A Context-Aware System for Bias Identification in Job Advertisements using Natural Language Processing (NLP)

Vogt Luise
Guan Zhize
Nasev Veselin
Guo Yonghui

Maastricht University

Agenda

- Introduction
  - Context and Motivation
  - Problem Statement and Research Questions
- Related Work
- Methodology
  - The Data&Data Prepcocessing
  - Data Annotation
    - Generic He/Generic She
    - Behavioral Stereotypes
    - Societal Stereotypes
    - Explicit Making of Sex
- Result & Analysis
  - Identification System
  - LIME Explaniability
- Future Work
  - Suggestion System
  - Other Bias Types
  - Modeling

Introduction

Context and Motivation

Bias Type:
Behavioral Stereotypes

Bias Type:
Generic He/Generic She

We are looking for a young and **driven** candidate who can bring innovation to our organization, and is a true team player for the rest within the organization. **He** needs to master relevant skills and techniques and be passionate about **his** job. We are still only 1% done at Facebook – this team is inventing every day and it takes tenacity, **bravery** and the ability to see the big opportunities to thrive.

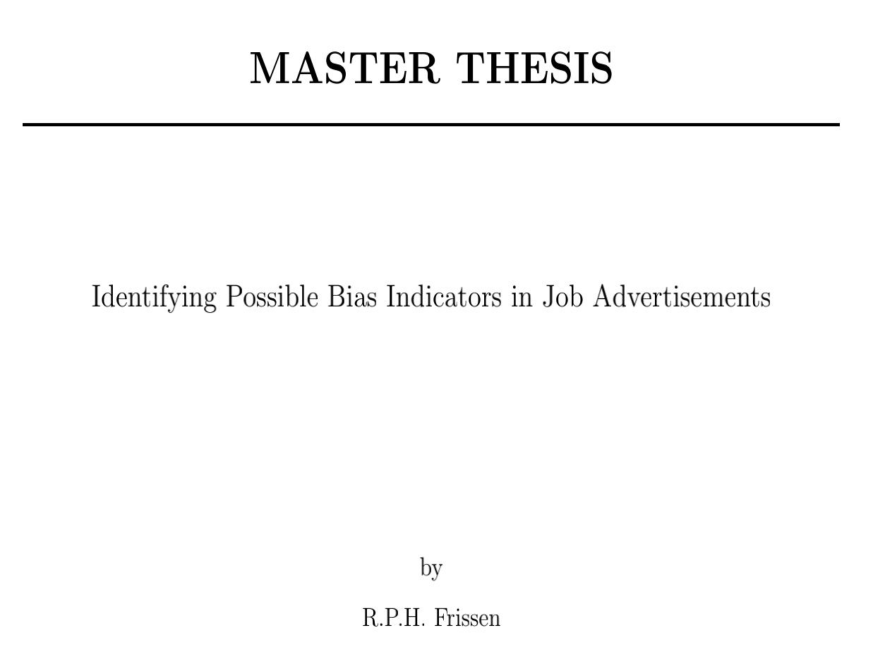
"Independent"
is a neutral word that you can use instead of "bavery"

Problem Statement and Research Questions

1. What kinds of biased language are commonly identified in job advertisements?
2. What words are related to the most common biases?
3. How can a context-aware natural language processing tool be used to classify different types of biases in job applications?
4. What are key predictors that explain the prediction for a particular class of bias?

Related Work

Related Work


MASTER THESIS

Identifying Possible Bias Indicators in Job Advertisements

by

R.P.H. Frissen

Figure 1: Thesis cover of Identifying Possible Indicators in Job Advertisments ([1])

Contributions:

- Five types of bias covering gender, rasical and other aspects
- Compared performance of different supervised machine learning classiers

Points to improve:

- The taxonomy of bias types
- Quality of annotations
- Context-aware model

Related Work

The Taxonomy of Gender Bias

- Five parent bias types
  - Generic Pronouns
  - Stereotyping Bias
  - Sexism
  - Exclusionary Terms
  - Semantic Bias

- Based on normal English text but not job advertisements

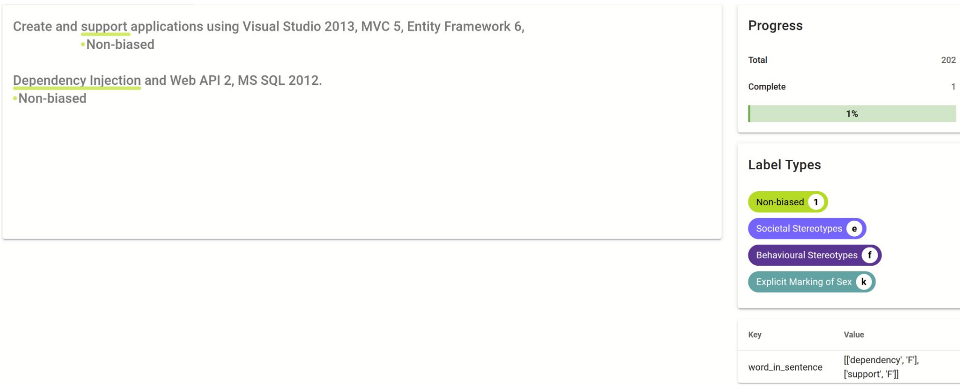
Bias Type	Bias Subtype	Example	Methodology
Generic Pronouns	Generic He	A programmer must carry his laptop with him to work.	Supervised Learning
	Generic She	A nurse should ensure that she gets adequate rest.	Supervised Learning
Stereotyping Bias	Societal Stereotypes	Senators need their wives to support them throughout their campaign.	Supervised Learning
	Behavioural Stereotypes	The event was kid-friendly for all the mothers working in the company.	Supervised Learning
Sexism	Hostile Sexism	Women are incompetent at work.	Supervised Learning
	Benevolent Sexism	They're probably surprised at how smart you are, for a girl.	Supervised Learning
Exclusionary Terms	Explicit Marking of Sex	Chairman, Businessman, Manpower, Cameraman	Lexicon-Based
	Gendered Neologisms	Man-bread, Man-sip, Man-tini	Lexicon-Based
Semantic Bias	Metaphors	"Cookie": lovely woman.	Supervised Learning
	Old Sayings	A woman's tongue three inches long can kill a man six feet high.	Supervised Learning

Figure 2: Overview of the taxonomy and link to detection methodology ([2])

Related Work

Data Annotation and Augmentation

- Doccano
  - Open source text annotation tool
  - Easy to use

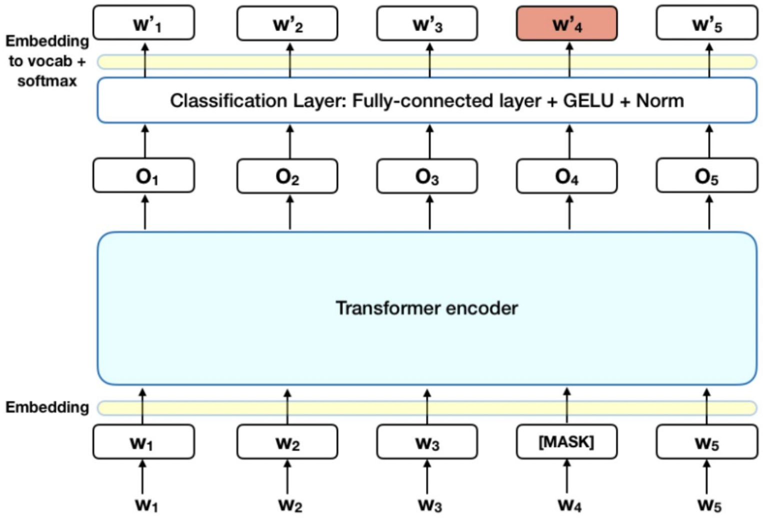


- FlashText
  - Python moudle
  - Replace keywords in sentences
  - Biased word list
- Masked language model with BERT
  - Replace random words in sentences
  - Not changing the context too much

Related Work

BERT model

- Bidirectional
- Pre-trained on two tasks
- Masked language modeling
  - Hide one word
  - Predict which word has been hidden based on context
- Next sentence prediction
  - Predict whether two given sentences have a logical, sequential connection



Methodology

Methodology - Data and Preprocessing

- Datasets
  - EMSCAD[3]
    - Employment Scam Aegean Dataset
    - Public available
    - 17,880 real-life job advertisements and 866 fake advertisements
  - Adzuna[4]
    - Dataset from Kaggle
    - 242,138 job adertisements

- Preprocessing
  - Remove white spaces and new rows
  - Remove HTML characters
  - Remove email addresses and phone numbers
  - Split job descriptions into sentences
  - Tokenize each sentence
  - Remove too long and too short sentences
  - Filter sentences that contain at least one biased word

Methodology - Generic He/She

- Found a dataset that addresses this type of gender bias
  - 1585 sentences that address (he, she, his, hers, himself, herself)

Example: Can you identify a programmer based on **his** code?

- Annotated 300 sentences from the job description dataset
  - Most of this sentences are not bias
  - Machine learning model can learn the difference

Example: John is our new programmer, **his** diligence and hard work….

Methodology - Behavioral Stereotypes

- Behavioral stereotypes attributes the behaviour of someone to its gender

Example:

Candidates must be **compassionate** and **understanding** to difficult situations while keeping focus on task at hand in a fast-paced collaborative environment.

Two methods to generate dataset:

- Annotated 1000 sentences from the EMSCAD dataset
- Repalce biased words on the list
- BERT + mask

- Manually labelled 500 sentences form Kaggle dataset
- BERT + mask

Methodology - Societal Stereotypes

- Not very common in job descriptions
  - Just a few examples from Jad Doughman's paper

Examples:

- Senators need their wives to support them throughout their campaign.
- The event was kid-friendly for all the mothers working in the company.

- Two methods to generate more examples
  - Combine different occupations with men or women together
  - Use techniques like fill-mask and GPT-2 to replace some parts of a sentence,

Examples:

- Professors are men.
- Doctors are women.
- Senators need their wives to support them and the country — and not try to take away their rights.
- The program was kid-friendly for all the mothers involved in the company.

Methodology - Explicit Marking of Sex

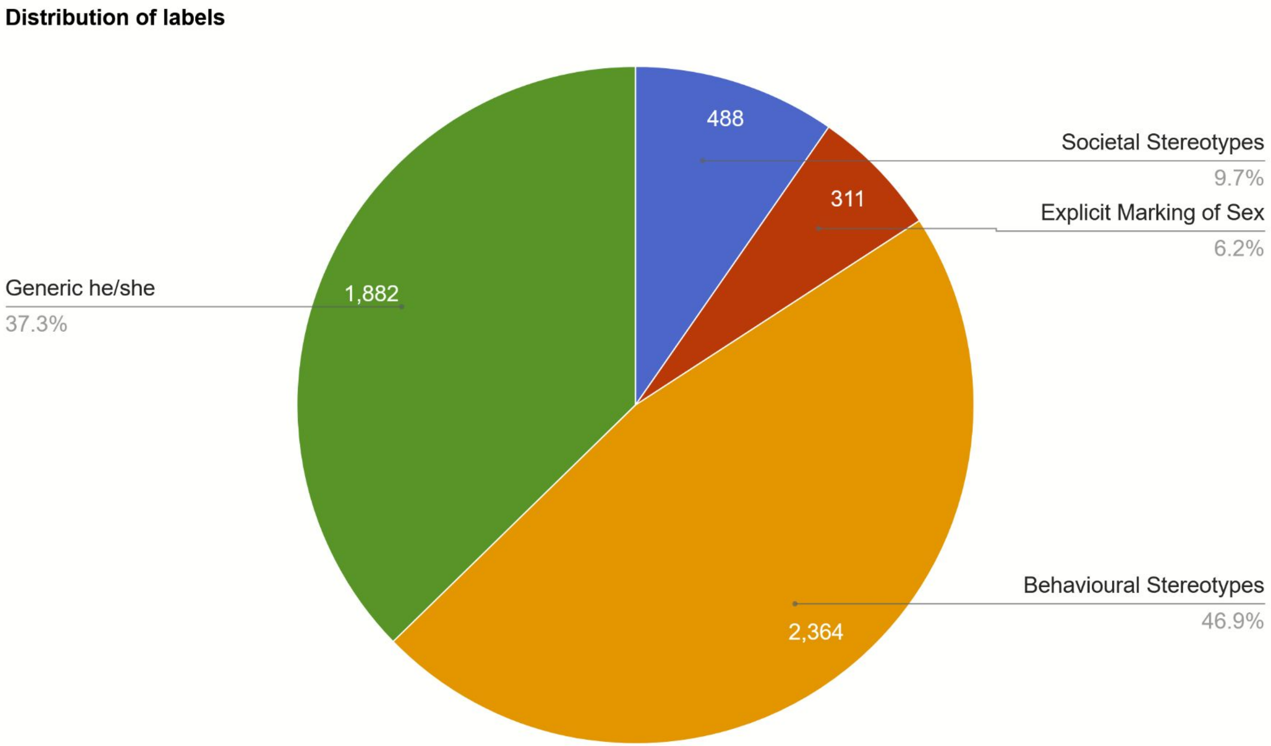
- mentions specific gender
- excludes other genders

- base lexicon
- word2vec
- find similar words
- manually check
- filter sentences

```
word
lady_fashion
gentleman
foreman
handyman
draughtsman
tradesman
manpower
chairman
landlord
manmanagement
acuman
man_management
```

Result & Analysis

Results and Analysis - Dataset



Maastri

Results and Analysis - Modeling

"It is necessary a dedicated and ambitious food development director to join our team."

⇩

Vector of length 768 for each token

⇩

Neural network with single input layer of 768 neurons

⇩

Output layer of length number of prediction classes

⇩

Softmax function on the output layer

Results and Analysis - Validation

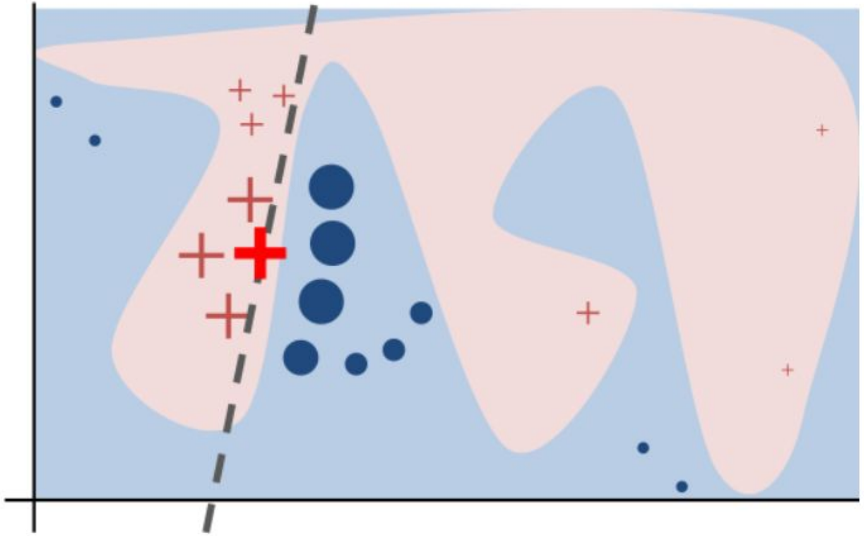
Precision	Recall	F1
0.98983	0.98983	0.98983

	Type	O	Behavioural Stereotypes	Explicit Marking of Sex	Societal Stereotypes	Generic he	Generic she
0	Precision	0.995627	0.861968	0.958333	1.000000	0.951456	0.970588
1	Recall	0.993589	0.931746	0.910891	0.991150	1.000000	0.990000
2	F1	0.994607	0.895500	0.934010	0.995556	0.975124	0.980198
3	Count	18561	630	101	113	196	200

20 % of dataset left for Validation

Results and Analysis — Explainability. LIME

- local interpretation
- local, more interpretable model
- approximates underlying model in sample neighborhood
- trained on corrupted copies of original sample
- detects important features in the data



Results and Analysis - Explainability. LIME

y=Generic She (probability **0.004**, score **-5.483**) top features

Contribution?	Feature
+7.914	<BIAS>
-13.397	Highlighted in text (sum)

The sheriff trusts the assistant because he was always honest.

y=O (probability **0.995**, score **5.484**) top features

Contribution?	Feature
+13.456	Highlighted in text (sum)
-7.971	<BIAS>

The sheriff trusts the assistant because he was always honest.

y=i-Societal Stereotypes (probability **0.001**, score **-7.434**) top features

Contribution?	Feature
-1.407	Highlighted in text (sum)
-6.028	<BIAS>

The sheriff trusts the assistant because he was always honest.

y=Generic She top features		**y=O** top features		**y=i-Societal Stereotypes** top features	
Weight?	Feature	Weight?	Feature	Weight?	Feature
+7.914	<BIAS>	+7.804	[6] he	-0.037	[4] assistant
+0.441	[1] sheriff	+1.960	[5] because	-0.097	[2] trusts
+0.123	[4] assistant	+1.907	[7] was	-0.101	[3] the
-0.292	[9] honest	+0.825	[8] always	-0.103	[9] honest
-0.325	[0] The	+0.469	[3] the	-0.119	[1] sheriff
-0.410	[2] trusts	+0.412	[2] trusts	-0.122	[0] The
-0.472	[3] the	+0.332	[0] The	-0.129	[8] always
-0.831	[8] always	+0.288	[9] honest	-0.349	[7] was
-1.895	[7] was	-0.114	[4] assistant	-0.350	[5] because
-1.946	[5] because	-0.428	[1] sheriff	-6.028	<BIAS>
-7.791	[6] he	-7.971	<BIAS>		

y=Behavioural Stereotypes (probability **0.991**, score **4.666**) top features

Contribution?	Feature
+9.540	Highlighted in text (sum)
-4.873	<BIAS>

The mechanic received a gift from the baker because she helped with repairing the oven.

y=Behavioural Stereotypes top features

Weight?	Feature
+7.632	[9] she
+0.387	[10] helped
+0.387	[8] because
+0.252	[11] with
+0.204	[0] The
+0.189	[14] oven
+0.179	[1] mechanic
+0.157	[12] repairing
+0.086	[6] the
+0.067	[13] the
-4.873	<BIAS>

Results and Analysis - Explainability. LIME

y=Generic He (probability **0.949**, score **2.918**) top features

Contribution?	Feature
+10.489	Highlighted in text (sum)
-7.570	<BIAS>

use those skills to provide services that are more supportive and comfortable for program participants.

y=Generic He top features

Weight?	Feature
+5.225	[9] supportive
+0.961	[8] more
+0.929	[10] and
+0.712	[12] for
+0.687	[14] participants
+0.551	[13] program
+0.495	[1] those
+0.417	[3] to
+0.349	[2] skills
+0.282	[4] provide
+0.215	[0] use
+0.161	[5] services
+0.099	[11] comfortable
-0.000	[6] that
-0.594	[7] are
-7.570	<BIAS>

y=Generic She (probability **0.000**, score **-9.149**) top features

Contribution?	Feature
+5.635	<BIAS>
-14.785	Highlighted in text (sum)

Supply chain coordinators are women.

y=Generic He (probability **0.001**, score **-6.564**) top features

Contribution?	Feature
-0.446	Highlighted in text (sum)
-6.118	<BIAS>

Supply chain coordinators are women.

y=i-Stereotyping Bias (probability **0.998**, score **9.121**) top features

Contribution?	Feature
+14.783	Highlighted in text (sum)
-5.663	<BIAS>

Supply chain coordinators are women.

y=Generic She top features

Weight?	Feature
+5.635	<BIAS>
-1.653	[2] coordinators
-1.815	[4] women
-2.150	[1] chain
-2.401	[0] Supply
-6.767	[3] are

y=Generic He top features

Weight?	Feature
-0.011	[1] chain
-0.096	[4] women
-0.339	[2] coordinators
-6.118	<BIAS>

y=i-Stereotyping Bias top features

Weight?	Feature
+6.725	[3] are
+2.384	[0] Supply
+2.131	[1] chain
+1.845	[4] women
+1.698	[2] coordinators
-5.663	<BIAS>

Results and Analysis - Explainability. LIME

y=i-Societal Stereotypes (probability **0.989**, score **4.514**) top features

Contribution?	Feature
+9.708	Highlighted in text (sum)
-5.194	<BIAS>

Forward landlord updates regarding events/activities: repairs, holiday/weekend/late night hours, security, parking, shuttle services etc.

y=i-Societal Stereotypes top features

Weight?	Feature
+7.681	[1] landlord
+0.623	[7] holiday
+0.620	[0] Forward
+0.418	[6] repairs
+0.366	[5] activities
+0.326	[14] shuttle
+0.152	[15] services
+0.118	[9] late
+0.098	[10] night
+0.058	[4] events
+0.056	[3] regarding
+0.024	[12] security
-0.082	[8] weekend
-0.732	[11] hours
-5.194	<BIAS>

Future Work

Future work - Suggestion System

1.The **biased words and phrases** that have already been identified in a job description.

⇩

2.**Masked-Language Modeling**: the model will **mask** the biased terms and then give possible **alternatives** according to neighboring words.

⇩

3.Filter out the biased alternatives in the **bias list** or are **classified as bias** in our identification system.

⇩

4.**Order** the unbiased alternatives based on some **criteria** such as cosine similarity.

⇩

5.Finally, the user will choose the most satisfactory alternative from the list.

Future work - Other Bias Types

- We only focused on five bias types:
  - Generic He/Generic She
  - Societal Stereotypes
  - Behavioural Stereotypes
  - Explicit Marking of Sex
- Other bias types from Jad Doughman's Gender Bias Taxonomy like Benevolent Sexism are also possible to appear in the job descriptions.
- In addition, it is necessary to read more related papers to expand the bias types.

Bias Type	Bias Subtype	Example	Methodology
Generic Pronouns	Generic He	A programmer must carry his laptop with him to work.	Supervised Learning
	Generic She	A nurse should ensure that she gets adequate rest.	Supervised Learning
Stereotyping Bias	Societal Stereotypes	Senators need their wives to support them throughout their campaign.	Supervised Learning
	Behavioural Stereotypes	The event was kid-friendly for all the mothers working in the company.	Supervised Learning
Sexism	Hostile Sexism	Women are incompetent at work.	Supervised Learning
	Benevolent Sexism	They're probably surprised at how smart you are, for a girl.	Supervised Learning
Exclusionary Terms	Explicit Marking of Sex	Chairman, Businessman, Manpower, Cameraman	Lexicon-Based
	Gendered Neologisms	Man-bread, Man-sip, Man-tini	Lexicon-Based

Figure 2: Overview of the taxonomy and link to detection methodology ([2])

Future work - Modeling

"It is necessary a dedicated and **ambitious** food development director to join our team."

dedicated and **ambitious** food development

- Experiment with different context size
- Use padding if the biased term is in the beginning or the end of the sentence
- Use weights for the context vs the biased term

References

[1] R.Frissen.A-machine-learning-approach-to-recognize-bias-and-discrimination-in-job-advertisements, 2021
[2] J.Doughman and W.Khreich.Gender bias in text:Labeled datasets and lexicons.CoRR, abs/2201.08675, 2022
[3] Employment Scam Aegean Dataset, 01 2020
[4] Text Analytics Explained-Job Description Data, 09 2018.https://www.kaggle.com/datasets/airiddha/trainrev1

Thank you for your attention