# Fairness in Natural Language Processing(Proposal)

Yewei Song

University of Birmingham, B17 0DA, UK
yxs987@student.bham.ac.uk

**Abstract.** This is a proposal about fairness in Natural Language Processing(NLP). This project focuses on the fairness of NLP technology. The author proposes to analyze and solve it from four levels: data, algorithm, evaluation, ethics and law. At the same time, based on his own research experience, he proposed to focus on the bias in automatic scoring.

**Keywords:** Fairness · Natural Language Processing · Bias Evaluate

## 1 Research Qusetion

In the process of using Natural Language Processing(NLP) tools, many fairness problems have arisen. As a technology that uses a lot of machine learning(ML), NLP usually aims to minimize training errors. When it involves more sensitive features such as gender or ethnicity, it often introduces discriminatory behavior.The bias of machine learning could brings disaster to some people. For example, machine learning bias in judicial decision-making will bring unfair judgments to the parties[2].

The author proposes the following typical problem scenarios in NLP bias.

*A. Inequality of dialect users and standard tone users when using automatic speech recognition technology.*

*B. The problem of accent bias in the automatic scoring system for standardized language proficiency tests.*

*C. Automatic translation system lacks understanding of special expressions of minority groups.*

*D. Does the recognition and filtering of comment content violate the freedom of speech?*

However, most of fairness can be formalize into the following categories : Demographic Parity, Equalized Odds, Predictive Rate Parity, Individual Fairness, Counterfactual fairness et al.[9]

## 2 Research Aim

In the case of unfairness prevalent in the use of NLP technology, the purpose of this project is to find out how to quantify this bias and how to eliminate it. Because there are so many types of NLP systems, system biases also exist in various groups. Consequently, based on personal experience and interests, the author chooses to pay attention to various biases in the NLP automatic scoring system, and extends to the two technical directions such as automatic speech recognition and information extraction. The results of the research will be applied to promote the fairness of NLP technology, provide feasible model optimization methods and bias assessment methods, and provide ethical and legal opinions on the fairness of NLP technology.

## 3 Review of the literature

### 3.1 What is fairness

Computer software is increasingly involved in tasks like translating or resume screening. But what if that seemingly neutral algorithm was unwittingly built with human bias baked in? Typical unfair is similar to stereotypes about gender and occupation, or results bias against different ethnicities and groups. The unfairness of this algorithm will lead to discriminatory behavior of the system, and then cause the injustice of social activities, especially when the algorithm is widely used in various fields.

One view is that by eliminating protected categories in the data, you can achieve fairness in model training[11]. But in fact, sensitive information can still be expressed through other characteristics related to it. For example, when we do not consider ethnicity, postal code and surname can also reflect ethnic characteristics.

However, when we remove sensitive information as much as possible, the reliability of the algorithm will cause great problems. For example, the loan is wrongly issued to applicants who do not meet the conditions. If the purpose of adopting the fairness criterion is to enhance or fairen the long-term well-being of all groups, then the Liu team's 2018 paper[13] showed that in certain scenarios, the fairness criterion actually violates this original intention.

The academic community is temporarily unable to reach a consensus on ”what is fairness”, so when we discuss fairness, we need to define it reasonably according to the scenario.

### 3.2 Fairness in NLP & ML

Fairness in machine learning is a hot topic nowadays. As one of the main fields of applied machine learning, NLP bears the brunt of fairness issues.

According to Dwork et al., discussing the connection of individual fairness and group fairness. Group fairness means that for individuals in two groups, A and non-A groups, the proportion of being classified into each category is the same. But the author points out that this method may cause the overall appearance to be fair but unfair to individuals. Individual fairness hopes to operate through a pairwise comparison. If $x_1$ and $x_2$ are very similar, then the results of the

classification decision need to be very similar.
Hardt et al. suppose equalized odds and equal opportunity on 2016[11]. Equalized odds refers to any grouping result made by a classifier, in which the protected features meet group fairness. A predictor $\widehat{Y}$ satisifies equalized odds with respect to a protected attribute A and outcome Y if $\widehat{Y}$ and A are independent conditional on Y. Which is equivalent to saying: $Pr\left\{\widehat{Y}=1|A=0,Y=y\right\} = Pr\left\{\widehat{Y}=1|A=1,Y=y\right\}, y \in \{0,1\}$. Equalized odds enforces that the accuracy is equally high in all demographics, punishing models that perform well only on the majority. So we often think of the outcome $Y=1$ as the "advantaged" outcome, the equal opportunity can be defined with: $Pr\left\{\widehat{Y}=1|A=0,Y=1\right\} = Pr\left\{\widehat{Y}=1|A=1,Y=1\right\}$.
Return to the topic about NLP, Mehrabi et al. investigated some common biases in machine learning, which covered many NLP applications[15]. For example, referring to gender discrimination in digestion, translation, and text generation[21]. May and her colleagues pointed out the human-like bias existing in the sentence embedding system. Blodgett and O'Connor reported some bias situation in social media African-American English[3]. Chang et al.'s presentation on EMNLP proposed to use Feature Invariant Learning and Adversarial Feature Learning to reduce the bias in image audio algorithms[6].

### 3.3 Data bias

Bias from data can be divided into two categories according to what the author reads. One is that the algorithm has learned the prejudice that humans have, and the other is the bias that is mined from the data. Caliskan et al.'s report in 2017 shows that if we build an intelligent system capable of understanding and generating language, then in the process, it will also obtain prejudiced historical and cultural associations from human corpus resources.[5] It is also mentioned in this article that the algorithm can mine the discrimination that is hidden in the corpus. The names of European Americans are significantly more easily associated with pleasant than unpleasant terms than African Americans.
Dixon and his colleagues found in their research using Wikipedia Talk that unbalanced data can lead to unexpected biases in text classifiers[8]. And Tatman et al.'s research proves that non-white accents still show a higher word error rate in automatic speech recognition technology[18]. Another noteworthy point is that, even with Wu language spoken by 70 million people, Siri does not provide ASR support. The neglect of the dialect also reflects the insufficient protection of the endangered language by the corpus, and also the lack of fairness of the NLP technology.

### 3.4 Model training

In the Sun and 2019 Literature Review, some methods were mentioned about how to reduce the gender bias of the NLP model. One type is the processing of the corpus, including Data Augmentation, Gender Tagging, Bias Fine-Tuning, Removing Gender Subspace, and Use Gender. -Neutral Word. The other is the optimization method in the algorithm, which mainly includes Constraining Predictions and Adversarial Learning[17]. Regarding relationship with Adversarial Learning and GAN, Xu et al. proposed FairGAN. One discriminator is trained to identify whether the sample is real or fake, and the other discriminator is trained to distinguish whether the resulting sample is from a protected group. By using these two discriminators for adversarial games, the generator can generate fair data with high practicality[19]. This idea suggests that we can consider adding similar constraints to model rules to promote fairness

### 3.5 Evaluating

Madnani and his colleagues developed RSMTool in 2017 to verify the fairness of the scoring system, and calculate the degree of bias based on the standardized mean difference between the real score and the machine score. This system can also be used for any machine learning task of numerical prediction[14]. In a 2016 article, Bolukbasi and his colleagues described some methods for calculating gender bias in Word Embedding. By computing the direction between representations of male and female word pairs from the Definitional List[4]. Gonen and Goldberg proposed the use of clustering algorithms to cluster word vectors in 2019 to discover gender biases contained in vocabulary[10].

### 3.6 Law & Ethics

Applying the principle of fairness to NLP technology is not only a subject of theoretical research, but also a response of enterprises to public expectations. Leins et al. raised some discussions on data ethics regarding automatic sentencing, and also pointed out the issue of double use[12]. In 2012, Davis mentioned in his book Ethics of Big Data that the abuse of data analysis technology may bring the creepy results[7]. The researchers found that the fairness of algorithms is largely related to privacy and other rights. In recent years, social media and search engines have widely used natural language processing technology, but some of its functions, such as content filtering, have been controversial. Whether this data processing technology affects freedom of speech is widely discussed[1]. Another cause of widespread concern is the legal risk caused by the deep mining of privacy by machine learning technology. In 2019, Price pointed out the catastrophic problems caused by data leakage in response to privacy issues in medical big data[16]. After the introduction of the EU GDPR Act and other privacy protection laws, how to legally use and collect data has become a huge challenge.

## 4 Research methods

Based on my understanding of the problem, I divided the main work to solve this topic into four levels:

**Data:** The goal is to design an evaluation standard that covers all stages of data collection. Including collection goals, collection process, and annotating. Our goals include removing sensitive information as much as possible[20], a more comprehensive and extensive data source, strengthening data for minority groups, and filtering pollution in samples.
**Algorithm:** Study the impact of different interventions on the fairness of the NLP system. Here we mainly study two methods, one is the influence of regular constraints on the fairness of the model in the training process, and the other is the effect of adjusting the training goals of the model on the fairness.
**Evaluate:** Investigate whether the existing methods for evaluating machine learning bias are applicable to NLP. If possible, try to propose an effective application and bias assessment method in the field of NLP. Initial ideas include recognition experiments, translation experiments, etc.
**Ethics and law:** An unavoidable part is the ethical and legal constraints of data collection. Discuss which data can be collected and which cannot. Also discuss the legal and moral dilemmas faced by NLP technology.
With the research target, I choose the NLP automated scoring system as the experiment objection. I plan to split the research into corresponding 4 directions:
The first part is to research how the training data affects the fairness of the NLP system.
The second part, investigate how to improve the fairness of NLP algorithms.
The third part, compare evaluate method in NLP, and make a common solution in model feedback evaluation.
The forth part, discuss ethics and law problem in NLP technology about fairness.

### 4.1 Part A: Pre-processing(Data)

A well-designed experiment is the best way to evaluate the bias of the data collecting against the NLP system. A feasible experiment can be briefly described as the following steps: choose the most popular ASR tool in the industry, collect different dialects and standard pronunciation speech in the same language, and train models with the same dialect in different proportions. Study the influence of dialect proportion and dialect recognition rate in training samples. It also studies the difference between dialects with larger deviations from standard sounds and smaller one on this issue. This experiment will demonstrate the effect of training data on the fairness of the automatic speech scoring system.

### 4.2 Part B: Algorithm processing

An experiment on the effect of using different Stemming or Lemmatisation strategies on bias. Experimental research on information extraction model training bias. For example, does the conversion of personal pronouns (with gender or personal characteristics) into neutral pronouns reduce the gender bias of the algorithm? At the same time, a feasible evaluation algorithm needs to be proposed to evaluate the NLP bias in different groups.

### 4.3 Part C: Evaluate and fixing

Design a comparative experiment, apply all kinds of methods proposed by the industry to evaluate the fairness of machine learning to NLP, and propose my improved evaluation method to join the comparison. A typical evaluable scenario is the evaluation of bias in English translation models, such as models that include gender bias or lack adaptability to the dialect. Compare the differences between various evaluation methods and discuss whether the evaluation results can be fed back to the model training process.

### 4.4 Part D: Ethics discussion

The problem is how to get information about the sensitive group during training time. There is a recent work along this line and optimizes accuracy parity using techniques from robust statistics without knowing the sensitive attribute. Also the ethics and law problem of data bias in NLP is a big topic to discussion, maybe a qualitative analysis is well-worthy research topic during whole research on fairness of NLP.

### 4.5 Product and contribution

On the whole, the research sub-project includes four parts, and each sub-part has the possibility of thesis output. Including how to deal with the NLP corpus to improve fairness, the relationship between the NLP model training process and fairness, a feedback method for evaluating the fairness of the NLP model, legal and ethical NLP technical dilemma demonstration. The integrated thesis can be used as a doctoral thesis for this topic.

From the perspective of personal contribution to the project, in-depth research in some areas can contribute to the project's measurable output, such as code, algorithm, and data. At the same time, personal interest in linguistics can also expand the language scope of the project.

## 5 Resources

For the experiment I supposed to do, the resource of data and calculate power is needed. The data collection is based on two sources, one is an open shared database, and the other is the Internet original data collection work. Voice data is mainly based on the former, and text data can make more use of public sources such as Twitter. Computing power requires some support from the supervisor, especially the high-performance server resources required for training models.

Also the whole project is a cooperative project, there are many of subdivision areas of topic. My personal knowledge area is focused on ASR and word vectors, and it does not cover all areas of NLP completely, and this topic requires a broad understanding and understanding of the entire field. At the same time, this subject involves a lot of ethical and legal knowledge, covering various subfields of social science and cognitive science.

Good academic training is very necessary, including research methods, cutting-edge fields of NLP, data analysis and evaluation methods.

## 6 Personal preparation

The two NLP research topic in my master degree, one is about ASR scoring system, another is text information extraction. Both are faced with fairness problems. In ASR scoring project, the bias in OpenSLR model cause a big unfair about accent. Also in the information extraction, with the steps about Named-entity recognition, the very poor performance of loanwords such as place names and person names make a big unfairness to non-native English users.
In these two projects, thanks to the guidance of Professor P.J. Hancox, I also learned how to conduct academic research in the field of NLP.

## References

1. Marvin Ammori. The new new york times: Free speech lawyering in the age of google and twitter. *Harv. L. Rev.*, 127:2259, 2013.
2. Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.
3. Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. *arXiv preprint arXiv:1707.00061*, 2017.
4. Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc., 2016.
5. Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.
6. Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, 2019.
7. Kord Davis. *Ethics of Big Data: Balancing risk and innovation.* " O'Reilly Media, Inc.", 2012.
8. Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 67–73, New York, NY, USA, 2018. Association for Computing Machinery.
9. Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.
10. Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *arXiv preprint arXiv:1903.03862*, 2019.

11. Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.
12. Kobi Leins, Jey Han Lau, and Timothy Baldwin. Give me convenience and give her death: Who should decide what uses of nlp are appropriate, and on what basis? *arXiv preprint arXiv:2005.13213*, 2020.
13. Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.
14. Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. Building better open-source tools to support fairness in automated scoring. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 41–52, 2017.
15. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019.
16. W Nicholson Price and I Glenn Cohen. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43, 2019.
17. Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.
18. Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *INTERSPEECH*, pages 934–938, 2017.
19. Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018.
20. Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.
21. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*, 2018.