actor, television, music, etc. However, there are observable differences between the art-female and art-male subgroups (Figure 3). For one, the male list tends to feature proper names, such as surnames and artist names (like Cortijo and Stasiuk), highlighting individual recognition, whereas the women's list relies more on generic descriptors and categories, suggesting a more generalized portrayal. Furthermore, commercial and sexualized terms like Playboy and playmate appear exclusively in the women's list, revealing a gendered framing that links female artists to sexualized or commercial contexts. No such terms can be observed in the male list.

6 Conclusion

To conclude, this analysis demonstrates that there is a presence of gender bias in the Wikipedia dataset, which is a common large-scale data source used for training NLP algorithms. While there are limitations to nPMI, it still reveals stark differences between males and female which could pose risks of perpetuating stereotypes, reinforcing unequal representations, and influencing downstream AI models in biased ways.

References

Hadas Kotek, Riley Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of the ACM Collective Intelligence Conference*, pages 12–24, November 2023.

Wikibioreformatted_balanced_100k dataset. Online, 2026. Accessed from https://maartensap.com/11830/files/WikiBioReformatted_balanced_100k.zip.

S. M. Watford, R. G. Grashow, V. Y. De La Rosa, R. A. Rudel, K. P. Friedman, and M. T. Martin. Novel application of normalized pointwise mutual information (npmi) to mine biomedical literature for gene sets associated with disease: use case in breast carcinogenesis. *Computational Toxicology (Amsterdam, Netherlands)*, 7:46–57, 2018. doi: 10.1016/j.comtox.2018.06.003. URL https://doi.org/10.1016/j.comtox.2018.06.003.

Tyler Zack, Eric Lehman, Mirac Suzgun, Juan A. Rodriguez, Leo Anthony Celi, Judy Gichoya, and Emily Alsentzer. Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study. *The Lancet Digital Health*, 6(1):e12–e22, 2024.

A Additional Details

A.1 Gender Based Analysis

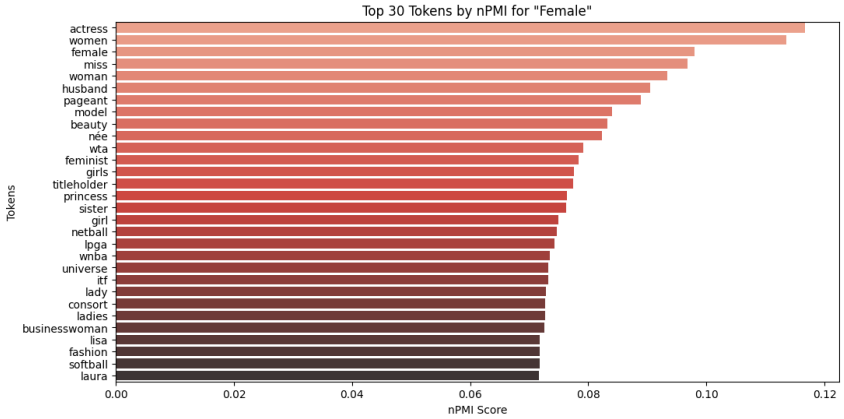

Figure 4: Top 30 tokens in Female group

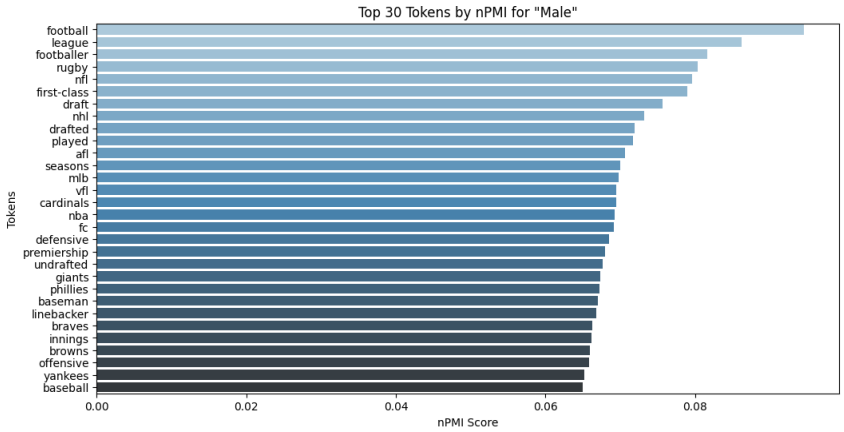

Figure 5: Top 30 tokens in Male group

A.2 Occupational Group Analysis

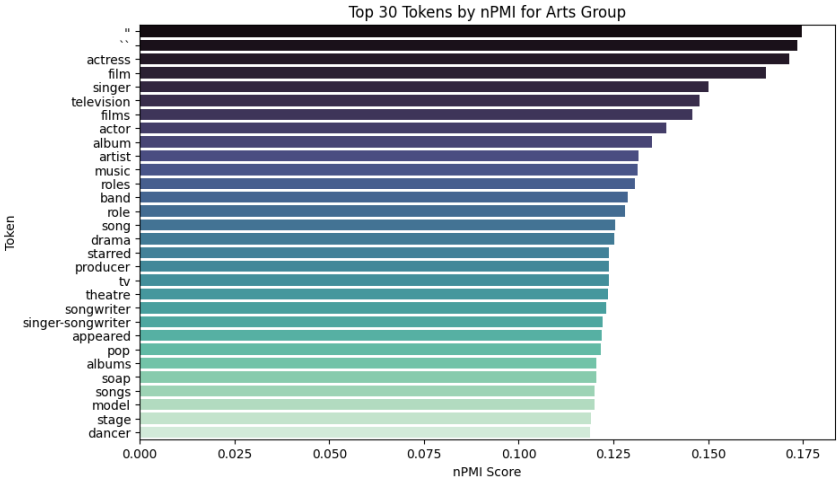

Figure 6: Top 30 tokens in Art group

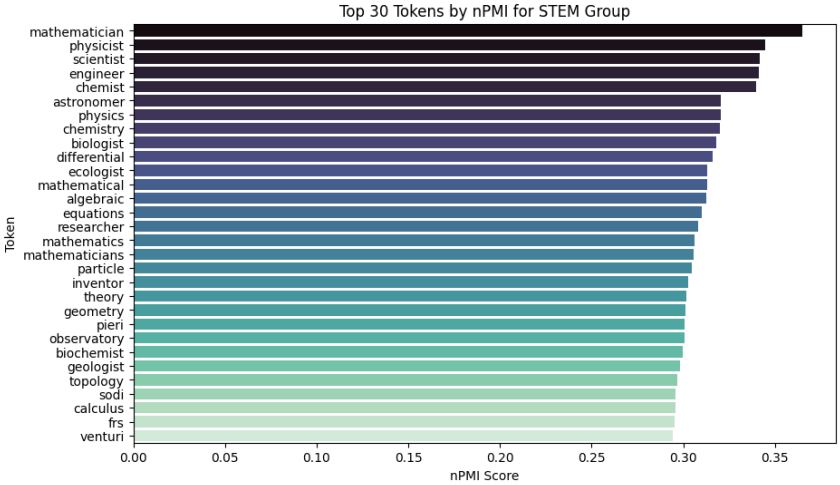

Figure 7: Top 30 tokens in STEM group

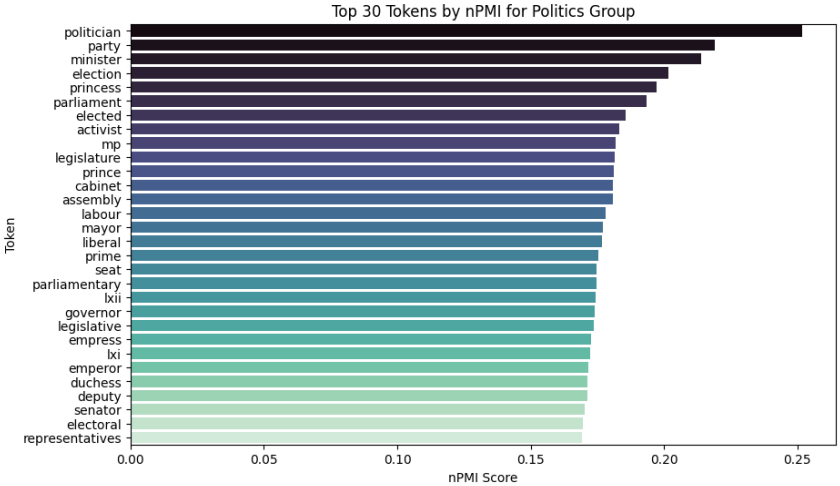

Figure 8: Top 30 tokens in Politics group

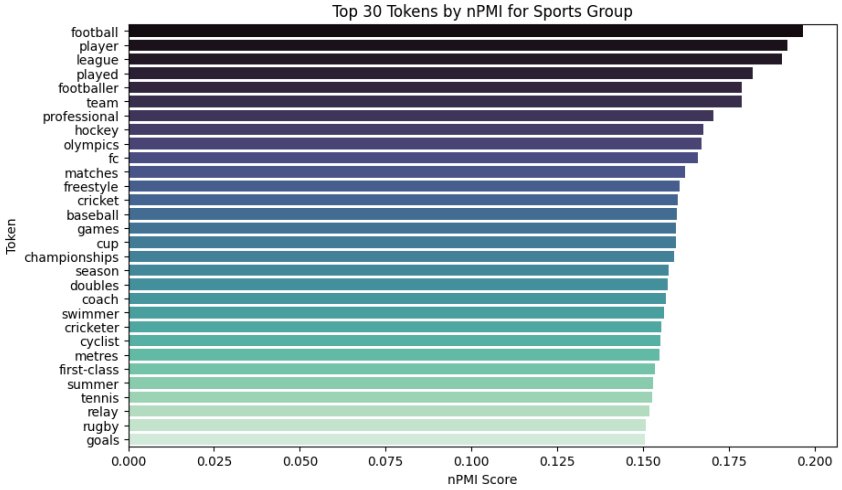

Figure 9: Top 30 tokens in Sports group

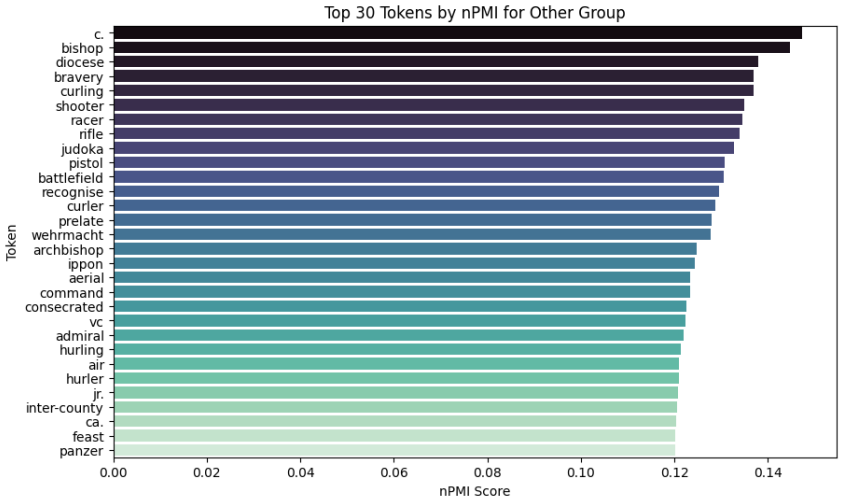

Figure 10: Top 30 tokens in Other group

A.3 Occupational Group - Gender Based Analysis

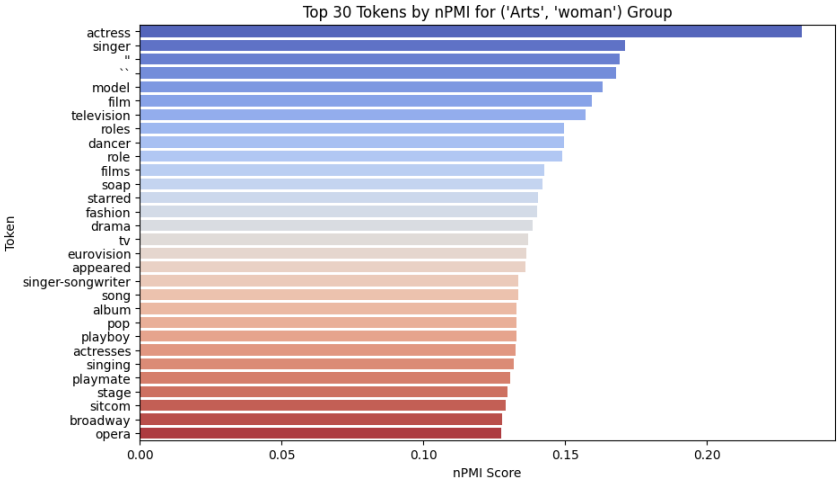

Figure 11: Top 30 tokens in Art-Female group

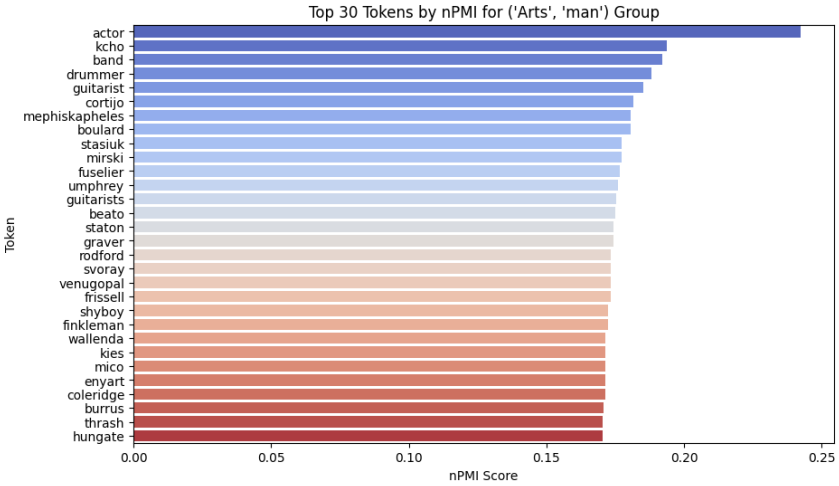

Figure 12: Top 30 tokens in Art-Male group

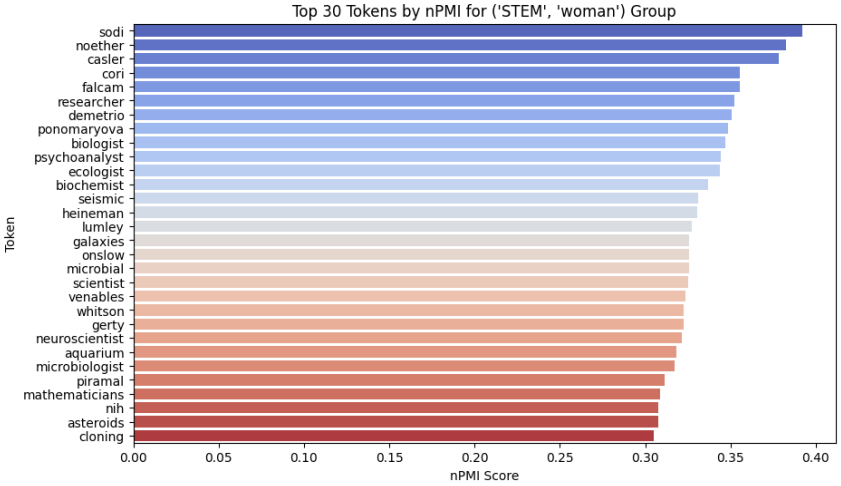

Figure 13: Top 30 tokens in STEM-Female group

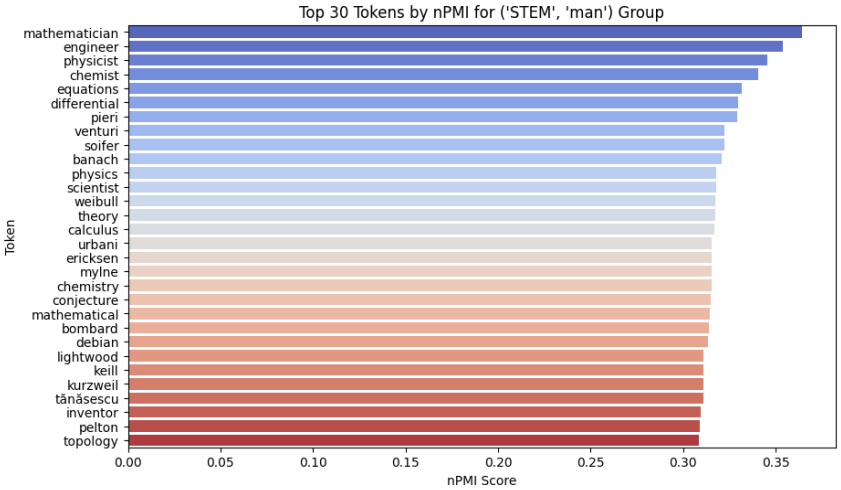

Figure 14: Top 30 tokens in STEM-Male group

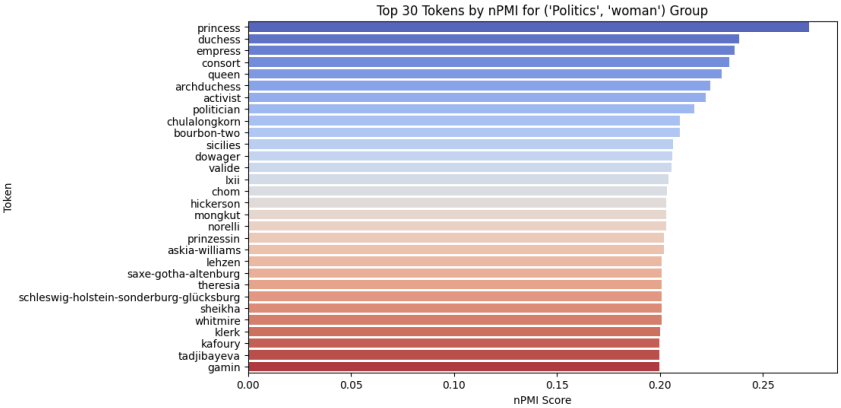

Figure 15: Top 30 tokens in Politics-Female group

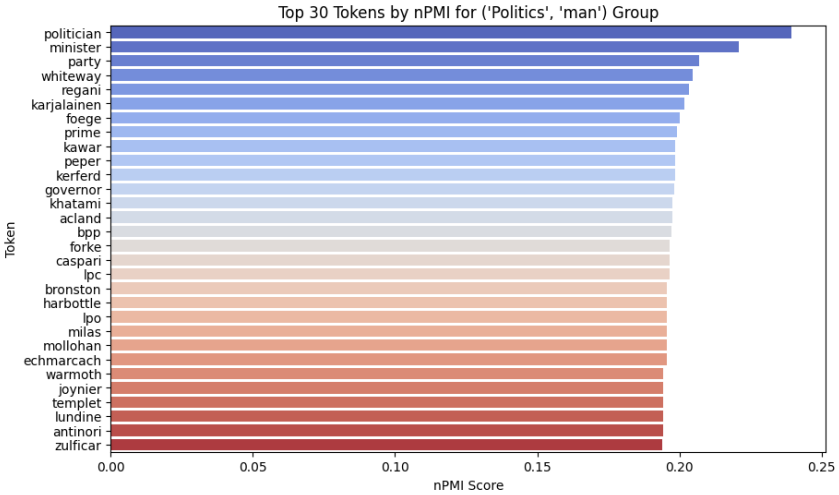

Figure 16: Top 30 tokens in Politics-Male group

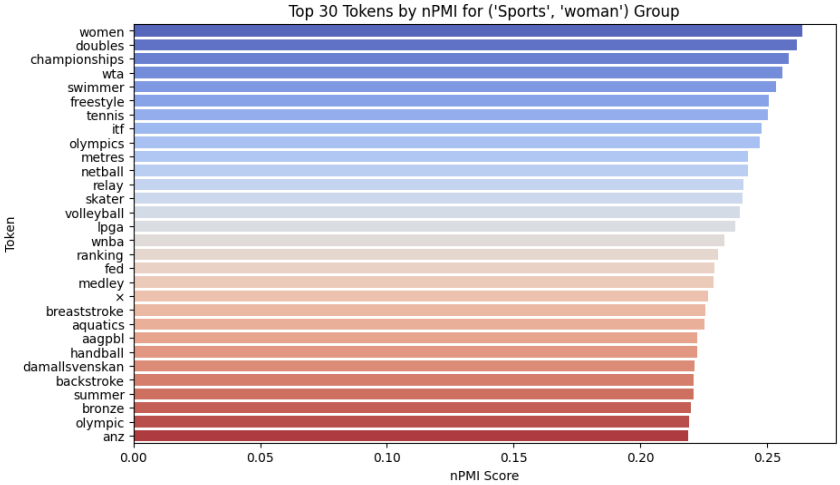

Figure 17: Top 30 tokens in Sports-Female group

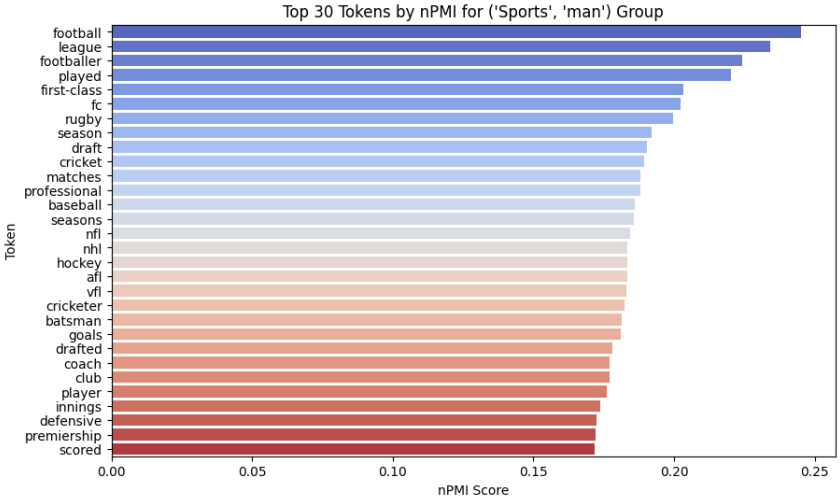

Figure 18: Top 30 tokens in Sports-Male group

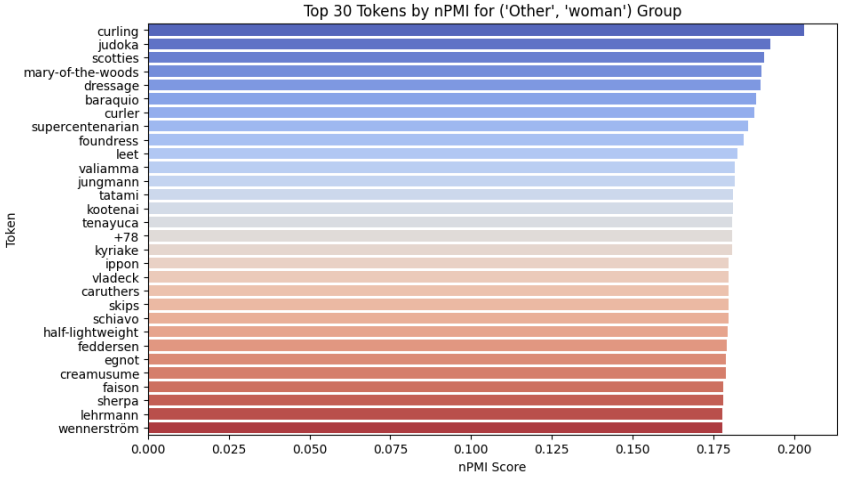

Figure 19: Top 30 tokens in Others-Female group

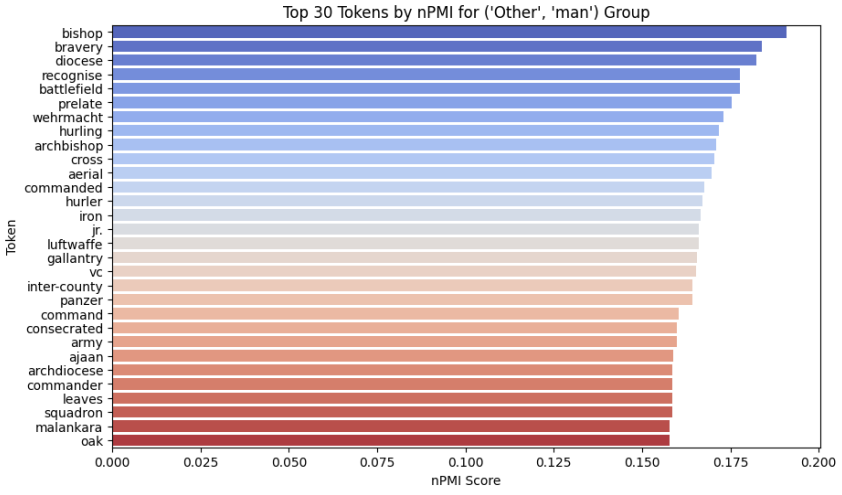

Figure 20: Top 30 tokens in Others-Male group