IntersectionRE: Mitigating Intersectional Bias in Relation Extraction Through Coverage-Driven Augmentation

Amirhossein Layegh[1], Amir H. Payberah[1] and Mihhail Matskin[1]

[1]KTH Royal Institute of Technology, Brinellvägen 8, Stockholm, 11428, Sweden

Abstract

Relation Extraction (RE) models are crucial to many Natural Language Processing (NLP) applications, but often inherit and deepen biases in their training data. The underrepresentation of certain demographic groups can lead to performance disparities, particularly when considering intersectional fairness, where biases intersect across attributes such as gender and ancestry. To address this issue, we present **IntersectionRE**, a framework to improve the representation of underrepresented groups by generating synthetic training data. **IntersectionRE** identifies gaps in demographic coverage and optimizes data generation, ensuring the quality of augmented data through Large Language Models (LLMs), perplexity scoring, and factual consistency validation. Experimental results on the NYT-10, and Wiki-ZSL datasets demonstrate that our approach effectively reduces intersectional representation and model performance disparities, particularly for historically underrepresented groups.

Keywords

Representation Bias, Synthetic Data Generation, Relation Extraction,

1. Introduction

Relation extraction (RE), a key task in natural language processing (NLP), identifies and classifies semantic relationships between entities [1]. It supports downstream tasks like knowledge graph construction [2], question-answering [3], and information retrieval [4]. Despite strong performance on benchmarks [5, 6, 7], modern neural RE models often exhibit biases across demographic groups [8, 9, 10].

Biases in RE models often stem from their training datasets, directly influencing model predictions [11, 12]. Poorly curated datasets may underrepresent certain populations due to biased data collection, historical inequalities, or sampling imbalances, leading to discriminatory outcomes and unreliable predictions [13, 14, 15]. For instance, an RE model trained mostly on data featuring male individuals may struggle with relationships involving female subjects. This systematic underrepresentation, known as *representation bias*, limits the model's ability to generalize across diverse populations [16].

Representation bias becomes more complex when multiple demographic attributes intersect, known as *intersectional fairness* [17]. Biases can arise within individual groups (e.g., gender or race) and intensify at their intersections. For example, a model may perform well for females and Asians separately, but struggle with Asian females due to underrepresentation [18]. These gaps can lead to systematic RE failures, reinforcing societal biases. Addressing them is crucial for equitable model performance and reducing errors for marginalized groups.

While bias mitigation strategies exist throughout the Machine Learning (ML) pipeline, addressing bias during pre-processing offers a fundamental solution by improving data distribution [15]. Prior work on analyzing biases in RE, such as [9] revealed gender-based performance disparities, and [10] expanded analysis to intersectional biases through cross-dataset comparisons. However, they do not propose methods to systematically address intersectional representation gaps.

To address these challenges, we present **IntersectionRE**, a framework for identifying and mitigating intersectional representation gaps in RE datasets. We use pattern-based coverage analysis to quantify demographic representation and identify *Maximal Uncovered Patterns (MUPs)* to highlight

Identity-Aware AI workshop at 28th European Conference on Artificial Intelligence, October 25, 2025, Bologna, Italy
amlk@kth.se (A. Layegh); payberah@kth.se (A. H. Payberah); misha@kth.se (M. Matskin)
https://AmirLayegh.github.io/ (A. Layegh)
0000-0002-3264-974X (A. Layegh); 0000-0002-2748-8929 (A. H. Payberah); 0000-0002-4722-0823 (M. Matskin)


key coverage gaps. We then apply an Integer Linear Programming (ILP) component to determine the minimal number of synthetic examples needed for balance. Finally, we generate high-quality synthetic data using an LLM-based generator, preserving data characteristics and feature distributions. This approach allows us to balance demographic representation across dimensions while maintaining data integrity. Our experiments demonstrate that this approach reduces intersectional coverage gaps and also improves model fairness and overall performance across underrepresented subgroups. Notably, we observe substantial gains in F1 and reductions in disparity across both NYT-10 [19] and Wiki-ZSL [20] datasets, with minimal augmentation overhead.

This study makes four key contributions: (1) **IntersectionRE**, a framework that detects and mitigates intersectional representation gaps in RE tasks through pattern-based gap analysis and synthetic data generation; (2) An ILP-based strategy and LLM-based synthetic data generator to enhance demographic representation while preserving data integrity; (3) Empirical evidence on the NYT-10, and Wiki-ZSL datasets showing effective bias mitigation and improved model performance across demographic groups; and (4) A practical method for enriching RE datasets with demographic attributes (gender and ancestry), enabling fine-grained fairness analysis previously noted as infeasible for RE datasets [9].

2. Background

ML models trained on biased datasets can amplify societal inequalities through unfair predictions [21]. In RE models, biased data often results in missing relationships for underrepresented groups. This section examines RE biases, their impacts, and introduces ways to quantify representation and address dataset gaps.

2.1. Relation Extraction and Patterns

Relation Extraction (RE) identifies and classifies semantic relationships between entities in text. Given a sentence x with subject s and object o, the goal is to predict their relation label $y \in \mathcal{Y}$, where $\mathcal{Y}$ is a set of predefined relation types, such as `founder` and `employer`. For example, in $x = \{\texttt{Steve Jobs is the founder of Apple}\}$, with $s = \{\texttt{Steve Jobs}\}$ and $o = \{\texttt{Apple}\}$, an RE model identifies $y = \{\texttt{founder}\}$.

A *pattern* P represents a subgroup of records sharing specific attribute values [22]. Formally, P is a vector of size d (number of attributes), where each element $P[i]$ is either a specific value from attribute i's domain or an unspecified value denoted as X. For example, in a dataset with three binary attributes $\{x_1, x_2, x_3\}$, the pattern $P = X01$ includes records with $x_2 = 0$, $x_3 = 1$, and any value for x_1. A record t *matches* pattern P (denoted as $Match(t, P)$) if for all i where $P[i] \neq X$, $t[i] = P[i]$.

To measure representation bias, we use *coverage* to quantify subgroup representation in a dataset $\mathcal{D}$: $Cov(P) = |\{t \in \mathcal{D} \mid Match(t, P)\}|/|\mathcal{D}|$. For example, if $|\mathcal{D}| = 100$ and 21 records match pattern $P = X01$, then $Cov(P) = 0.21$. A pattern P is *uncovered* if $Cov(P) < \tau$, where τ is the minimum required coverage.

Coverage gaps occur when patterns in a dataset are uncovered, leading to potential biases and unfair predictions for these subgroups. Given a dataset $\mathcal{D}$ and a coverage threshold τ, the coverage gap for a pattern P is: $Gap(P) = (\tau - Cov(P)) \times |\mathcal{D}|$. This represents the minimum number of additional records needed to meet the threshold. For example, if $|\mathcal{D}| = 100$, $Cov(P) = 0.21$, and $\tau = 0.3$, the gap is $(0.3 - 0.21) \times 100 = 9$, meaning nine more records are needed for adequate representation of P.

Two patterns are related through a *parent-child* relationship based on their specified attributes. Pattern P_1 is a *parent* of P_2 ($P_1 \in parent(P_2)$) if it can be formed by replacing exactly one specified value in P_2 with X. Conversely, P_2 is a *child* of P_1 ($P_2 \in child(P_1)$). A pattern can have multiple parents and children. For example, for $P = 101$, its parents are $parent(P) = \{X01, 1X1, 10X\}$, each created by replacing one value with X.

In analyzing coverage gaps, we identify the most general uncovered patterns, called *Maximal Uncovered Patterns (MUPs)*. A pattern P is an MUP if: (1) it is uncovered ($Cov(P) < \tau$) and (2) all its parents

have adequate coverage ($\forall P' \in parent(P) : Cov(P') \geq \tau$). MUPs capture broad underrepresented subgroups without redundancy from more specific child patterns.

2.2. Fairness Definition

In this work, we focus on demographic attributes of gender $\mathcal{G}$ and ancestry $\mathcal{A}$. Our primary fairness objective is to improve the representation of underrepresented groups in $\mathcal{G} \times \mathcal{A}$. However, as RE models are ultimately judged by predictive behavior, we also assess fairness in model predictions to ensure consistency across demographic groups [9]. We evaluate fairness from two complementary perspectives: (1) representation in the data and (2) equitable model performance across demographic groups.

Representation Fairness Metrics. To assess demographic representation, we evaluate four normalized metrics in the range $[0, 1]$, where higher values indicate better balance: (1) *Balance Score*: normalized female-to-male ratio, $\min\big(\frac{Cov(\texttt{FemaleX})}{Cov(\texttt{MaleX})}, \frac{Cov(\texttt{MaleX})}{Cov(\texttt{FemaleX})}\big)$, (2) *Gender Gap*: indicates the absolute coverage difference between genders $\big|Cov(\texttt{FemaleX}) - Cov(\texttt{MaleX})\big|$, (3) *Ancestry Gap*: the complement of the standard deviation over coverage of each ancestry, $1 - \mathrm{std}\big(Cov(Xa) \mid a \in \mathcal{A}\big)$, and (4) *Intersectional Gap*: the complement of the standard deviation over all subgroup coverages, $1 - \mathrm{std}\big(Cov(ga) \mid g \in \mathcal{G}, a \in \mathcal{A}\big)$.

Performance Fairness Metrics. To evaluate consistency in model behavior, we adopt two metrics from [9]: the *Disparity Score* (DS) and the *Performance Parity Score* (PPS). DS quantifies performance variation across groups. We compute the average pairwise absolute difference in F1 scores across all demographic subgroups in $\mathcal{G} \times \mathcal{A}$. Let $\{g_1, g_2, \ldots, g_m\}$ denote the set of demographic groups, and F_{g_i} the F1 score for group g_i. Then:

$$DS = \frac{2}{m(m-1)} \sum_{1 \leq i < j \leq m} |F_{g_i} - F_{g_j}|,$$

where $m = |\mathcal{G} \times \mathcal{A}|$ representing the number of gender and ancestry combinations. A lower DS indicates more uniform model performance across groups. On the other hand, PPS combines accuracy and fairness into a single measure. It is defined as the difference between the macro-averaged F1 score across all subgroups and the DS. A higher PPS reflects models that are both accurate on average and consistent across demographic groups.

We adopt these fairness metrics because most existing notions, such as *Demographic Parity* and *Equalized Opportunity*, are originally defined for binary classification, where a single positive or negative prediction is made [23]. However, RE is inherently a multi-label task. As noted by Liu et al. [24], directly applying binary fairness metrics to multi-label settings is problematic due to label imbalance and co-occurrence patterns. This imbalance leads to unreliable or unstable fairness estimates, especially for infrequent relations. Our chosen metrics are tailored to operate at the group level across all demographic groups and account for representation bias in data and variation in model behavior.

2.3. Problem Definition

Given a RE dataset $\mathcal{D}$ with triples (subject s, relation r, object o) and demographic attributes (gender $\mathcal{G}$, ancestry $\mathcal{A}$), the goal is to mitigate representation biases from coverage gaps, especially intersectional ones, that affect model performance for underrepresented groups. We analyze intersectional representation using patterns P and identify MUPs to address gaps without redundant subpattern analysis.

Improving MUP coverage is crucial because MUPs represent the broadest underrepresented subgroups, and by increasing coverage for these general patterns, we automatically improve the coverage of all their more specific child patterns. For each MUP M, at least $Gap(M)$ additional records are needed to meet the coverage threshold τ. This process balances fairness across gender and ancestry while minimizing synthetic data to preserve data quality.

2.4. Synthetic Data Generation

Synthetic data generation is a key approach to addressing representation bias in ML datasets, where imbalanced demographics can lead to discriminatory model behavior [25]. It helps mitigate biased predictions by balancing demographic attributes while minimizing generated records [26]. However, balancing representation is challenging, especially with intersectional attributes [15], due to the difficulty of ensuring proportional representation across dimensions (e.g., gender, race) while addressing coverage gaps [27]. For example, if Black females are underrepresented compared to Black males or Asian females, data generation must fill this gap without disrupting other balances. Overcompensation can create new biases, making it hard to maintain fairness and data integrity.

To address these challenges, we optimize synthetic data generation to minimize records while meeting representation goals [28]. Traditional greedy algorithms often yield suboptimal results and struggle to maintain demographic balance [29, 30]. To overcome this, we use ILP [31] to define coverage and intersectional balance constraints, minimizing synthetic records while ensuring fair representation across all intersections [32]. This approach is especially effective for MUPs, providing globally optimal solutions that satisfy all gaps and constraints. The next section details our ILP formulation and implementation.

3. IntersectionRE

This section presents **IntersectionRE** for addressing intersectional representation bias in RE datasets, consisting of five components: (1) a data enrichment pipeline adding demographic attributes, (2) a pattern identification algorithm detecting underrepresented groups via MUP analysis, (3) an ILP-based planner minimizing required records while ensuring balance, (4) an entity collection module sourcing data from knowledge bases, and (5) an LLM-based generator producing synthetic factual samples. The following sections detail each component's role in mitigating bias.

3.1. Data Enrichment Pipeline

Analyzing intersectional fairness in RE datasets requires demographics (e.g., gender, ancestry), which are often missing [10]. For example, a record like (`Steve Jobs`, `Founder`, `Apple`) lacks demographic details. To address this, we developed a data enrichment pipeline using Wikidata to extract demographic attributes, consisting of two stages: First, for each record, we focus on relation labels, such as `founder`, `place_of_birth`, `profession`, and `nationality` that involve human entities, excluding records without them to ensure relevant demographic analysis. Then, for identified human entities, we retrieve attributes like gender and citizenship from Wikidata. We map each country to a broader ancestry group (e.g., `African`, `Asian`, `European/Western`, `Latino/Caribbean`, `Middle Eastern`) using a curated country-to-ancestry mapping, enabling meaningful aggregation to identify representation patterns and coverage gaps.

3.2. Pattern Identification

After enriching the dataset with demographic attributes, we identify underrepresented groups by analyzing coverage patterns based on gender and ancestry, focusing on MUPs that represent the broadest coverage gaps. Since identifying all MUPs is computationally intensive [15], we propose an algorithm inspired by DEEPDIVER [22]. DEEPDIVER uses a hybrid strategy combining downward traversal with immediate upward verification, checking all ancestor patterns to confirm MUPs, but our approach simplifies this by verifying only the immediate parent during downward traversal and deferring full maximality checks to a post-processing phase. We apply two pruning strategies: (1) *Coverage-Based Pruning*, where patterns meeting or exceeding the threshold have their children explored as potential MUPs, and (2) *Parent-Based Pruning*, where patterns below the threshold are pruned if their immediate parent is also uncovered. This reduces verification overhead, with post-processing ensuring only maximal patterns are retained.

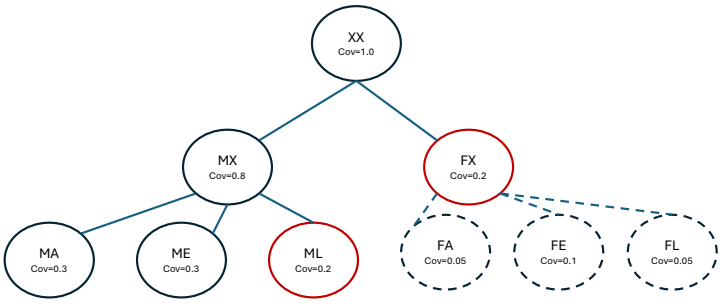

Figure 1: Tree structure illustrating DFS-based MUP discovery with pruning. Red nodes represent identified MUPs, while dashed nodes and edges indicate pruned patterns and paths.

As shown in Figure 1, consider a dataset with attributes Gender: {`Male` (M), `Female` (F)} and Ancestry: {`Asian` (A), `European` (E), `Latino` (L)}, and a threshold $\tau = 0.3$. Starting from the root XX (coverage 1.0), its children MX and FX are explored since XX exceeds the threshold. MX (coverage 0.8) is not a MUP, so its children MA, ME, and ML are explored. FX (coverage 0.2) is a potential MUP, and *Coverage-Based Pruning* skips its children (FA, FE, FL) since their parent is already uncovered. For ML (coverage 0.2), our algorithm checks only its immediate parent (MX), unlike DEEPDIVER, which checks both MX and XX. This streamlined approach flags ML as a potential MUP, with maximality verified during post-processing.

3.3. ILP-based Generation Plan

After identifying MUPs, we use an ILP-based planner to minimize synthetic records while ensuring coverage. Unlike greedy algorithms [30], which require iterative MUP recalculations and struggle to maintain demographic balance, our ILP ensures global optimality in a single step. It minimizes synthetic records under two constraints: (1) generating at least $Gap(M)$ records per MUP to meet coverage thresholds and (2) maintaining balanced gender ratios within each ancestry group. This prevents addressing gaps for one group (e.g., Asian females) from creating imbalances in others.

Let $\mathcal{G}$ = {`Female`, `Male`} and $\mathcal{A}$ = {`African`, `Asian`, `European/Western`, `Latino/Caribbean`, `Middle Eastern`}. To avoid new biases, we track for each ancestry $a \in \mathcal{A}$ the number of female records (F_a), total records (T_a), and female ratio ($R_a = F_a/T_a$). Simply adding new records can skew the balance. For example, for MUP $M_1 = \{\texttt{Female}, \texttt{Asian}\}$ with 0.01 coverage in a dataset of 1000 records and threshold $\tau = 0.05$ where the pattern $P = \{\texttt{X}, \texttt{Asian}\}$ has the coverage of 0.05, the gap $Gap(M_1) = 40$ requires 40 more records. Adding only female records would skew gender balance, so the ILP determines how many male Asian records to add to maintain fairness. To formulate this as an ILP, we define decision variables ($x_{g,a} \geq 0, \forall g \in \mathcal{G}, a \in \mathcal{A}$) indicating the number of synthetic records to generate for each gender and ancestry combination. These variables are only active for demographic combinations linked to MUPs, minimizing unnecessary data generation.

The next step in the ILP formulation is defining the objective function. Our primary goal is to minimize the total number of synthetic records required to meet demographic coverage and intersectional balance requirements: minimize $\sum_{g\in\mathcal{G}} \sum_{a\in\mathcal{A}} x_{g,a}$. This minimization ensures efficient data generation by creating only the necessary records to address coverage gaps identified by MUPs.

Then, we need to define the constraints. Our ILP formulation includes two constraints to ensure adequate coverage and intersectional balance: (1) *coverage constraints* and (2) *gender balance constraints*. To satisfy the *coverage constraints*, for each MUP ($M \in \mathcal{M}$), we ensure coverage gaps are filled: $\sum_{g\in\mathcal{G}_M} \sum_{a\in\mathcal{A}_M} x_{g,a} \geq Gap(M)$, where $\mathcal{G}_M$ and $\mathcal{A}_M$ represent the gender and ancestry sets specified in MUP M. For specified attributes (e.g., `Female`), the set contains only that value, and for unspecified attributes (X), it includes all possible values.

To satisfy the *gender balance constraints*, for each ancestry group $a \in \mathcal{A}_M$ we implement adaptive bounds: $\text{min\_}R_a \leq \frac{F_a + x_{\text{female},a}}{T_a + x_{\text{female},a} + x_{\text{male},a}} \leq \text{max\_}R_a$. We set $(\text{min\_}R_a, \text{max\_}R_a) = (\min(\alpha_1 R_a, 0.5), \max(\beta_1 R_a, 0.5))$ when $R_a < 0.33$ (severe), and $(\min(\alpha_2 R_a, 0.45), \max(\beta_2 R_a, 0.55))$ otherwise, where F_a and T_a are current female and total counts,

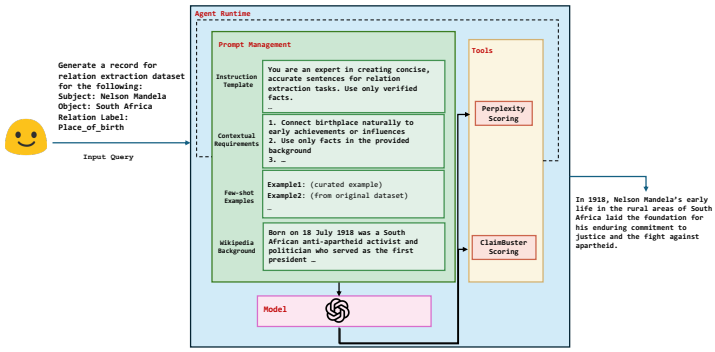

Figure 2: LLM-based record generation architecture, featuring prompt management providing relation-specific prompts for the model, with generated sentences validated by Perplexity Scoring and ClaimBuster tools.

and $x_{\text{female},a}, x_{\text{male},a}$ are decision variables. The threshold 0.33 reflects a 2:1 male-to-female ratio [10], and $\alpha_1, \alpha_2, \beta_1, \beta_2$ are control adjustment rates. We then obtain an ILP plan with variables $x_{g,a}$ ($g \in \mathcal{G}_M$, $a \in \mathcal{A}_M$) specifying the minimal records required per intersection to close coverage gaps.

3.4. Entity Collection from Knowledge Bases

To generate realistic records, we convert the ILP-based plan into synthetic data using Wikidata (structured) and Wikipedia (unstructured). For each gender–ancestry pair, we map ancestry to countries (Section 3.1), distribute entities accordingly, and apply per-citizenship limits for diversity. SPARQL queries retrieve Wikidata entities with matching gender and citizenship, along with biographical details (e.g., `founder`, `employer`, `place_lived`, `religion`, `profession`, `nationality`). To enrich context, we also fetch Wikipedia introductions via the WikiMedia API. This blend of structured and unstructured data ensures factual accuracy while meeting demographic requirements.

3.5. LLM-based Record Generation

Our framework's final stage uses a GPT-4-powered AI agent to generate synthetic records. Figure 2 illustrates the agent's architecture, which consists of prompt management components, a core generation model, and validation tools. The agent uses GPT-4 to map each relation-specific prompt p to synthetic records x. Each prompt p is tailored to capture the unique traits of a relation label $y \in \mathcal{Y}$, ensuring the generated sentence x accurately reflects the entity relationship. In this process, the agent addresses two validation challenges: (1) *distribution alignment*, ensuring x matches the linguistic and structural patterns of the original dataset $\mathcal{D}$, and (2) *factual consistency*, ensuring x accurately reflects input relationships. It uses a Perplexity Scoring Tool for language alignment and ClaimBuster [33] for factual consistency.

To guide GPT-4, we design relation-specific prompts p with the following components (Figure 2): (1) a *system prompt & instruction template* tailored to each relation y, defining constraints and guidelines, (2) *contextual requirements*, focusing on verified facts, achievements, or relevance (e.g., `lived_in` for locations, `employer` for roles), and (3) *few-shot examples*, combining curated and dynamic samples for diverse, in-context guidance.

The agent incorporates two key validation mechanisms to ensure the quality of generated records: *distribution alignment* and *factual consistency*. For distribution alignment, we measure perplexity per relation using a pre-trained model (e.g., GPT-2), where lower perplexity indicates better fluency and alignment. Specifically, we ensure that the perplexity of any generated sentence does not exceed the mean plus two standard deviations of perplexity values calculated for existing sentences of the same relation label. This method confirms that generated sentences maintain a consistent quality and style with the dataset's typical variability.

For factual consistency, the agent uses ClaimBuster with dynamic thresholding. Let $\phi(x)$ be the ClaimBuster scoring function assigning a factuality score in [0,1]. For each relation y, we set the threshold θ_r

at the 25th percentile of original dataset scores: $\theta_r = \text{percentile}_{25}(\{\phi(x) \mid x \in \mathcal{D}, \text{relation}(x) = y\})$, ensuring generated sentences are at least as factual as 75% of the original dataset. Sentences must meet $\phi(x) \geq \theta_r$; those below are refined with stricter prompts and re-evaluated. Only sentences passing after either stage are accepted, ensuring high factual consistency. The agent iteratively refines and regenerates sentences using adjusted prompts until they meet both distributional and factual standards or reach a retry limit, ensuring the generation of high-quality, realistic synthetic records that effectively address representation gaps.

4. Experimental Results and Analysis

This section focuses on evaluating our framework for addressing intersectional representation bias in RE datasets, specifically: (1) improving demographic representation, (2) impact on model performance across subgroups, and (3) efficient synthetic data generation via ILP.

4.1. Experimental Setup

Datasets. We conduct our experiments on two RE benchmarks: NYT-10 [19] and Wiki-ZSL [20]. NYT-10 is a widely used benchmark with 70,339 records and 52 relation labels, collected from the New York Times corpus via distant supervision using Freebase. To enable demographic analysis, we filtered for records containing at least one human entity, resulting in 30,818 records across 15 human-centric relation types. The most frequent relations include `place_of_birth` (21.3%), `nationality` (18.7%), `employer` (15.4%), and `place_lived` (14.2%). We enriched the dataset with demographic attributes like gender and ancestry (Section 3.1), revealing a skewed distribution (12.4% `females` vs. 87.6% `males` and ancestry disparities (`European/Western` 71.1%, `Middle Eastern` 11.7%, `Asian` 9.2%, `Latino/Caribbean` 4.9%, and `African` 3.1%). These imbalances highlight the need to address intersectional coverage gaps for equitable representation.

Wiki-ZSL is a RE dataset constructed from Wikipedia via distant supervision, containing 113 relation types in total [20]. For our experiments, we selected a fixed subset (SEED) comprising five distinct relations of {`employer, place of birth, religion, country of citizenship, residence`} and used it for training and testing. Similar to the NYT-10 setup, we filtered the dataset to retain only instances involving at least one human entity, resulting in 8,827 records. We randomly split this subset into 80% training and 20% testing data (7,061 training records). The same demographic enrichment procedure (Section 3.1) was applied to annotate each record with gender and ancestry attributes. The resulting training set revealed a skewed distribution (12.1% `females` vs. 87.9% `males`) and ancestry disparities (`European/Western` 84.8%, `Asian` 5.3%, `Latino/Caribbean` 5.1%, `Middle Eastern` 2.7%, `African` 2.1%).

Implementation. We queried demographic attributes from Wikidata using SPARQL, optimized via `SPARQLWrapper`, and pre-designed citizenship to ancestry mappings. The ILP was formulated using Gurobi, applying dynamic gender balance constraints based on R_a (stricter when $R_a < 0.33$: $\alpha_1 = 1.5, \beta_1 = 2$; relaxed otherwise: $\alpha_2 = 0.9, \beta_2 = 1.1$). Synthetic records were generated with GPT-4 (200-token limit, temperature 0.0) and validated via GPT-2 perplexity scoring [34] for fluency and ClaimBuster [33] for factual consistency. We fine-tuned the REBEL-large model [5] (a seq2seq BART-based RE model [35]) on the respective datasets, training for 3 epochs with AdamW (learning rate $2e-4$, batch size 4). The implementation and experimental artifacts for **IntersectionRE** are available at the project GitHub page[1]. To validate our method, we include a naive oversampling baseline, where instances from underrepresented gender–ancestry subgroups are duplicated until each group meets the target coverage threshold [36]. This allows us to compare our ILP-based approach against a simple yet commonly used strategy for addressing representation bias.

[1] https://github.com/AmirLayegh/IntersectionalRE

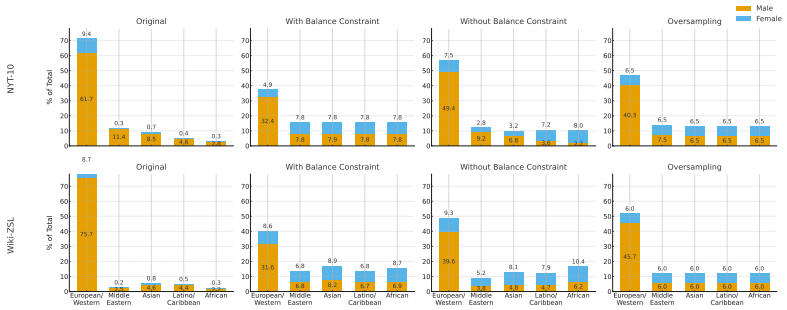

Figure 3: Intersectional representation across the original dataset, generated dataset with intersectional balance constraint, generated dataset without intersectional balance constraint, and oversampled dataset.

4.2. Representation Fairness

We analyzed intersectional representation in the original dataset and our proposed solutions. To evaluate our ILP-based constraints, we conducted experiments in four settings: (1) the original dataset, (2) our framework with intersectional balance constraints, (3) our framework with constraints off, focusing on the MUP coverage threshold, and (4) a naive oversampling baseline.

Baseline Coverage Gap in Original Dataset. To assess intersectional gaps, we used a coverage threshold of 0.15, representing the minimal expected coverage per group, given five ancestry groups and two genders (ideally 10% each if evenly distributed). This value balances real-world demographic imbalances with meaningful targets for underrepresented groups. Figure 3 (Left) shows demographic representation in the original datasets, with rectangles indicating the percentage of each gender-ancestry combination and color intensity reflecting coverage (darker = higher). In both original datasets, `European/Western males` dominate (61.7%, 75.7%), while female representation is minimal, peaking at 9.4% and 8.7%, and dropping to 0.3% for `African females`. Groups like `African males` (2.8%, 2.2%) and `Latino/Caribbean females` (0.4%, 0.5%) fall well below the threshold, highlighting systemic biases and the need for targeted augmentation.

Coverage Improvements with Augmentation. Figure 3 also illustrates the impact of augmentation strategies across both datasets. In the *With Balance Constraint* setting, representation becomes substantially more uniform, with most groups reaching or exceeding the 0.15 threshold. Female representation improves across ancestries, and coverage becomes more equitably distributed. The *Without Balance Constraint* and *Oversampling* conditions show uneven results. While some underrepresented groups see improvements, subgroups like `European/Western males` benefit disproportionately, increasing to 49.4% and 45.7%, whereas groups such as `Middle Eastern` and `Asian females` remain below the 0.15 threshold. This highlights the need for intersectional balance constraints to achieve fair coverage.

Gender Ratios Across Approaches. To evaluate gender equity across ancestry groups, we computed the female-to-male ratio for each group. An ideal ratio of 0.5 indicates equal representation, yet in the original dataset, ratios are skewed, with females comprising less than 0.15 in most groups, showing severe under-representation. Applying intersectional balance constraints achieves near-parity across ancestries, effectively addressing these imbalances. For example, `African` and `Middle Eastern` groups reach ratios close to 0.5 from near-zero. In contrast, the absence of such constraints leads to partial improvements but fails to ensure consistent gender equity. The *Oversampling* method also results in broadly similar gender ratios to the *With Balance Constraint* approach, reflecting improved parity across most subgroups.

Intersectional Representation Fairness. Figures 4a and 4b show the metrics defined in Section 2.2 for assessing intersectional representation fairness. The original datasets exhibit substantial disparities,

Figure 4: Intersectional fairness metrics for (a) Wiki-ZSL and (b) NYT-10 across *Original*, *Without Constraints*, and *With Constraints*; higher is better for all four metrics.

Table 1
Model Performance Comparison Across Demographic Groups on NYT and Wiki-ZSL Datasets. Constrained indicates our ILP-based optimization applied for data generation.

		NYT Original		NYT Constrained		NYT Oversampled		Wiki Original		Wiki Constrained	
Gender	Ancestry	F1	FPR	F1	FPR	F1	FPR	F1	FPR	F1	FPR
Female	African	0.000	1.000	1.000	0.000	1.000	0.000	1.000	0.000	1.000	0.000
	Asian	0.773	0.074	0.941	0.111	0.630	0.370	0.444	0.556	0.857	0.143
	European/Western	0.795	0.199	0.890	0.199	0.660	0.340	0.774	0.226	0.729	0.271
	Latino/Caribbean	0.889	0.200	1.000	0.000	0.800	0.200	0.800	0.200	1.000	0.000
	Middle Eastern	0.870	0.020	0.950	0.015	0.000	1.000	1.000	0.000	0.800	0.200
Male	African	0.756	0.179	0.923	0.143	0.607	0.393	0.600	0.400	0.737	0.263
	Asian	0.902	0.178	0.861	0.200	0.756	0.244	0.795	0.205	0.807	0.193
	European/Western	0.805	0.323	0.755	0.228	0.607	0.393	0.737	0.263	0.776	0.224
	Latino/Caribbean	0.911	0.116	0.911	0.163	0.837	0.163	0.769	0.231	0.757	0.243
	Middle Eastern	0.890	0.025	0.950	0.018	0.811	0.189	0.826	0.174	0.731	0.269
Disparity Score (DS)		**0.226**	–	**0.080**	–	**0.272**	–	**0.186**	–	**0.113**	–
Performance Parity Score (PPS)		**0.533**	–	**0.838**	–	**0.399**	–	**0.589**	–	**0.706**	–

with consistently low scores (≤ 0.15) across all metrics, while the *Without Intersectional Balance Constraints* approach shows moderate, uneven improvements (0.35–0.45). In contrast, the *With Intersectional Balance Constraints* method achieves the highest scores across both datasets, with ancestry gap reaching 0.75 and intersectional gap up to 0.68, effectively mitigating representation biases. The balance score and gender gap improve substantially from 0.124 (original) to 0.569 (constrained) in NYT-10, and from 0.105 (original) to 0.51 (constrained) in Wiki-ZSL, successfully reducing gender disparities while maintaining ancestry balance.

4.3. Model Fairness Evaluation

Table 1 shows significant variations in the REBEL model's performance when fine-tuned on the original NYT-10 dataset versus the demographically augmented version. The augmented model's F1 score improves from 0.782 to 0.845, reflecting better overall performance, though gains are uneven across demographic groups. Notably, underrepresented groups like `African males`, `African females`, `Middle Eastern females`, and `Latino/Caribbean females` see substantial improvements, indicating the augmentation effectively addresses representation gaps. While dominant groups such as `European/Western males` show a slight F1 decrease (-0.050), this is offset by improvements across underrepresented groups. The augmented model also reduces false positive rates (FPR) across most demographics while maintaining strong performance for `Middle Eastern` groups. However, these results are influenced by demographic imbalances in the test set, potentially affecting metric reliability for smaller groups. This highlights the need for evaluation methods that account for distributional biases in both training and testing phases.

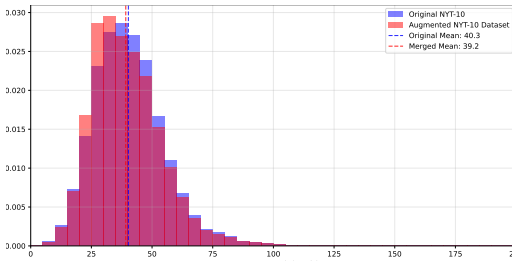

Figure 5: Sentence length distributions.

A similar pattern is observed on the Wiki-ZSL dataset, where overall F1 improves from 0.745 to 0.772 with constrained augmentation. Performance gains are most notable for groups that initially showed lower scores, such as `Asian females` and `Latino females`, both of whom show considerable improvement in F1 and FPR. Dominant groups like `European/Western males` maintain consistent performance with minimal variation. These findings suggest that the proposed augmentation strategy generalizes well across datasets, improving both effectiveness and fairness.

In contrast, the oversampling approach applied to the NYT dataset results in a drop in overall F1 to 0.642, with uneven changes across demographic groups. While some underrepresented groups see slight improvements, others (e.g., `Middle Eastern females`) experience degraded performance, including complete failure in recall. F1 scores for dominant groups like `European/Western males` also decrease significantly, indicating that duplicating data without structural balance introduces noise and redundancy.

The additional fairness metrics provide a clearer picture. The DS drops from 0.226 to 0.080 with our constrained augmentation on NYT, and from 0.186 to 0.113 on Wiki-ZSL, showing a measurable reduction in performance gaps. The PPS also increases across both datasets, reflecting more consistent outcomes across demographic groups. In contrast, oversampling results in the highest DS (0.272) and lowest PPS (0.399), confirming that it introduces new performance imbalances rather than resolving the existing ones.

4.4. Statistical Consistency

We compared sentence length distributions between the original NYT-10 and the augmented dataset to assess stylistic consistency. Figure 5 shows closely aligned density curves, supported by a low Jensen-Shannon Divergence (0.0411) and KS test statistic (0.0491, $p < 0.0001$). Sentence length statistics confirm this: the original dataset has a mean of 40.95, a median of 39.00, and a standard deviation of 78.92, while the augmented dataset shows a mean of 39.81, a median of 37.00, and a standard deviation of 75.55, indicating minimal deviation.

For quality assessment, the vocabulary size grew from $37,168$ to $42,862$, showing that the augmented dataset introduces new vocabulary while maintaining a reasonable growth rate. This suggests the generated text preserves the domain-specific language of the original dataset. The *Type-Token Ratio* (TTR), measuring lexical diversity as the ratio of unique words to total words, rose slightly from 0.0349 to 0.0378 ($+8.3\%$), maintaining diversity without excessive repetition. The *Hapax Percentage*, indicating the proportion of words appearing only once, increased from 24.71% to 27.60% ($+11.7\%$), reflecting more unique terms, likely from new entity names. These results demonstrate that our augmentation approach effectively enhances coverage and diversity while preserving linguistic and structural integrity.

4.5. Discussion

As shown by the results, our approach effectively reduces representation bias in the dataset, consequently enhancing intersectional representation fairness. Notably, this is achieved by improving demographic balance in the data and also by producing more equitable model predictions across demographic groups.

We acknowledge that our method introduces complexity through the use of an ILP-based framework, MUP analysis, and intersectional constraints. However, this complexity is justified by the nature of the problem. As highlighted by Asudeh et al. [22], achieving a globally optimal solution that minimizes the

number of synthetic records while satisfying strict multi-dimensional coverage constraints is inherently difficult. Simpler alternatives, such as naive oversampling strategies or group-level balancing, are inadequate for addressing fine-grained intersectional gaps and often lead to over-augmentation in some groups while still neglecting others [15]. These imbalances can negatively affect model fairness, as our experiments with naive oversampling demonstrate.

The problem of identifying MUPs has been shown to be NP-hard [9], making it infeasible to solve using standard polynomial-time algorithms. While heuristic methods may provide partial improvements, they do not offer control over which subgroups are affected or how many synthetic examples are generated. In contrast, our ILP formulation allows us to target specific coverage deficits, ensuring that synthetic records are only added where needed. This is especially important because synthetic data generation is computationally expensive. Generating unnecessary records not only increases the cost but can also distort the dataset and reintroduce bias. Our method avoids this by finding the minimal feasible augmentation that meets all fairness constraints. Although the optimization layer adds complexity, it ultimately reduces redundancy and helps produce a more balanced and efficient dataset. We believe this trade-off is warranted given the gains in both representation and model fairness.

While our ILP formulation is tailored to optimize coverage gaps based on gender and ancestry, the underlying principle is generalizable. Our focus on these two attributes was motivated by the strong demographic imbalances observed in RE datasets and the fact that they are the only attributes we could reliably extract from external sources such as Wikidata. In principle, additional attributes such as age or occupation could be integrated by extending the ILP constraints to support higher-dimensional demographic groups. However, doing so would introduce challenges in data extraction, attribute sparsity, and scalability. We view our current work as a foundational step toward addressing intersectional bias in RE, and plan to explore how our method can scale to more complex demographic structures in future research.

5. Conclusion

This work tackles intersectional fairness in relation extraction (RE) datasets, addressing representation bias that leads to disproportionate model errors for underrepresented groups. We propose **IntersectionRE** to identify and mitigate demographic coverage gaps, ensuring balanced representation across gender and ancestry while preserving linguistic and factual integrity. Empirical results show that our augmentation strategy improves demographic representation, reduces performance disparities, and enhances the REBEL model's F1 score, especially for underrepresented groups. Our findings demonstrate the effectiveness of structured augmentation in mitigating demographic bias. Future work should extend this framework to include more attributes (e.g., age, profession), diversify demographic sources beyond Wikidata, and move beyond binary gender classifications. Our approach offers a scalable, adaptable method for promoting demographic fairness in RE, supporting more equitable AI systems.

6. Related Work

Bias in RE has been widely studied, particularly in terms of gender disparities. [9] introduced WikiGenderBias, showing that RE models exhibit gender-based performance gaps, particularly in occupation and spouse-related relations. Beyond gender, [37] highlighted entity-level biases, where RE models overly depend on entity mentions rather than textual context, proposing counterfactual inference (CORE) to mitigate bias at inference time. While their approach aims at debiasing predictions, it does not address bias in the training data itself. Additionally, [38] pointed out systematic biases in distantly supervised datasets, arguing that traditional held-out evaluation methods misrepresent model fairness due to label noise. More recently, [10] conducted a cross-dataset bias analysis, revealing that RE datasets often underrepresent non-Western nationalities and female entities, leading to skewed model behavior.

While these studies primarily analyze and detect bias, our work takes a proactive approach by mitigating bias at the data level through coverage-driven augmentation. Unlike prior debiasing techniques that either mask entity bias or adjust model inference, our method identifies and fills demographic gaps in the dataset using ensuring a balanced, high-quality dataset for fairer RE models.

References

[1] R. Bunescu, R. Mooney, A shortest path dependency kernel for relation extraction, in: EMNLP, 2005, pp. 724–731.

[2] I. Muhammad, A. Kearney, et al., Open information extraction for knowledge graph construction, in: DEXA, 2020, pp. 103–113.

[3] D. Luo, J. Su, S. Yu, A bert-based approach with relation-aware attention for knowledge base question answering, in: 2020 IJCNN, IEEE, 2020.

[4] C. Khoo, S. H. Myaeng, Identifying semantic relations in text for information retrieval and information extraction, in: The semantics of relationships: An interdisciplinary perspective, 2002, pp. 161–180.

[5] P.-L. H. Cabot, R. Navigli, Rebel: Relation extraction by end-to-end language generation, in: Findings of the ACL: EMNLP, 2021.

[6] W. Tang, B. Xu, Unirel: Unified representation and interaction for joint relational triple extraction, in: Proceedings of the EMNLP 2022 conference, 2022, pp. 7087–7099.

[7] R. Orlando, P.-L. H. Cabot, E. Barba, R. Navigli, Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget, arXiv preprint arXiv:2408.00103 (2024).

[8] L. Li, X. Chen, H. Ye, Z. Bi, On robustness and bias analysis of bert-based relation extraction, in: Knowledge Graph and Semantic Computing, CCKS 2021, Guangzhou, China, Proceedings 6, Springer, 2021, pp. 43–59.

[9] A. Gaut, T. Sun, S. Tang, et al., Towards understanding gender bias in relation extraction, arXiv preprint arXiv:1911.03642 (2019).

[10] M. Stranisci, et al., Dissecting biases in relation extraction: A cross-dataset analysis on people's gender and origin, in: Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP), ACL, 2024. doi:10.18653/v1/2024.gebnlp-1.12.

[11] S. Barocas, A. D. Selbst, Big data's disparate impact, Calif. L. Rev. 104 (2016) 671.

[12] J. Stoyanovich, B. Howe, H. V. Jagadish, Responsible data management, Proceedings of the VLDB Endowment 13 (2020).

[13] I. Chen, F. D. Johansson, D. Sontag, Why is my classifier discriminatory?, Advances in neural information processing systems 31 (2018).

[14] D. Firmani, L. Tanca, R. Torlone, Ethical dimensions for data quality, Journal of Data and Information Quality (JDIQ) 12 (2019) 1–5.

[15] N. Shahbazi, Y. Lin, A. Asudeh, H. Jagadish, Representation bias in data: A survey on identification and resolution techniques, ACM Computing Surveys 55 (2023) 1–39.

[16] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 77–91.

[17] J. R. Foulds, et al., Bayesian modeling of intersectional fairness: The variance of bias, in: Proceedings of the SIAM International Conference on Data Mining, SIAM, 2020.

[18] Z. Jin, M. Xu, C. Sun, A. Asudeh, H. Jagadish, Mithracoverage: a system for investigating population bias for intersectional fairness, in: Proceedings of the ACM SIGMOD ICMD, 2020.

[19] S. Riedel, L. Yao, A. McCallum, Modeling relations and their mentions without labeled text, in: Machine Learning and Knowledge Discovery in Databases: ECML PKDD, Springer, 2010.

[20] C. Y. Chen, C.-T. Li, Zs-bert: Towards zero-shot relation extraction with attribute representation learning, in: NAACL, 2021, pp. 3470–3479.

[21] H. Suresh, J. Guttag, A framework for understanding sources of harm throughout the machine learning life cycle, in: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, 2021, pp. 1–9.

[22] A. Asudeh, Z. Jin, H. Jagadish, Assessing and remedying coverage for a given dataset, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 554–565.

[23] P. Czarnowska, Y. Vyas, K. Shah, Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics, Transactions of the ACL (2021).

[24] T. Liu, H. Wang, Y. Wang, X. Wang, L. Su, J. Gao, Simfair: A unified framework for fairness-aware

multi-label classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, 2023, pp. 14338–14346.
[25] B. Draghi, Z. Wang, P. Myles, A. Tucker, Bayesboost: Identifying and handling bias using synthetic data generators, in: Third International Workshop on Learning with Imbalanced Domains: Theory and Applications, PMLR, 2021, pp. 49–62.
[26] K. Wang, J. Zhu, M. Ren, Z. Liu, S. Li, Z. Zhang, C. Zhang, X. Wu, Q. Zhan, Q. Liu, et al., A survey on data synthesis and augmentation for large language models, arXiv preprint arXiv:2410.12896 (2024).
[27] A. Fournier-Montgieux, M. Soumm, A. Popescu, B. Luvison, H. L. Borgne, Fairer analysis and demographically balanced face generation for fairer face verification, arXiv preprint arXiv:2412.03349 (2024).
[28] N. Micheletti, R. Marchesi, N. I.-H. Kuo, S. Barbieri, G. Jurman, V. Osmani, Generative ai mitigates representation bias and improves model fairness through synthetic health data, medRxiv (2023) 2023–09.
[29] N. Shahbazi, M. Erfanian, A. Asudeh, Coverage-based data-centric approaches for responsible and trustworthy ai., IEEE Data Eng. Bull. 47 (2024) 3–17.
[30] M. Erfanian, H. V. Jagadish, A. Asudeh, Chameleon: Foundation models for fairness-aware multi-modal data augmentation to enhance coverage of minorities, Proc. VLDB Endow. 17 (2024) 3470–3483. URL: https://doi.org/10.14778/3681954.3682014. doi:10.14778/3681954.3682014.
[31] Y. Nandwani, R. Ranjan, P. Singla, et al., A solver-free framework for scalable learning in neural ilp architectures, Advances in Neural Information Processing Systems 35 (2022) 7972–7986.
[32] C. Dwork, K. Greenewald, M. Raghavan, Synthetic census data generation via multidimensional multiset sum, arXiv preprint arXiv:2404.10095 (2024).
[33] D. Jimenez, C. Li, An empirical study on identifying sentences with salient factual statements, in: 2018 International Joint Conference on Neural Networks (IJCNN), IEEE, 2018, pp. 1–8.
[34] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.
[35] M. Lewis, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).
[36] V. Iosifidis, E. Ntoutsi, Dealing with bias via data augmentation in supervised learning scenarios, Jo Bates Paul D. Clough Robert Jäschke 24 (2018).
[37] Y. Wang, et al., Should we rely on entity mentions for relation extraction? debiasing relation extraction with counterfactual analysis, in: Proceedings of the 2022 Conference of the North American Chapter of the ACL, 2022.
[38] P. Li, X. Zhang, W. Jia, W. Zhao, Active testing: An unbiased evaluation method for distantly supervised relation extraction, in: Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 204–211.