Analyzing Gender Bias in Sentiment Analysis Algorithms Using a Large Corpus of Gendered Language Data

Ali Khan
University of California, Berkeley
alikhan447@berkeley.edu

Abstract

Resolving gender bias in pronouns remains a formidable challenge in algorithmic coreference resolution. Existing models grapple with the intricacies of ambiguous pronouns, which often result in unclear text interpretation and compromise the integrity of historical and contemporary data records. Such ambiguity can lead to misinterpretation and the propagation of unreliable data sources. This study addresses the challenge by employing the Gender Ambiguous Pronouns (GAP) Coreference Dataset to develop a logistic regression model that accurately identifies gender pronouns, achieving a 77.85% success rate and labeling 8,908 instances of ambiguous pronouns within the corpus. The model emerges as a pivotal tool for evaluating gender bias, sets a new standard for gender sensitivity in NLP systems, and offers a scalable solution for automated labeling in future linguistic research.

1 Introduction

The field of Natural Language Processing (NLP) has witnessed remarkable strides in enabling algorithms to decipher human language intricacies. However, the resolution of gendered pronouns remains a significant hurdle, with far reaching implications for the perpetuation of gender biases and stereotypes. Pronoun resolution—linking pronouns to their corresponding entities in text—becomes particularly complex when these pronouns carry gender implications. This complexity arises from the need to understand not only the grammatical rules but also the social and cultural contexts that inform gender references. Missteps in this area not only skew data interpretation but also reinforce societal stereotypes.

The urgency to rectify gender bias in NLP transcends academic discourse, touching the core of ethical AI deployment in critical sectors such as job recruitment and healthcare diagnostics. Gendered pronoun resolution is emblematic of the broader challenge of gender bias in AI, where inaccuracies can cascade into systemic inequities.

Despite the wealth of research on gender bias within textual data and machine learning models, focused studies on gendered pronoun resolution in NLP remain scarce. This gap is significant; as pronouns are fundamental to language comprehension, their misresolution can distort information retrieval, machine translation, and question-answering systems, further entrenching gender biases (Falenska and Çetinoğlu, 2021).

As AI systems increasingly influence critical aspects of society, from jurisprudence to medical advice, the ethical integrity of these technologies is paramount. The implications of gender bias in pronoun resolution extend beyond mere technical inaccuracies, threatening the fairness and inclusivity of AI applications (O'Connor and Liu, 2023).

While strides have been made in identifying and mitigating gender bias in NLP (Reagle and Rhue, 2011), the specific challenges of gendered pronoun resolution demand focused attention. This study bridges this research gap by undertaking a thorough examination of gender-ambiguous pronouns within the Wikipedia corpus.

Confronting the complexities of gendered pronoun resolution, this paper addresses the critical question: How can machine learning algorithms be refined to enhance the precision of gendered pronoun resolution while actively reducing gender bias?

2 Background

The quest to develop algorithms with a nuanced understanding of human language is ongoing, and within this pursuit, the resolution of gendered pronouns stands as a persistent challenge. Addressing gender bias in AI technologies, especially in the realm of NLP, has garnered significant attention. A seminal work contributing to this discourse is "Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities" by Sinead O'Connor and Helen Liu (2023). They highlight the ethical quandaries posed by AI, contending that while AI has the potential to either reinforce or alleviate societal biases, the current trajectory often skews toward perpetuation. Their insights underscore the imperative for conscientious AI and NLP methodologies that actively avoid entrenching gender biases, thereby laying the groundwork for this study.

Coreference resolution, the task of linking pronouns to corresponding entities within a text, is a cornerstone of language comprehension and is emblematic of the challenges in understanding nuanced human language (Attree, 2019). The introduction of gendered pronouns into this equation introduces a layer of complexity, intersecting with broader societal issues concerning gender representation and bias. The GAP dataset, a gender-balanced corpus crafted by (Webster et al., 2018) to assess coreference resolution systems, has become a touchstone for evaluating gender bias in pronoun resolution. Their findings spotlight the gender bias endemic to extant systems and advocate for balanced training datasets as a remedy. It is this dataset that our study's analysis model targeting biased pronouns will employ.

2.1 Dataset with Gender Ambiguous Pronoun

The persistent issue of gender bias within NLP is exemplified by the inherent biases in word embeddings. (Sun et al., 2019) highlighted that these embeddings, which are a common input feature for NLP models, frequently contain gender biases. This is a critical concern as models trained on biased embeddings could perpetuate or even amplify these biases in a range of real-world applications, from search engines to voice-activated assistants. Furthermore, (Rudinger et al., 2018) demonstrated that coreference resolution systems often exhibit poorer performance on examples involving female pronouns and entities, pointing to a systemic issue within AI systems that necessitates gender-balanced training data and bias-aware modeling techniques.

In response to these challenges, the Gender Ambiguous Pronouns dataset, introduced by (Webster et al., 2018) , fills a critical void in NLP resources. It offers a robust collection of instances where pronoun gender is not immediately clear, requiring models to rely on contextual clues for accurate resolution. The GAP dataset is notable for its explicit focus on gender balance, aiming to counteract the observed trend of coreference resolution systems performing better with masculine pronouns—a tendency that can contribute to the entrenchment of gender biases.

Several key contributions of the GAP study include:

1. **Gender Balance**: The GAP dataset was meticulously curated to ensure equal representation of genders, countering biases in existing systems.

2. **Diverse Challenges**: It encompasses a wide variety of pronouns and names from Wikipedia articles, presenting a range of real-world scenarios to more authentically challenge coreference resolution systems.

3. **Benchmarking**: It established benchmarks for coreference resolution system performance using the GAP dataset, with the F1 score as a critical metric combining precision and recall measuring accuracy.

4. **Bias Measurement**: They introduced a novel "Bias Factor" metric, derived from comparing the F1 scores for feminine versus masculine pronouns, to quantify and address bias within coreference systems.

The aforementioned studies collectively underscore the importance of using balanced datasets like GAP for training and evaluating NLP systems, ensuring that the advancement of AI technologies does not come at the cost of perpetuating gender bias.

3 Ethical AI and Gender Pronoun Resolution: A Logistic Regression Approach

This project centers on developing a logistic regression model trained on the GAP dataset, aiming to tackle the persistent issue of gender bias in pronoun resolution within AI. The choice of logistic regression, known for its simplicity and interpretability, reflects a commitment to ethical AI by prioritizing transparency and fairness in machine learning practices.

Employing a balanced dataset is a crucial ethical consideration, ensuring that the model does not inherit the biases often present in unbalanced data. This approach is in line with the current AI research ethos that stresses accountability and interpretability in machine learning (Doshi-Velez and Kim, 2017).

The developed model is adept at discerning gender pronouns accurately while actively minimizing bias. Its interpretability aligns with the advocacy for AI that can be rigorously examined for biases. The practical implications of this model are far-reaching, extending to applications where gender-neutral language processing is essential, such as in voice-activated assistants, recommendation systems, and automated text analysis.

Furthermore, the model stands as a benchmark for subsequent research, offering a solid baseline for the development of advanced algorithms that further the cause of ethical AI. By effectively classifying gender-ambiguous pronouns and illuminating biases, it enhances NLP tasks and contributes to the ongoing effort to imbue AI technologies with ethical standards.

In sum, this project achieves a dual milestone: it provides a technical resolution to a complex NLP issue and represents a stride towards the ethical imperative of reducing bias in AI. The model's 77.85% accuracy rate in classifying gender biases demonstrates its value as a tool for both researchers and practitioners, signaling a move towards more equitable AI systems.

4 Data Collection

For this study, we utilized the Gendered Ambiguous Pronouns (GAP) dataset, a gender-balanced corpus designed explicitly to evaluate coreference resolution systems. The GAP dataset, released by Google AI Language, comprises 8,908 coreference-labeled pairs of ambiguous pronouns and their antecedent names, meticulously sampled from Wikipedia. This dataset is publicly available and can be accessed through GAP Coreference Dataset.

4.1 GAP Dataset Characteristics

The GAP dataset is divided into three subsets: test, development, and validation, each consisting of 4,000, 4,000, and 908 pairs, respectively. These subsets are intended for official evaluation, model development, and parameter tuning. The data is structured into eleven columns within .tsv (tab-separated values) files. These columns contains:

- **ID**: Unique identifier for an example (two pairs)
- **Text**: Text containing the ambiguous pronoun and two candidate names. About a paragraph in length.
- **Pronoun**: The pronoun, text.
- **Pronoun-offset**: Character offset of Pronoun in Column 2 (Text).
- **A^**: The first name, text.
- **A-offset**: Character offset of A in Column 2 (Text).
- **A-coref**: Whether A corefers with the pronoun, TRUE or FALSE.
- **B^**: The second name, text.
- **B-offset**: Character offset of B in Column 2 (Text).
- **B-coref**: Whether B corefers with the pronoun, TRUE or FALSE.
- **URL^^**: The URL of the source Wikipedia page.

4.2 Data Source

The GAP dataset's design is a conscious effort to mitigate the gender bias observed in coreference systems. By ensuring gender balance, it addresses the problem of gender bias at the data level, which is crucial for training unbiased AI models. This dataset not only forms the basis of our analysis but also serves as a testament to the importance of ethical considerations in AI development, a theme that is recurrent in the literature.

The structured balance in gender representation within the dataset provides a vital and realistic benchmark for evaluation. In our study, we utilize this balanced resource to build a logistic regression model that aims to accurately assess and correct gender biases in pronoun resolution. By doing so, we ensure that our model's training is free from gender imbalance, reflecting the real-world complexity of text and enhancing the model's applicability and fairness.

4.3 Existing Solution of GAP

The GAP dataset, while a central resource in our study for its gender-balanced composition, also provides pre-established benchmarks that have informed the broader discourse on fairness and accountability in AI. These benchmarks, detailed in the seminal "Mind the GAP" paper, offer a vital reference for our work, specifically in evaluating the effectiveness and fairness of coreference resolution models.

Task Setting	M	F	B	O
Snippet-context	69.4	64.4	0.93	66.9
Page-context	72.3	68.8	0.95	70.6

Table 1: Performance benchmarks on the GAP dataset.

From the **Table 1**; Task Setting indicates the context in which the coreference resolution model is evaluated. There are two settings mentioned:

1. *Snippet-context:* The model uses only the snippet of text provided in the dataset for resolving coreferences.

2. *Page-context:* Page-context The model can use the entire content of the source Wikipedia page (accessible via the URL) to resolve coreferences.

The benchmark metrics are defined as follows:

1. **M (Masculine):** The F1 score for examples with masculine ambiguous pronouns.

2. **F (Feminine):** The F1 score for examples with feminine ambiguous pronouns.

3. **B (Bias factor):** A metric derived by dividing the F1 score for feminine examples by that for masculine examples. A value approaching 1 suggests less bias, as it indicates comparable performance across genders.

4. **O (Overall):** The aggregate F1 score across all examples, irrespective of the pronoun's gender.

Our study utilizes these benchmarks as a comparative backdrop for our logistic regression model. While the established benchmarks highlight the persistent challenges in resolving gender ambiguity in pronouns, our model seeks to further understand and mitigate bias, offering a fresh perspective on ambiguous pronoun resolution. By comparing our model's performance against these benchmarks, we aim to quantify improvements in fairness and to contribute meaningfully to the ongoing efforts to reduce bias in AI.

4.4 Preprocessing for GAP

To enhance the accuracy of resolving gender-ambiguous pronouns, our preprocessing pipeline incorporated several critical steps, including tokenization, which encompasses:

1. **Case Normalization**: We began by converting all text to lowercase to treat pronouns uniformly, regardless of their case (e.g., 'He' vs. 'he'). This step is essential for ensuring the model does not differentiate the same word based on capitalization, which is particularly relevant for gender pronouns.

2. **Punctuation Removal**: Next, we removed punctuation to eliminate syntactic elements that do not contribute to understanding gender references, thus streamlining the model's focus on textual content.

3. **Stop-word Filtering**: Utilizing the NLTK corpus, we filtered out stop-words. These are common words that add little informational value to gender pronoun resolution. This refinement enables the model to focus on content-rich terms that are more likely to be informative in the context of gender.

4. **Lemmatization:** We applied lemmatization to condense words to their base or dictionary form (e.g., 'running' and 'ran' to 'run'). This step helps preserve the semantic consistency of gendered actions or descriptions, which is crucial for interpreting the context surrounding gender pronouns.

Finally, to prepare our data for the machine learning process, we transformed the preprocessed text into a numerical representation using the TfidfVectorizer. This tool computes numerical scores that reflect the relevance of each word in the corpus, based on term frequency (TF)—the occurrence rate of a word in a particular document—and inverse document frequency (IDF)—the rarity of the word across all documents. By integrating TF with IDF, the TfidfVectorizer discerns the significance of words, giving more weight to those that are pertinent to gender pronouns and context, while mitigating the influence of common but less informative terms. This weighting is crucial for our analysis, as it highlights linguistic patterns that may indicate gender biases.

These preprocessing steps are instrumental not only in building a capable pronoun resolution model but also in shedding light on gender representation within Wikipedia. By leveraging a balanced dataset and sophisticated text analysis, we contribute to the ongoing efforts to forge NLP systems that are both technically sound and ethically aware, thus advancing the dialogue on gender fairness in AI.

4.5 Annotation

The annotation process, as delineated in **Table 2**, was a critical step in preparing the data for our logistic regression model. Pronouns traditionally associated with female gender were labeled with '0', while those associated with male gender received a '1'. This binary labeling aligns with our objective to examine and reduce gender bias in pronoun resolution.

Gender Bias	Pronouns	Label
Female	*she*, *her*	0
Male	*he*, *his*, *him*	1

Table 2: Annotation of pronouns

In crafting a balanced training dataset, we ensured an equal representation of labels for both female and male pronouns, thus mitigating potential biases in the model's predictions. The dataset encompasses two principal features: the textual excerpts from Wikipedia articles and the corresponding gender bias labels.

Building upon the preprocessing steps detailed earlier, we introduced the "Gender Bias" column as a pivotal element for the logistic regression model. It provides a definitive target variable, derived from the gender associations present in the "Text" column, ensuring that the model can effectively learn and predict gender bias.

The final dataset, optimized for the logistic regression model, comprises these two columns: "Text" and "Gender Bias". The "Text" column presents the pronouns in their linguistic context from the GAP dataset, while the "Gender Bias" column imparts the gender association cues essential for the model's learning process.

5 Modeling GAP

Following the meticulous preprocessing of our dataset, we embarked on the critical phase of model training. We elected to use logistic regression, a widely endorsed statistical method for binary classification tasks. This model is particularly apt for dichotomous outcomes, such as our gender bias labeling task, due to its efficiency, interpretability, and well-established performance on baseline binary classifications.

5.1 Training the Model

The training regimen commenced with a normalization of pronouns, a fundamental step to ensure consistent representation critical for a model's pattern recognition. A balanced training dataset was crafted, comprising an equal distribution of labels for female and male pronouns to circumvent potential prediction biases. This dataset featured two principal components: the textual content from Wikipedia articles and the corresponding gender bias labels, the latter serving as the target variable.

5.2 Vectorization

To render the textual data amenable to the logistic regression model, we transformed each text entry into a numerical vector utilizing the TfidfVectorizer. This pivotal transformation translates text into numerical values, enabling the model to process the data effectively. Moreover, it amplifies the significance of each term in relation to the corpus. The application of the TfidfVectorizer was especially beneficial in our context, as it underscored gendered pronouns and their associated contextual terms, which are integral to our analysis.

5.3 Results and Implications

Our logistic regression model achieved an accuracy of 77.85% on the test dataset, which is indicative of its robust ability to identify and classify gender biases based on pronouns. The performance of the model reaffirms the efficacy of our preprocessing and vectorization techniques in creating a feature set that is both representative of the textual nuances and sensitive to the gender cues present within the data.

Interpreting the Results The accuracy metric, while substantial, opens the discussion for potential improvements and the necessity for more nuanced metrics in evaluating gender bias. Given the complexity of language and the subtleties involved in gender pronouns, a deeper analysis of false positives and false negatives would be instrumental in understanding the model's limitations and biases.

Implications for NLP and AI Ethics The implications of our work extend into the realms of ethical AI and fair machine learning practices. By successfully training a model to detect gender bias, we contribute to the ongoing discourse on eliminating prejudicial inclinations in NLP systems. Our research underscores the importance of balanced datasets and rigorous preprocessing to prevent the perpetuation of stereotypes and biases in automated systems.

Future Directions Building upon the foundation laid by this research, future work could explore multi-class gender classification to encompass a broader spectrum of gender identities. Additionally, integrating contextual embeddings from models like BERT or GPT could enhance the understanding of pronoun antecedents, thereby refining the accuracy and fairness of coreference resolution systems.

The scalability of our approach to other datasets promises a pathway to creating more inclusive and equitable NLP applications. Our methodology, with further refinement and adaptation, has the potential to serve as a benchmark for assessing and mitigating biases in a multitude of language processing tasks.

6 Conclusion

This study's logistic regression model was subjected to a comprehensive data preprocessing, normalization, and vectorization regimen, ensuring a refined approach to the detection and classification of gendered pronouns. Tested against an unseen dataset, the model demonstrated a commendable accuracy of 77.85%. This performance not only testifies to the model's capabilities but also draws attention to the inherent intricacies of gender bias detection in language processing.

The attained accuracy is a testament to the model's robustness yet simultaneously sheds light on the complexities and challenges that persist in detecting and understanding gender biases. Our model's outcomes are a promising indicator of the role that machine learning can play in identifying and addressing biases within textual data, furthering the pursuit of more equitable NLP applications.

In sum, the work presented herein goes beyond the demonstration of methodological effectiveness in gender bias classification; it actively contributes to the critical conversation around ensuring fairness and accountability in AI. As the field progresses, the insights gleaned from our research will serve as valuable guides in the quest to develop NLP systems that are as ethically conscious as they are technologically advanced.

References

Sandeep Attree. 2019. Gendered ambiguous pronouns shared task: Boosting model confidence by evidence pooling. *arXiv preprint arXiv:1906.00839*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Agnieszka Falenska and Özlem Çetinoğlu. 2021. Assessing gender bias in wikipedia: Inequalities in article titles. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 75–85.

Sinead O'Connor and Helen Liu. 2023. Gender bias perpetuation and mitigation in ai technologies: challenges and opportunities. *AI & SOCIETY*, pages 1–13.

Joseph Reagle and Lauren Rhue. 2011. Gender bias in wikipedia and britannica. *International Journal of Communication*, 5:21.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *arXiv preprint arXiv:1804.09301*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.