Project Plan

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 Data Science and Knowledge Engineering

Maastricht University

March 17, 2022

Contents

1 Introduction

When companies start a hiring process, the first step to get in contact with a possible future employee is often a job advertisement. Job advertisements are written in natural language and are therefore often biased in terms of the words that are used. Many words in our spoken language are sensitive to people from minority classes, e.g. race and gender. For example, the word "bravery" is a word men can often identify with whereas women struggle to call themselves brave. The unconscious use of such words in job advertisements may discourage members of the mentioned minority classes from applying to the job, thus reducing the diversity of candidates.

There has been research that found that a gender fair and inclusive language attracts a more diverse selection of possible future employees in the recruiting process with a wider range of background [1]. Furthermore, Nikhil Garg et al. [4] investigated gender and ethnic bias in natural language using word embedding. However, there has not been much research in linguistic and semantic analysis of job advertisements using a context-aware system to identify biased language in job advertisements.

2 Problem Statement and Research Questions

This project will focus on investigating approaches to identify and replace biased language in job advertisements using a context-aware system. In particular, we will explore if a context-aware system can be used to identify different types of biased and discriminatory language and if machine learning techniques can be used to implement a suggestion system that provides an alternative unbiased term once a biased term has been classified.

The major research question addressed in this project is: How can state-of-the-art NLP and machine learning (ML) techniques be utilized to develop a context-aware biased language filter and augmentation system? To answer this research question, we decided to break it down to the following five 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?

5. How can a NLP machine learning model be used to establish a suggestion system to replace biased language?

3 Related Work

The project was executed once before by a master's student Richard Frissen(MSc Business Intelligence and Smart Services). He wrote a thesis that covers his research on the topic [3]. In his paper, he researched the following five types of bias: masculine and feminine bias, exclusive language, LGBTQ-coloured language, and Demographic/Racial language. Throughout his research, Richard Frissen found papers related to the above-mentioned biases and created a dataset with words for each of them.

In his work, Richard explores several supervised state-of-the-art approaches for word classification. He achieved the highest accuracy with FastText word embeddings combined with a Random Forest classifier. He tested other alternative models, for example: named entity recognition models(NER) and Bidirectional Encoder Representations from Transformers(BERT).

Richard Frissman's thesis will serve as a starting base for our project, but the goal will be to improve on the limitations of his work. The main goal will be to explore how context will influence the prediction. The second improvement will be to build a suggestion system to replace the biased language with neutral-sounding alternatives.

3.1 Data Source

The Employment Scam Aegean Dataset (EMSCAD) is a publicly available dataset containing 17,880 real-life job ads. The dataset contains 17,014 legitimate and 866 fraudulent job ads published between 2012 to 2014. In our project, we will use only legitimate ads.

Richard Frissman used this dataset in his experiments, and we think we can utilize the data to create a dataset with biased words and the context they are used in.

4 Concepts and Approach

4.1 Concepts

An important concept in our project is the taxonomy of bias, which contains the categories of biased terms and the relationships of these biases. Jad Doughman[2]has studied gender bias through word embedding for a long time and provided a version of the taxonomy of gender bias. We will mainly refer to his work in our project.

The figure below shows his updated version of the gender bias taxonomy.

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

4.2 Approach

To guarantee that the outcome of the project will be of good quality, we will focus on improving the training dataset. We will use the words that Richard Frissen found in his research and construct a new dataset with sentences that include those words. Having the context will help us manually label the data and prevent incorrect labels.

The second step will be to focus on research and find a suitable model for classification. We will test the model, and when satisfactory results are achieved, we can continue with the next step.

The third step will be to build a suggestion system that replaces the biased language. To achieve higher quality results we want to include the context, instead of using synonyms of the word. With this approach, we can find a more suitable replacements.

After we have the key components ready, we can combine them into one system. In this system, users will be able to enter a job advertisement and receive a corrected version of it with an explanation of the different types of biases found in the text.

Machine learning models are becoming sophisticated and hard to explain. They start to feel like a black box. To prevent that feeling, and increase the reliability of our results, we will focus on explainability. We will use popular tools like SHAP [5] and LIME [6] to find the key predictors/features.

5 Deliverables

At the end of the project, there will be two deliverables to be provided:

- A demo that presents a context-aware system to identify biased language and suggests unbiased alternative words using natural language processing and machine learning
- A report that explains the design of the demo and solutions for our research questions

Apart from what was mentioned above, the result of each phase would be presented by a PowerPoint presentation. For phrase one and two, a project plan and a Layman's description would also be delivered respectively.

6 Planning

Setting up milestones and designing Gantt chart allows us to have a broader understanding of overall project goals and process. The next two phases of this project have different focus.

6.1 Milestones

- Propose a taxonomy of bias

  Creating a taxonomy of bias can provide an easy way to organize and show the biases. With taxonomy, biased words can be grouped based on their meanings. It can also reduce the difficulty to label new words.

- Facilitate data annotation based on the taxonomy of bias

  Based on the reference of previous research of Frissen [3], we will carry out further data annotation by the taxonomy of bias for Employment Scam Aegean Dataset (EMSCAD). The EMSCAD dataset can serve as a significant testbed for the development of new methods for combating employment scams.

- Design a context-aware system

  To have a more accurate classification of biased language, context should be taken into account. Some terms can be recognized as biased words in certain contexts but can also have natural meaning in a different context. It helps to identify the word that needs to be replaced more precisely by highlighting the context in assessing if the word is biased.

- Design a suggestion system

  After identifying the biased language, a suggestion system can assist in providing alternative terms which are semantically and grammatically appropriate. It can help minimize unconscious biases in job advertisements and boost the probability of employing a diverse team.

6.2 Gantt Chart

Figure 2 below shows the main tasks and time arrangement for this project.

7 Risk analysis and contingency plans

The following list shows the identified risks of the project and their contingency plans.

- **A team member becomes unavailable.**

  Description – While we are a group of 4 people now, it might be the case that one of our colleagues might need to drop the course or find himself/herself unavailable or sick.

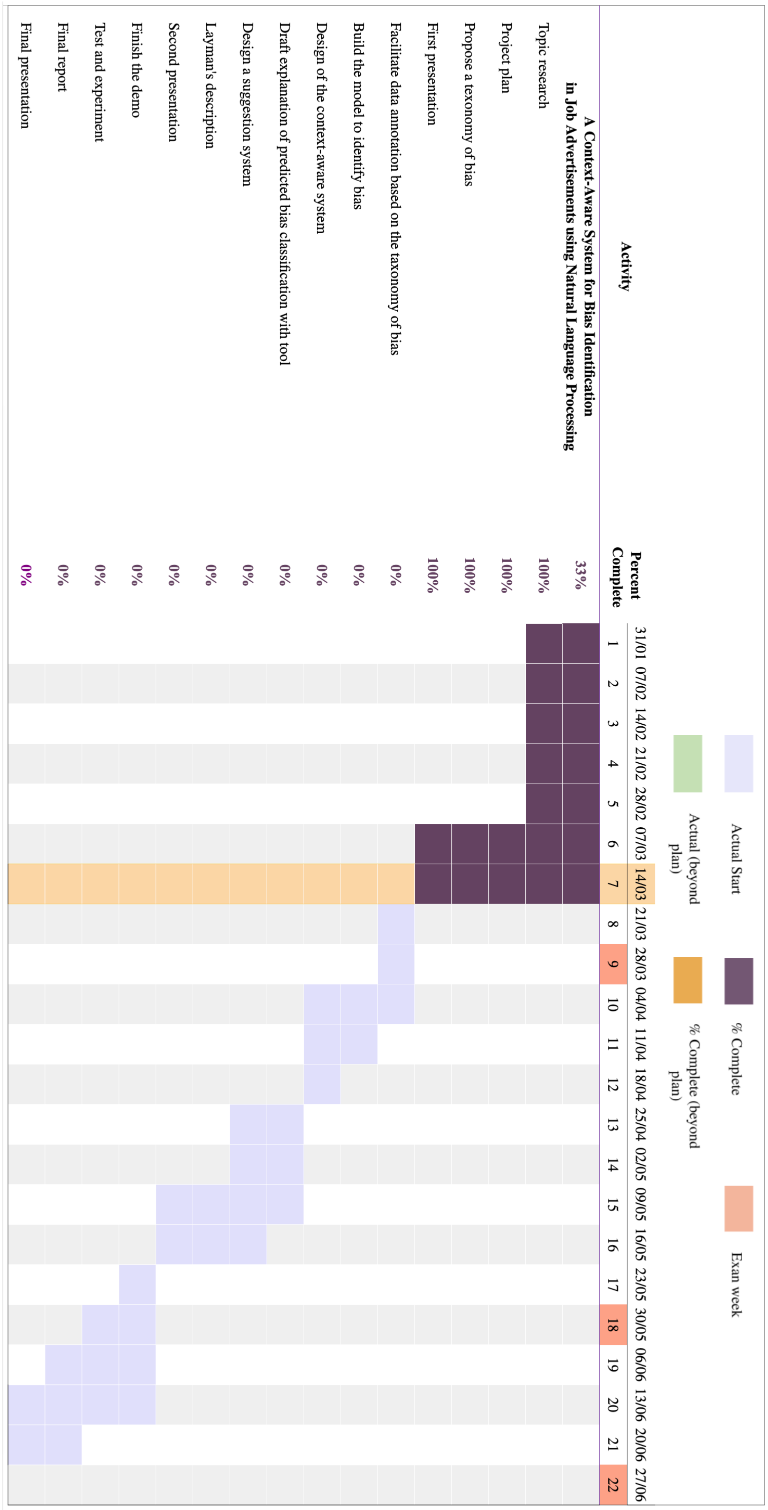

Figure 2: Timetable for the whole project

Identification strategy – By having active communication in different channels (WhatsApp group, emails, weekly meetings) we can increase the probability of anticipating the identification of this problem.

Contingency plan – We have an internal meeting almost every week to fully communicate the progress of everyone's tasks. At the same time, we upload individual achievements to the Github repository to share with others. That way we can ensure that each member is familiar with others' achievements and we reduce the impact of a sudden loss of a teammate. The creation of a Github repository for this project will ensure that all changes are logged.

- **Delay in agreement over project direction and methodology**

  Description – The disagreement between different team members over the direction of the project-i.e. focus on the leading NLP techniques first or data annotation supplement first- could delay the project and hinder the achievements.

  Identification strategy – Regular internal meetings, specially at the beginning of the project, could show if such disagreement exist on a timely manner.

  Contingency plan – In case that disagreement persists, we will discuss it with our supervisor first to get his suggestions. Then if we still can't reach an agreement, we will separate the group into two parts and pursue the ideas independently with a plan to merge them at the end of the project phase.

- **Implementing of NLP techniques is troublesome**

  Description – At the beginning of our project, only one teammate has learned Natural Language Processing before, and other members are all beginners in this field. So these members may have to spend a lot of time studying and could sometimes have trouble with this field. This is a big uncertainty and it will affect the progress of the project.

  Identification strategy – This happens to all of us when we study NLP, or when we read the literature about our project.

  Contingency plan – To deal with this, in every internal meeting and supervisor meeting, we will feedback questions in time, ask others for help, and share related learning resources with each other to make sure that our project is on track without too much delay.

References

[1] 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.

[2] J. Doughman and W. Khreich. Gender bias in text: Labeled datasets and lexicons. *CoRR*, abs/2201.08675, 2022.

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

[4] N. Garga, L. Schiebingerb, D. Jurafskyc, and J. Zoue. Word embeddings quantify 100 years of gender and ethnic stereotypes, Apr 2018.

[5] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[6] M. Tulio Correia Ribeiro. Lime: Explaining the predictions of any machine learning classifier, 2021.