Research Project Report

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

Contents

Abstract: Language, as a tool that people use every day, has a great impact on our lives and work. The biased language in job advertisements is one of the manifestations of this. To better identify biased words in job advertisements, we propose an updated version of the bias taxonomy, construct a labelled database, and develop a model. Experiments show that, depending on the context, this model can identify Generic He/Generic She, Behavioral Stereotypes, Societal Stereotypes, and Explicit Making of Sex in job advertisements.

Keywords: Biased language, Bias taxonomy, Job advertisements, Context-aware

1 Introduction

Natural language as we use it every day is biased, meaning that bias is prevalent in every aspect of live where natural language is used. This also includes job advertisements as Gaucher et al. [9] show. When companies start a hiring process, the first step to get in contact with a possible future employee is often a job advertisement. There has been research that has shown that especially women but also ethnic minorities found job advertisements less appealing when exclusive language is used [9]. For example, the word "bravery" is a word men can often identify with whereas women struggle to call themselves brave. Furthermore, it was shown that gender fair and inclusive language attracts a more diverse selection of possible future employees in the recruiting process with a wider range of background [3]. In addition to that, Jad Doughman [5] provide a taxonomy of gender bias as well as examples of each class to identify gender bias in written text. However, there has not been much research about how to use NLP techniques to identify and classify bias automatically in written text. This project aims on identifying and classifying gender bias in job advertisements using a context-aware system. In this report, (a shortened version of) the taxonomy of gender bias proposed by [5] has been used to provide a dataset that consists of labeled sentences from job advertisements according to the taxonomy. Furthermore, a fine-tuned BERT model [4] was applied to identify and classify gender bias in job advertisement in a context-aware manner. The following research questions are going to be discussed in this report:

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?

The main contribution of this project is going to be the dataset while the identification system is going leave room for improvement. The focus will be on generating a dataset of good quality with a taxonomy that supports the data annotation.

2 Related Work

This project is an extension of a previous project by a master's student Richard Frissen(MSc Business Intelligence and Smart Services). He wrote a thesis covering his study on the subject[8].In his research, he analyzed five forms of bias: Masculine and Feminine bias, Exclusive language, LGBTQ-colored language, and Demographic/Racial language. He achieved to identify biased languages in job advertisements based on machine learning methods and compared the performance of different models.

2.1 Taxonomy

Bias does not exist independently and discretely. The initial stage in identifying biased language is to classify the different types of bias while keeping a distinct separation between the resulting groupings[5]. To identify bias more comprehensively and precisely, we develop a taxonomy of bias applicable to job advertisements based on previous work. Jad Doughman developed a full taxonomy of gender bias types and their following subclasses, including Generic Pronouns, Sexism, Occupational Bias, Exclusionary Bias, Semantics[6]. Their work is based on various kinds of English texts but does not focus on job advertisements. In our work, we give an updated version of the taxonomy, which has been abridged according to the characteristics of the job advertisement.

2.2 Data Annotation and Augmentation

Identifying bias in text requires a classified dataset. To create the dataset, we used different methods and tools for annotation and augmentation. Doccano is a manual annotation tool that we use[15]. It is a user-friendly and lite tool for people to create labelled data for sentiment analysis and other tasks. Manually labelling is a time-consuming task, so we also adopt other data augmentation tools like Flashtext and Masked language with BERT. Flashtext is used to replace keywords in sentences to expand our dataset[17]. Masked language model with BERT is used to replace random words in sentences while not changing the context too much[20]. With the two-step process of masking and replacing, we can generate new sentences that share a similar context with the original sentences.

2.3 BERT Model

BERT is a deep bidirectional masked language model, and it can read text from right-to-left and left-to-right[4]. This model has the ability to understand context and ambiguity in text. By considering the full context of words, BERT can better grasp the intent of sentences, which helps identify bias.

3 Methodology

The following section will describe the data that has been used as well as the methods used for generating the dataset. The presented dataset contains of sentences that are assigned to five out of ten gender bias types introduced by Jad Doughman.[5], namely Generic He, Generic She, Behavioral Stereotypes, Societal Stereotypes and Explicit Marking of Sex. The five types of bias has been chosen on the frequency of their appearance in the datasets [7] and [18]. More details about the datasets as well as about the generation of the presented dataset by this report follow in the next subsections.

3.1 The Data

Two different job description datasets have been used to generate the gender bias dataset. EMSCAD [7] 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 [18].

3.2 Data Preprocessing

Various preprocessing techniques were applied to the datasets to achieve a cleaner, structured, and anonymized dataset. The following list of steps was executed to achieve the above-mentioned results:

- Remove white spaces and new rows
- Remove HTML characters
- Remove email addresses and phone numbers
- Split job descriptions into sentences
- Tokenize each sentence
- Remove sentences comprised of less than 5 tokens or 50 characters
- Remove sentences comprised of more than 25 tokens or 150 characters

- Filter sentences that contain at least one biased word from the list of biased words we have collected

3.3 Data Annotation

In this section we demonstrate our approach for each biased class from the taxonomy presented in the beginning. Different approaches were needed for each class, based on the presence of the bias in the data.

3.3.1 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.

3.3.2 Behavioral Stereotypes

Behavioral stereotypes attributes the behaviour of someone to its gender[5]. 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.

3.3.3 Societal Stereotypes

It is the assumption that certain characteristics or behaviors are unique to a specific social group[13]. They depict traditional gender roles that reflect social norms[11].

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[5].

3.3.4 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. [5] has been used and adapted to the datasets [18] and [7]. Firstly, a base lexicon was manually build using sources like the gender sensitive language guidelines published from UNESCWA [10] and other sources ([1], [12], [2]). Secondly, a Word2Vec model [14] has been trained on the EMSCAD [7] and

Adzuna [18] 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 ([7], [18]), 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.

4 Results & Analysis

In the last chapter, we explained the steps we took to collect, annotate and create a single dataset to recognize different types of gender bias. In this section, we show one approach to model the problem as a classification problem. Furthermore, we will dive into the validation metrics to access the quality of the model and the data. Next, we will try to find key predictors that explain the prediction of the model.

4.1 Classification Performance

The visualization below 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. The rest of this chapter will show one approach to model the problem and discuss the performance of it. The

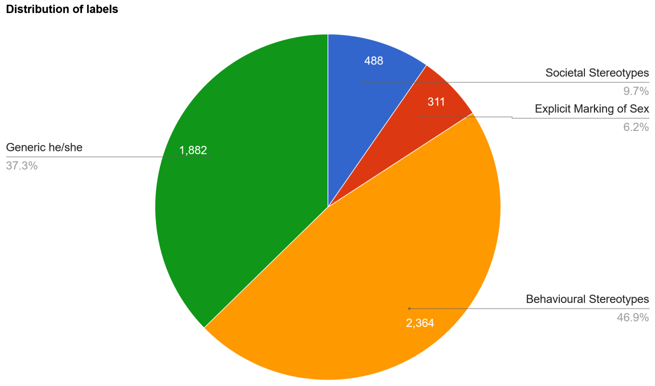


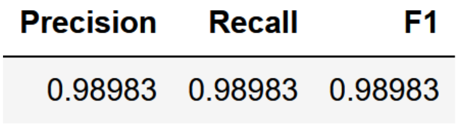
Figure 1: Class distribution

following steps describe our modeling process:

- Input a sentence to BERT to get an embedding
- Get an embedding per token
- Use single layer neural network with input size 768(length of BERT embedding)
- Use Sigmoid [19] activation function to find class probabilities

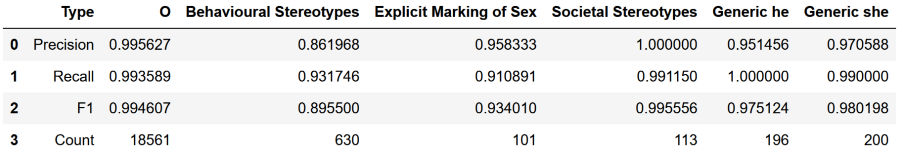
The problem was approached as a Named Entity Recognition(NER) problem. The input is a single token and the output is one of the biased classes or a non-biased class. The context is taken into account only through the BERT embedding, which might be a limitation but further research is needed to prove or disprove that observation.

The model was trained with 80% of the data and the rest was left for validation. The two tables below show the precision, recall, and F1 score. The figure 2 shows the average score of all the classes. The figure 3 presents a more detailed view of the evaluation metrics divided per class.



Precision	Recall	F1
0.98983	0.98983	0.98983

Figure 2: Validation metrics - averaged



	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

Figure 3: Validation metrics per class

Based on the results we decided to explore the inferior performance of the model on the Behavioural Stereotypes. Our findings show that whenever several adjectives refer to the same subject the cosine similarity of their BERT embeddings is high. Therefore, the model classifies adjectives that are not biased as biased. Let's look at the following example from the test dataset: "Our client, a multinational manufacturing company, is seeking an assertive, team-oriented individual to join their team.". The only biased term in the sentence is "assertive" but the model recognized "assertive" and

”team-oriented”. That mistake could be eliminated by modeling the problem differently. In the last section Future Work 4 we provide an alternative approach. Further validation and comparison are needed to find the optimal modeling technique but that is out of the scope of this paper.

4.2 Model Explainability

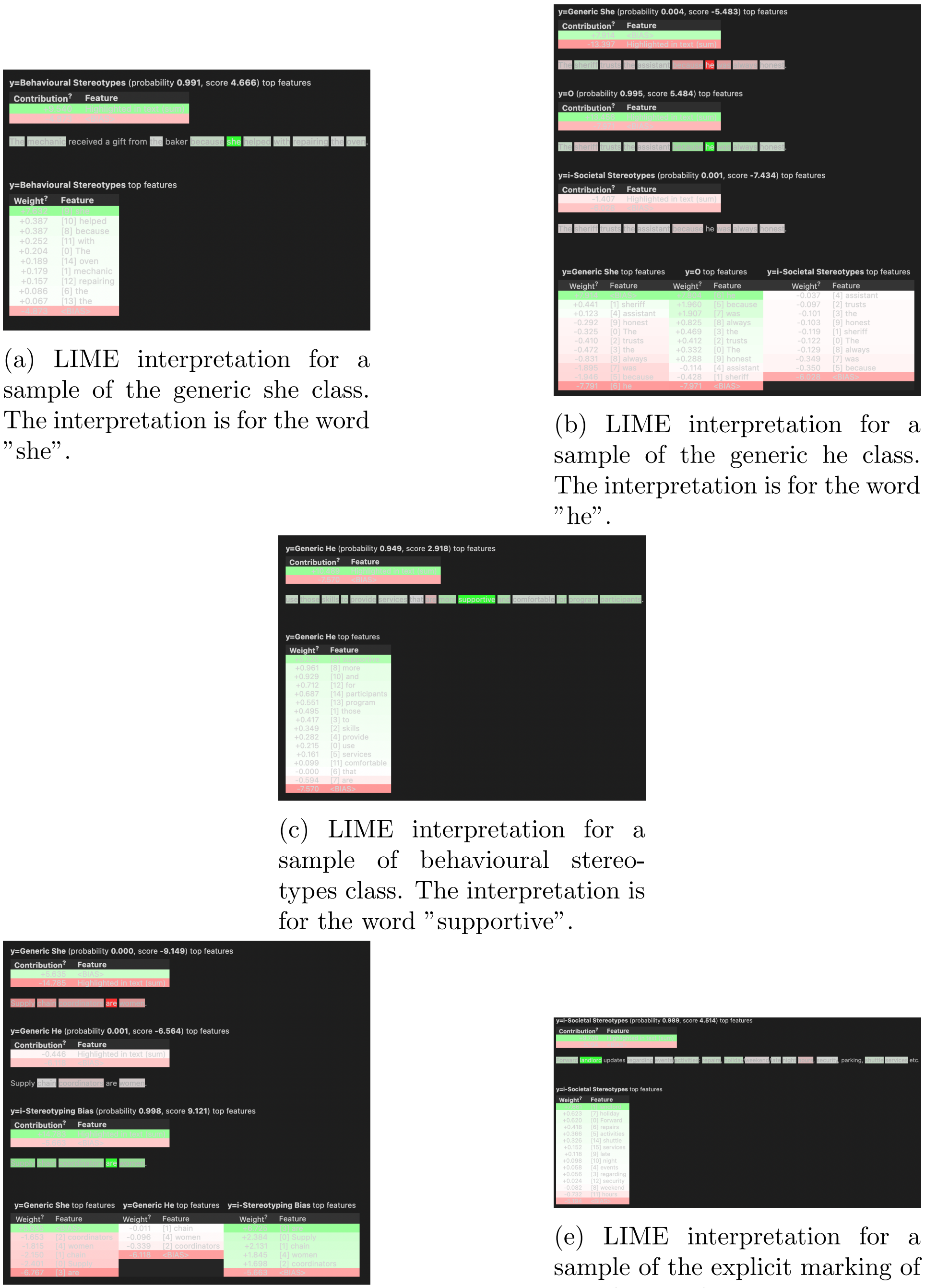
In order to understand the approach presented above better, LIME [16] has been used to find key predictors that explain the prediction of the model for a particular class of bias.

LIME is an interpretation technique that aims on understanding the model by changing the input data and understanding how the prediction of the model changes. It modifies single data samples by removing features, observes the resulting output and gives therefore a local interpretation of the single input sample. This technique gives an insight about what features in the data are important for the final prediction rather than analyzing the internal components of the model itself. The explanation of a data sample is achieved by training a local, more interpretable model, e.g. linear models, that approximates the underlying model. This model is trained on the modified data and approximates the underlying model only locally in the neighborhood of the data sample. In the case of textual input, the original sentence is changed by removing words from it. [16]

Since the presented approach is classifying every word in the input sentence and not the whole sequence, but LIME can only approximate the underlying model for a single prediction, the interpretation of key features is on word level. The local model was trained on 5000 modified sentences for each original input sample and gives an interpretation for the first biased word in each sample. Figure 4 shows one example interpretation for each type of bias. The first line always shows the predicted class for the selected word in the sentence of the internally trained model by LIME and the probability of that prediction. Words marked in green indicate that those words were important for the prediction, meaning that the probability for the predicted class decreased when the word was missing while words marked in red increased the probability when they were missing. The weights show the rate by how much the probability changes when that word was present in the sample sentence. LIME outputs the interpretation for classes with the most change in probability when words are missing or present while it was limited to three outputs hence, figure 4 displays from one to three interpretations per sentence.

All interpretations in figure 4 show results as somewhat expected. The he or she in a sentence of the generic he or generic she class, respectively, was the main key feature for the decision of the LIME model. An interesting observation is that in both cases the object the pronoun was referring to is influencing the probability for a predicted class in the opposite way as the pronoun, which leads to the conclusion that only the pronoun itself is important. In a sentence of the behavioural stereotypes class the describing adjective plays the main role for the final prediction as subfigure 4c shows. Furthermore, it can be seen that almost every single word in the sentence increased the probability for the predicted class which indicates that the context is important for this type of bias. This can also be observed for the societal stereotypes in subfigure 4d. The whole sentence is important for the decision of the model, meaning that the context is important for this type of bias as well. Having a look at subfigure 4e shows that the key feature for the explicit marking of sex bias is the word that contains the gender. The context does not seem to be very significant for the prediction, since most of the words do not influence the probability of a class and the words that do, only marginally.

However, figure 4 shows also the limitations of LIME. The predicted class with the highest probability never matches the actual label. This is probably because the underlying BERT model is too complex to approximate with a linear model, even locally. Considering this, LIME gives a first notion of what the main features for a prediction in each class are but there is room for improvement and interpretation left.



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>

(a) LIME interpretation for a sample of the generic she class. The interpretation is for the word "she".

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>		

(b) LIME interpretation for a sample of the generic he class. The interpretation is for the word "he".

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>

(c) LIME interpretation for a sample of behavioural stereotypes class. The interpretation is for the word "supportive".

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		**y=Generic He** top features		**y=i-Stereotyping Bias** top features	
Weight?	Feature	Weight?	Feature	Weight?	Feature
+5.635	<BIAS>	-0.011	[1] chain	+6.725	[3] are
-1.653	[2] coordinators	-0.096	[4] women	+2.384	[0] Supply
-1.815	[4] women	-0.339	[2] coordinators	+2.131	[1] chain
-2.150	[1] chain	-6.118	<BIAS>	+1.845	[4] women
-2.401	[0] Supply			+1.698	[2] coordinators
-6.767	[3] are			-5.663	<BIAS>

(d) LIME interpretation for a sample of the societal stereotype class. The interpretation is for the first word in the sentence

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
+7.661	[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>

(e) LIME interpretation for a sample of the explicit marking of sex class. The interpretation is for the word "PLACE HERE".

Figure 4: LIME interpretations for each type of bias.

5 Conclusion

In this paper, we propose an updated version of gender bias taxonomy in job advertisement. The taxonomy includes Generic He/Generic She, Behavioral Stereotypes, Societal Stereotypes, and Explicit Making of Sex. In light of prior research, we collect a list of phrases that related to gender bias. We also create a labelled dataset that contains information about biased language in job description and develop a bias identification model. Based on the results section, the performance of our model is good but further validation and comparison are still required.

6 Future Work

6.1 Suggestion System

The idea of the suggestion system is that we've already identified the biased words and phrases, and then the suggestions system will provide a list unbiased alternatives for users to replace the biased ones. For this system, we mainly want to use a model called Masked-Language Modeling with Bert. First, the model will mask the biased terms and then give possible alternatives according to neighboring words. Next, we will filter out the biased alternatives which are in our bias list or are classified as bias in our identification system. After that, we will order the unbiased alternatives based on some criteria such as cosine similarity. Finally, the user will choose the most satisfactory alternative from the list.

6.2 Other Bias Types

In our project, we only focused on five bias types, that were Generic He, Generic She, Societal Stereotypes, Behavioural Stereotypes and Explicit Marking of Sex. But other kinds of bias types from Jad Doughman's Gender Bias Taxonomy[5] like Benevolent Sexism are also possible to appear in the job descriptions even if there are fewer cases. In addition, it is necessary to read more related papers to expand the bias types and enrich the dataset continuously.

6.3 Training Other Classifiers

Another way to validate the dataset is to try different modeling techniques. We discussed one way of modeling the problem but here we would like to suggest a second option that can be explored. The previously suggested model considers the context indirectly, by making use of BERT embeddings but using only a single word as an input. The context might play an important role and make the model more robust. Therefore, we suggest the problem be modeled in the following way:

- word-1, word-2, BIASED TERM, word-4, word-5

The input to the model will be not only the biased term but the surrounding context. Different sizes of the context can be taken and further experimentation is needed to find the optimal size. If the biased term is at the beginning or the end of the sentence then it can be padded to achieve an uniform length.

References

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

[2] Md_gender_bias dataset.

[3] D. Collier and C. Zhang. Can we reduce bias in the recruiting process and diversify pools of candidates by using different types of words in job descriptions?, Oct 2016.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[6] J. Doughman, W. Khreich, M. El Gharib, M. Wiss, and Z. Berjawi. Gender bias in text: Origin, taxonomy, and implications. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 34–44, Online, Aug. 2021. Association for Computational Linguistics.

[7] Employment Scam Aegean Dataset, 01 2020.

[8] R. Frissen. A-machine-learning-approach-to-recognize-bias-and-discrimination-in-job-advertisements, 2021.

[9] D. Gaucher, J. Friesen, and A. Kay. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of personality and social psychology*, 101:109–28, 03 2011.

[10] Gender-sensitive language guidelines.

[11] 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.

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

[13] 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.

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

[15] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang. doccano: Text annotation tool for human, 2018. Software available from https://github.com/doccano/doccano.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.

[17] V. Singh. Replace or Retrieve Keywords In Documents at Scale. *ArXiv e-prints*, Oct. 2017.

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

[19] Wikipedia. Sigmoid Function — Wikipedia, the free encyclopedia. `https://en.wikipedia.org/wiki/Sigmoid_function#References`, 2022.

[20] X. Wu, T. Zhang, L. Zang, J. Han, and S. Hu. " mask and infill": Applying masked language model to sentiment transfer. *arXiv preprint arXiv:1908.08039*, 2019.