Codebook

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

Luise Vogt
Zhize Guan
Veselin Nasev
Yonghui Guo

Supervisor - Rohan Nanda

Department of Advanced Computing Sciences

Maastricht University

June 22, 2022

1 Brief Overview of the Dataset

The dataset contains information about sentences in job descriptions which have biased words or phrases. For each sentence, there is a bias label belonging to it. There are five bias types in total and they are Generic He/Generic She(n=1882), Behavioral Stereotypes(2364), Societal Stereotypes(488), and Explicit Making of Sex(311).

1.1 Data Source

Two different job description datasets have been used to generate the gender bias dataset. EMSCAD [4] is a publicly available dataset that contains 17,880 real-life job advertisements and 866 fake advertisements. For the use case of creating the gender bias dataset, only the 17,880 real-life job advertisements were used. The second dataset was published by Adzuna for a salary prediction competition and can be found here [10].

1.2 Data Distribution

The Figure 1 shows the distribution of the classes. The dataset is imbalanced but it is an accurate representation of which classes are dominant in the job descriptions dataset.

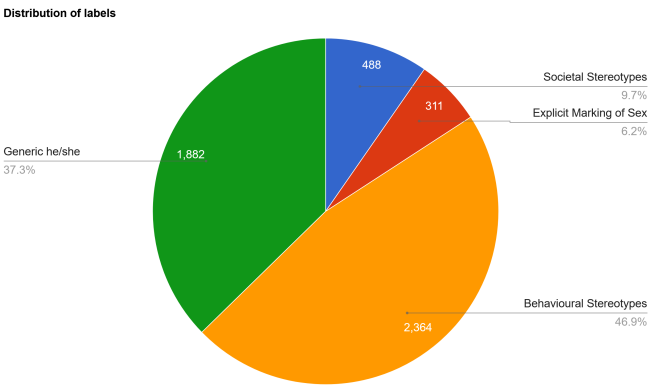

Figure 1: Class distribution

2 Example of the Dataset

In the original annotation dataset, there are only two variables. one is the sentence variable and the other is the label variable, including the starting and ending points and label.

Scientists are women.	[[0,19, 'Societal Stereotypes']]
The program was kid-friendly for all the mothers involved in the company .	[[16,53, 'Societal Stereotypes']]
We're looking for a young and stubborn new recruit to join our team.	[[30, 39, 'Behavioural Stereotypes']]
You must be taught to work independently, work, and exercise sound!	[[22, 40, 'Behavioural Stereotypes']]

Table 1: Example of the Dataset.

The Table 1 above shows some instances in the dataset.

3 Annotation Method

In this section we demonstrate our annotation method for each biased class from the taxonomy. Different approaches are needed for each class, based on the presence of the bias in the data.

3.1 Generic He / Generic She

Generic He/She is the choice of using a sex-defined pronoun to reference a specific job title. For example:

- A teacher is expected to be a good role model in all areas of his life.

In the sentence above, "his" refers to the profession teacher. This automatically implies that a male candidate is expected to full fill the position. Two techniques were used to assemble the dataset for this bias class. A research was carried out to find an open-source dataset which contains generic he/she gender. The research was successful and we found a dataset with 1585 biased sentences. The dataset was created in 2021 and can be found at Gender Bias Evaluation Set. The following step was to extract sentences that contain generic pronouns from the job description dataset. We extracted 300 sentences, part of them were not biased and part of them were biased. The extracted sentences were manually annotated. The main goal was to find sentences that contain generic pronouns but are used in a non-biased way. Thus, once a model is trained it will be able to learn the difference between biased and non-biased generic pronouns.

3.2 Behavioral Stereotypes

Behavioral stereotypes attributes the behaviour of someone to its gender[3]. For example:

- Mary must love dolls because all girls like playing with them.

This is a behavioral bias sentence. In this sentence, the act of girls liking dolls is directly applied to Mary, and it is actually uncertain whether Mary likes dolls. In our case, if a sentence contains a feminine biased word, and this sentence is used to describe the behaviour of certain person, the term in this sentence should be classified as biased. To generate the dataset of behavioral bias, we used two methods. The first one is that we manually labelled 1000 sentences that contain Behavioral biased words. These sentences were the original data points of behavioral bias type. Next, we replaced the biased words in these sentences with other biased terms. Then we went through the generated sentences and removed meaningless sentences and sentences that are not grammatically correct. After that, we used masked language model with BERT to create more sentences. In this step, the biased keywords in the sentence are retained.

The second method is extracting more sentences that contain biased terms from another dataset. These sentences were labelled manually (500 sentences). In order to expand the amount of data, the same masked language model with BERT approach to augmentation was applied to these sentences.

3.2.1 Societal Stereotypes

It is the assumption that certain characteristics or behaviors are unique to a specific social group[8]. They depict traditional gender roles that reflect social norms[6].

Below are several examples that explain the concept of societal stereotypes:

- 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.

Because this kind of bias is not very common in job descriptions, most of the training instances were created by ourselves manually. The method is simple, and we just combined different occupations with men or women together. For the rest several instances, we just used some techniques like fill-mask and GPT-2 to replace some parts of a sentence, which was collected from Jad Doughman's paper[3].

3.3 Explicit Marking of Sex

Explicit marking of sex is a type of gender bias that describes words that explicitly mention a gender and therefore excluding all genders not mentioned like "manpower" or "man hours". The first step to generate sentences for this type of bias was to create a lexicon of words that belong to this class. For generating such a lexicon, the method proposed by Jad Doughman. [3] has been used and adapted to the datasets [10] and [4]. Firstly, a base lexicon was manually build using sources like the gender sensitive language guidelines published from UNESCWA [5] and other sources ([1], [7], [2]). Secondly, a Word2Vec model [9] has been trained on the EMSCAD [4] and Adzuna [10] datasets, respectively. After the Word2Vec model was trained, the first 100 most similar words for every entry in the base lexicon were queried and the results that contained "man", "men", "woman", "women", "lord" or "lady" as sub string were selected to expand the lexicon. Since also words like "manage" would be chosen by this method, the final step was to manually check the entries in the enlarged lexicon. Different form for example the behavioral stereotypes, are words corresponding to the explicit marking of sex type of bias always biased and not dependent on the context. Hence, it was sufficient to filter for sentences that contain words form the lexica. From in total 717,880 samples ([4], [10]), approximately 330 sentences were found containing lexicon words. 19 sentences were discarded, because of bad quality which leads to 311 sentences for this type of bias.

References

[1] guidelines for non-sexist use of language - the american philosophical.

[2] Md_gender_bias dataset.

[3] J. Doughman and W. Khreich. Gender bias in text: Labeled datasets and lexicons. *arXiv preprint arXiv:2201.08675*, 2022.

[4] Employment Scam Aegean Dataset, 01 2020.

[5] Gender-sensitive language guidelines.

[6] Y. Hitti, E. Jang, I. Moreno, and C. Pelletier. Proposed taxonomy for gender bias in text; a filtering methodology for the gender generalization subtype, Aug. 2019.

[7] R. Kelly. A list of offensive (exclusionary) words used in job descriptions [2020 update], Oct 2018.

[8] A. Locksley, C. Hepburn, and V. Ortiz. Social stereotypes and judgments of individuals: An instance of the base-rate fallacy. *Journal of experimental social psychology*, 18(1):23–42, 1982.

[9] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space, 2013.

[10] Text Analytics Explained-Job Description Data, 09 2018.