Hamna Mustafa

Professor Karen Mazidi

Human Language Technologies

13 November 2022

Summary of “A Survey of Race, Racism, and Anti-Racism in NLP”

Authors	Affiliations	Citations
Anjalie Field	Carnegie Mellon University	352
Su Lin Blodgett	Microsoft Research	1063
Zeerak Waseem	University of Sheffield	-
Yulia Tsvetkov	University of Washington	4531

The problem that this paper focuses on is the race-related bias that exists in all stages of NLP model development and is encoded in NLP systems. Natural Language Processing is rooted in language and linguistics, and despite the inextricable ties between race and language, there has been very little work on the effects of those ties on NLP. The paper talks about the need of proactive consideration about how NLP systems uphold racial hierarchies and how the voices of historically marginalized people are barely present in NLP literature. The paper brings this issue to light by surveying 79 papers from the ACL anthology and calls for inclusion and racial justice in future NLP research.

The prior work that this paper considers is 79 ACL papers that contain the terms ‘*racial*’, ‘*racism*’, or ‘*race*’. The team that worked on this research first found 165 papers that contained the terms mentioned above but after removing the papers that didn’t actually engage on the topic, their final analysis consisted of 79 papers. In analyzing these papers, the team found that racial bias exists in all parts of the NLP pipeline: data, data labels, models, model outputs, and social analyses of outputs. The ACL papers showed that NLP systems, especially word embeddings and language models, can absorb and amplify social biases in data sets. Not only can model biases occur from raw data, but several of the ACL papers also identified biases in the way researchers categorize or obtain data annotations and labels. Furthermore, the papers also found evidence that the racial biases of outputs produced by a model can be changed by model instances or architectures. Additionally, they discovered that NLP systems could amplify bias if some model outcomes were deployed. For example, classifiers for abusive language are more likely to label text containing identity terms like ‘black’ as offensive, which propagates racial bias. Lastly, the papers deduced that racial bias also existed in the interpretation of results produces by analysis models.

The authors of this article noted that although there has been a lot of research on gender bias concerns in NLP, the presence of racial bias in NLP is barely looked it. Thus, this article made a unique contribution by bringing this issue to light and proving through their research that it is a very real and serious problem. Not only that, but the writers also uniquely identified that the way that racism is examined is limited, has marginalized research, and requires deeper evaluation.

Since the authors based this work on the analysis of previous ACL papers, they evaluated their work by finding evidence of the problem within those papers. It was a qualitative analysis, not quantitative so there are no quantifiable results or evaluations, aside from a table which organizes their analysis of the ACL papers.

Anjali Field is the first author listed in the article. She is a PhD student at Carnegie Mellon University with 352 citations on Google Scholar. Field focuses on social-oriented natural processing. Her work involves developing machine learning models to examine social issues like propaganda, stereotypes, and prejudice in complex real-world data sets, as well as exploring their amplification and ethical impacts in AI systems. Her work, including this article, is very important because of the severity of existing social issues and their presence in AI.

Su Lin, the second author of this article, is a senior researcher in the Fairness, Accountability, Transparency, and Ethics in AI group at Microsoft Research. Lin has 1063 citations on Google Scholar and focuses on the social and ethical implications of NLP technologies. Like Field, Lin's work is also very important because it deals with the social aspects of NLP, how these issues are embedded in NLP systems, and develops approaches to mitigate these issues.

Zeerak Waseem is the third author of this article and works as an academic researcher from University of Sheffield. Waseem did not have any citations on Google Scholar, but their research is based on voice activity detection, test cases, language identification, and prejudice. Waseem has also done research on racism, including this paper. Waseem's work is important because it covers a variety of important topics, some of which are crucial for the advancement of the AI field while others are crucial to create awareness about racial issues in NLP.

Yulia Tsvetkov, the final author of this paper and the advisor of Anjalie Field, has the most citations on Google Scholar, a total of 4531 citations. Tsvetkov is an assistant professor in the Paul G. Allen School of Computer Science & Engineering, at the University of Washington. She focuses her research on the intersection of machine learning and theoretical or social linguistics. The goal of her research is to extend the capabilities of human language technology beyond individual populations and across language boundaries, thereby enabling NLP for diverse and disadvantaged users, the users that need it most. Her research is very important because it is based on progressing NLP so that it’s demographic can be expanded. Research like this lay the foundation of the future of NLP.

References

https://aclanthology.org/2021.acl-long.149.pdf

https://www.clsp.jhu.edu/events/anjalie-field-carnegie-mellon-university/#.Y3LBSC1h3s1

https://sblodgett.github.io

https://typeset.io/authors/zeerak-waseem-pflh95ppt4

https://homes.cs.washington.edu/~yuliats/