TDDE09 – PROJECT

Algorithmic Bias in Automated Recruitment

Identifying and Mitigating Gender Bias in AI-Driven Hiring Systems Using Transformer-Based NLP Pipelines

NLP · Transformer Models · Fairness in AI · DistilBERT · DistilRoBERTa · BoW

Presentation Outline

01 **Background**
Why algorithmic bias in hiring matters

02 **Dataset**
What data was used

03 **Pipeline Architecture**
Two-stage transformer workflow

04 **PII anonymization**
DistilBERT NER

05 **Methods**
All experiments

06 **Results**
Key findings

07 **Discussion**
Implications and limitations

08 **Conclusion**
Summary and future work

Background

- LLMs inherit societal biases from human-generated training data
- AI can encode and amplify bias, not just replicate it (Sheng et al., 2019)
- 15.46% of resumes in experimental datasets showed gender bias effects (Gagandeep et al., 2024)

Sheng et al. (2019). The Woman Worked as a Babysitter. EMNLP.
Gagandeep et al. (2024). Evaluating and mitigating gender bias in machine learning based resume filtering. Multimedia Tools and Applications.

Dataset

DATASET

Kaggle: yaswanthkumary/ai-recruitment-pipeline-dataset

`Columns: Resume, Job_Description, decision`

Dataset size: 10,174 resumes

Labels: ‘select’ (hired), ‘reject’ (not hired)

`Binary classification target`

Limitations

Ground Truth Bias

Decision is also dependant on interviews

Lacks real-world representation

Pipeline Architecture

ANONYMIZED PIPELINE

INPUT
Raw Resumes

1
Baseline Classifier (DistilRoBERTa)

2
PII Anonymize (DistilBERT NER)

3
Re-Classify (DistilRoBERTa)

OUTPUT
Bias-Evaluated Predictions

PARALLEL ANALYSIS (Bag of Words)

INPUT
Original / Anonymized Resumes

BoW
Logistic Regression on word frequency

OUTPUT
Word weights by gender

PII Anonymization

Names

DistilBERT NER + Name Dictionary
[PERSON]

Identify gender

Probabilistic name-based lookup
M/F/unisex/unknown

Emails

Contextual Regex detection
[EMAIL]

Gender words

Gendered noun identification
[GENDER]

Usernames

Name-variant handle scrubbing
[USERNAME]

Gender job titles

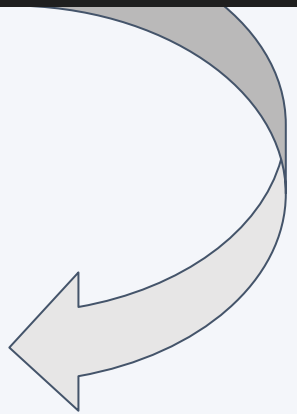
Neutral mapping (Waiter -> Server)
Neutral Equivalent

Example:

```
John Walker
UX Designer

Contact Information:

* Email: [john.walker@email.com](mailto:john.walker@email.com)
* Phone: 555-555-5555
* LinkedIn: linkedin.com/in/johnwalkerux
* Portfolio: johnwalker.design
```



PII Removal

```
[PERSON]
UX Designer

Contact Information:

* Email: [[EMAIL]@email.com](mailto:[EMAIL]@email.com)
* [PERSON]: 555-555-5555
* LinkedIn: [USERNAME]
* Portfolio: [USERNAME]
```

Method - Resume classification

ab-ai/pii_model_based_on_distilbert

PII Anonymization (NER)

- DistilBERT fine-tuned for NER; detects FIRSTNAME, LASTNAME, PHONENUMBER, DATE
- Chunked inference (400 tokens, 50 overlap) handles long resume texts
- Replaces names with [PERSON], emails with [EMAIL], usernames with [USERNAME]
- Gender inferred from first name via probability lookup in common names list
- Gendered job titles replaced with neutral equivalents (e.g. 'stewardess' → 'flight attendant')

DistilRoBERTa-base

Hiring Re-Classification

- RoBERTa architecture distilled for efficiency and performance
- Fine-tuned for binary hiring prediction: hired / not hired
- Optimized for F1-score to balance precision and recall
- Applied to anonymized text output from the NER pipeline
- Outputs compared against Stage 1 baseline to quantify gender bias

Method - Training Configuration

HYPERPARAMETERS

Base model	`distilroberta-base`
Epochs	`5 (early stopping, patience=2)`
Train batch size	`16 / Eval batch size: 32`
Learning rate	`3e-5`
Weight decay	`0.01`
Warmup steps	`200`
Max grad norm	`1.0`
Precision	`fp16 (mixed precision)`
Best model metric	`F1-score (higher = better)`
Eval/save strategy	`Per epoch (save_total_limit=1)`

INPUT FORMAT & TOKENIZATION

Input template:

```
"JOB DESCRIPTION: <jd>" [SEP]
"RESUME: <resume>"
```

`max_length`	384 tokens
`stride`	64 tokens (sliding window for long texts)
`overflow`	return_overflowing_tokens=True
`padding`	Dynamic via DataCollatorWithPadding
`chunks`	Each chunk inherits the sample's label

Method - BoW Classifier

- Count frequencies
- Disguards grammar, uninteresting words
- Logistic-Regression

```
=== BoW results: BEFORE anonymization | WOMEN ===
Words that increase hiring chances:
  recruitment                    0.293
  mobile                         0.295
  strategies                     0.300
  inventory                      0.300
  power                          0.307
  robotics                       0.316
  hardware software              0.318
  maintain                       0.320
  airflow                        0.327
  manager                        0.327
  section                        0.331
  hernandez                      0.332
  hardware                       0.335
  danielle                       0.393
  provide                        0.395
  nlp                            0.401
  app                            0.434
  jessica                        0.448
  security                       0.677
  karen                          0.829
```

Result - BOW Classifier

Bag-of-words classifier

The words associated with applications - gender comparison

```
=== BoW results: BEFORE anonymization | MEN ===
Words that increase hiring chances:
  hyperparameter tuning          0.290
  hyperparameter                 0.290
  michael                        0.295
  andrew                         0.301
  kevin                          0.302
  robinson                       0.302
  automation                     0.304
  architect                      0.305
  data solutions                 0.313
  excel                          0.315
  jeremy                         0.319
  technical                      0.331
  tuning                         0.334
  dale                           0.351
  marketing                      0.359
  data analyst                   0.377
  ux                             0.436
  gregory                        0.453
  christopher                    0.598
  data                           0.897
```

```
=== BoW results: BEFORE anonymization | WOMEN ===
Words that increase hiring chances:
  recruitment                    0.293
  mobile                         0.295
  strategies                     0.300
  inventory                      0.300
  power                          0.307
  robotics                       0.316
  hardware software              0.318
  maintain                       0.320
  airflow                        0.327
  manager                        0.327
  section                        0.331
  hernandez                      0.332
  hardware                       0.335
  danielle                       0.393
  provide                        0.395
  nlp                            0.401
  app                            0.434
  jessica                        0.448
  security                       0.677
  karen                          0.829
```

Female names mentioned: 3

Male names mentioned: 8

Result - BOW Classifier

Bag-of-words classifier

The words associated with applications - gender comparison

```
=== BoW results: BEFORE anonymization | WOMEN ===
Words that increase hiring chances:
  recruitment                    0.293
  mobile                         0.295
  strategies                     0.300
  inventory                      0.300
  power                          0.307
  robotics                       0.316
  hardware software              0.318
  maintain                       0.320
  airflow                        0.327
  manager                        0.327
  section                        0.331
  hernandez                      0.332
  hardware                       0.335
  danielle                       0.393
  provide                        0.395
  nlp                            0.401
  app                            0.434
  jessica                        0.448
  security                       0.677
  karen                          0.829
```

```
=== BoW results: AFTER anonymization | WOMEN ===
Words that increase hiring chances:
  data science                   0.290
  impact                         0.295
  mobile                         0.298
  experiences                    0.300
  strategies                     0.300
  bi                             0.302
  power bi                       0.302
  inventory                      0.312
  robotics                       0.317
  power                          0.321
  manager                        0.322
  airflow                        0.325
  maintain                       0.326
  hardware software              0.329
  section                        0.335
  hardware                       0.348
  provide                        0.406
  nlp                            0.415
  app                            0.459
  security                       0.683
```

Results - Training Performance

Biased results - without removing PII

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.6238	0.6244	0.5632	0.7582	0.1638	0.2629
2	0.6093	0.6271	0.5684	0.5503	0.2365	0.3494
3 ★	**0.5992**	**0.6700**	**0.5671**	**0.5321**	**0.9611**	**0.6581**
4	0.5821	0.6448	0.5711	0.5493	0.8099	0.6498
5	0.5744	0.6534	0.5788	0.5411	0.7885	0.6411

★ *Best model (highest F1, selected for inference)*

Unbiased results - After removing PII

Epoch	Train Loss	Val Loss	Accuracy	Precision	Recall	F1
1	0.6263	0.6157	0.5731	0.8009	0.1551	0.2599
2	0.6995	0.6468	0.5674	0.5327	0.8516	0.6554
3 ★	**0.6291**	**0.6439**	**0.5630**	**0.5263**	**0.9546**	**0.6786**
4	0.6062	0.6192	0.5689	0.5317	0.9038	0.6695
5	0.5757	0.6458	0.5818	0.5496	0.4763	0.6330

★ *Best model (highest F1, selected for inference)*

0.6581
Best F1

0.5788
Accuracy

0.7582
Precision

0.9611
Recall

Discussion – Key findings

- Resumes that include specific words tend to advantage men and women differently unrelated to job application
- Job applications rely heavily on the job interview – not just the application
- The marginal difference in employment rates between women (49.1%) and men (50.2%) suggests a lack of clear bias against women in the final hiring decision

Discussion – Limitations

- Training on the full dataset (10,171 samples) strained Google Colab's compute and memory limits, causing difficulties saving model checkpoints mid-training

- Current datasets used includes transcripts from interviews which plays a role in acceptance rate of the applicant –> bias

Discussion – Limits of project

- **Coded Language:** Gendered word choices—such as "competitive" (masculine-coded) vs. "collaborative" (feminine-coded)—can trigger unconscious bias in recruiters.

- **Context Loss:** Hiding gender-specific achievements (e.g., a "[GENDER] in STEM" scholarship) will tip of the recruiter of the actual gender.

- **Interview Threshold:** Anonymization only works during the initial screening; biases often resurface once the candidate meets the recruiter in person.

Conclusion

- LLMs inherit societal biases from human-generated training data
- Further investigation: different countries, job markets and education
- Investigate other biases such as age, sexual identity, ethnicity, religion and disability

Questions?

Thank you for listening!