Gender Bias Correction in NLP

Layla Shalabi[*]
Razi Khawaja[*]
lshalabi@hawk.iit.edu
rkhawaja@hawk.iit.edu
Illinois Institute of Technology
Chicago, Illinois, USA

ABSTRACT

Natural language processing (NLP) generally involves extensive data preparation since natural language can become ambiguous outside of its native context. Our project aims to find and correct gender biased instances of text.

CCS CONCEPTS

• **Software and its engineering**; • **Natural language processing**;

KEYWORDS

Natural, Language, Processing, NLP, Gender, Bias

1 INTRODUCTION

The aim of this report is to outline a proposal for our CS 577 - Deep Learning class. The team consists of two members: Layla Shalabi and Razi Khawaja. The goal of the project is to detect gender bias in natural language inputs and datasets and offer corrections as output. Our approach to achieve this will implement deep learning models and traditional artificial intelligence models e.g. Naive Bayes and compare results.

2 PROBLEM STATEMENT

The characteristic ambiguity of natural language makes robust definitions of bias hard to define. Biases can exist in the domains of sex, gender, age, race, ethnicity, and more either independently or at the same time. The objectives of this project will be defining gender bias, identifying gender bias, and then correcting gender bias in text.

3 LITERATURE REVIEW

Not much literature research has been conducted by our team thus far but we did begin the process with a paper by Stanczak and Agenstein [3]. Some notable takeaways from our reading include:

- Definitions of bias mechanisms in NLP including structural bias and contextual bias.
- A definition of hostile and benevolent sexism.
- Several gendered lexica and databases from which we can use in our project.
- Mathematically defined probability equations to use for baseline traditional artificial intelligence (AI) approaches.
- Many references to read in the coming weeks.
- Ideas for directions we could take to solve our correction problems including database manipulation and adjusting algorithms to 'unbias' results.

There are many papers in the area of bias in NLP with many focusing on gender. While our team did not read many of the papers yet, a cursory glance at headlines and paper counts suggests that the issue is currently unsolved. A description of what would make our project different from existing papers requires more research to say with accuracy. However, a cursory assessment of the research in the space seems concerned with ratings and metrics and not creating a model that can be applied to fix an identified bias problem. Our work aims to offer replacement text that has been 'cleaned,' for lack of a better phrase, of gender bias. This has applications in business environments such as resume degendering for application review, headline cleaning to avoid gendered tropes in news, and script review for creative works.

4 MILESTONES

4.1 Research ∼ 2 Weeks

The team needs to conduct a thorough literature review. We intend to begin with the references of Stanczak and Agenstein [3] and expand out from there. This phase of the project has to be aggressive in order to better narrow our problem statement.

4.2 Datasets ∼ 2 Weeks

This search will run concurrently with our literature review. Outside of the aforementioned datasets nestled inside papers, we have begun to find candidate datasets for our project [1, 2]. Obviously, dataset selection will depend on our final problem statement which is dependent on our research.

4.3 Pseudocode ∼ 1 Week

Once we know what we're doing, it'll be easier to get it done. Our project program flow will then be defined. At this stage, there is no reason to believe our flow will be remarkable though the specifics are vague. We will seek to define at least two models, one made with deep learning (DL) and the other with traditional artificial intelligence (AI) models e.g. Naive Bayes (NB). We will then define metrics by which to rate the models. Afterwards, we will have to test the outputs of both models and summarize our results. Seeing as we aim to change text to clean the model output, we will have to think on where exactly the respective model will construct its output text. This is not explicitly what NB is designed to do as it's a classification model and not a conversation model.

4.4 Project Summary ∼ 1 Week

At this stage, we will have to compose an initial progress report. We ought to have everything settled at this point though we may not have actually coded anything. The intention, and hope, is that

[*] Both authors contributed equally to this research.

we can compare and use models that are relatively off the shelf such that our job will come down to tuning hyperparameters as opposed to defining network structures explicitly. Working with natural language is tedious as a field and so we are aiming to avoid defining lexica and dictionaries wherever possible.

4.5 Project Development ~ 4 Weeks

We intend to work on the project to get something useful from this point onwards. Model tuning and comparison metrics will have to be compiled to make it in to the final report.

REFERENCES

[1] The Devastator. [n. d.]. *Women in Headlines: Bias*. https://www.kaggle.com/datasets/thedevastator/women-in-headlines-bias

[2] Ibrahimmazlum. [n. d.]. *MBTI Personality Type Twitter Dataset.* https://www.kaggle.com/datasets/mazlumi/mbti-personality-type-twitter-dataset

[3] Karolina Stanczak and Isabelle Augenstein. [n. d.]. A Survey on Gender Bias in Natural Language Processing. arXiv:2112.14168 [cs] http://arxiv.org/abs/2112.14168

Received 1 October 2023; accepted NA