|  |
| --- |
| A white square with a blue border  Description automatically generated |
|  |

|  |  |
| --- | --- |
| **Synthetic Data for Bias Mitigation (BIAS Project)** | |
| Type of work: **Term Project 2** | |
|  | |
| Degree programme: | **Master of Science in Engineering (MSE)**  **specialisation Data Science** |
| Specialisation: | **Data Science** |
| Student: | **Shilpi Garg (gargs1)** |
| Project: | **PA2** |
| Project Advisor: | **Dr. Kurpicz-Briki Mascha** |
| Co-Advisor: | **Dr. Alexandre Puttick** |
| Date: | **25 July 2025** |

**Contents:**

1. **Abstract**
2. **Introduction**
   1. **Related work:**
      1. **Gender Disparities in Training Data**

**2.1.2. Synthetic Data for Bias Mitigation**

1. **Materials and Methods**
   1. **BiasBios Dataset**
   2. **Models**
      1. **Model LLaMA-2-7B-Chat (Meta AI, 2023)**
      2. **Model LLaMA-3.2-3B-Instruct (Meta AI, 2024)**
   3. **Hardware specifications**
   4. **Experimental setups**
   5. **Gender classifiers**
      1. **Gender Classifier Without Masking**
      2. **Gender Classifier with lowercase text and without Masking**
      3. **Gender Classifier with masking gender pronouns and lowercase text**
      4. **Gender Classifier with masking Names and pronouns**
      5. **Gender Classifier with masking Social and Workplace roles**
      6. **Gender Classifier with masking Personality traits**
   6. **Synthetic Data Generation**
      1. **Job Advertisement Generation**
      2. **CV and Cover Letter Generation (Generic Application)**
   7. **Evaluation experiments**
      1. **Gender classifier on generated synthetic data (without lower casing and masking)**
      2. **Classifier with lowercasing the text and stop words, and masking names, pronouns, social and workplace proxies**
2. **Results**
   1. **Analysis on BiasBios dataset**
   2. **Gender Classifier** 
      1. **Classifier without masking or lower text**
      2. **Classifier with lowercase text and without masking**
      3. **Classifier with masking gender pronouns and lowercase text**
      4. **Classifier with Masking Names and pronouns**
      5. **Classifier with masking Social and Workplace roles**
      6. **Classifier with masking Personality traits**
   3. **Evaluation experiments**
      1. **Classifier on generated synthetic data (without lower casing and masking)**
      2. **With lowercasing and stop words, and masking names, pronouns, social and workplace proxies**
   4. **Analysis based on Accuracy metrics on Test data**
   5. **Analysis based on Low confidence of gender predictions**
3. **Discussion**
4. **Bibliography**
5. **Declaration of Authorship**
6. **Abstract:**

The Horizon Europe project BIAS explores how societal biases manifest in AI systems and large language models, particularly within the context of the labor market. As part of this initiative, this subproject focuses on bias detection and mitigation in the generation and evaluation of synthetic CVs and cover letters situated within the broader context of AI recruitment systems. The primary objective is to develop a modular synthetic data generation framework that enables the systematic creation of job application documents—including job ads, CVs, and cover letters—while carefully controlling for demographic variables, linguistic patterns, and skill profiles. This synthetic dataset is intended to serve as a benchmark for evaluating and improving fairness in downstream natural language processing (NLP) applications.

In this context, synthetic data for bias mitigation refers to the deliberate generation of artificial application materials that allow for fine-grained manipulation of attributes such as gender, ethnicity, age, and socio-economic background. This is a part of an effort to ensure that models trained or tested using such data do not perpetuate or exacerbate existing societal biases. The framework supports the injection of sensitive terms (e.g., “female,” “disabled”) and proxy terms (e.g., “parental leave,” “youth leader”) into the documents to test the model’s predictive behavior in response to these variables. These terms are often correlated with demographic markers and may trigger unintended discriminatory outcomes, making them critical for bias stress-testing and mitigation strategies.

**Keywords**- NLP, Synthetic Data, RAG, data augmentation, bias mitigation, sensitive attributes, proxy words, machine learning, LIME

1. **Introduction:**

The integration of artificial intelligence (AI) into recruitment processes has transformed how organizations evaluate job applicants. While AI offers scalable tools for resume screening and candidate evaluation, growing evidence suggests that these systems risk perpetuating or amplifying **gender biases** embedded in training data and model architectures ([Fabris et al., 2025).](https://dl.acm.org/doi/10.1145/3696457) This is particularly critical during the early stages of hiring, where automated models interpret CVs and cover letters—documents that inherently reflect individual self-presentation strategies shaped by societal norms ([Marti Marcet, 2023](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/44294/Applied_Data_Science_Thesis_Sara.pdf)). AI systems trained on real-world data are vulnerable to perpetuating historical societal biases, including gender biases embedded in language and occupational roles. This study is part of the broader Horizon Europe BIAS project, which aims to examine and mitigate such biases in AI systems deployed within labor market contexts.

This project proposes a modular synthetic data generation framework for constructing balanced and bias-aware training and evaluation datasets. Specifically, our framework focuses on the systematic generation of job application materials. We generate artificial CVs and cover letters that allow controlled variation of:

* Demographic attributes (e.g., gender markers like he, she, her, his; names like Jennifer, Piyush)
* Linguistic features (eg. leader, mentor)
* Sensitive/proxy terms (e.g. director, support, assistant, leader,”): **Proxy words** are terms that can be used by the model as indirect signals (often unintended) to predict sensitive attributes (like gender, race, or biases).

Related detailed work and experiments are discussed in Materials and Methods section.

The objectives of this study are:

* How much gender information can be inferred from the data (bios/CV cover letters)? What can be done to mask that information?
* How does the presence of gender information affect the generation of synthetic data? How susceptible is the generated data to gender bias?
* Evaluate the behaviour of existing NLP models against this synthetic dataset using LIME classifier to assess and mitigate their bias responses.
* Is there potential to use the explored methods to help mitigate unfair discrimination in recruitment?

**2.1. Related work:**

**2.1.1 Gender Disparities in Training Data**

One of the foundational challenges in mitigating bias in AI hiring systems is the representational imbalance in training datasets. The BiasBios dataset is scraped from the internet and likely reflects gender disparities that are also in the internet-scraped pre-training data of LLMs. As visualized in Figure 1 and 2, a disproportionate representation exists between male and female entries, with 53.9% identified as male (213,543) and 46.1% as female (182,646). While this may seem modest, such imbalances become more problematic when they intersect with gender-stereotyped occupational labels, where roles like “nurse” or “yoga teacher” show over **80% female representation**, while positions such as “software engineer” or “rapper” are male-dominated ([Mansouri et al., 2024](https://www.mdpi.com/2673-2688/5/1/19); [Mihaljević et al., 2022](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274312))

Further disaggregation by profession (Figure 3) illustrates that certain occupations are strongly gender-coded in the dataset. For instance, female representation is overwhelmingly dominant in roles such as:

* Dietitian (93%)
* Nurse (91%)
* Yoga Teacher(85%)
* Interior Designer (81%)

In contrast, male-dominated professions include:

* Rapper (90%)
* Dj(86%)
* Surgeon (85%)
* Software Engineer (85%)

These imbalances are more than representational; they shape model behavior. For instance, ([Njoto et al.,2022](https://aclanthology.org/2022.nlpcss-1.15/)) illustrate that pretrained NLP models frequently learn semantic associations between gendered terms and specific job roles, leading to disparate treatment of candidates in algorithmic hiring pipelines. These disparities reinforce stereotypical gender-role associations and pose serious risks when used in training AI systems for job screening or candidate recommendation. Without intervention, models trained on such data are likely to learn and reproduce biased mappings between gendered language cues and professional competence.

The analysis of the BiasBios dataset identified gender imbalances both in overall representation and gender-stereotyped occupations (e.g., 90% of “rappers” male; 93% of “dietitians” female). This supports the detection of societal and occupational gender biases in training data commonly used by NLP models.

A screenshot of a computer

AI-generated content may be incorrect. A graph with a blue and pink bar chart

AI-generated content may be incorrect.

**Figure 1:** Gender distribution on whole dataset. **Figure 2:** Gender distribution in BiasBios dataset

A graph with different colors and text

AI-generated content may be incorrect.

**Figure 3:** Gender Representations in various Professions

**2.1.2. Synthetic Data for Bias Mitigation:**

To counteract the aforementioned biases in training data, synthetic data generation has emerged as a promising solution ([Peña et al., 2023](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)),([Frazzetto, 2025](https://tesidottorato.depositolegale.it/bitstream/20.500.14242/202137/1/tesi_Paolo_Frazzetto_fin.pdf)). The synthetic data are designed to vary along critical demographic dimensions while preserving semantic and professional coherence. Proxy terms themselves are not inherently biased; however, their strong associations with protected attributes make them crucial for evaluating model bias. This is consistent with recent research that leverages counterfactual generation to assess and reduce biased model behavior ([Kumar et al., 2023](https://ceur-ws.org/Vol-3490/RecSysHR2023-paper_7.pdf)).

By leveraging controlled synthetic data and fairness-driven evaluation protocols, this work contributes to the development of transparent, accountable, and equitable AI tools in recruitment pipelines. This approach is aligned with emerging principles (fairness, transparency, accountability and scalability) of **Human-Centered AI** ([Bartl et al., 2025](https://dl.acm.org/doi/full/10.1145/3700438)) and contributes to a growing body of work advocating for **accountable AI in employment technologies** ([Chaturvedi & Chaturvedi, 2025](https://arxiv.org/abs/2504.21400)),([Serna et al., 2023](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)). Moreover, it provides a scalable blueprint for testing bias sensitivity in downstream NLP applications, especially in high-stakes domains.

1. **Materials and Methods:**

**3.1 BiasBios Dataset:**

BiasBios is a large-scale, human-labeled dataset developed to facilitate the study and mitigation of gender bias in machine learning systems, particularly in automated occupation classification tasks. Introduced by [De-Arteaga et al., (2019).](https://arxiv.org/abs/1901.09451) in their seminal work “Mitigating Unwanted Biases with Adversarial Learning” (AIES 2019), the dataset comprises approximately 400,000 short biographies scraped from English-language publicly available professional platforms. It covers 28 diverse occupations including healthcare, tech, education, and law. It includes binary gender annotations inferred from first names or pronouns. BiasBios provides a realistic benchmark for evaluating fairness-aware algorithms and debiasing techniques. Due to the inherent occupational and gender imbalances within the dataset—mirroring real-world labor distributions—it is widely used in fairness research to analyze, detect, and mitigate societal biases perpetuated by machine learning classifiers and natural language processing (NLP) models.

* 1. **Models:**

We use the large language models **LLaMA-2-7B** and **LLaMA-3.2-3B** to automatically generate professional documents: **CVs**, **job advertisements**, and **cover letters**. These models are prompted with structured data and sample bios from the BiasBios dataset to produce tailored, coherent, and contextually appropriate content. By leveraging the models’ ability to understand and generate human-like language, we can create customized outputs that reflect specific job roles, skills, and candidate profiles while also enabling scalable document generation for hiring or application pipelines.

* + 1. **Model LLaMA-2-7B-Chat (Meta AI, 2023):**

The LLaMA-2-7B-Chat model is a 7-billion parameter instruction-tuned large language model released by Meta AI as part of the LLaMA-2 family. It is specifically fine-tuned for dialogue-based interactions and chat-style completions. Built on a transformer architecture, LLaMA-2 was trained on 2 trillion tokens of publicly available data and fine-tuned using Reinforcement Learning from Human Feedback (RLHF). Its context window is 4096 tokens. It performs competitively with other open-source models like Falcon-7B and Mistral-7B on instruction-following tasks and is particularly suited for structured generation tasks such as CV/cover letter synthesis and bias analysis ([Touvron et al., 2023](https://arxiv.org/abs/2307.09288); [Shumer et al., 2023](https://mistral.ai/news/announcing-mistral-7b))

* + 1. **Model LLaMA-3.2-3B-Instruct (Meta AI, 2024)**

The LLaMA-3.2-3B-Instruct is a lightweight, instruction-tuned language model from the LLaMA 3 family developed by Meta AI. As a 3.2 billion parameter variant, it offers a strong trade-off between computational efficiency and linguistic capability, making it particularly well-suited for low-latency inference, real-time evaluation, and structured generation tasks under resource-constrained environments. Unlike its larger counterparts (e.g., LLaMA-3 7B and 70B), the 3.2B model emphasizes efficiency over scale, allowing for high-throughput generation across applications such as CV/cover letter synthesis, prompt probing, and fairness stress-testing. Despite its smaller size, it achieves competitive instruction-following performance through fine-tuning on a combination of supervised instruction datasets and synthetic reasoning prompts, including chain-of-thought, multi-turn dialogues, and system-level instructions. ([Meta AI.,2024](https://ai.meta.com/blog/meta-llama-3/))

* 1. **Hardware specifications:**

The models were hosted and run on a **local workstation/server** with the following **GPU and CPU specifications**: **CPU**: AMD Ryzen (32 Cores, 64 Threads, 3.5 GHz base clock). High-core-count CPU is used for efficient token pre-processing and parallel sampling. **GPU**: NVIDIA A100 40GB (PCIe version). CUDA Capability: 8.0. FP16/TF32 optimization for faster matrix multiplications during transformer forward passes.

* 1. **Experimental Setup:**

Data:

The extracted dataset used for the experiments consists of 1,000 ‘accountant’ bios, each represented as a row containing two columns:

* text: The biographical text, describing the individual's professional background.
* labels: The corresponding label (gender) associated with each bio .

All text entries were lowercased, which simplifies normalization and text comparison. The bios contain common English stop words, many of which are not semantically useful for downstream NLP tasks. Stop word removal was implemented by splitting each bio into tokens (words), filtering out those matching the stop word list, and rejoining the filtered tokens into a cleaned string.

A new column named clean\_text was added to the dataset. Each row now includes the original lowercase bio (text), the filtered version (clean\_text), and the associated labels.

This preprocessing step ensures the model is trained on more informative content, reducing noise and improving generalization in learning tasks.

Classification:

Classification is carried out using a transformer-based architecture for text classification using the BERT (Bidirectional Encoder Representations from Transformers) model ([Jacob Devlin et al., 2018](https://arxiv.org/abs/1810.04805)). Specifically, we utilized the bert-base-uncased variant available through the Hugging Face Transformers library ([Wolf et al., 2020](https://aclanthology.org/2020.emnlp-demos.6/)), implemented via the AutoModelForSequenceClassification class. The input data is pre-processed using the associated tokenizer with a fixed max\_length (by default 512) padding strategy to ensure uniform sequence length across all samples. The dataset was partitioned into training and test sets using a standard 80/20 train–test split. For training the model, we leveraged the Trainer API along with TrainingArguments from the Transformers framework, which allows streamlined training and evaluation workflows.

Evaluation:

After training the classifier to predict gender from the input text (bios, cv/cover letters), explanations were obtained using LIME (discussed in the section of Gender classifiers). Evaluation of model performance was conducted using a suite of classification metrics from the scikit-learn package, including accuracy, precision, recall, and F1-score, to comprehensively assess the predictive performance. Additionally, a confusion matrix was computed to visualize the distribution of classification results, which was rendered using a heatmap generated with the seaborn visualization library.

* 1. **Gender classifiers:**

To ensure interpretability of the model predictions, we employed LIME (**Local Interpretable Model-agnostic Explanations**) ([Ribeiro et al., 2016](https://arxiv.org/pdf/1602.04938)), a post-hoc explanation technique designed to provide insight into individual predictions made by black-box models. It explains the model’s behavior only around a single instance being analyzed, not globally. This is achieved by generating perturbed samples about the instance and observing the black-box model's predictions on these samples. The coefficients of this surrogate model indicate the importance of different input features (e.g., words in a sentence) for the prediction.

When explaining the model's prediction for an instance (e.g. CV or coverletter), LIME highlights the words that contributed most to the model’s decision. By analyzing multiple instances where the model predicts a particular label (e.g., “male” or “female”), one can observe if certain words consistently have high importance. The classifier’s performance is an indicator of how much sensitive gender information is present in the data. Proxy words may often appear as highly weighted features in the local explanations, revealing that the model may be relying on these words as shortcuts or proxies for sensitive attributes. This identification helps in auditing and mitigating bias, for example by removing or masking proxy words or adjusting training data.

To identify test samples with **low prediction confidence**, we use the absolute difference between the top two predicted class probabilities.

*diffs = np.abs(pred\_probs[:, 0] - pred\_probs[:, 1])*

As the experiment unfolds, a smaller diff value indicates that the model found it harder to distinguish between class 0 and class 1 — i.e., it was **less confident** in its prediction. The top rows in the table have the **lowest diff values** (e.g., 0.038370), meaning the model was almost equally likely to assign the input to either class.

These low-confidence predictions are valuable for:

**-**Investigating potential patterns in ambiguous or hard-to-classify bios.

**-**Flagging samples for manual inspection or labeling.

**-**Prioritizing these cases in active learning or re-training.

In addition to analyzing low-confidence predictions, we also examined the **most confident predictions** made by the model. This is useful for:

-Verifying the correctness of confident predictions.

-Building a **high-precision subset** for downstream tasks.

-Comparing feature patterns in clear vs. ambiguous cases.

* + 1. **Gender Classifier Without Masking:**

In our setup for gender classification using unprocessed text—without applying any masking, lowercasing, or stop word removal—the model was trained to predict gender labels, where "female" is represented as 1 and "male" as 0. The classifier relied on the full text of CVs and cover letters, making predictions based on the language patterns present. Certain words appeared to influence the model’s decisions more strongly toward one gender than the other. These included direct gender indicators (e.g., pronouns, titles, and names) as well as indirect cues, such as professional roles or cultural references. This suggests leaving text unprocessed (i.e., not masking identity terms or normalizing stylistic features) enables models to learn and act upon the signals of direct gender indicators (e.g., *he*, *Ms.*, *Jessica*) or indirect or proxy indicators (e.g., *nurse*, *CEO*)—leading to potential gender bias in classification or downstream decisions (e.g., resume screening).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A screenshot of a graph  AI-generated content may be incorrect. | | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('He', 3.34961242228)  ('he', 1.9166196645)  ('his', 0.9823642110)  ('him', 0.59896187373)  ('Mr', 0.212177381359)  ('For', 0.15553675710)  ('an', 0.139449532486)  ('services', 0.127056438325)  ('help', 0.12677291047)  ('Penticostal', 0.126314986395) | ('She', 20.334972885)  ('her', 6.5985962079)  ('she', 5.3634170559)  ('Her', 2.318591337)  ('Ms',-1.21109984780)  ('Mrs', 0.54779522537)  ('Jessica', 0.54219316489)  ('herself', 0.51449886288)  ('Their', 0.488303965837)  ('CPAs', 0.47497754571) | | |
| **Figure 4:** Gender classifier using LIME (without masking) for ‘accountant’ Bio. | **Figure 5:** List of top 10 Gender words without masking. | |

* + 1. **Gender Classifier with lowercase text and without Masking:**

The dataset text was converted to **lowercase** before being analyzed, which ensures consistency and avoids discrepancies due to capitalization.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('he', 18.6270818434)  ('his', 5.95132019243)  ('has', 2.3456170134)  ('him', 1.4558856212)  ('is', 0.89128831843)  ('mr', 0.71970626260)  ('and', 0.66137833001)  ('of', 0.49258934339)  ('was', 0.47370002206)  ('with', 0.334369473813)) | ('she', -19.308946400)  ('her', -6.01606865461)  ('ms', -1.34748517974)  ('wendy', -0.482439262318)  ('jessica', -0.469753149833)  ('catherine', -0.39792333042)  ('carla', -0.395147152900)  ('women', -0.36915380735)  ('heidi', -0.353715836909)  ('helen', -0.295214132876) | |
| **Figure 6:** Gender classifier using lowercase text and without masking | **Figure 7:** List of top 10 Gender words(lowercase) |

* + 1. **Gender Classifier with masking gender pronouns and lowercase text:**

We introduce a gendered pronoun masking list, defined as GENDERED\_PRONOUNS\_LIST. This technique replaces explicitly gendered pronouns with neutral equivalents (e.g., "he" → "they", "her" → "their", "himself" → "themselves"). This is implemented as part of a de-biasing pipeline to reduce the model’s dependence on overt pronouns. (Figure 8)

|  |  |
| --- | --- |
|  |  |
| **Figure 8:** Gendered\_Pronouns\_List | **Figure 9:** Confusion Matrix when masked pronouns |

The confusion matrix (Figure 9) predicted all correct predictions, 121 for male and 79 for female, without any discrepancies.

The map is applied to all input texts, replacing binary pronouns and titles with gender-neutral forms to mask explicit gender signals. A probabilistic binary classifier (e.g., fine-tuned BERT) is trained to predict author gender based on masked and normalized text. We integrate LIME to provide token-level feature attribution to show which words led to higher probability toward prediction “male” or “female” and to particularly observe how gendered tokens, both pronouns and topic-specific terms, influence the model.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('has', 1.17898188322)  ('and', 1.17655255983)  ('with', 0.77307225342)  ('clients', 0.69116933936)  ('of', 0.63496592557)  ('is', 0.62597390435)  ('the', 0.55393978037)  ('chairman', 0.39864502513)  ('firm', 0.350447666888)  ('an', 0.32305872052) | ('their', -5.4682621306)  ('they', -4.3065325870)  ('cpas', -2.88792374435)  ('jennifer', -0.90215375421)  ('ms', -0.87324952951)  ('helen', -0.86212453973)  ('lynette', -0.76605312065)  ('she', -0.73431637606)  ('vinita', -0.72458977815)  ('dorothy', -0.70042032430) | |
| **Figure 10:** Gender classifier with masking gender Pronouns | **Figure 11:** List of top 10 Gender words with Masking gender pronouns |

* + 1. **Gender Classifier with masking Names and pronouns:**

We introduce a de-biasing pipeline that combines name masking and gendered pronoun masking. Names are replaced with neutral placeholders [NAME] to remove identity-specific cues, while a predefined GENDERED\_PRONOUNS\_LIST (Figure 8) is used to substitute explicitly gendered pronouns with neutral equivalents (e.g., "he" → "they", "her" → "their", "himself" → "themselves". This dual-masking approach aims to reduce the model’s reliance on overt gender indicators and mitigate bias in downstream predictions. To enhance the quality of text for modeling, a manual list of stop words was used for filtering. These words are typically non-informative for tasks like classification or text generation. The following stop words were removed from dataset:

["the", "a", "an", "and", "has", "of", "to", "in", "as", "is", "was", "were", "are", "for"]

Figure 13 presents a confusion matrix reflecting the model’s classification performance after implementing a different masking strategy—likely one that still includes gendered information or proxy terms. Here, the model correctly predicts 106 instances as class 1 (likely "Male") and 63 instances as class 0 (likely "Female"). However, 16 female samples were misclassified as male, and 15 male samples were misclassified as female. Compared to the previous matrix, this result shows a slight increase in misclassifications for both classes. While the overall accuracy remains relatively high, the rise in false negatives for male samples and false positives for female samples may suggest a residual bias or less effective masking. This highlights the importance of comprehensive de-biasing techniques to ensure balanced model performance across gender labels.

|  |  |
| --- | --- |
|  |  |
|  | **Figure 13:** Confusion Matrix when masked Names and pronouns |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('their', 14.6906762877)  ('also', 1.47572660242)  ('with', 1.1655265219)  ('mr', 1.1490734879)  ('firm', 0.96374543498)  ('been', 0.81750617225)  ('member', 0.68123684162)  ('chief', 0.68081897801)  ('led', 0.64416844677)  ('joined', 0.60580479599) | ('them', -12.4820850622)  ('they', -6.3435352564)  ('cpas', -2.17668464134)  ('ms', -1.648035163)  ('catherine', -0.81104826285)  ('lisazamparo', -0.78613007418)  ('johanna', -0.77123226101)  ('eithne', -0.71688879392)  ('mrs', -0.68312876278)  ('husband', -0.60467560403) | |
| **Figure 14:** Gender classifier with Masking Names and Pronouns | **Figure 15:** List of top 10 Gender words) with Masking Names and pronouns |

* + 1. **Gender Classifier with masking Social and Workplace roles:**

Figure 15 shows a curated list of proxy gendered terms categorized by perceived gender associations used for Gender Classification task with Masking Social and Workplace roles (Accountant). The words — such as "husband," "executive," and "investor" for males, and "wife," "mother," and "support" for females — reflect social, familial, and workplace roles that often carry gender-coded connotations. These proxy terms were systematically replaced with neutral alternatives like " " to minimize gender bias during model training. To enhance the quality of text for modeling, a manual list of stop words was used for filtering. These words are typically non-informative for tasks like classification or text generation. The following stop words were removed from dataset:

["the", "a", "an", "and", "their", "they", "them", "has", "of", "to", "in", "as", "is", "was", "were", "are", "for"]

Figure 16 displays a confusion matrix highlighting the model’s performance after applying both name and pronoun masking as part of the gender de-biasing process. The matrix reveals that the model correctly identified 106 instances as class 1 (likely "Male") and 49 instances as class 0 (likely "Female"), while it misclassified 30 female samples and only 15 male samples. This suggests a relatively strong predictive performance post-masking, with a notable reduction in false positives for male-labeled samples, indicating improved balance in gender classification accuracy.

|  |  |
| --- | --- |
|  |  |
| **Figure 15:** Proxy Words\_List | **Figure 16:** Confusion Matrix when masked Social and workplace roles |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('cfo', 0.77153954520)  ('mr', 0.69781285287)  ('NAME', 0.61565521068)  ('clients', 0.55813310238)  ('nathan', 0.54272420036)  ('africa', 0.50768297750)  ('ceo', 0.49910892680)  ('palm', 0.476064840504)  ('married', 0.39430602224)  ('well', 0.380954417607) | ('ms', -3.2044305602)  ('mrs', -1.20517361042)  ('jennifer', -0.85710581796)  ('catherine', -0.83606492114)  ('she', -0.83071808506)  ('eithne', -0.82835549021)  ('chairwoman', -0.80797391132)  ('amy', -0.80607115835)  ('jane', -0.77951221925)  ('shermanetta', -0.71213493092) | |
| **Figure 17:** Gender classifier with Masking Social and workplace roles | **Figure 18:** List of top 10 Gender words) with Masking Names and pronouns |
|  |  |
|  |  |

* + 1. **Gender Classifier with masking Personality traits:**

The Personality\_LIST is a dictionary that groups gender-related personality traits under "male" and "female". Traits (like "ambitious", "competitive", "confident") are paired with the neutral symbol " " under ‘Male’ category and Traits (like "empathetic", "supportive", "responsible") are paired with the neutral symbol " " under ‘Female’ category (Figure 19). This setup is used to test how well a gender prediction model performs when these traits are masked. This helps reduce bias and hides obvious gender signals during testing.The confusion matrix shows how well the gender model works when personality traits are masked. The model correctly predicts male (class 1) very well — it gets 105 right and 16 wrong. But for female (class 0), it’s less accurate — it gets 48 right but makes 31 mistakes, wrongly labeling them as male (Figure 20).This shows the model still finds, male-linked traits more detectable, even after masking, while it struggles more with female-linked traits when the gender signal is hidden.

|  |  |
| --- | --- |
|  |  |
| **Figure 19:** PersonalityTraits\_List | **Figure 20:** Confusion Matrix when masked Personality Words |
|  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('firm', 1.12320166568)  ('cfo', 1.01663746955)  ('mr', 0.79497674884)  ('business', 0.75258890695)  ('clients', 0.74351335758)  ('management', 0.6305861748)  ('officer', 0.59233758302)  ('president', 0.58675362988)  ('on', 0.54873321321)  ('time', 0.53368911700) | ('ms', -2.53728012150)  ('cpas', -1.11125453694)  ('mrs', -0.97005281594)  ('eithne', -0.87057661819)  ('accountant', -0.85611043580)  ('catherine', -0.83720843904)  ('jennifer', -0.8327113712)  ('amy', -0.8200314079)  ('jane', -0.78240314218)  ('ava', -0.75021025728) | |
| **Figure 21:** Gender classifier with Masking Social and workplace roles | **Figure 22:** List of top 10 Gender words) with Masking Names and pronouns |
|  |  |

* 1. **Synthetic Data Generation**

To investigate potential biases in large language models (LLMs) when generating job application materials and job advertisements, we constructed a synthetic dataset comprising biographical inputs (bios), CVs, cover letters (CLs), and job ads, each having explicit gender info. Our experiment leverages two generative LLMs of varying scale — **LLaMA-2-7B** and **LLaMA-3.2-3B** — and follows a multi-phase pipeline that systematically simulates the job application process. The setup is designed to examine how identity and background features in bios influence the language, tone, and content of model-generated outputs, as well as how models adapt to targeted job descriptions.

We curated a set of biographical sketches (bios), each describing a hypothetical candidate with varying identity attributes (e.g., gender, ethnicity), educational background, work experience, and professional goals. These bios serve as the foundational inputs across all stages of generation.

* + 1. **Job Advertisement Generation:**

Each bio was used to condition the generation of job advertisements using both LLaMA 2.8B and LLaMA 3.2B models. We employed a consistent prompt template:

*“Generate a job ad for the following profession. {profession}”*

This step produced a total of **10 job ads**, each for every set of bios. These ads reflect how the models interpret and emphasize different features from a candidate’s profile when generating a corresponding job opportunity.

Focusing on the “Accountant” role as a case study, we extracted all job ads generated for accountant-related bios and summarized them into a single unified job description for each model. This summarization was performed using the prompt detailed in **prompt\_summarizedJobAD\_Accountant.txt**

This step produced two standardized job ads (one per model) for the accountant role to be used in downstream application generation.

* + 1. **CV and Cover Letter Generation (Generic Application):**

For each bio, we generated both a CV and a cover letter using each model and the generated job ad. Separate prompts were used for CVs and CLs to better reflect real-world application document structures:

* CV Prompt:

*“Generate a professional CV based on the following biography:\n\n{bio}\n\n*

*Generated CV:”*

* CL Prompt:

*"Write a personalized cover letter for the following biography:\n\n{bio}\n*

*Generated CoverLetter:”*

This resulted in a total of 2 x10 x 2 documents (10 CV and 10 CL per model).

To simulate a scenario where candidates apply for a specific job, we selected a representative ‘accountant’ bio and asked each model to generate a targeted CV and CL using the summarized job description. Subsequently, the same prompts for CV and cover letter generation were reused to produce a unified batch of outputs, which were compiled into a consolidated CSV file to facilitate downstream analysis and further experimental evaluation.

* 1. **Evaluation experiments:**

**3.7.1 Gender classifier on generated synthetic data (without lower casing and masking):**

To interpret model predictions and identify influential features within the synthetic application documents, we applied a LIME classifier directly to the consolidated CSV containing generated CVs and cover letters. The input text was used in its original form—without applying any preprocessing such as stop word removal, lowercasing, or word masking—to preserve the semantic and lexical nuances in the generated content. This approach allowed us to assess how surface-level features and identity-linked language might contribute to classifier behavior.

**A screenshot of a computer

AI-generated content may be incorrect.**

**Figure 23:** Gender classifier on generated CV/CL without masking

|  |  |
| --- | --- |
| Top 10 Male words | Top 10 Female words |
| ('Boca', 0.0144123908891)  ('problem', 0.000467150125930)  ('about', -0.002215660556)  ('gain', -0.0149808646086)  ('Dear', -0.024377636960)  ('provide', -0.025381619710)  ('notary', -0.033302472296)  ('considering', -0.044621557865)  ('I', -0.0528930293781)  ('request', -0.057881901255) | ('childhood', -0.0491804092233)  ('cultural', -0.045383766627)  ('number', -0.042381719462)  ('Thank', -0.040983936058)  ('a', -0.040670710436)  ('look', -0.0401269232593)  ('proven', -0.03898721714)  ('implemented', -0.035664336821)  ('free', -0.0338793104387)  ('dedicated', -0.0314995503925) |

**Figure 24:** List of top 10 Gender words on generated CV/CL without masking

* + 1. **Classifier with lowercasing the text and stop words, and masking names, pronouns, social and workplace proxies:**

To reduce surface-level lexical bias and examine deeper structural and semantic patterns in gender prediction, we conducted a second round of classification using a pre-processed version of the synthetic CV and cover letter dataset. This version applied the following text normalization and masking steps:

* Lowercasing: All text was converted to lowercase to neutralize case-based variations.
* Stop Word Removal: Common English stop words (e.g., *the*, *and*, *in*) were removed to emphasize meaningful content terms.
* Entity and Proxy Masking:
  + Names
  + Pronouns (e.g., *he*, *she*, *his*, *her*), Figure 8
  + Social identifiers (e.g., *mother*, *brother*), Figure 15
  + Workplace and positional cues (e.g., *manager*, *assistant*, *executive*)

Each identified token was replaced with a neutral placeholder (e.g., [NAME],or [ ]) to minimize the influence of direct identity cues on the classification process.

This setup was designed to test whether gender prediction could still be achieved when overt linguistic signals were stripped, thereby isolating subtler semantic or stylistic patterns. The resulting model performance and LIME explanations were compared against the original, unprocessed classifier to evaluate the persistence of bias even under content obfuscation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | |  |  | | --- | --- | | Top 10 Male words | Top 10 Female words | | ('nbahaa', 0.011897478872)  ('main', 0.0025193954913)  ('certifications', 0.002089670896)  ('knowledge', 0.0009344210469)  ('not', 0.000357532212618)  ('high', -0.00045574813184)  ('prepared', -0.011131784865)  ('strong', -0.0123455308287)  ('sincerely', -0.012858732555)  ('expand', -0.01677243165) | ('excited', -0.01627939078)  ('dear', -0.01273802769)  ('empowering', -0.01150273659)  ('memories', -0.01079713000)  ('sincerely', -0.01037179952)  ('name', -0.01029350643)  ('look', -0.01017772986)  ('i', -0.00896300982)  ('rica', -0.00822006863)  ('years', -0.00819463317) | |
| **Figure 25:** Gender classifier on generated CV/CL with lowercasing and stop words, and masking names, pronouns, social and workplace proxies | **Figure 26:** List of top 10 Gender words on generated CV/CL with lowercasing and stop words, and masking names, pronouns, social and workplace proxies |

**ALEX: Can you include the experiment where you generate the CV/CL with as much gender info as possible masked from the bio? It would be interesting to compare the classifier results. It would also be interesting to do a further experiment with gender-masked bios, but injecting proxy words.**

1. **Results:**
   1. **Analysis on BiasBios dataset**

As shown in Figures 1 and 2, the dataset is 53.9% male (213,543) and 46.1% female (182,646). While this difference may seem small, it becomes more serious when combined with gender-stereotyped job roles. For example:

* Over 80% of “nurses” or “yoga teachers” are female.
* Jobs like “software engineer” or “rapper” are mostly male.
  1. **Gender Classifier** 
     1. **Classifier without masking or lower text:**

The LIME classifier predicts gender based on text, where female is predicted as 1 and male as 0 (Figure 4). Certain words seemed to push the model more strongly toward one gender than the other. For example, male-related words like *he*, *his*, *him*, *Mr.*, *Penticostal*, *aspects*, and *grant* made the model more likely to predict "male". On the other hand, female-related words such as *she*, *her*, *Ms.*, *Mrs.*, *Jessica*, *CPAs*, *McGinty*, and *financial* increased the chances of predicting "female" (Figure 5). This shows that the model may be picking up on indirect gender clues, or proxy words, which are often tied to cultural roles or job titles.

* + 1. **Classifier with lowercase text and without masking:**

In our setup, the model predicts gender based on text, where female is predicted as 1 and male as 0 (Figure 6). Certain words clearly influenced the model’s predictions more than others. For instance, male-associated words like *he*, *his*, *him, post, reporting*, made the model more likely to predict "male." In contrast, female-associated words such as *she*, *her*, *ms*, *women*, *joining*, *financial*, increased the likelihood of predicting "female" (Figure 7). These patterns highlight how text-based models can learn and reproduce gender biases based on word usage, job roles, and social cues embedded in the data.

* + 1. **Classifier with masking gender pronouns and lowercase text:**

The LIME classifier interpretation reveals distinct word associations driving gendered predictions. (Figure 10) When prediction probabilities are skewed with 'Female' = 0.84 and 'Male' = 0.16, the top words influencing a female classification include terms like "*cpas*," "*accounting*," "*vinita*," and "*responsible*", which may reflect personal names or organizational references linked to gender. In contrast, words such as "*meara*," "*clients*," "*chairman*," and "*firm*" dominate male-associated predictions, indicating a professional or hierarchical bias (Figure 11). These patterns suggest that the model leverages both lexical cues and contextual gender stereotypes, reinforcing the need for careful mitigation strategies to ensure fairness and neutrality in classification outcomes.

* + 1. **Classifier with Masking Names and pronouns:**

The LIME classifier revealed stark differences in the prediction probabilities and feature attributions for gender classification when masked for Names and pronouns both. In a representative case, the model predicted 'Male' with 0.99 probability and 'Female' with only 0.01, indicating strong gendered bias in favor of the male label (Figure 14). Words most associated with the 'Female' prediction included "angel", "cpas", "husband"—terms often stereotypically linked with traditionally female-associated roles. In contrast, 'Male'-associated terms like "kingdom", "director", "companies", "kenya", "chief" -- reflect high-ranking, leadership-oriented positions (Figure 15). This contrast highlights how occupational and contextual word choices can reinforce gendered stereotypes in model outputs.

* + 1. **Classifier with masking Social and Workplace roles:**

In this instance, the LIME classifier assigns a prediction probability of 1 to 'Female' and 0 to 'Male' for an accountant biography where social and workplace roles have been masked (Figure 17). The top influential words contributing to the 'Female' classification include a*ccountant*, *financial*, *responsible*, *mrs*, *chairwoman*, along with several masked name tokens and the top influential words contributing to the Male classification include: *marries*, *senior africa*, *clients*, *ceo*, *cfo* (Figure 18). These results highlight that even in the absence of overt gender indicators, the classifier still latched onto semantic, regional, or role-related language to infer gender. For instance, terms like *ceo* and *cfo* — even when role-masked — may have been incompletely filtered or emerged through context. Similarly, *chairwoman* and *mrs* might indicate partial leakage or structural patterns retained in the generation process.

This finding underscores the challenge of fully anonymizing application texts and suggests that bias may be embedded in structural phrasing, word associations, and occupational framing, rather than just explicit identity markers.

* + 1. **Classifier with masking Personality traits:**

In this instance, the LIME classifier outputs a prediction probability of 0.01 for ‘Male’ and 0.99 for ‘Female’, indicating a strong bias toward predicting the bio as male (Figure 21). The classifier also provides insights into the top words influencing each prediction. For the ‘Female’ class, words like “preparation,” “accountant,” “cpas,” “experience,” “accounts,” were influential. These terms are largely domain-specific and related to accounting and finance, suggesting that the model still finds professional language partially associated with female predictions. On female side many names also appeared even if the experiment was masked for names. On the other hand, for the ‘Male’ class, words such as “senior,” “firm,” “management,” and “business” were highlighted terms that are more male oriented. (Figure 22).

The lack of distinctive male-coded features influencing the prediction, despite the high male probability, may reflect underlying model bias or a tendency to default to male classification when gender cues are weak or ambiguous. This outcome underscores that even with neutral or technical language, classifiers can lean heavily on subtle biases, possibly rooted in training data distributions or historical associations within job descriptions.

* 1. **Evaluation experiments:** 
     1. **Classifier on generated synthetic data (without lower casing and masking):**

The classifier yielded the following average prediction probabilities:Female class: 0.81andMale class: 0.19 (Figure 23)

These results suggest a stronger model confidence in identifying documents associated with male-coded bios, potentially indicating asymmetric linguistic cues across gendered outputs.LIME further highlighted the most influential words contributing to each gender prediction. Top tokens associated with *female* classification included:*excited*, *sincerely*, *quickly*, *cultural*, *proven.* Top tokens associated with *male* classification included:*problem*, *gain*, *notary*, *request* (Figure 24).These findings reflect subtle stylistic and semantic differences in LLM-generated documents that may reinforce or reflect stereotypical gender-coded language. The prominence of affective or interpersonal terms in female-associated outputs versus action- or task-oriented terms in male-associated outputs aligns with prior literature on gendered communication in professional writing.

* + 1. **With lowercasing and stop words, and masking names, pronouns, social and workplace proxies:**

Following the initial classification on raw, unprocessed CVs and cover letters, we re-ran the LIME classifier on a version of the synthetic dataset where all text was lowercased, stop words removed, and key identity-related entities — such as names, pronouns, social descriptors, and workplace role proxies — were masked. This preprocessing aimed to neutralize overt lexical gender markers and evaluate the classifier’s reliance on deeper linguistic signals.

Prediction Probabilities resulted to Female class: 0.52 and Male class: 0.48 (Figure 25)

These near-equal probabilities indicate a significant reduction in model confidence compared to the unmasked version suggesting that identity-specific language played a substantial role in initial predictions. The drop in discriminatory power reflects successful obfuscation of key gender-linked features.

The Top influential tokens identified by LIME for Female-associated words are *accounting*, *sincerely*, *experimenting*, *references*, *memories*, *empowering,* and for Male-associated words are *certifications*, *knowledge*, *prepared*, *strong* (Figure 26). Interestingly, even after identity masking, certain semantic and stylistic cues persisted in influencing gender classification. Words like *sincerely* and *empowering* may reflect interpersonal tone, while *strong* and *certifications* suggest professional self-presentation — echoing subtle gender-coded communication norms observed in real-world data. These findings demonstrate that bias may persist even in anonymized settings, revealing the importance of structural and thematic language in shaping algorithmic outcomes.

* 1. **Analysis based on Accuracy metrics on Test data:**

A white sheet with black text and numbers

AI-generated content may be incorrect.**Figure 27**: Table of accuracy metrics for all classifier experiments

To evaluate how linguistic obfuscation and preprocessing impact gender classification performance, the performance metrics of Accuracy, Recall, Precision, and F1 Score on all classifiers under varying levels of text masking and normalization are conducted. The baseline classifier used the original, unprocessed dataset (with full names, pronouns, and casing intact), and additional variants introduced targeted masking strategies. The report in the Figure 27 is interpreted as follows:

* The baseline model (no masking, original casing) achieved perfect scores across all metrics (Accuracy = 1, F1 = 1), demonstrating that direct gender markers (e.g., names, pronouns) strongly influence classification.
* Applying lowercasing or masking only pronouns did not affect performance; scores remained at 100%, indicating these signals alone are not critical when other identity cues are present.
* Masking both names and pronouns led to a moderate drop in performance (Accuracy = 0.85, F1 = 0.87), showing these cues are important but not solely responsible for gender prediction.
* Masking proxy words (e.g., gender-associated job roles, social descriptors) further reduced accuracy to 0.77, and masking personality-related terms (e.g., words associated with affect or behavior) yielded similar results (Accuracy = 0.76, F1 = 0.82).

These results indicate that proxy features and stylistic language also contribute significantly to gender inference, beyond explicit identity terms.

* When applying the classifier to LLM-generated application text without masking or lowercasing, performance dropped sharply (Accuracy = 0.75, Recall = 0), reflecting misalignment between synthetic and training distributions.
* In the most obfuscated setting—masking all proxy signals in generated text—the classifier was effectively neutralized, with accuracy dropping to 0.5 (random chance) and a notably low precision of 0.33. This suggests that comprehensive proxy masking can effectively disrupt gender prediction, highlighting its potential as a fairness-preserving intervention.
  1. **Analysis based on Low confidence of gender predictions:**

A white sheet with black text

AI-generated content may be incorrect. **Figure 28**: Table of Low confidence while gender predictions

To better understand the certainty with which the gender classification model makes its predictions, the proportion of low-confidence predictions (i.e., predictions close to the decision boundary) under different levels of text masking and preprocessing was measured. A higher low-confidence score indicates greater model uncertainty, while a lower value implies stronger confidence in gender inference.

When only lowercasing was applied, low-confidence predictions increased sharply to 0.45, suggesting that capitalization of certain identity-related terms plays a role in boosting model confidence. Masking pronouns alone also increased uncertainty (0.70), indicating that even simple lexical gender cues significantly influence the classifier’s certainty. Masking both names and pronouns led to a sharp drop in confidence (0.14), while further masking of proxy words (e.g., job roles, identity indicators) caused low-confidence predictions to rise to 0.98 (Low Confidence = 0.02), indicating the classifier was no longer able to confidently detect gender. Similarly, masking personality-associated terms raised low confidence to 0.93 (Low Confidence = 0.07), showing that stylistic language also contributes to certainty.

The classifier also showed very low confidence when applied to LLM-generated CVs and cover letters:

* + Without masking or lowercasing: Low Confidence = 0.03
  + With all proxies masked: Low Confidence = 0.04

These values suggest that synthetic text, even when unmasked, introduces uncertainty likely due to domain shift or lack of familiar lexical signals. Once proxies are removed, the model becomes nearly agnostic to gender, unable to confidently assign label.

1. **Discussion:**

The study and experiment shown here set out to examine the persistence and mechanics of gender bias in text-based classification models within the hiring context, using both real-world data (BiasBios) and synthetic CVs and cover letters generated by large language models (LLMs). Across all experiments, the results consistently revealed that gender inference is deeply embedded in both explicit identity markers (e.g., names, pronouns, job titles) and more subtle semantic, stylistic, and structural cues.

The imbalances in the dataset such as BiasBios (Section 2.1.1 Gender Disparities in Training Data) don’t just reflect social trends — they directly affect how LLMs learn. Research shows that NLP models trained on such data start linking gendered language with specific jobs, which lead to biased or unfair treatment of candidates. For instance, women may be less likely to be recommended for technical roles, simply because the model has learned those roles are usually associated with men in the data. These representational skews serve as a foundation upon which gender biases in NLP models are likely learned and amplified.

Our experiments show that a significant amount of gender information is embedded in both real and synthetic documents—even when direct indicators such as names and pronouns are removed. Classifiers trained on unprocessed bios or CVs achieved near-perfect accuracy and confidence, with LIME revealing reliance on not only gendered terms (e.g., *she*, *Mr.*, *Jessica*), but also proxy words (e.g., *nurse*, *director*, *chairwoman*) and stylistic elements (e.g., tone, verb choices).

Masking explicit identity markers (names, pronouns, social/professional roles) reduces the classifier’s ability to confidently infer gender, but it does not eliminate gender signals completely. Our results showed that even with full masking and text normalization, classification probabilities only approached random (50/50) when all proxy features were also removed. This implies that gender is redundantly encoded in language through both vocabulary and structure.

Thus, effective masking should combine:

* Lexical masking (names, pronouns, identity terms)
* Contextual masking (job roles, personality traits)
* Stylistic normalization (removal of affective cues)
* Advanced techniques, such as adversarial perturbation or fair representation learning

When using LLMs to generate application materials (CVs, cover letters) from gendered bios, we observed that the generated text inherits and amplifies existing gender cues. LIME-based analysis revealed:

* Female-coded outputs often contained emotional or interpersonal language (e.g., *sincerely*, *excited*, *empowering*)
* Male-coded outputs emphasized assertiveness and achievements (e.g., *problem*, *gain*, *prepared*, *notary*)

Even when not explicitly asked to include gender-specific content, models tend to mirror real-world stereotypes found in their training data. This suggests that LLMs **are highly susceptible to reproducing gender bias**, especially in domains like hiring where occupational stereotypes are strong.

Careful prompt design, data augmentation with counter-stereotypical examples, and post-generation filtering are potential strategies to reduce this risk in synthetic document generation.

LIME (Local Interpretable Model-agnostic Explanations) has proved to be a valuable tool in:

* Identifying key lexical features that drive gender predictions
* Comparing models’ reliance on explicit vs. implicit gender cues
* Assessing the effectiveness of masking strategies across multiple classification setups

By visualizing token importance, LIME helps researchers pinpoint which features contribute to biased outcomes, thereby enabling more targeted interventions. Furthermore, combining LIME analysis with confidence scores and probability shifts provides a multi-dimensional understanding of model behavior under different data conditions. However, LIME still identified gender-skewed features such as “*sincerely*,” “*empowering*” (female) and “*certifications*,” “*strong*” (male), suggesting that stylistic tone and thematic framing remain as residual signals for gender inference, even under full anonymization.

The layered masking strategies, combined with diagnostic tools like LIME, offer a promising framework for auditing real hiring datasets for gender imbalance and bias-prone language, for de-biasing AI-generated application materials, and for training or fine-tuning fairer NLP models for downstream tasks such as resume screening, job recommendation, or applicant ranking.

While masking alone may not fully eliminate gendered patterns, it significantly weakens the link between text and gender prediction, particularly when combined with structural and semantic normalization. This opens the door to developing fairer recruitment tools that focus on skill and merit rather than linguistic proxies of identity.

Future work could explore:

* Adversarial debiasing techniques
* Inclusion of fairness constraints during model training
* Human-in-the-loop frameworks to guide fair evaluation and intervention

Ultimately, these methods provide a scalable and interpretable pathway toward reducing algorithmic discrimination in hiring and promoting equitable access to opportunity.

1. **Bibliography:**

 [De-Arteaga et al](https://arxiv.org/abs/1901.09451) (2019) . Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. [arxiv.org/abs/1901.09451](https://arxiv.org/abs/1901.09451)

 Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. <https://dl.acm.org/doi/full/10.1145/3700438>

 Chaturvedi, S., & Chaturvedi, R. (2025). Who Gets the Callback? Generative AI and Gender Bias. <https://arxiv.org/abs/2504.21400>

 Fabris, A., Baranowska, N., Dennis, M. J., Graus, D. et al. (2025). Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. <https://dl.acm.g/doi/10.1145/3696457>

 Frazzetto, P. (2025). Leveraging Deep Learning in Human Resources: Graph Neural Networks for Candidate-Job Matching. [Thesis PDF](https://tesidottorato.depositolegale.it/bitstream/20.500.14242/202137/1/tesi_Paolo_Frazzetto_fin.pdf)

 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-Art Natural Language Processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. <https://aclanthology.org/2020.emnlp-demos.6>

 Kumar, D., Greif, E., Rekabsaz, N., & Schedl, M. (2023). Identifying Words in Job Advertisements Responsible for Gender Bias via Counterfactual Learning. CEUR-WS. [PDF](https://ceur-ws.org/Vol-3490/RecSysHR2023-paper_7.pdf)

 Mansouri, T., Alameer, A., & Albaroudi, E. (2024). AI Techniques for Addressing Algorithmic Bias in Job Hiring. AI, 5(1), 19. <https://www.mdpi.com/2673-2688/5/1/19>

 Marti Marcet, S. (2023). Natural Language Processing in Resume Data: The Interplay Between Gender and Occupation on Resume Writing Style. Utrecht University. [PDF](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/44294/Applied_Data_Science_Thesis_Sara.pdf)

 Njoto et al. (2022). *Professional Presentation and Projected Power: A Case Study of Implicit Gender Information in English CVs.* <https://aclanthology.org/2022.nlpcss-1.15/>

 Mihaljević, H., Müller, I., Dill, K., & Yollu-Tok, A. (2022). Towards Gender-Inclusive Job Postings: A Data-Driven Comparisonof augmented writing technologies. PLOS ONE, 17(12). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274312>

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., ... & Scao, T. L. (2023).  
*LLaMA 2: Open Foundation and Fine-Tuned Chat Models. Meta AI.* <https://arxiv.org/abs/2307.09288>

Shumer, B., Jernite, Y., Papazian, A., Sun, Z., Roberts, A., & Raffel, C. (2023). *Mistral 7B.* Mistral AI. <https://mistral.ai/news/announcing-mistral-7b/>

Meta AI. (2024). *LLaMA 3: Open Foundation and Instruction Models.* <https://ai.meta.com/blog/meta-llama-3/>

 Peña, A., Serna, I., Morales, A., Fierrez, J., & Ortega, A. (2023). Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI‑Based Recruitment. SN Computer Science, 4(5). [s42979-023-01733-0.pdf](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)

 Ribeiro, M., Singh, S., & Guestrin, C. (2016). *Why should I trust you?": Explaining the predictions of any classifier.* [1602.04938v3.pdf](https://arxiv.org/pdf/1602.04938)

 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* <https://arxiv.org/abs/1810.04805>

1. **Declaration of Authorship**

I hereby certify that I composed this work completely unaided, and without the use of any other sources or resources other than those specified in the bibliography. All text sections not of my authorship are cited as quotations and accompanied by an exact reference to their origin.

Place, date: BFH, Biel

Signature: Shilpi Garg