Investigating Raciolinguistic Biases in Natural Language Processing Systems and their Perpetuation of Social Hierarchies

Benjamin Mckitterick
Computing & Communications
Lancaster University
Lancaster, UK
b.mckitterick@lancaster.ac.uk

Andreea Burcin
Computing & Communications
Lancaster University
Lancaster, UK
a.burcin@lancaster.ac.uk

Ziling Xie
Computing & Communications
Lancaster University
Lancaster, UK
z.xie3@lancaster.ac.uk

Mei Lu
Computing & Communications
Lancaster University
Lancaster, UK
lum8@lancaster.ac.uk

***Abstract* —** We investigate how raciolinguistic biases that exist within natural language processing systems affect marginalized communities, specifically focusing on the population of African-American English speakers. The paper aims to provide insight into whether these systems are maintaining social hierarchies by reproducing language ideologies. Critical areas of existing research are highlighted, and alternate ways of investigating how bias in said systems can be understood and reduced are discussed. Most current work has been done from the perspective of a technologist; instead, this paper adopts a sociolinguistic approach, recognizing greater meaning in the relationship of language and social hierarchies. A clear set of objectives and hypotheses are identified, following a small controlled experiment to gather measurable evidence to support or disprove these. Surveys, questionnaires and semi-structured interviews will be conducted to gain a new perspective into if members of the said community face emotional difficulties with such technology. We hope this study will shed light on this area and make room for future work to be done further to mitigate this bias in natural language processing systems.

I. Introduction

As technological development increases exponentially [1] and societal domestication becomes widespread, thought must be given towards how such advances impact people. The amount of data generated globally is rapidly increasing [2], and a widely utilized and effective tool used to control this influx of information is artificial intelligence: a system capable of performing tasks that require human intelligence and discernment [3]. Natural language processing (NLP) technology is a branch of artificial intelligence associated with human-computer interaction through natural language; particularly, how to program machines to read, decipher, and comprehend human language [4].

Regarding NLP, the subject of language can take on a social meaning; it can be utilized to accentuate beliefs and stereotypes of social groups, reinforcing social inequalities. Group labels identify a category of people, thus, conveying categorical boundaries and a position within a hierarchical taxonomy [5]. Social hierarchies refer to the ranking of members within social groups, based on the dominance, power, or influence they exhibit [6]. Studies have shown that NLPs can be insensitive to differences in dialect and suppressing already marginalized voices [7]. Since language and social hierarchies are deeply intertwined [5], many groups have sought to bring about social change through language changes, disrupting these patterns of oppression and marginalization [8]. Understanding the role of language in social hierarchical maintenance is critical towards analyzing bias in NLP systems.

As NLP technologies are mostly reliant on human-generated data, the bias that exists within individuals is transferred to the systems and in some cases, even amplified [9]. Bias can originate from various sources and can be of different types [10, 11], although no matter the source, all consequently have discriminatory assessment built on demographic features. The term bias in the context of a system refers to computer systems which generate results that differ for certain people in comparison to others [12] Recent work that explores language bias exposes

proof of systematic racial bias when the tested classifiers predicted the tweets containing elements of African American English (AAE) as more offensive compared to the tweets that did not have those elements [13]. To better understand raciolinguistic biases [5] we plan on investigating the foundations of the matter, starting with the relations between technology and racialized communities[12].

A large quantity of work already exists that attempts to eliminate or reduce bias in NLP systems: handling race-talk [14], improving training data [13, 7], introducing new data sets [15, 16], utilizing distinguishable feature engineering techniques and training classifiers with pre-trained language models [17], introducing data statement schemata and standards for documentation and publication [18]. Most existing work has been from the perspective of a technologist; in that, they criticize the design process and data sets. In contrast, we attempt to address the root of the issue by adopting a more sociolinguistic approach. Has enough assessment been done into the adverse impacts of how these systems affect sub-populations? Could they be maintaining systematic social hierarchies?

A brief overview of the subsequent sections of this proposal are as follows: research questions will be identified, and clear, measurable objectives will be set within section II. Section III will describe and justify the design approach and methodologies used to gain an empirical characterization of NLPs supposed maintenance of social hierarchies. In section IV, we formalize a feasible work plan with achievable milestones and a detailed description of how to implement them.

II. Objectives

The purpose of this research proposal is to investigate how bias in NLP systems that are reproduced through linguistic ideologies; a set of beliefs about language as used within their social contexts [19], impact marginalized communities and if they contribute to the preservation of social hierarchies. Whorf says that "language is not merely the mirror of our society it is the major force in constructing what we perceive as reality" [20], language can be used as a tool for control and communication, fabricating linguistic forms that are implicit in the manifestation of a societal scale [21]. For feasibility and pragmatism, this study will focus on the population of AAE; a dialect of Standard-American English (SAE), due to its widespread use, well-established history in the sociolinguistic literature, and demographic associations [22]. Thus, the research questions to be answered are as follows:

- Are members of the AAE community made to adhere to existing linguistic, ideological assumptions embedded in NLP systems?
- If so, what emotional costs will be elicited by the AAE community if NLP systems uphold these linguistic ideologies?

The research objectives of this study are thus to provide a response to the aforementioned questions:

- Gain empirical quantitative data that evidence whether an NLP system upholds linguistic, ideological assumptions that cause members of the AAE community to adapt their language practices.
- Collect both qualitative and quantitative data that identifies the emotional issues that AAE communities face when using NLP systems.

We will use the following hypotheses to guide our answers to the research questions:

- H1 - NLP systems will cause members of the AAE community to adapt their linguistic ideologies.
- H2 – NLP systems upholding linguistic ideologies will elicit a negative emotional response from the AAE community.

Our study will contribute to the subject field by primarily focusing on delivering valuable apprehension over how discriminative referred technologies can be towards AAE speaking communities. Subsequently, we hope to highlight how getting input from marginalized groups can often be understudied but of great importance on the future development of fair systems. Furthermore, we intend on gathering empirical data to evidence whether NLP systems maintain social hierarchies and if AAE communities are negatively affected by this, more specifically, discovering in what ways members of the AAE community must adapt their linguistical behavior to cater to these technologies and how their emotions are affected by this.

The standards of today's society require a positive and openminded view towards different social backgrounds. Not allowing discriminative attitudes to be built into the technology that people use is an unquestionable fact. In recent years, much research has been done around AI bias that focuses on data provenance [23, 24], and techniques [25] to mitigate this fairness concern. Unfortunately, there is not enough research that delves further into the consequences of this issue, and there is no empirical evidence showing whether said systems maintain social hierarchies. As Blodgett et al., highlight in their analysis [22], there is a need for studies that go beyond machine learning and instead detail the links among language and social hierarchy. It has been proven that even algorithms from a mathematical perspective would be viewed as fair, are certainly not seen this way by general society as they do not fit into social beliefs of equality [26]. We plan to connect with the communities that are directly impacted by these algorithmic biases.

III. Methodology

We plan to conduct our experiment online through Amazon Mechanical Turk (AMT), giving us the capabilities to acquire a global accessible population alongside consent approval of each participant through AMT. The population we aim to generalize to are individuals that communicate using AAE dialect. Our sample frame size would ideally be 50 participants with a high acceptance rate on AMT; any collected data containing irrelevant responses would be discarded, i.e., participants that provide short answers to open-ended questions. Resultingly, this would give us a fair randomized sample selection that justifiably represents our target population. Depending on the timeframe and feasibility, we would conduct repeated trails to gain further evidence; accentuating the studies external validity and allowing us to defend better against criticisms of generalization.

To study the effects that NLP systems have on the AAE community we run a small controlled experiment on AMT to limit confounding factors; due to the ongoing pandemic, we are limited to controlled experimentation online; thus, we view this as a limitation of our study. The experiment will begin with participants being shown a short introduction to the study. Next, existing AAE tweets that have been labelled as offensive by NLP systems trained on two corpora of tweets widely used in hate speech detection [7] are shown to the participants. Afterwards, a short online survey containing a set of Likert scales on emotional semantic adjectives and a non-verbal pictorial assessment technique; the self-assessment manikin [27], will be used to collect ordinal data. Then, an online questionnaire will be handed to the participants; the order of questionnaire and the emotional survey will be counterbalanced to minimize effects of carry-over. The questionnaire will contain both open-ended questions, i.e., "Do NLP systems impact you negatively? If so, in what ways?" and dichotomous questions, i.e., "Would you change your language practices to avoid negative perception online?" Finally, a semi-structured interview would be conducted to gather further data based on participant views. This way, we can accurately and thoroughly collect information from the AAE communities, while controlling the question order – see the appendix for an example of our interview structure. The said experimental procedures would allow us to collect both quantitative data and qualitative data for analysis.

Before analyzing the data, the collected quantitative data will be inspected for any missing points or outliers. Aggregated mean responses for each metric from the emotional survey will be calculated, and correlative Spearman Rho p-values will be plotted in a correlation matrix. T-tests will be used to infer the probability of the difference, by comparing the significance between two difference averages of AAE users. More information on data analysis will be covered in the implementation section.

For the qualitative data, the interviews that will be conducted with the participants will be transcribed. The results will then be explored through thematic analysis to detect common topics that repeatedly occur within the data. In this way, we will first familiarize ourselves with the data, identify initial codes, and then search for frequent themes based on the these. After defining and naming those themes, the results will be accessible for discussion. By categorizing the data collected from the interviews with this method, it will allow us to determine whether the emotional response of the AAE community is a negative one.

IV. Implementation

In figure 1, the overall work plan and research activities of this project are shown. To manage the

workload, we have divided the project into five phases: the research proposal, preparation and design, data collection, data analysis, and evaluation & conclusion. The phases will be as follows:

PHASE	START	END	DURATION (weeks)
Research Proposal	20/11/2020	11/12/2020	4
Preperation & Design	11/01/2021	17/02/2021	1
Data Collection	18/01/2021	31/01/2021	2
Data Analysis	01/02/2021	28/02/2021	4
Evaluation & Conclusion	01/03/2021	21/03/2021	3

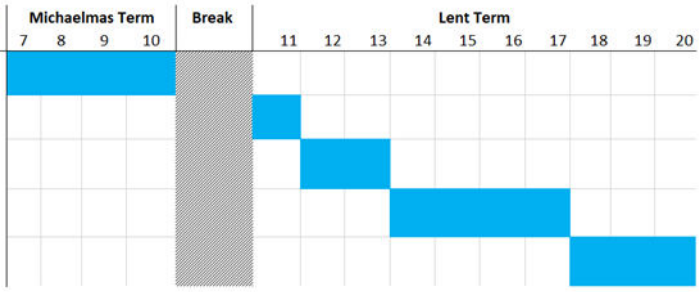

Figure 1 – Gantt Chart to illustrate overall work plan for research project and time periods for each project phase.

The preparation and design phase will entail formalizing the emotional semantic and SAM survey categories, along with finalizing the questionnaire and interview structure. The surveys and questionnaires will be constructed through Qualtrics. Due to our lack of knowledge in the field, discussions with sociolinguistic experts and further research into the state-of-the-art would be conducted.

For data collection, we will gather our participants through AMT and provide them with links to the Qualtrics survey. Only participants with a Human Intelligence Task (HIT) score above 98% will be accepted. We aim to have the survey length no longer than 5 minutes while interviews will be 20 to 30 minutes and conducted through Microsoft Teams. The set times are necessary to eliminate any confounds that would be introduced if participants began to lose interest. Once all data has been collected, we will download the data from Qualtrics and convert it to .csv format for the data analysis stage.

Python libraries such as scipy, numpy, pandas and sklearn along with SPSS software, will be utilized to statistically analyze our data set. The Spearman rank-order correlation coefficient would provide us with the means to analyze our gathered nonparametric metrics, the strength of the direction of association between every possible combination of metrics would then be plotted in a correlation matrix. Then we can then study the relationship of all possible combinations of emotional adjective rankings, allowing us to group positive and negative emotions, thus, helping identify the direction of emotional response from AAE communities. This will provide us with relevant data to either accept or reject our hypothesis H2. To test our hypothesis H1, we will use a one-sample t-test to check for statistical significance between the average number of users who adapted their linguistic ideologies and the number that did not.

The evaluation and conclusions would represent the final phase of our project, in which we will be able to determine whether our results fit within the expected outcome and if our research succeeded in bringing a valuable contribution to the field of study. Here, we would extrapolate meaning from our analyzed data, explain what our results mean, and reflect on our findings. At this stage, we will discuss and evaluate any limitations that we may encounter while conducting research, and to establish if our results can be generalized to a larger population.

V. References

[1] R. R. Schaller, "Moore's law: past, present and future," in *IEEE spectrum*, 1997.

[2] W. A. Qader, M. M. Ameen and B. I. Ahmed, "Big Data Characteristics, Architecture, Technologies and Applications," Journal of Computer Science, Erbil, 2020.

[3] "artificial intelligence," Oxford Reference, [Online]. Available: https://www.oxfordreference.com/view/10.1093/oi/authority.20110803095426960.. [Accessed 26 November 2020].

[4] K. R. Chowdhary, "Natural language processing: Fundamentals of Artificial Intelligence," Springer, New Delhi, 2020.

[5] S. Blodgett, "Language (Technology) is Power: A Critical Survey of" Bias" in NLP.," arXiv, 2020.

[6] H. X. &. I. R. O. Jessica E. Koski, "Understanding social hierarchies: The neural and psychological foundations of status perception," Routledge Taylor & Francis Group, 2015.

[7] M. Sap, D. Card, S. Gabriel, Y. Choi and N. Smith, "The risk of racial bias in hate speech detection.," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[8] J. Hill, The everyday language of white racism, John Wiley & Sons, 2009.

[9] E. F. Ntoutsi, Bias in data-driven artificial intelligence systems - An introductory survey., Wiley Interdisciplinary Reviews, 2020.

[10] S. Jia, T. Meng, J. Zhao and K.-W. Chang, "Mitigating Gender Bias Amplification in Distribution by Posterior Regularization," in *Association of Computational Linguistics*, 2020.

[11] L. K, Bias amplification in artificial intelligence systems, arXiv, 2018.

[12] B. Friedman and H. Nissenbaum, "Bias in computer systems," ACM Transaction on Information Systems, 1996.

[13] T. Davidson, D. Bhattacharya and I. Weber., "Racial bias in hate speech and abusive language detection datasets.," arXiv, 2019.

[14] A. Schlesinger, K. P. O'Hara and S. T. Alex, "Let's Talk About Race: Identity, Chatbots, and AI," CHI, 2018.

[15] A. Asudeh, Z. Jin and H. V. Jagadish, "Assessing and remedying coverage for a given dataset," in *IEEE 35th International Conference on Data Engineering*, Michigan, 2019.

[16] M. Wiegand, J. Ruppenhofer and T. Kleinbauer, "Detection of abusive language: the problem of biased datasets," in *NAACL-HLT* , 2019.

[17] M. Mozafari, R. FarahBakhsh and N. Crespi, "Hate speech detection and racial bias mitigation in social media based on BERT model," PLOS ONE, 2020.

[18] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," Association for Computational Linguistics, 2018.

[19] B. Schieffelin, K. Woolard and P. Kroskrity, " Language ideologies: Practice and theory," Oxford University Press, 1998.

[20] B. L. Whorf, The Relation of Habitual Thought and Behaviour to Language, London, 1956.

[21] S. Resta, "Words and social change. The impact of power and ideology on the language of Economics and Law," OpenEdition Journals, 1998.

[22] S. L. Blodgett, L. Green and B. O'Connor, "Demographic Dialectal Variation in Social Media: A Case Study of African-American English," arXiv, 2016.

[23] G. Zhang, B. Bai, J. Liang, K. Bai, S. Chang, M. Yu, C. Zhu and T. Zhao, "Selection Bias Explorations and Debias Methods for Natural Language Sentence Matching Datasets," arXiv, 2019.

[24] J. Huang, H. Oosterhuis, M. d. Rijke and H. V. Hoof, "Keeping Dataset Biases out of the Simulation: A Debiased Simulator for Reinforcement Learning based Recommender Systems," ACM, 2020.

[25] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez and C. Xiong, "Double-Hard Debias: Tailoring Word Embeddings for Gender Bias Mitigation," arXiv, 2020.

[26] M. K. Lee and S. Baykal, "Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division," in *ACM conference on computer supported cooperative work and social computing*, 2017.

[27] M. M. Bradley and P. J. Lang, "Measuring Emotion: The Self-Assessment Manikin and The Semantic Differential," Journal of behaviour therapy and experimental psychiatry, 1994.

Appendix?