Project Proposal Template

First Author	Second Author	Third Author	Fourth Author
G01375933	G01399256	G01379939	G01377914
Hruthik Bojja	Sumanth Bejugam	Anthony Mary Thalapaneni	Lokeshwar Sai
sbojja4@gmu.edu	sbejugam@gmu.edu	athalapa@gmu.edu	lsreeram@gmu.edu

1 Introduction

1.1 Task / Research Question Description

The paper "NLPositionality: Characterizing Design Biases of Datasets and Models" by Santy et al. (2023) seeks to develop a general and robust framework for characterizing the design biases of datasets and models across a variety of languages and NLP tasks. This work is motivated by the observation that datasets and models often reflect the positionality of their creators, which can lead to design biases that can impact the performance of NLP systems. The paper proposes a framework called NLPositionality, which can be used to measure the positionality of datasets and models and to identify potential areas of bias. The paper also evaluates the NLPositionality framework on a variety of NLP tasks, and finds that it is effective in identifying design biases.

1.2 Motivation and Limitations of existing work

There are other works that attempt to tackle this problem, but most of the existing efforts suffer serious limitations. As an illustration, several works have looked into biases in specific data sets or models, but a more generic way of describing design bias is required. Moreover, most research has been centered on English language databases and models where biases can be studied more thoroughly too.

To this end, This research attempts to provide a broad schema for conceptualizing prejudices in datasets as well as models regarding different languages. Secondly, This will discuss the effects of the bias in the system's design on the efficiency of NLP systems.

1.3 Proposed Approach

The paper proposes to develop a framework for characterizing design biases in datasets and models based on the following steps:
Identify the relevant dimensions of positionality (e.g., demographics, expertise, cultural background).
Develop methods for measuring the positionality of datasets and models on each dimension.
Analyze the relationship between the positionality of datasets and models and the performance of NLP systems.
The paper plans to use a variety of methods to measure the positionality of datasets and models, including:
Demographic analysis of annotators and developers
Analysis of the content of datasets and models
Expert surveys
The paper plans to evaluate the performance of its framework on a variety of NLP tasks, such as sentiment analysis, machine translation, and question answering.

1.4 Likely challenges and mitigations

One of the biggest challenges in this work is developing methods for measuring the positionality of datasets and models. Positionality is a complex concept, and it is difficult to quantify. Additionally, the positionality of datasets and models can vary depending on the specific NLP task.

To address these challenges, the paper plans to use a variety of methods to measure positionality and to triangulate its results. The paper also plans to work closely with experts in the field to develop and refine its methods.

Another challenge is that collecting and annotating data for NLP tasks can be expensive and time-consuming. To address this challenge, the

paper plans to use existing datasets and crowdsourcing platforms. The paper also plans to develop methods for reducing the amount of data needed for annotation.

Finally, it is possible that the paper's experiments will not go as planned. If this happens, the paper will adjust its approach as needed. The paper is also committed to open science, and it will share its code and data with the community so that others can reproduce its work.

Overall, the paper is confident that its proposed approach is feasible and that it will make a significant contribution to the field of NLP.What is hard about this task / research question? What are your contingency plans if things turn out to be harder than expected or experiments do not go as planned?

2 Related Work

Paper 1: Title: Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey (Hort et al., 2023)
Description: This paper offers a comprehensive survey of 341 publications on bias mitigation methods in Machine Learning, categorizing them by intervention procedures and techniques. The insights gathered aim to assist practitioners in informed decision-making when developing and evaluating new bias mitigation methods, including fairness metrics and dataset choices.

Paper 2: Title: Identifying and mitigating spurious correlations for improving robustness in nlp models: (Wang et al., 2023)
Description: NLP models have made significant strides but are criticized for lacking robustness due to exploiting spurious correlations between training data and task labels. This paper introduces an automated method to identify and mitigate such shortcuts, enhancing model robustness across various applications efficiently.

Paper 3: Title: "A survey on bias and fairness in machine learning: (Mehrabi et al., 2023)
Description: The importance of addressing fairness and mitigating biases in AI systems has grown significantly with their widespread use in sensitive domains, emphasizing the need to ensure non-discriminatory decision-making. This survey explores real-world applications, sources of biases, and fairness definitions while highlighting ongoing efforts to address bias-related challenges, aiming to inspire further research in various AI domains.

Paper 4: Title: Fairness-aware class imbalanced learning: (Subramanian et al., 2023)
Description: Class imbalance is a common challenge in many NLP tasks, and has clear connections to bias, in that bias in training data often leads to higher accuracy for majority groups at the expense of minority groups. However there has traditionally been a disconnect between research on class-imbalanced learning and mitigating bias, and only recently have the two been looked at through a common lens. In this work we evaluate long-tail learning methods for tweet sentiment and occupation classification, and extend a margin-loss based approach with methods to enforce fairness. We empirically show through controlled experiments that the proposed approaches help mitigate both class imbalance and demographic biases.
How the work in this paper is related to the work in the four papers:
Paper 1: The work in this paper builds on the work of Mehrabi et al. by providing a more comprehensive and robust framework for characterizing and mitigating design biases in NLP datasets and models. My framework considers a wider range of dimensions of positionality, and it also takes into account the impact of bias on the performance of NLP systems. Paper 2: The work in this paper is related to the work of Wang et al. because I am also interested in developing methods for identifying and mitigating bias in NLP datasets and models. However, my work focuses on characterizing the design biases of datasets and models, while Wang et al. focus on developing methods for improving the robustness of NLP models to spurious correlations. Paper 3: My work is related to the work of Chouldechova et al. because I am also interested in understanding the different types of bias that can occur in NLP systems, as well as the various definitions of fairness. However, my work focuses on developing methods for characterizing and mitigating design biases in NLP datasets and models, while Chouldechova et al. provide a more general survey of bias and fairness in machine learning. Paper 4: My work is related to the work of Borkan et al. because I am also interested in developing methods for mitigating bias in NLP datasets and models. However, my work focuses on characterizing the design biases of datasets and models, while Borkan et al. focus on developing methods for mitigating bias in class-imbalanced

Demographic	country_longest	education	ethnicity	gender	native_language	age	country_residence	religion
African-Islamic	0.47	0.40	0.46	0.40*	0.26	0.43*	0.32	0.25
Baltic	0.71*	0.71*	0.69*	0.32	0.10	0.37	-0.03	0.12
Catholic-Europe	0.69*	0.58*	0.73*	0.39	0.11	0.34	0.18	0.18
Confucian	0.69*	0.53*	0.70*	0.50*	0.24	0.44	0.20	0.42
English-Speaking	0.77*	0.63*	0.72*	0.70*	0.36*	0.60*	0.38*	0.43*
Latin-America	0.46	0.42	0.48	0.41	0.18	0.35	0.17	0.28
Orthodox-Europe	0.58*	0.60*	0.67*	0.35	0.20	0.38	0.12	0.18
Protestant-Europe	0.65*	0.51*	0.67*	0.49*	0.37*	0.38*	0.39*	0.37*
West-South-Asia	0.63*	0.60*	0.59*	0.35	0.23	0.37	0.37	0.33
College	0.73*	0.64*	0.70*	0.67*	0.34*	0.58*	0.39*	0.40*
Graduate School	0.73*	0.56*	0.67*	0.59*	0.31*	0.53*	0.32*	0.40*
High School	0.68*	0.54*	0.67*	0.60*	0.34*	0.54*	0.35*	0.36*
Phd	0.64*	0.51*	0.64*	0.51*	0.27	0.46*	0.28	0.40*
Pre-High School	0.56*	0.46*	0.59*	0.36	0.19	0.40	0.18	0.33
Professional School	0.53*	0.46*	0.49*	0.59*	0.05	0.36*	0.08	0.20
Asian	0.66*	0.55*	0.62*	0.58*	0.38*	0.49*	0.34*	0.38*
Black	0.61*	0.50*	0.57*	0.51*	0.27	0.33*	0.29	0.35*
Latino/Latina	0.62*	0.53*	0.52*	0.44*	0.30	0.44*	0.25	0.29
Native American	0.59*	0.52*	0.64*	0.14	0.27	0.26	0.28	0.28
Pacific Islander	0.65*	0.63	0.62	0.25	0.55	0.50	0.46	0.53
White	0.74*	0.61*	0.70*	0.71*	0.33*	0.60*	0.36*	0.40*
Man	0.74*	0.62*	0.71*	0.66*	0.35*	0.61*	0.35*	0.37*
Non-Binary	0.60*	0.52*	0.56*	0.59*	0.19	0.37*	0.25	0.33*
Woman	0.76*	0.59*	0.75*	0.67*	0.37*	0.56*	0.38*	0.37*
English	0.78*	0.64*	0.72*	0.71*	0.35*	0.61*	0.39*	0.43*
Not English	0.64*	0.52*	0.66*	0.52*	0.40*	0.50*	0.42*	0.44*
10-20	0.69*	0.60*	0.69*	0.63*	0.37*	0.59*	0.39*	0.34*
20-30	0.73*	0.66*	0.70*	0.66*	0.35*	0.56*	0.37*	0.43*
30-40	0.68*	0.51*	0.62*	0.49*	0.28*	0.43*	0.32*	0.36*
40-50	0.62*	0.54*	0.64*	0.65*	0.29	0.57*	0.29	0.35*
50-60	0.66*	0.52*	0.58*	0.53*	0.35*	0.46*	0.31	0.26
60-70	0.58*	0.44*	0.52*	0.41	0.39	0.45	0.14	0.14
70-80	0.56*	0.52*	0.56*	0.49	0.36	0.35	0.25	0.85*
80+	0.52	0.48	0.60	0.63	0.01	0.45	-0.09	0.43
African-Islamic	0.40	0.40	0.45	0.34	0.14	0.23	0.16	0.26
Baltic	0.68*	0.59*	0.48	0.44	0.20	0.66	0.23	0.20
Catholic-Europe	0.57*	0.42*	0.66*	0.38	0.15	0.45*	0.20	0.24
Confucian	0.69*	0.62*	0.72*	0.56*	0.35	0.53*	0.46	0.42
English-Speaking	0.77*	0.65*	0.73*	0.70*	0.36*	0.62*	0.40*	0.43*
Latin-America	0.55*	0.57*	0.60*	0.30	0.12	0.26	-0.04	0.17
Orthodox-Europe	0.52*	0.60*	0.60*	0.25	0.29	0.29	0.23	0.23
Protestant-Europe	0.63*	0.52*	0.62*	0.51*	0.32	0.36*	0.32	0.33*
West-South-Asia	0.61*	0.57*	0.53*	0.62	0.18	0.53	0.09	0.19
Buddhist	0.64*	0.58*	0.55*	0.48	0.18	0.22	0.16	0.38
Christian	0.74*	0.54*	0.72*	0.53*	0.37*	0.49*	0.35*	0.35*
Hindu	0.65*	0.60*	0.58*	0.50*	0.29	0.36	0.31	0.35
Jew	0.66*	0.60*	0.60*	0.70*	0.31	0.48*	0.28	0.28
Muslim	0.62*	0.60*	0.68*	0.47	0.22	0.36	0.18	0.25
Spiritual	0.61*	0.60*	0.72*	0.47	-0.12	0.38	0.26	nan0

Table 1: Our Results

DATASETS: SocialChemistry, DynaHate MODELS: GPT-4, Delphi, PerspectiveAPI, RewireAPI, ToxiGen RoBERTa

Demographic	Social Acceptability					Toxicity & Hate Speech						
	#	α				#	α					
Country (Lived Longest)												
African Islamic	316	0.20	0.54*	0.49	0.47	234	0.22	0.39	0.29	0.39	0.27	0.25
Baltic	140	0.41	0.73*	**0.72***	0.71*	54	0.50	0.38	-0.08	0.20	0.05	0.46
Catholic Europe	452	0.28	0.64*	0.59*	0.68*	183	0.41	0.32	0.12	0.32	0.21	0.21
Confucian	528	0.42	0.75*	0.58*	**0.74***	154	0.24	0.47	0.28	0.51*	0.12	**0.52***
English-Speaking	8289	0.51	**0.76***	0.61*	**0.74***	4025	0.40	**0.70***	**0.33***	**0.58***	**0.37***	0.41*
Latin American	281	0.33	0.45	0.41	0.47	65	0.20	0.39	0.10	0.28	0.09	0.17
Orthodox Europe	426	0.39	0.56*	0.58*	0.67*	139	0.32	0.36	0.18	0.47	0.15	0.13
Protestant Europe	706	0.48	0.65*	0.57*	0.67*	387	0.37	0.40*	0.32	0.23	0.29	0.34
West South Asia	413	0.40	0.63*	0.60*	0.59*	116	0.21	0.34	0.20	0.33	0.30	0.21
Education Level												
College	4489	0.48	**0.74***	**0.66***	**0.69***	2383	0.39	**0.66***	**0.34***	**0.56***	**0.38***	**0.39***
Graduate School	1116	0.53	0.72*	0.54*	**0.69***	604	0.36	0.59*	0.28*	0.51*	0.25	0.38*
High School	2183	0.49	0.67*	0.54*	0.64*	908	0.41	0.60*	0.25	0.49*	0.30*	0.37*
PhD	709	0.46	0.65*	0.55*	0.61*	359	0.45	0.48*	0.19	0.43*	0.26	0.31
Pre-High School	406	0.40	0.56*	0.46*	0.59*	116	0.26	0.37	0.24	0.45*	0.25	0.38
Professional School	460	0.40	0.53*	0.46*	0.49*	195	0.09	0.61*	0.10	0.35	0.09	0.19
Ethnicity												
Asian, Asian American	1160	0.55	0.66*	0.55*	0.63*	644	0.45	0.57*	0.35*	0.47*	0.33*	0.39*
Black, African American	465	0.52	0.61*	0.50*	0.57*	287	0.34	0.56*	0.32	0.36*	0.31	0.37*
Latino / Latina, Hispanic	314	0.57	0.62*	0.52*	0.54*	239	0.36	0.43*	0.39*	0.46*	0.31	0.31
Native American, Alaskan Native	103	0.64	0.59*	0.52*	0.64*	65	—	0.23	0.31	0.31	0.32	0.33
Pacific Islander, Native Australian	38	0	0.65*	**0.63**	0.62	27	—	0.36	**0.65**	0.54	**0.64**	**0.57**
White	3102	0.55	**0.73***	0.61*	**0.70***	1831	0.44	**0.69***	0.29*	**0.56***	0.32*	0.38*
Gender												
Man	4082	0.45	0.73*	**0.63***	0.69*	1798	0.37	**0.65***	**0.34***	**0.56***	0.34*	0.36*
Non-Binary	858	0.41	0.60*	0.51*	0.55*	329	0.48	0.57*	0.21	0.37*	0.27	0.31*
Woman	4368	0.55	**0.74***	0.60*	**0.73***	2357	0.39	0.63*	**0.34***	0.53*	**0.38***	**0.37***
Native Language												
English	7338	0.51	**0.76***	**0.64***	**0.71***	3622	0.40	**0.70***	**0.33***	**0.60***	**0.39***	**0.42***
Not English	2157	0.40	0.62*	0.54*	0.64*	1020	0.27	0.46*	0.32*	0.39*	0.32*	0.36*
Age												
10-20 yrs old	3360	0.50	0.70*	0.61*	0.69*	1615	0.39	0.61*	0.32*	0.55*	0.36*	0.36*
20-30 yrs old	4066	0.47	**0.74***	**0.66***	**0.70***	2114	0.39	**0.65***	0.34*	0.56*	**0.38***	0.42*
30-40 yrs old	870	0.51	0.66*	0.52*	0.61*	419	0.28	0.48*	0.14	0.41*	0.24	0.29
40-50 yrs old	655	0.44	0.62*	0.55*	0.63*	256	0.28	0.63*	0.29	**0.57***	0.31	0.37*
50-60 yrs old	308	0.49	0.69*	0.53*	0.60*	199	0.39	0.57*	0.26	0.41*	0.20	0.25
60-70 yrs old	204	0.48	0.64*	0.49*	0.60*	19	—	0.57	**0.42**	0.46	0.05	-0.18
70-80 yrs old	68	—	0.56*	0.52*	0.56*	24	—	0.50	0.35	0.36	0.24	**0.85***
80+ yrs old	24	—	0.52	0.48	0.48	12	—	0.63	0.01	0.45	-0.09	0.43
Country (Residence)												
African Islamic	164	0.27	0.49	0.48	0.46	116	0.21	0.35	0.23	0.29	0.15	0.16
Baltic	53	0.02	0.65	**0.65**	0.33	14	0.00	0.42	0.14	0.52	0.35	**0.75**
Catholic Europe	406	0.33	0.53*	0.41*	0.64*	172	0.37	0.32	0.11	0.38	0.15	0.22
Confucian	268	0.42	0.68*	0.55*	**0.77***	83	0.17	0.41	**0.36**	0.45	0.33	0.48
English-Speaking	7315	0.50	**0.76***	**0.65***	0.73*	3819	0.40	0.72*	0.34*	**0.60***	**0.38***	0.42*
Latin American	166	0.43	0.54*	0.56*	0.59*	53	0.15	0.30	0.12	0.26	-0.04	0.17
Orthodox Europe	264	0.38	0.47	0.57*	0.60*	90	0.31	0.25	0.28	0.37	0.29	0.17
Protestant Europe	736	0.46	0.63*	0.57*	0.61*	387	0.36	0.45*	0.31	0.23	0.31	0.31
West South Asia	166	0.44	0.61*	0.57*	0.53*	21	—	**0.77**	0.22	0.57	0.07	0.16
Religion												
Buddhist	189	0.33	0.64*	0.58*	0.55*	69	0.40	0.48	0.10	0.25	0.19	**0.50**
Christian	1969	0.50	**0.73***	0.55*	**0.73***	1080	0.29	0.56*	**0.34***	**0.49***	**0.36***	0.34*
Hindu	201	0.75	0.65*	**0.60***	0.58*	109	0.46	0.63*	**0.34**	0.41	0.30	0.38
Jewish	204	0.50	0.66*	**0.60***	0.60*	144	0.45	**0.64***	0.29	0.43*	0.29	0.33
Muslim	319	0.36	0.63*	0.59*	0.72*	89	0.33	0.42	0.16	0.29	0.14	0.31
Spritual	88	0.48	0.61*	**0.60***	0.72*	13	—	0.35	-0.16	0.15	0.20	—

Table 2: **Positionality of NLP datasets and models** quantified using Pearson's r correlation coefficients. # denotes the number of annotations associated with a demographic group. α denotes Krippendorff's alpha of a demographic group for a task. * denotes statistical significance ($p < 2.04e-05$ after Bonferroni correction). For each dataset or model, we denote the minimum and maximum Pearson's r value for in demographic category in red (X) and blue (X) respectively.

Figure 1

datasets.

3 Experiments

3.1 Datasets

I plan to use the following datasets for my work: NLPositionality dataset (Santy et al., 2023): This dataset is publicly available and contains preprocessed data for two NLP tasks: social acceptability and hate speech detection. The dataset also contains information about the positionality of the annotators. For each task, we have two versions of datasets: raw: demographic information unprocessed. processed: demographic information processed in the same way as in the NLPositionality paper and website.

3.2 Implementation

The reimplemented code was uploaded in the below repository https://github.com/sumanthnani10/nlp-final-project
We referred the below code from the original paper authors git repository https://github.com/liang-jenny/NLPositionality

3.3 Results

The results for the two tables are given in the previous page. Table 1: The results of our Execution was presented in the table 1
Table 2: The results presented in the paper published in ACL 2023 are given in table 2 The values in the both tables are not same. There is a little difference in the values.

3.4 Discussion

The hardware issue is a common problem when training large language models. These models require a lot of computational resources, and even with the best hardware, it can take a long time to train them. In this case, We took time to troubleshoot the hardware issue and identify the missing libraries. This is a common process when working with large language models, and it is important to be patient and persistent when troubleshooting hardware issues.

The difference between the generated results and the published ones is likely due to the fact that We fine-tuned the model with different parameters. Fine-tuning a model involves adjusting the hyperparameters of the model to improve its performance on a specific task. .

3.5 Resources

The cost of reproducing the results of the "NLPositionality: Characterizing Design Biases of Datasets and Models" paper in terms of resources is as follows: Computation: The NLPositionality framework requires a significant amount of computational resources to train the BERT model.As of now We used Google collaboratory But to reduce the computational time we planned to use ORC for the 2nd checkpoint. Time:The time spent on searching/researching for the project is 6 hours and the time taken to troubleshoot and fine-tune the model is 9 hours. In total it took around 2 hours for the model with different hyper parameter combinations to get the best result. The NLPositionality framework is not as time-consuming to run.It takes reasonable time to run the framework on the NLPositionality dataset using a single GPU. People: The 3 of us in the group have contributed to the project in different aspects like researching for the paper, reproducing the results, reading the paper, creating the tables for results and developing a report. Development effort: The NLPositionality framework is relatively complex to implement. We worked hard on improvising the libraries used for model, finding better hyperparamets which produce the maximum result and studied on different related works. Communication with the authors: We have not yet communicated with the authors of the NLPositionality paper. However, I may need to do so if I have any questions about the framework or the dataset.

3.6 Error Analysis

Here are two examples of error analysis that we ran on the model:
Example 1: Failed instance: The model predicts that a tweet is offensive, when in fact it is not. Analysis: The model may have made this mistake because it was trained on a dataset that contains a lot of offensive tweets. As a result, the model may have learned to associate certain words or phrases with offensiveness, even when they are used in a non-offensive way.
Example 2: Failed instance: The model predicts that a sentiment analysis task is a positive sentiment, when in fact it is a negative sentiment. Analysis: The model may have made this mistake because it was trained on a dataset that is not representative of the real world. For example, the dataset may contain more positive tweets than neg-

ative tweets. As a result, the model may have learned to be biased towards positive sentiment

4 Conclusion

In conclusion, this paper has proposed a new framework for characterizing the design biases of datasets and models. The framework is comprehensive and robust, and it considers a wider range of dimensions of positionality than previous work. The authors have also made the code and data used to train and evaluate the model publicly available, and they have clearly described the methodology used in the paper. The computational resources required to train and evaluate the model are also reasonable.

This paper is reproducible. However, the reproducibility of the paper could be improved by performing hyperparameter optimization which we planned to do it in checkpoint 2 .

Overall, this paper makes a significant contribution to the field of NLP by addressing the important issue of design bias in datasets and models. The proposed framework is a valuable tool for developing more fair and inclusive NLP systems.

References

https://aclanthology.org/2023.acl-long.505/
https://github.com/liang-jenny/NLPositionality
https://aclanthology.org/2021.emnlp-main.155/
https://arxiv.org/abs/1908.09635
https://aclanthology.org/2022.findings-naacl.130/
https://arxiv.org/abs/2207.07068
https://github.com/sumanthnani10/nlp-final-project