UNIVERSITY OF COPENHAGEN
DEPARTMENT OF COMPUTER SCIENCE

Ph.D. Thesis

Anna Katrine van Zee

NLP Across Social Groups

Advisor: Professor Anders Søgaard, PhD, dr.phil.

This thesis has been submitted to the PhD School of The Faculty of Science, University of Copenhagen on February 8, 2024.

Abstract

The widespread adoption of Natural Language Processing (NLP) tools has become increasingly prevalent. From customer service, social media and news reporting to healthcare, education and virtual assistants. However, the uneven performance of these tools across different social groups raises significant concerns and amplifies the potential detrimental impact on society. This thesis delves into the investigation of this pertinent issue by systematically analyzing the performance of NLP tools across a spectrum of tasks, ranging from syntactic analysis to speech recognition, with a particular emphasis on their implications for various social groups.

Our exploration extends to a comprehensive examination of the concept of fairness within the context of NLP. We scrutinize the intricate relationship between bias and fairness, and illustrate how these are indeed separate phenomena. A critical aspect of our inquiry involves considering the Rawlsian fairness principle and its prevalence in the current research field of NLP, and we propose a nuanced discussion on alternative definitions of fairness, emphasizing the necessity of evolving conceptual frameworks to address the pressing issue of increasing performance disparities within the field of NLP. By doing so, we aim to contribute to the ongoing conversation about inclusive and equitable NLP technologies and their application across diverse social groups.

Resumé

Sprogteknologiske værktøjer bliver stadig mere udbredte i vores dagligdag. Fra kundeservice, sociale medier og nyhedsrapportering til sundhedssektoren, uddannelse og virtuelle assistenter. Disse værktøjer klarer sig dog til tider ujævnt, når de anvendes på tværs af forskellige sociale grupper. Dette er bekymrende og kan have en potentielt skadelig effekt på samfundet. Denne afhandling dykker ned i denne relevante problemstilling ved systematisk at analysere ydeevne af sprogteknologiske værktøjer på et bredt spektrum af opgaver, lige fra syntaktisk analyse til talegenkendelse, med særlig vægt på deres implikationer for forskellige sociale grupper.

Afhandlingen inkluderer derudover et omfattende studium af begrebet retfærdighed inden for moderne sprogteknologi. Vi gransker det komplekse forhold mellem forudindtagethed (bias) og retfærdighed (fairness) og illustrerer, hvordan disse rent faktisk er separate fænomener. En væsentlig del af denne afhandling involverer en diskussion af det Rawlsianske retfærdighedsprincip samt princippets udbredelse i forskning inden for sprogteknologi. Vi fremsætter desuden en nuanceret diskussion om alternative definitioner af retfærdighed, hvor vi lægger vægt på nødvendigheden af at udvikle konceptuelle rammer for at tackle det presserende problem med stigende forskelle i ydeevne inden for sprogteknologi. På denne måde sigter vi mod at bidrage til den igangværende diskurs om inkluderende og retfærdige sprogteknologier og deres anvendelse på tværs af forskellige sociale grupper.

Acknowledgments

Not only have I received years of free, excellent education in a safe society that expects women to participate, I have also been surrounded by remarkable people. I have had resourceful, helpful parents, and friends and family who have supported and encouraged me.
I hope my fortunate upbringing becomes the standard.

Anders: Thank you for your companionship.
My parents: Thank you for encouraging curiosity and conscientiousness.
Henrik: Thank you for taking the seriousness out of life.
Hans en Wilma: Thank you for teaching me that knowledge is lived.
Most importantly, Marc: Thank you for the freedom to create.

Månesol and Blå Anemone: May your lives be as fanciful as your dreams.

Publications

This thesis is structured around a series of articles. The content of these articles remains largely the same as in their initial publication, with only minor modifications such as typo corrections and adjustments to tables and figures for consistency.

1. Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of Studying and Processing Dialects in Social Media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China, July 2015. Association for Computational Linguistics.

2. Anna Jørgensen and Anders Søgaard. A Test Suite for Evaluating POS Taggers Across Varieties of English. In *25th International World Wide Web Conference*, pages 615–618, Montreal, Canada, April 2016. International World Wide Web Conferences Steering Committee.

3. Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of NAACL-HLT 2016*, pages 1115–1120, San Diego, CA, United States, June 2016. Association for Computational Linguistics.

4. Anna Jørgensen and Anders Søgaard. Evaluation of Summarization Systems across Gender, Age, and Race. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 51–56, Online and in the Dominican Republic, November 2021. Association for Computational Linguistics.

5. Anna Katrine van Zee, Marc van Zee, and Anders Søgaard. Group Fairness in Multilingual Speech Recognition Models. Manuscript under review, 2024.

6. Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the Independence of Association Bias and Empirical Fairness in Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 370–378, Chicago, IL, USA, June 2023. Association for Computing Machinery.

7. Anna Katrine Jørgensen and Anders Søgaard. Rawlsian AI fairness loopholes. *AI and Ethics*, 3(2):1185–1192, 2022.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The advancement of technology, especially in artificial intelligence and natural language processing (NLP), has seen exponential growth in recent years. However, like with many other technological advancements, there is a risk of unintentional biases and exclusions. This dissertation is motivated by the belief that technology should be for everyone.

NLP systems have become the backbone for numerous applications: from virtual assistants to chatbots, from content analyzers to real-time translators. But how do these systems fare when faced with diverse dialects, regional nuances, and non-standard pronunciation? Are they adept at comprehending the variation of language used in everyday communications?

The importance of equitable and inclusive NLP goes beyond just understanding language. It is about the inclusivity in the digital age. Discrepancies in the performance of NLP tools across different user groups not only hinder efficient communication but also perpetuate existing inequalities.

This dissertation delves deep into the intricacies of NLP performance and fairness, by identifying gaps and shortcomings in current systems and discussing what equitable and inclusive NLP is. It seeks to offer solutions that ensure these tools are just as adept at understanding a local dialect as they are the mainstream and resourceful languages.

If our advanced NLP systems consistently neglect linguistic diversity, we are not only denying speakers the full benefit of modern digital tools, but we are also perpetuating a sense of invisibility around their lived realities. The ramifications of this oversight extend beyond technology, touching the areas of culture, identity, and societal inclusivity.

Linguistic variation does not exist in isolation, but is dependent upon and

influenced by identity markers such as race, gender, socio-economic status, and more. As such, linguistic variation becomes a proxy for social identity. Linguistic variation is therefore not random but is situated in the cultural and social identities of the speaker and follows a repeatable pattern of rules or comme-il-fauts (Sankoff, 1980; Johnstone, 2004). For instance, an individual might speak a regional dialect flavored by other markers of their social identity such as gender, age, race and/or class. Think of a young woman's language use, now change her race, then her age and lastly her gender. The resulting language use will differ from the original language use of the young woman, exemplifying the influence of social identity on language use and variation.

Intersectionality, a term coined by Kimberlé Crenshaw (Crenshaw, 1989), provides a lens through which we can better understand the multifaceted nature of identities, their impact on language variation, and how social identities interplay in systems of power and discrimination. Intersectionality acknowledges the interconnected nature of social categorizations and other identity factors and emphasizes that individuals' experiences of discrimination and privilege are shaped by the intersection of multiple aspects of their identity. These intersections create unique and often complex forms of oppression or privilege that cannot be fully understood by examining individual identity markers in isolation. Intersectionality promotes a more inclusive understanding of social issues, advocating for an approach that considers the overlapping and interrelated systems of power and discrimination that individuals may face based on various aspects of their identity. Intersectionality provides us with a framework for investigating the AI-driven discrimination or inclusion of certain social groups according with respect to their language use.

Imagine a scenario where an NLP system is adept at understanding the mainstream dialect but falters when processing an ethnolect predominantly spoken by a marginalized community. Now, imagine that within this community, there are further layers of intersectional identities — women, LGBTQIA+ individuals, or economically disadvantaged groups, and that performance drops further when the speaker belongs to one of these, ie. racially and economically marginalized from the mainstream. The failure of the NLP system in this case does not just sideline a dialect or ethnolect; it perpetuates a cycle of invisibility and marginalization for these intersectional groups. They face a double or even triple bind, where their linguistic identity and other aspects of their personhood are not recognized or are misrepresented. When discussing performance disparities in NLP, it is therefore

important to consider these through an intersectional lens to fully grasp the implications of unfair performance.

Incorporating intersectionality into NLP development and research means recognizing that language is not just a means of communication; it is a reflection of complex, intertwined identities. It calls for a broader understanding of representativity in datasets and evaluation and for a heightened attention to the broader socio-cultural context of language use. This involves actively seeking out diverse data sources, understanding the historical and cultural contexts of language use by social groups, and continuously iterating on models to ensure they are both technologically sound and socially equitable. It means changing the objective of NLP from highest performance to equitable performance and to ensure that technology uplifts rather than oppresses the variety of social groups.

It is worth noting that exclusivity in NLP systems does not arise spontaneously. The systems mirror existing linguistic homogeneity in the datasets on which they are trained, and they inherit the architectural and developer biases which arise from the structure and design of algorithms that may inherently favor certain linguistic patterns and cultural preferences.

Fairness in NLP hinges on the idea that every user, regardless of linguistic background or identity, should experience consistent and unbiased treatment from NLP systems. As such, fairness dictates how technology interacts with diverse populations and the question of fairness in NLP is deeply ethical and societal. An NLP system that misinterprets or misrepresents certain social groups does not just provide skewed information; it can inadvertently sideline voices and marginalize communities.

The objective of this dissertation is two-layered. While we delve deep into the technological challenges and solutions for NLP, we equally probe the societal ramifications of these technologies. Together, the chapters in this dissertation interweave to form a narrative that is as much about the pursuit of technological excellence as it is about fostering a society where diversity in language use is supported and encouraged.

1.1 Contributions

The chapters in this dissertation reflect the two-pronged nature of our exploration. We systematically address the performance disparity across NLP tasks, focusing on intersectional groups in particular, while concurrently un-

dertaking a thorough discussion of fairness, thereby contributing to a comprehensive understanding of the intricate relationship between technological advancements and the imperative for societal equity. Our main contributions are listed below, and each point corresponds to a chapter:

1. We perform a large-scale dialectal study of African-American Vernacular English (AAVE) on social media and show how both newswire- and Twitter-adapted state-of-the-art POS taggers perform significantly worse on AAVE tweets, hindering research on language variation beyond the surface level as well as the deployment of equitable downstream NLP tools (Chapter 2).

2. We develop a testsuite for evaluating variations of English and include traditionally marginalized language variations and domains (Chapter 3).

3. We develop a new algorithm that improves POS tagger performance for a specific social group, indicating that inclusive and equitable NLP can be achieved with a change in optimization (Chapter 4).

4. We show that the social identities of human raters and data creators bias summarization models and lead them to create summaries that support the preference of certain social groups over others (Chapter 5).

5. We show that larger and better ASR models have larger performance disparities between social groups, and we show that this is true independent of language (Chapter 6).

6. We provide a critique of the assumed dependence between bias and fairness, and we demonstrate how these in fact are distinct phenomena (Chapter 7).

7. We dissect the widely adopted Rawlsian fairness and show how loopholes in this definition may inadvertently cause larger inequalities. We argue for the need to adopt a more equalitarian definition of fairness such as that in the work by Kai Nielsen (Chapter 8).

1.2 Relating our Contributions to the Broader Research Landscape

Below, we offer a more comprehensive overview of our contributions and contextualize them within the broader research landscape.

POS Tagger Performance for AAVE Speakers

The first part of this dissertation is dedicated to addressing the performance disparity in NLP systems between speakers of standard and non-standard language varieties. We focus on African-American Vernacular English (AAVE), a dialect deeply rooted in socio-cultural, ethnic, and geographic contexts and with a history marked by cultural and economic stigmatization and discrimination (Rickford et al., 2015; Labov, 1972a; Wolfram, 1969; Baugh, 2005).

In Chapter 2, we examine the unique presence of dialects in social media and the associated challenges faced by contemporary NLP tools. Our extensive study explores well-established sociolinguistic hypotheses related to phonological features of AAVE and their reflections as spelling variations on social media. Several studies have shown that besides syntactic structures and lexical entities, phonological variation is also present in tweets (Eisenstein, 2013b, 2015; Eisenstein et al., 2010; Jørgensen et al., 2015). While confirming several of these hypotheses, our contributions extend beyond the sociolinguistic realm. We engage in a comprehensive discussion on biases inherent in social media data, and we demonstrate that state-of-the-art POS taggers exhibit subpar performance on dialects in social media, particularly when dialect markers are present.

A number of related papers have explored social media with respect to sociolinguistic and dialectological questions (Rao et al., 2010; Eisenstein, 2013a; Doyle, 2014; Hovy et al., 2015a; Volkova et al., 2015; Johannsen et al., 2015; Hovy and Søgaard, 2015; Eisenstein, 2015). On Twitter, Volkova et al. (2013) identified several distinctions in subjective language usage between male and female users especially in hashtag and emoticon usage, and Rao et al. (2010) are able to classify the gender, age, regional origin, and political orientation of users based on their language use. Learning models for AAVE-like language and other language varieties is often complicated by the absence of standard writing systems (Boujelbane et al., 2013; Bernhard and Ligozat,

2013; Duh and Kirchhoff, 2005). Closely related to our work, Eisenstein (2013a) and Doyle (2014) both found evidence to support the notion that dialectal phonological variation directly influences spelling on Twitter.

As a contribution to ensure more equitable performance for social groups in POS taggers, we present a suite of 12 datasets designed to benchmark POS taggers against variations in the English language in Chapter 3. We create three new AAVE-heavy datasets collected from hip-hop lyrics, subtitles and Twitter. We provide a thorough description of the data and how it differs from standard POS tagging datasets. We also provide a benchmark evaluation to serve as an evaluation reference for researchers and developers. The datasets are the first of their kind. No lyrics or subtitles datasets have, to our knowledge, previously been published, and although several part-of-speech annotated datasets from Twitter are available, this is likewise the first that has been created specifically for AAVE.

Typically in NLP and much of AI, researchers have focused on one or a few specific domains for evaluation, and have thereby run the risk of overfitting their models to these datasets (Hovy et al., 2014). We publish the testsuite with variations of English with the aim of increasing the awareness of the importance of linguistic variation in evaluation sets and to aid in the efforts against over-fitting models to particular domains or user groups.

In Chapter 4, we delve into the task of enhancing the performance of a POS tagger specifically tailored for AAVE speakers. Our approach entails learning from a combination of randomly sampled and manually annotated Twitter data, along with unlabeled data that we automatically and partially annotate using mined tag dictionaries.

The efficacy of word representations learned from representative unlabeled data, such as word clusters or embeddings, has been known to improve the accuracy of NLP tools for languages and domains with limited resources (Owoputi et al., 2013; Aldarmaki and Diab, 2015; Gouws and Søgaard, 2015). This strategy leverages the similarity in labels assigned to analogous words, providing the model with support for words not encountered during the training phase. For instance, Aldarmaki and Diab (2015) significantly enhances the performance of a POS tagger for Arabic without additional training data by incorporating features carrying morphological information.

To aid in both the automatic labeling of training data and during inference, we employ automatically created tag dictionaries. In our first model, our labeling strategy draws inspiration from work on domain adaptation (Li

et al., 2012; Wisniewski et al., 2014; Hovy et al., 2015b; Plank et al., 2014a) and cross-lingual transfer (Wisniewski et al., 2014; Das and Petrov, 2011; Täckström et al., 2013). We use tag dictionaries as an incomplete-yet-sufficient annotation scheme, where entities are labelled *ambiguously*. In our second model, in a semi-supervised setting (Garrette and Baldridge, 2013; Plank et al., 2014a), the tag dictionaries are used to impose type constraints on the model during inference. The most successful model is obtained through the first approach, namely by ambiguously labeling the training data.

Our leading POS tagger achieves a 55% reduction in errors compared to a state-of-the-art newswire POS tagger and 15-25% error reductions compared to a prominent Twitter POS tagger. These results underscore how subtle adjustments in the data pool can contribute to diminishing performance disparities between social groups.

Performance in Summarization Systems across Demographics

Moving beyond POS tagging, we investigate performance disparity for social groups in more complex NLP systems in the following two chapters.

In Chapter 5, we investigate the impact of evaluator demographics on summarization systems. Traditionally, summarization systems are evaluated with the input of humans. Either the model output is compared to a human summary using some text similarity metric (Lin, 2004; Nenkova and Passonneau, 2004) or human raters are employed to rank the system output against another system output or across aspects such as completeness, fluency and relevance. While using similarity metrics is controversial (Liu and Liu, 2008; Graham, 2015; Schluter, 2017), the standard way to evaluate summarization systems is a combination of both. In Liu and Liu (2008), for example, the human subjects are five undergraduate students in Computer Science. Undergraduate students in Computer Science are not necessarily representative of end users of the technologies we develop. We too observe that human annotators and raters in summarization systems often do not reflect the demographics of end-users. In this first study of its kind, we evaluate two very different summarization systems, namely TextRank (Mihalcea and Tarau, 2004) and MatchSum (Zhong et al., 2020). TextRank, influenced by PageRank, is an extractive model, while MatchSum is abstractive and generates

a summary based on the most salient information in the input data. We take the position that the closer an evaluation is to the real use case of a system, the more indicative the evaluation becomes of the system. Including human preference variation in quality evaluation is a necessary bi-product of evaluating with humans in the loop. We disagree with the idea that quality evaluations should be done by experts, where unanimity can be achieved with well-defined criteria and trained specialists (Gillick and Liu, 2010). To this, we question whose preferences are catered for in the unanimous opinion? We find that the preference of summaries is indeed dependent on demography, and note that looking at preference intersectionally offers insights into how the training data and algorithmic setup may cater more to one demographic than to others. The two main findings in our study indicate these differences and highlight more well-founded issues with summarization systems.

First, negation is a well-known error in summarization systems (Fiszman et al., 2006), but we find that young female or black participants in our study prefer TextRank over the more modern and complex MatchSum system. TextRank summaries in our study contain more negation than the MatchSum comparables, and the users preferring these summaries may view negated sentences as more important than other demographic groups (Kaup et al., 2013).

Second, our study shows that also informativeness and cohesion play a role with regards to which model the demographic groups prefer. While, MatchSum and TextRank – overall – are rated more or less equally informative and useful across demographics, it appears that MatchSum is preferred by older men, especially for white participants, in cases where the MatchSum summaries have been rated more informative than those produced by TextRank. For the American Indian/Alaska Native participants, MatchSum summaries are rated higher than those of TextRank, when they include pronouns. This suggests that the abstractive generation in MatchSum aids in overcoming another known error, namely referencial clarity and cohesion across coreference chains. TextRank may suffer from the extractiveness of the model extracting sentences with pronouns without breaking coreference chains (Pitler et al., 2010; Durrett et al., 2016).

Fairness in ASR Systems across Intersectional Demographic Groups

Next, we turn our attention to the intersectional disparities in ASR models in Chapter 6.

Performance disparity in ASR systems for speakers across a range of demographic groups and protected attributes is well-known, and includes sound evidence for disparities dependent on socio-demographic variables such as gender, race, ethnicity, age, accent and nativeness (Adda-Decker and Lamel, 2005; Tatman and Kasten, 2017; Koenecke et al., 2020; Feng et al., 2024). The main reason given for these disparities seems to be an underrepresentation in the training data by the worst-off groups and thus a too homogeneous representation of language use. See Ngueajio and Washington (2022) for a thorough overview of performance disparity in ASR systems.

Our study in Chapter 6 goes beyond these demographic groups and look at performance of ASR models for *intersectional* groups. We find significant performance disparities across binary genders for adolescents across languages. We look closer at Spanish and find – in a three-tonged intersection with age, dialect and gender – that the performance disparity for the intersectional groups is more pronounced for certain age groups.

We investigate the performance disparity between the demographic groups across model sizes, and find, perhaps discouragingly, that performance disparity correlates negatively with model size, ie. the larger the model, the more disparate performance.

Finally, we look at performance disparity over time, since language changes over time. We do find signs of temporal variation in performance disparity, although the temporal span of our dataset does not allow us to investigate this in sufficient depth. The primary focus in ASR fairness research has centered on evaluating fairness within a static framework, without taking the temporal drift of language into consideration. This is a first attempt to shed light on a consistent problem, especially in deployed models.

On the Independence of Bias and Fairness in NLP

In Chapter 7, we disentangle the relationship between representational biases and fairness in language models, showing that these can be independent. The chapter includes a thought experiment and empirical evidence supporting the lack of correlation between bias metrics and fairness metrics.

Approximately 6% of the NLP papers published in 2022 are concerned with exploring social biases in pre-trained language models. The focus has primarily been on addressing representational bias, also known as association bias, wherein certain demographic groups are portrayed in a discriminatory manner within NLP models (Crawford, 2017; Chaloner and Maldonado, 2019). Association bias has often been conflated with performance disparity (Hashimoto et al., 2018) or empirical fairness (Shen et al., 2022). It is also frequently noted that diminishing association bias will improve empirical fairness (Chen et al., 2020; Friedrich et al., 2021; Cao et al., 2022; Dayanik and Padó, 2020; Castelnovo et al., 2022; Reddy et al., 2021).

In our study, drawing on a thought experiment and a series of experiments, demonstrates that association bias and empirical fairness are frequently independent matters.

Rawlsian Fairness in NLP

Finally, Chapter 8 critiques the uniform adoption of a Rawlsian definition of fairness in NLP, where fairness is defined as maximizing performance for the least advantaged. NLP papers often cite Rawls when mentioning fairness, indicating a NLP-wide adoption of this notion of fairness (Larson, 2017; Vig et al., 2020; Ethayarajh and Jurafsky, 2020; Li et al., 2021; Chalkidis et al., 2022). While the NLP community has discussed how best to evaluate Rawlsian fairness (Williamson and Menon, 2019; Hedden, 2021), the framework as such has not been discussed, nor has it been problematized that Rawlsian fairness does not actually ensure equality. In our study, we discuss the use of Rawlsian fairness in NLP and argue that the Rawlsian definition of fairness has loopholes and the wide adaptation of his notion can contribute to social and economic inequalities. We also offer an alternative more egalitarian view of fairness in the works of Kai Nielsen.

The chapter, finalizing the dissertation, invites a critical discussion on the principles of fairness that drive AI and NLP development, urging a reevaluation of what constitutes equitable technology.

Chapter 2

Challenges of Studying and Processing Dialects in Social Media

Abstract

Dialect features typically do not make it into formal writing, but flourish in social media. This enables large-scale variational studies. We focus on three phonological features of African American Vernacular English and their manifestation as spelling variations on Twitter. We discuss to what extent our data can be used to falsify eight sociolinguistic hypotheses. To go beyond the spelling level, we require automatic analysis such as POS tagging, but social media language still challenges language technologies. We show how both newswire- and Twitter-adapted state-of-the-art POS taggers perform significantly worse on AAVE tweets, suggesting that large-scale dialect studies of language variation beyond the surface level are not feasible with out-of-the-box NLP tools.

2.1 Introduction

Dialectal and sociolinguistic studies are traditionally based on interviews of small sets of speakers of each variety. The Atlas of North American English (Labov et al., 2005) has been the reference point for American dialectology since its completion, but is based on only 762 speakers. Dallas is represented by four subjects, the New York City dialect by six, etc. Data is costly to collect, and, as a consequence, scarce.

Written language was traditionally used for formal purposes, and therefore differed in style from colloquial, spoken language. However, with the rise of social media platforms and the vast production of user generated content, differences between written and spoken language diminish. A number of recent papers have explored social media with respect to sociolinguistic and dialectological questions (Rao et al., 2010; Eisenstein, 2013a; Volkova et al., 2013; Doyle, 2014; Hovy et al., 2015a; Volkova et al., 2015; Johannsen et al., 2015; Hovy and Søgaard, 2015; Eisenstein, 2015). Emails, chats and social media posts serve purposes similar to those of spoken language, and consequently, features of spoken language, such as interjections, ellipses, and phonological variation, have found their way into this type of written language. Our work differs from most previous approaches by investigating several phonological spelling correlates of a specific language variety.

The 284 million active users on Twitter post more than half a billion tweets every day, and some fraction of these tweets are geo-located. Eisenstein (2013a) and Doyle (2014) studied the effect of phonological variation across the US on spelling in Twitter posts, and both found some evidence that dialectal phonological variation has a direct impact on spelling on Twitter. Both authors note various methodological problems using Twitter as a source of evidence for dialectal and sociolinguistic studies, including what we refer to as User Population Bias and Topic Bias below.

In this paper, we collect Twitter data to test eight (8) research hypotheses originating in sociolinguistic studies of African-American Vernacular English (AAVE). The hypotheses relate to three phonological features of AAVE, namely derhotacization, interdental fricative mutation, and backing in /str/. Some of our findings shed an interesting light on existing hypotheses, but our main focus in this paper is to identify the methodological challenges in using social media for testing sociolinguistic hypotheses.

Almost all previous large-scale variational studies using social media have

focused on *spelling variation* and *lexical markers* of dialect. Ours is no exception. However, dialectal variation also manifests itself at the morpho-syntactic level. To investigate this variation, we also annotate some data with part-of-speech (POS) tags, using two NLP systems. This approach reveals a severe methodological challenge: sentences containing AAVE features are associated with significant drops in tagger performance.

This result challenges large-scale variational studies on social media that require automated analyses. The observed drops in performance are prohibitive for studying syntactic and semantic variation, and we believe the NLP community should make an effort to provide better and more robust dialect-adapted models to researchers and industry interested in processing social media. The findings also raise the question of whether NLP technology systematically disadvantages groups of non-standard language users.

2.1.1 Contributions

- We identify eight (8) research hypotheses from the sociolinguistic literature. We test them in a study of the distribution of three phonological features typically associated with AAVE in Twitter data. We test the features' correlations with various demographic variables. Our results falsify the hypothesis that AAVE is male-dominated (but see §3.1).
- We identify five (5) methodological problems common to variational studies in social media and discuss to what extent they compromise the validity of results.
- Further, we show that state-of-the-art newswire and Twitter POS taggers perform much worse on tweets containing AAVE features. This suggests an additional limitation to large-scale sociolinguistic research using social media data, namely that it is hard to analyze variation beyond the lexical level with current tools.

2.1.2 Sociolinguistic hypotheses

AAVE is, in contrast to other North American dialects, not geographically restricted. Although variation in AAVE does exist, AAVE in urban settings has been established as a uniform system with suprasegmental norms (Ash and Myhill, 1986; Labov et al., 2005; Labov, 2006; Wolfram, 2004). This

paper considers the following eight (8) hypotheses from the sociolinguistic literature about AAVE as a ethnolect:

H1: AAVE is an urban ethnolect (Rickford, 1999; Wolfram, 2004)

H2: AAVE features are more present in the Gulf states than in the rest of the United States (Rastogi et al., 2011)

H3: The likelihood of speaking AAVE correlates negatively with income and educational level, and AAVE is more frequently appropriated by men (Rickford, 1999, 2010).

H4: Derhotacization is more frequent in African Americans than in European Americans (Labov et al., 2005; Rickford, 1999).

H5: Derhotacization is negatively correlated with income and educational level (Rickford, 1999).

H6: Interdental fricative mutation is more frequent in AAVE than in European American speech (Pollock et al., 2001; Thomas, 2007a).

H7: Interdental fricative mutation is predominantly found in the Gulf states (Rastogi et al., 2011).

H8: Backing in /str/ (to /skr/) is unique to AAVE (Rickford, 1999; Thomas, 2007a; Labov, 2006).

Hypotheses 1–8 are investigated by correlating the distribution of phonological variants in geo-located tweets with demographic information. Our method is similar to those proposed by Eisenstein (2013a) and Doyle (2014), lending statistical power to sociolinguistic analyses, and circumventing traditional issues with data collection such as the Observer's Paradox (Labov, 1972b; Meyerhof, 2006). Our work differs from previous work by studying phonological rules associated with specific dialects, as well as considering a wide range of actual sociolinguistic research hypotheses, but our main focus is the methodological problems doing this kind of work, as well as assessing the limitations of such work.

2.1.3 Methodological problems

One obvious challenge relating social media data to sociolinguistic studies is that there is generally not a one-to-one relationship between phonological variation and spelling variation. People, in other words, do not spell the way they pronounce. Eisenstein (2013a) discusses this challenge ((1) Writing Bias), but shows that effects of the phonological environment carry over to social media, which he interprets as evidence that there is at least is at least some causal link between pronunciation and spelling variation.

A related problem is that non-speakers of AAVE may cite known features of AAVE with specific purposes in mind. They may use it in citations, for example:

(1) My 5 year old sister texted me on my mums phone saying "why did you take a picher in da bafroom" lool okay b (Twitter, Feb 21 2015)

or in meta-linguistic discussions:

(2) Whenever I hear a black person inquire about the location of the "bafroom"... (Twitter, Jan 20 2015)

We refer to these phenomena as (2) Meta-use Bias. This bias is important with rare phenomena. With "bafroom", it seems that about 1 in 20 occurrences on Twitter are meta-uses. Meta-uses may also serve social functions. AAVE features are used as cultural markers by Latinos in North Carolina (Carter, 2013), for example.

Some of the research hypotheses considered (**H3** and **H5**) relate to demographic variables such as income and educational levels. While we do not have socio-economic information about the individual Twitter user, we can use the geo-located tweets to study the correlation between socio-economic variables and linguistic features at the level of cities or ZIP codes. [1]

Eisenstein et al. (2011) note that this level of abstraction introduces some noise. Since Twitter users do not form representative samples of the population, the mean income for a city or ZIP code is not necessarily the mean income for the Twitter users in that area. We refer to this problem as the User Population Bias.

[1] Unlike many others, we rely on physical locations rather than user-entered profile locations. See Graham et al. (2014) for discussion.

Another serious methodological problem known as (4) Galton's Problem (Naroll, 1961; Roberts and Winters, 2013), is the observation that cross-cultural associations are often explained by geographical diffusion. In other words, it is the problem of discriminating historical from functional associations in cross-cultural surveys. Briefly put, when we sample tweets and income-levels from US cities, there is little independence between the city data points. Linguistic features diffuse geographically and do not change at random, and we can therefore expect to see more spurious correlations than usual. Like with the famous example of chocolate and Nobel Prize winners, our positive findings may be explained by hidden background variables. A positive correlation between income-level and a phonological pattern may also have cultural, religious or geographical explanations.

Reasons to be less worried about Galton's Problem in our case, include that a) we only consider standard hypotheses from the sociolinguistics literature and not a huge set of previously unexplored, automatically generated hypotheses, b) we sample data points at random from all across the US, giving us a very sparse distribution compared to country-level data, but more notably, c) location is an important, explicit variable in our study. Galton's Problem is typically identified by clustering tests based on location (Naroll, 1961). Obviously, the phonological features considered here cluster geographically, as evidenced by our geographic correlations in Table 2.2, but since our studies explicitly test the influence of location, it is not the case for most of the hypotheses considered here that geographic diffusion is the underlying explanation for something else.

In § 2.3, we discuss whether these four methodological problems compromise the validity of our findings. One other methodological problems that may be relevant for other studies of dialect in social media, is almost completely irrelevant for our study: It is often important to control for topic in dialectal and sociolinguistic studies (Bamman et al., 2014), e.g., when studying the lexical preferences of speakers of urban ethnolects. We call this problem (5) Topic Bias. Using word pairs with equivalent meanings for our studies, we implicitly control for topic (but see § 2.3.1).

2.2 Data and Method

We focus on derhotacization, backing in /str/, and interdental fricative mutation. Specifically, we collect data to study the following four phonological

Feature	Positive	Negative	Count
/r/ → /Ø/ or /ə/	brotha	brother	9528
	foreva	forever	3673
	hea	here	4352
	lova	lover	1273
	motha	mother	4668
	ova	over	3441
	sista	sister	5325
	wateva	whatever	2974
	wea	where	5153
	total		40,387
/str/ → /skr/	skreet	street	1226
	skrong	strong	1629
	skrip	strip	1101
	total		3659
/ð/ → /d/ or /v/	brova	brother	3715
	dat	that	2610
	deez	these	4477
	dem	them	3645
	dey	they	2434
	dis	this	2135
	mova	mother	2462
	total		21,478
/θ/ → /t/ or /f/	mouf	mouth	3861
	nuffin	nothing	2861
	souf	south	1102
	teef	teeth	1857
	trough	through	2804
	trow	throw	1090
	total		13,575
All tweets			79,396

Table 2.1: African-American Vernacular English (AAVE) and General American (GA) word pairs. *Positive* and *Negative* refer to the presence of the AAVE feature.

variations (the latter two are both instances of interdental fricative mutation): *a*) derhotacization: /r/ → /Ø/ or /ə/, *b*) /str/ → /skr/, *c*) /ð/ → /d/ or /v/ and, *d*) /θ/ → /t/ or /f/. .

In non-rhotic dialects, /r/ is either not pronounced or is approximated as a vocalization in the surface form, when /r/ is in a pre-vocalic position. This can result in an elongation of the preceding vowel or in an off-glide schwa /ə/, e.g., *guard* → /gɑ:d/, *car* → /ka:/, *fear* → /fiə/ (Thomas, 2007a).

Backing in /skr/ denotes the substitution of /str/ for /skr/ in word-initial positions resulting in pronunciations such as /skrit/ for *street*, /skrɑŋ/ for *strong* and /skrɪp/ for *strip*. Backing in /str/ has been reported to be a unique feature in AAVE, as it is unheard in other North American dialects (Rickford, 1999; Labov, 1972a; Thomas, 2007a).

The two interdental fricative mutations relate to substitutions of/ð/ and /θ/ by /d/, /v/ and /t/, /f/ in words such as *that* and *mother* or *nothing* and *with*. It has been reported that mutations of /ð/ and /θ/ are more common among African Americans than among European Americans and that the

frequency of the mutations is inversely correlated with socio-economic levels and formality of speaking (Rickford, 1999).

We follow Eisenstein (2013a) and Doyle (2014) in assuming that spelling variation may be a result of phonological differences and select 25 word pairs for our study (Table 2.1). For each word pair, we collect positive (e.g., "skreet") and negative occurrences (e.g., "street"), resulting in a total number of 79,396 tweets. The word pairs were chosen based on the unambiguity, frequency and representability of the phonological variations. Uniquely, backing in /str/ is represented by three word pairs of high similarity, which is due to phonological restrictions on the variation of /str/ to /skr/ and to the fact that backing in /str/ is a very rare phenomena.

The Twitter data used in the experiments was gathered from May to August 2014 using TwitterSearch.[2] We only collected tweets with geo-locations in the contiguous United States, from users reporting to tweet in English, and which were also predicted to be in English using *langid.py*.[3] The demographic information was obtained from the 2011 American Community Survey from the United States Census Bureau, as was information about population sizes in US cities. We linked each tweet in our data to demographic information using the geo-coordinates of the tweet and its nearest city in the following way.

For the 110 US cities of $\geq$ 200,000 inhabitants, we gathered information about: *a*) percentage high school graduates, *b*) percentage below poverty level, *c*) population size, *d*) median household income, *e*) percentage of males, *f*) percentage between 15 and 24 years old, *g*) percentage of African Americans and *h*) unemployment rate.

The overall geographical distribution of our data is shown in Figure 2.1. The map shows that we see more tweets with AAVE features in the Gulf states, in particular Louisiana and Mississippi and Georgia. This lends preliminary support to **H2**.

2.3 Results with phonological features

Occurrences of the phonological variations related to AAVE were correlated with the geographic and demographic variables using Spearman's ρ (Table 2.2–2.3), at the level of individual tweets. From the correlation coef-

[2]https://pypi.python.org/pypi/TwitterSearch/
[3]https://pypi.python.org/pypi/langid

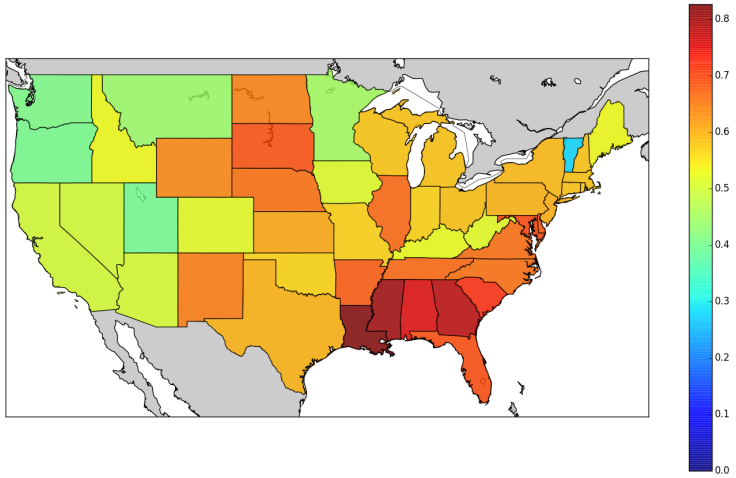

Figure 2.1: The ratio of AAVE examples across US states.

ficients we see that the distributions of the three chosen AAVE rules are best explained by longitude, the distinction between the Gulf states and the rest of the US, and by the distribution of African Americans (with explained variances in the range of 0.03–0.05).

Our data suggests that **H2**, namely that AAVE is more prevalent in the Gulf states, is probably true. Hypothesis **H1**, that AAVE is an urban ethnolect, lends some support in our data, but the correlation with urbanicity is weaker (and negatively correlated or non-significant in half of the cases).

Our data only lends limited support to the first half of hypothesis **H3**. While derhotacization and /str/ correlate (negatively) significantly with income levels, we see no significant correlations within /ð/ and a positive correlation within /θ/. However, our data does not suggest that **H3** is false, either. Our data does lend support to the more specific hypothesis **H5**, namely that derhoticization is sensitive to income level, while the strong correlation with the distribution of African Americans lends support to **H4**.

More interestingly, our data suggests that *women use AAVE features more often than men*, i.e., there is a negative correlation between male gender and AAVE features, contrary to the second half of **H3**, namely that AAVE is more frequently appropriated by men. Note, however, that our

Feature	Word Pairs	Latitude	Longitude	Urban	Gulf
/r/	brotha/brother	***	***	***	***
	foreva/forever	***	**	–	***
	hea/here	***	***	*	***
	lova/lover	***	***	**	***
	motha/mother	–	–	***	–
	ova/over	***	–	–	***
	sista/sister	–	***	**	***
	wateva/whatever	***	***	**	***
	wea/where	***	***	-	***
	total	***	***	***	***
/str/	skreet/street	***	–	***	***
	skrong/strong	***	*	***	***
	skrip/strip	***	–	***	***
	total	***	**	–	***
/ð/	brova/brother	***	***	***	***
	dat/that	***	*	–	***
	deez/these	*	***	–	–
	dem/them	***	***	–	***
	dey/they	***	***	–	***
	dis/this	***	-	–	***
	mova/mother	*	***	***	***
	total	***	***	***	***
/θ/	mouf/mouth	***	–	–	***
	nuffin/nothing	***	***	***	***
	souf/south	***	***	***	***
	teef/teeth	**	–	**	***
	trough/through	–	***	–	–
	trow/throw	***	**	-	***
	total	*	***	***	***

Table 2.2: Geographic correlations with AAVE-features. Shading corresponds to negative correlations, and the asterisks correspond to alpha value cut offs: * = $0.05 > p \geq 0.01$, ** = $p \leq 0.01$, *** = $p \leq 0.0001$.

gender ratios are aggregated for city areas, and with the demographic bias of Twitter, these correlations should be taken with a grain of salt. Considering the small gender ratio differences, we also compute correlations between our linguistic features and gender using the Rovereto Twitter N-gram Corpus (RTC) (Herdagdelen and Baroni, 2011).[4] The RTC corpus contains information about the gender of the tweeter associated with n-grams. While there is too little data in the corpus to correlate gender and backing in /str/, derhotacization and both interdental fricative mutations (/ð/ → /d/ or /v/ and /θ/ → /t/ or /f/) correlate significantly with women. Out of our words, 10 correlate significantly with female speakers; seven with male. The correlations are found in Table 2.4. For each feature, certain words correlate significantly with female speakers, while others correlate significantly with male speakers. Consequently, neither our Twitter data nor the Twitter data in the RTC suggest that AAVE is more often appropriated by men. We discuss whether our data provides a basis for falsifying the second half of **H3** in §3.1.

[4] http://clic.cimec.unitn.it/amac/twitter_ngram/

Feature	Word Pairs	Male	Black	15-24	Urban	Edu.	Income	Poverty	Unempl.
/r/ → /Ø/ or /ə/	brotha/brother	***	**	–	–	**	–	–	–
	foreva/forever	**	***	–	–	–	–	**	–
	hea/here	–	***	**	***	***	***	***	*
	lova/lover	–	–	–	–	***	*	**	–
	motha/mother	-	**	–	*	-	**	–	–
	ova/over	***	***	–	–	–	***	***	–
	sista/sister	*	***	–	–	**	–	–	–
	wateva/whatever	***	***	–	–	–	***	***	–
	wea/where	**	***	***	***	***	***	***	*
	total	***	***	***	***	***	***	***	–
/str/ → /skr/	skreet/street	–	–	–	**	*	**	*	**
	skrong/strong	**	***	–	*	**	**	**	*
	skrip/strip	*	–	*	***	***	–	***	***
	total	***	***	–	***	***	***	–	–
/ð/ → /d/ or /v/	brova/brother	***	***	***	***	***	–	***	***
	dat/that	-	***	–	–	–	**	**	–
	deez/these	–	–	–	**	***	–	**	***
	dem/them	*	***	**	**	***	–	-	-
	dey/they	***	***	**	*	**	**	***	–
	dis/this	–	***	**	–	–	–	*	*
	mova/mother	***	***	***	–	***	***	–	***
	total	***	***	***	–	***	-	***	***
/θ/ → /t/ or /f/	mouf/mouth	**	–	–	–	–	–	–	–
	nuffin/nothing	***	***	***	***	***	***	–	***
	souf/south	***	–	**	–	**	–	***	***
	teef/teeth	–	–	–	-	**	–	–	–
	trough/through	–	–	–	-	**	–	*	*
	trow/throw	*	–	–	***	**	*	**	**
	total	***	***	***	***	–	**	–	*

Table 2.3: Demographic correlations with AAVE-features. Shading corresponds to negative correlations, and the asterisks correspond to alpha value cutoffs: * = $0.05 > p \geq 0.01$, ** = $p \leq 0.01$, *** = $p \leq 0.0001$.

The high correlation between mutations of /ð/ and longitude supports the presence of these mutations of /ð/ in non-standard northern varieties (Rickford, 1999). The mutation of /θ/ is also correlated with longitude, and with latitude, suggesting an Eastern American feature rather than a distinct Southern feature (Rickford, 1999). The variation in mutations could possibly be explained by both geography as well as the distribution of African Americans.

There is evidence in our data that backing in /str/ (to /skr/) is appropriated more often by AAVE speakers than by speakers of other dialects (**H8**). There is also a negative correlation between latitude and backing in /str/ as well as a strong positive correlation with the Gulf states, suggesting that backing in /str/ is a feature primarily seen in this region. The data thereby suggests that the feature is appropriated significantly more by African Americans than by speakers of the Southern dialect.

In sum, while our data lends support to several of the common hypotheses from the sociolinguistics literature, we found one unexpected tendency, going against the second half of **H3**, namely that AAVE features were found more often with females. We now discuss this finding in light of the methodological

problems discussed in §1.2.

Feature	Word Pairs	Male
/r/ → /Ø/ or /ə/	brotha/brother	**
	foreva/forever	**
	hea/here	*
	lova/lover	–
	motha/mother	**
	ova/over	**
	sista/sister	–
	wateva/whatever	–
	wea/where	**
ð → /d/ or /v/	brova/brother	*
	dat/that	**
	deez/these	**
	dem/them	**
	dey/they	**
	dis/this	**
	mova/mother	–
θ → /f/ or /t/	mouf/mouth	**
	nuffin/nothing	**
	souf/south	**
	teef/teeth	–
	trough/through	**
	trow/throw	**

Table 2.4: Gender correlations in the RTC dataset. Shading corresponds to negative correlations, and the asterisks correspond to alpha value cutoffs: * = $0.05 > p \geq 0.01$, ** = $p \leq 0.01$, *** = $p \leq 0.0001$.

2.3.1 Is AAVE *not* male-dominated?

We now discuss whether our data falsifies the second half of **H3**, one methodological problem at a time (see § 2.1.3). If Written Bias were to bias our conclusions, one gender should be more likely to exhibit more phonologically motivated spelling variation. This may actually be true, since it is well-established that women tend to be more linguistically creative and have larger vocabularies (Labov, 1990). Whether women are also more meta-linguistic (Meta-use Bias), has to the best of our knowledge not been studied. Since

genders are almost equally geographically distributed, and since Twitter is generally considered gender-balanced, neither User Population Bias nor Galton's Problem is likely to bias our conclusions. Topic Bias, on the other hand, may. While our semantically equivalent pairs control for topic, the pragmatics sometimes differ. Just like code-switching is a strategy for bilinguals, using the spelling *motha* instead of *mother* could mean something, say irony, which one gender is more prone for. In sum, while we do believe that our data should lead sociolinguists to question whether AAVE is male-dominated, our findings may be biased by Written Bias.

2.4 POS tagging

We need automated syntactic analysis to study morpho-syntactic dialectal variation. We ran a state-of-the-art POS tagger trained on newswire[5] (Stanford), as well as two state-of-the-art POS taggers adapted to Twitter, namely Gate[6] and Ark[7], on our data. We had one professional annotator manually annotate 100 positive (AAVE) and 100 negative (non-AAVE) sentences using the coarsegrained tags proposed by Petrov et al. (2011). We map the tagger outputs to those tags and report tagging accuracies. See Table 2.5 for results, with $\Delta(+,-)$ being the absolute difference in performance from non-AAVE to AAVE.

While Gate is certainly better than Stanford on our data, performance is generally poor and prohibitive of many downstream applications and variational studies. We also note that both the best and worst tagger perform significantly worse on AAVE tweets than on non-AAVE tweets. What are the sources of error in the AAVE data? One example is the word *brotha*, which is tagged as a both an adverb, a verb, and as X (foreign words, mark-up, etc.). Contractions like *finna* ("fixing to" meaning "going to") and *gimme* ("give me") are often tagged as particles, but annotated as verbs or, as in the case of *witchu* ("with you"), as a preposition. Another interesting mistake is tagging adverbial *like* as a verb.

[5]https://nlp.stanford.edu/software/
[6]https://gate.ac.uk/wiki/
[7]http://www.cs.cmu.edu/ ark/

	Stanford	Gate	Ark
AAVE	61.4	**79.1**	77.5
non-AAVE	74.5	**83.3**	77.9
$\Delta(+,-)$	13.1	4.2	0.4

Table 2.5: POS-tagging accuracies on AAVE and non-AAVE (%).

2.5 Conclusion

Large-scale variational studies of social media can be used to question received wisdom about dialects, lending support to some sociolinguistic research hypotheses and questioning others. However, we caution that our results were biased by several factors, including the representativity of the social media user bases. We also show how state-of-the-art POS taggers are more likely to fail on dialects in social media. The performance drops may be considered prohibitive of studying morph-syntactic patterns across dialects and as a challenge to us as a community.

Chapter 3

A Test Suite for Evaluating POS Taggers across Varieties of English

We present a suite of 12 datasets for evaluating POS taggers across varieties of English to enable researchers to evaluate the robustness of their models. The suite includes three new datasets, sampled from lyrics from black American hip-hop artists, southeastern American Twitter, and the subtitles from the TV series *The Wire*. We present an example evaluation of an off-the-shelf POS tagger across these datasets.

3.1 Introduction

Most off-the-shelf POS taggers for English have been induced from and evaluated on manually annotated newspaper corpora such as the English Penn Treebank. This has led to community-wide overfitting to a very specific instantiation of newspaper English, and it is well-established that, unsurprisingly, these off-the-shelf taggers perform significantly worse on other domains and varieties of English. Many researchers have tried to obtain better performance across domains or varieties by using heavier model regularization or by learning from mixtures of labeled and unlabeled data. This also means that researchers have gone from evaluating their models on newspaper English, specifically the Wall Street Journal sections 22–24 in the English Penn Treebank, to using other datasets. Typically, however, researchers have focused on one or a few specific domains, again running the risk of over-fitting to these datasets (Hovy et al., 2014). In this paper, we present a suite of 12 datasets for evaluating POS taggers across varieties of English, including three new datasets of subtitles, tweets, and hip-hop lyrics. In the machine learning community, it is an important rule of thumb to evaluate new classification algorithms across at least a dozen datasets (Demsar, 2006), and this is an attempt to make a dozen datasets available for English POS tagging, an important NLP task for many downstream applications. We also present an example evaluation of the Stanford POS tagger across all these datasets. The three new domains – hip-hop lyrics, southeastern American Twitter and subtitles from the TV series *The Wire* – differ with varying degrees syntactically and lexically from newswire as well as in the communicative functions of the domains. In our suite, we ignore datasets that have been used for decades, such as the Brown Corpus, the Switchboard Corpus, and the English Penn Treebank, to avoid community-wide over-fitting, as well as fitting our models to 20^{th} century varieties of English. The 12 datasets are presented in Table 3.1.

3.2 New Datasets

In this paper, we present three new English datasets from domains containing non-canonical language use reflecting aspects of African American Vernacular English (AAVE) such as syntactic variation, lexical items, abbreviations and phonologically-motivated spellings.

Domain	Reference	Sentences	Stanford
Answers	EWT	1,744	91.5%
Emails	EWT	2,450	90.7%
Newsgroups	EWT	1,195	91.6%
Newspaper	Dundee Treebank	51,502	93.2%
Reviews	EWT	1,906	92.5%
Spoken (child-directed)	CHILDES	5,222	94.0%
Spoken (interviews)	LOWLANDS	500	85.0%
Twitter	LOWLANDS	500	87.0%
Weblogs	EWT	1,015	92.1%
Lyrics	This work	509	77.7%
Subtitles	This work	1,074	83.3%
Twitter (AAVE)	This work	374	61.4%

Table 3.1: Datasets in the evaluation suite. EWT = English Web Treebank.

The three domains were chosen because of the heavy presence of AAVE features as well as for the individual characteristics of the domains, which too vary from those of the commonly-used evaluation datasets in NLP. All three domains are low in formality and high in individualistic expression, which is a desired quality for analyzing vernacular language uses. Much annotated NLP data, such as legal documents and newswire are formal in writing style and subject to high demands of standardization because of topic and the public distribution of texts from these domains. Lyrics, subtitles and tweets, while public, are not subject to as strict demands of standardization. Lyrics and subtitles showcase a certain subculture in a narrative form and wish to establish certain personas through language use (and mise-en-scène), while tweets are examples of phatic communication.

All three datasets were annotated by a trained linguist with experience in African American Vernacular English. We used the Universal Google tag set (Petrov et al., 2011) with 12 categories[1] to ensure higher usability for POS tagger evaluation. Hovy et al. (2014) has shown that with this tag set, annotation guidelines are not necessary to obtain high-quality POS

[1]ADJ (adjectives), ADP (adpositions), ADV (adverbs), CONJ (conjunctions), DET (determiners and articles), NOUN (nouns), NUM (numerals), PRON (pronouns), PRT (particles), "." (punctuation marks), VERB (verbs) and X (miscellaneous).

annotations, and we reach an inter-annotator agreement of 93.6%.

In the following, we present the three data sets with an explanation of the collection process and the preprocessing of the data. We further describe the individual characteristics of the datasets and compare these with the commonly-used evaluation sets in NLP. Statistics on the three new datasets are provided in Table 3.2 with example sentences and annotations from each dataset in Table 3.4.

Domain	Sentences	Types	Av. Length
Lyrics	509	1314	9.1
Subtitles	1074	1519	8.9
Tweets (AAVE)	374	1606	9.7

Table 3.2: Statistics on the new AAVE datasets

3.2.1 Hip-hop Lyrics

The annotated hip-hop lyrics dataset consists of lyrics from black American hip-hop artists[2] produced between 1993–2012. We collected the hip-hop lyrics dataset using the *Rap Genius* depository.[3] *Rap Genius* is a database of lyrics from various musical genres, where contributors can annotate semantic meanings of the lyrics. These contributions were used to annotate entities and constructions in the hip-hop lyrics which were unknown to the annotator.

The hip-hop lyrics dataset contains many non-canonical lexical entities such as *tecks* (Teck 9, a type of firearm[4]) and *loc'd out* (loco, crazy), phonologically-motivated spellings *fo', pimpin'* and *betta* and domain-specific entities used for rhythm and rhyme such as *badonkadonk-donk*.

The most characteristic trait of the hip-hop lyrics dataset is the absence of punctuation marks. As can be see in Table 3.5, only 6% of the tokens in the hip-hop lyrics dataset are punctuation marks, and this figure is even inflated, since we split rhyming words on hyphens (such as *badonkadonk-donk*). Excluding commas and hyphens, the percentage of end punctuation

[2] 2Pac, 50Cent, Birdman, COOLIO and the Gang, Lil Wayne, Missy Elliott and Precious Paris.

[3] http://rap.genius.com

[4] http://www.urbandictionary.com/define.php?term=teck

marks is a mere 1.4% of the tokens in this dataset. This lack of sentence separating punctuation marks poses a challenge to a POS tagger trained on newswire, where a punctuation mark is typical at the end of every sentence. It also posed a problem for us in determining where one sentence ends and the next begins.

The poetic nature of the domain plays on the ambiguity created by sentences flowing into each other, but for simplicity and to eliminate the fear of wrongly interpreting the intended meaning of the lyrics, we declare each line in the lyrics a sentence. Not only end punctuation is important for correct syntactic analysis, however. Consider the change in syntactic function of the word *this* (from determiner to pronoun) introduced by commas encircling *bitch* in the following sentence, again from Lil Wayne's "Gangstas and Pimps", *I'm in this bitch hold your pictures* (500 Degreez, 2002).

3.2.2 Subtitles from The Wire

The television series *The Wire*[5] is a fictional narrative portraying various criminal institutions in Baltimore, MD as well as the local law enforcement departments.

The dataset presented here is the full first episode of the first season of the show, which is concerned with the Baltimore drug scene. The dataset consists of dialogues between the characters on the show, and features both white and black police officers, prostitutes, politicians, gangsters, drug dealers, judges and junkies. While the vast majority of the characters are male, there is a handful of women in the episode and they too represent a variety of sociolinguistic categories.

Non-canonical language can be used as a resource in cultural products for creating personas and exhibiting social and cultural characteristics and affiliations. Trotta and Blyahher (2011), who perform a linguistic analysis of the use of AAVE in *The Wire*, note that much of the dialog in *The Wire* is marked by AAVE features such as copula deletion, habitual *be*, completive *done* and continuative *steady*. They further remark that characters on the show can be seen as representatives of "genuine AAVE" speakers (Trotta and Blyahher, 2011, p. 38).

We collected the subtitles using *opensubtitles.org*,[6] where subtitles are

[5]HBO, 2002–2008. http://www.hbo.com/the-wire

[6]http://www.opensubtitles.org/en/search

available for various television series and movies in a variety of languages. We manually controlled that the subtitles collected indeed reflected the dialogues in the episode. The data format used by *opensubtitles.org* contains time stamps for each utterance, which means that sentences can be split over several time stamps, especially if the utterance is long. Since we are interested in sentences rather than utterances, we join utterances together across time stamps, until end punctuation marks the end of the utterance. We split utterances that contain several sentences into separate sentences and tokenize these, separating all punctuation from words except apostrophes not used as discourse markers (e.g., *'em*).

The spoken origin of subtitles means that the data contains quintessential elements of spoken language such as interjections, cut-off sentences, contractions and exclamations. However, because *The Wire* is scripted, not all aspects of natural speech are seen in this data set. There are for instance no hesitations, corrections or false starts in our dataset. The form of dialogue introduces a higher quantity of certain syntactic structures than present in the commonly-used CMU Twitter corpus (Owoputi et al., 2013). The *The Wire* dataset presented here contains, as example, ∼28% questions, where the CMU Twitter corpus only contains 11.4%.[7] Comparatively, the AAVE tweets dataset contains 8.5% questions while only ∼2% of the sentences in the hip-hop lyrics dataset are questions. The AAVE tweets dataset is described below. This skew in the distribution of syntactic forms is evidence of the necessity of using training data with a high degree of variation as well as evaluating on data from various domains.

We believe that subtitles can be a rich source of data for various NLP tasks because of the high degree of linguistic variation and syntactic structures in dialogues while being in a readily manageable format.

3.2.3 AAVE Tweets

Several part-of-speech annotated datasets from Twitter are currently available, but to the best of our knowledge, this is the first that has been geographically restricted to increase the degree of AAVE features. The tweets for the AAVE tweets dataset were collected through the Twitter API using the Python package, TwitterSearch,[8] between 2015-02-28–2015-03-03. The

[7] These figures are measured by counting sentences ending in a question mark, and they are therefore not the result of a discourse analysis of the data.

[8] https://pypi.python.org/pypi/TwitterSearch/

Domain	Lyrics	Subtitles	Tweets
Sentences	509	1074	374
Av. length	9.1	8.9	9.7
Types	1314	1519	1606
OOV types	9.8%	6.2%	11.0%
OOV tokens	12.1%	10.4%	22.2%

Table 3.3: Dataset statistics

geo-coordinates provided in the metadata were used to exclude all tweets not posted within the American Gulf Coast states.[9] We chose to only include tweets from these states because of the higher percentage of black Americans in the population of this region,[10] and because it has been show that AAVE features are more prevalent in tweets from this area than from elsewhere in the United States (Jørgensen et al., 2015). Since we did not want to limit the data to urban language use, we chose states rather than cities as our search frame. This also enables comparative linguistic analyses of urban and rural uses of AAVE in this dataset. Five linguistic challenges of non-standard variation are endemic to online social media data, namely punctuation, capitalisation, spelling, vocabulary and syntactic structures (Eisenstein, 2013c). These variations are also present in our AAVE tweets dataset along with emoticons, which can be seen in the higher frequency of the miscellaneous category, *X*, in this dataset as shown in Table 3.5.

Several studies have shown that besides syntactic structures and lexical entities, phonological variation is also present in tweets (Eisenstein, 2013b, 2015; Eisenstein et al., 2010; Jørgensen et al., 2015).

Jørgensen et al. (2015), who focus on phonologically-motivated spellings of AAVE on Twitter, show that AAVE features such as interdental fricative mutation, derhotacization and backing in [str][11] is present on Twitter and

[9]North Carolina, South Carolina, Louisiana, Georgia, Mississippi, Tennessee, Arkansas, Alabama and Florida.

[10]U.S. Census Bureau, Census 2000 Redistricting Data (Public Law 94-171) https://www.census.gov/prod/cen2010/briefs/c2010br-06.pdf,http://www.indexmundi.com/facts/united-states/quick-facts/all-states/black-population-percentage#chart

[11]Uniquely to AAVE, word-initial [str] can be substituted by [skr] in words such as "street", "strip" and "strawberry"(Jørgensen et al., 2015; Pollock et al., 2001; Rickford,

that these features partially correlate with demographic information about the location of the tweeter.

Dataset	Sentences
Lyrics	2Pac/NOUN cares/VERB ,/. if/CONJ don't/VERB nobody/PRON else/ADJ care/VERB
	I'm/PRON a/DET loc'd/ADJ out/PRT gangsta/NOUN set/NOUN trippin'/VERB banger/NOUN
Subtitles	Life/NOUN just/ADV be/VERB that/DET way/NOUN ,/. I/PRON guess/VERB ./.
	Couldn't/VERB help/VERB hisself/PRON ./.
Tweets (AAVE)	Dat/DET highlight/NOUN and/CONJ contour/NOUN doe/ADV URL/NOUN
	Aww/X damn/X I/PRON can't/VERB believe/VERB dat/PRON lol/X

Table 3.4: Example sentences from new datasets

3.2.4 Dataset characteristics

There are notable differences in what tags are likely to follow each other across the three new datasets. In the AAVE tweets dataset, for example, an instance of the miscellaneous class, X, is most often followed by another instance of the same category, whereas in the other two datasets, a word belonging to the miscellaneous class is most often followed by a punctuation mark. On Twitter, several punctuation marks can be used after each other, while this is less common in the other domains. These differences in syntactic constructions in the four domains illustrate the necessity of using multiple and varied data sets for evaluation of a POS tagger.

Table 3.3 shows the sizes of the new datasets, the number of word types and average tokens per sentence as well as the percentage of out-of-vocabulary (OOV) types and tokens in the three new datasets compared to the CMU Twitter corpus.

The highest percentage of both OOV types and tokens in both categories is not surprisingly in the AAVE tweets dataset, followed by the hip-hop

1999; Thomas, 2007b)

lyrics dataset on both accounts, while the lowest in both categories is in the subtitles dataset.

While the hip-hop lyrics dataset has almost as high a percentage of OOV types as the AAVE tweets dataset, they are used less frequently. In the AAVE tweets dataset almost every 4$^{\text{th}}$ token is an out-of-vocabulary word.

Twitter is known for its short messages, but of the new datasets presented here, the AAVE tweets have the longest average sentence length. While this is interesting given Twitter's 140-characters restriction and the issues with developing NLP tools for Twitter (Foster et al., 2013; Eisenstein, 2013c), it is not surprising. Hip-hop lyrics are essentially rapped poems to a beat, and the lyricists therefore also have strict restrictions on length as well as on rhythm and rhyme[12], and subtitles represent conversation and contains cut-off sentences, exclamations and short remarks[13].

The distributions of part-of-speech tags in the CMU Twitter corpus and the three new datasets are presented in Table 3.5. Two observations related to the distribution of the 12 POS tags are worth briefly mentioning here, as these observations points to deviations in the new datasets from the CMU Twitter corpus.

Firstly, it is clear that the amounts of punctuation marks and of the miscellaneous, *X* category vary from domain to domain. About 10% of both the CMU Twitter corpus and our AAVE tweets dataset are punctuation marks, while every 5$^{\text{th}}$ token in the subtitles dataset and only 6% of the tokens in the hip-hop lyrics dataset are punctuation marks.

Secondly, the CMU Twitter corpus contains fewer determiners, pronouns and verbs than the three new datasets. The difference in the frequencies of nouns and determiners is smaller in the three new test sets($\sim$10%) than in the CMU Twitter corpus($\sim$14%). The highest distributions of verbs and pronouns are found in the hip-hop lyrics dataset, which is possibly due to the concise, poetic writing style, while the CMU Twitter corpus set contains a much smaller amount of pronouns than all the other data sets.

In closing, we believe the variation present in these three new datasets is relevant for testing the robustness of POS tagging systems, and we encourage researchers to include these datasets in the evaluation of their systems.

[12]e.g., *Cadillac smoke **dro** just me and the **ho***. Lil Wayne feat. Birdman. "Gangstas and Pimps". *500 Degreez*, 2002

[13]e.g., *Yo.*. Example is taken from the subtitles dataset.

		New datasets		
POS	CMU	Subtitles	Lyrics	Tweets (AAVE)
.	12%	21%	6%	10%
ADJ	5%	4%	5%	5%
ADP	9%	8%	9%	7%
ADV	5%	5%	4%	5%
CONJ	2%	2%	5%	2%
DET	6%	9%	11%	10%
NOUN	20%	18%	20%	22%
NUM	1%	1%	1%	1%
PRON	7%	14%	15%	12%
PRT	5%	2%	2%	1%
VERB	15%	17%	21%	17%
X	13%	1%	1%	8%
Total tokens	34.3k	4.2k	4.5k	5k

Table 3.5: Tag distribution per data set

3.3 Example Evaluation

We present a very simple evaluation of the Stanford POS tagger in the rightmost column in Table 3.1. We observe that results on the three new datasets are significantly lower than for any of the other datasets. This, in our view, demonstrates the need for evaluation data representing minority varieties of English such as African American Vernacular English.

Chapter 4

Learning a POS tagger for AAVE-like language

Abstract

Part-of-speech (POS) taggers trained on newswire perform much worse on domains such as subtitles, lyrics, or tweets. In addition, these domains are also heterogeneous, e.g., with respect to registers and dialects. In this paper, we consider the problem of learning a POS tagger for subtitles, lyrics, and tweets associated with African-American Vernacular English (AAVE). We learn from a mixture of randomly sampled and manually annotated Twitter data and unlabeled data, which we automatically and partially label using mined tag dictionaries. Our POS tagger obtains a tagging accuracy of 89% on subtitles, 85% on lyrics, and 83% on tweets, with up to 55% error reductions over a state-of-the-art newswire POS tagger, and 15-25% error reductions over a state-of-the-art Twitter POS tagger.

4.1 Introduction

Modern part-of-speech (POS) taggers perform well on what some consider canonical language, as found in domains such as newswire, for which sufficient manually-annotated data is available. For many domains, such as subtitles, lyrics, and tweets, however, labeled data is scarce, if existing, and the performance of off-the-shelf POS taggers is prohibitive of downstream applications. Furthermore, subtitles, lyrics and tweets are very heterogeneous. Subtitles span from Shakespeare to The Wire, and the lyrics of Elvis Costello are very different from those of Tupac Shakur. Twitter can be anything from teenagers discussing where to go tonight, to researchers discussed the implications of new findings. All three sources of data exhibit a very high degree of linguistic variation, some of which is due to the dialects of the speakers or authors.

In this paper, we use a corpus of POS-annotated tweets recently released by CMU,[1] consisting of semi-randomly sampled US tweets. We want to use this corpus to learn a POS tagger for subtitles, lyrics, and tweets, which are typically associated with African-American Vernacular English (AAVE). We believe our POS tagger can broaden the coverage of NLP tools, and serve as an important tool for large-scale sociolinguistic analyses of language use associated with AAVE (Jørgensen et al., 2015; Stewart, 2014), which relies on the accuracy of these NLP tools.

We combine several recent trends in domain adaptation, namely word embeddings, clusters, sampling, and the use of type constraints. Word representations learned from representative unlabeled data, such as word clusters or embeddings, have been proven useful for increasing the accuracy of NLP tools for low-resource languages and domains (Owoputi et al., 2013; Aldarmaki and Diab, 2015; Gouws and Søgaard, 2015). Since similar words receive similar labels, this can give the model support for words not in the training data. In this paper, we use word clusters and word embeddings in both our baseline and system models.

Using unlabeled data to estimate a target distribution for importance sampling, or for semi-supervised learning (Søgaard, 2013), as well as wide-coverage, crowd-sourced tag dictionaries to obtain more robust predictions for out-of-domain data have been succesfully used for domain adaptation (Das and Petrov, 2011; Hovy et al., 2015b; Li et al., 2012). In this paper, we use

[1] https://github.com/brendano/ark-tweet-nlp/tree/master/data/twpos-data-v0.3

automatically-harvested tag dictionaries for the target variety(/-ies) in two different settings: for labeling the unlabeled data using a technique elaborating on previous work (Li et al., 2012; Wisniewski et al., 2014; Hovy et al., 2015b), and for imposing type constraints at test time in a semi-supervised setting (Garrette and Baldridge, 2013; Plank et al., 2014a). Our best models are obtained using partially labeled training data created using tag dictionaries.

Our contributions We present a POS tagger for AAVE-like language, mining tag dictionaries from various websites and using them to create partially labeled data. Our contributions include: (i) a POS tagger that performs significantly better than existing tools on three datasets containing AAVE markers, (ii) a new domain adaptation algorithm combining ambiguous and cost-sensitive learning, and (iii) an annotated corpus and trained POS tagger made publicly available.

4.2 Data

For historical reasons, most of the manually annotated corpora available today are newswire corpora. In contrast, very little data is available for domains such as subtitles, lyrics and tweets — especially for language varieties such as AAVE. Learning robust models for AAVE-like language and other language varieties is often further complicated by the absence of standard writing systems (Boujelbane et al., 2013; Bernhard and Ligozat, 2013; Duh and Kirchhoff, 2005).

In this paper, we use three manually annotated data sets, consisting of subtitles from the television series *The Wire*, hip-hop lyrics from black American artists and tweets posted within the south-eastern corner of the United States. We do not use this data for training, but only for evaluation, so our experiments use unsupervised (or weakly supervised) domain adaptation.

Although the language use in the three domains vary, they have several things in common: the register is very informal, and the subtitles, lyrics and tweets contain slang terms such as "loc'd out", "cheesing with" and "po'", spoken language features such as "uh-hum, huh" and "oh", phonologically-motivated spelling variations such as "dat mouf" and "missin'", and contractions such as "we'll" and "I'd". These features are infrequent in or absent from most commonly used training corpora for NLP.

The data was annotated by two trained linguists with experience in analyzing AAVE, using the Universal Part-of-Speech tagset Petrov et al. (2011). They obtained an inter-annotator agreement score of 93.6%. The test sections consist of 528 sentences (subtitles), 509 sentences (lyrics), and 374 sentences (tweets). In addition, we had 546 sentences of subtitles annotated for development data. Note that we only use one domain for development to avoid overly optimistic performance estimates.

For all experiments, we use a publicly available implementation of structured perceptron[2] and train on the 1827 tweets from the CMU Twitter Corpus (Gimpel et al., 2011). Note that despite the fact that the training data also comes from an informal domain, the distribution of POS tags in this data set is different from those of the test sets. For instance, the percentage of determiners in the CMU Twitter corpus is on average 4% lower than in our test domains, and there are 7% more pronouns in the test sets than in the CMU Twitter corpus.

We also create a large unlabeled corpus of data that is representative of our test sets. This corpus, consisting of 4.5M sentences, is created using subtitles from the TV series *The Wire* and *The Boondocks*, English hip-hop lyrics, and tweets from the south-eastern states of the US. None of the unlabeled data overlaps with our evaluation datasets. We use this corpus for two purposes: to induce word clusters and embeddings, and to partially annotate a portion of it automatically, which we include in the training data of our ambiguous supervision model (see § 4.3 below).

4.3 Robust learning

Word representations To learn word embeddings from our unlabeled corpus, we use the Gensim implementation of the word2vec algorithm (Mikolov et al., 2013a,b). We also learn Brown clusters from a large corpus of tweets (Owoputi et al., 2013),[3] and add both as additional features to our training and test sets. The word representations capture latent similarities between words, but more importantly enable our tagging model to generalize to unseen words.

Partially labeled data Model performance generally benefits from additional data and constraints during training (Hovy and Hovy, 2012; Täckström

[2] https://github.com/coastalcph/rungsted
[3] http://www.cs.cmu.edu/~ark/TweetNLP/

et al., 2013). We therefore also use the unlabeled data and tag dictionaries as additional, partially labeled training data. For this purpose, we extract a tag dictionary for AAVE-like language from various crowdsourced online lexicons. Partial constraints from tag dictionaries have previously been used to filter out incorrect label sequences from projected labels from parallel corpora (Wisniewski et al., 2014; Das and Petrov, 2011; Täckström et al., 2013). We use a combination of a publicly available dump of Wiktionary (Li et al., 2012),[4] entries from Hepster's Glossary of Musical Terms,[5] a list of African-American names,[6] and Urban Dictionary (UD).[7] We augment our tag dictionary by scraping UD for all words in our unlabeled corpus and extracting the part-of-speech information where available. See an example entry for the word "hooch" below, which has five possible parts of speech in our tag dictionary: VERB, NOUN, ADJ, PRON, ADV.

> Hooch: Chewing tobacco commonly placed in the lower lip region. Hooch can be used as a verb, noun, adjective, pronoun, or an adverb.

We use the tag dictionary to label the unlabeled corpus. For instance, for the word "hooch", we assign it the label VERB/NOUN/ADJ/PRON/ADV. We present two ways of using this data for learning better POS models: one where the tag dictionaries are used in an ambiguously supervised setting, and one where they are used as type constraints at prediction time in a self-training setup.

Ambiguous supervision Our algorithm is related to work in cross-lingual transfer (Wisniewski et al., 2014; Das and Petrov, 2011; Täckström et al., 2013) and domain adaptation (Hovy et al., 2015b; Plank et al., 2014a), where tag dictionaries are used to filter projected annotation. We use the tag dictionaries to obtain partial labeling of in-domain training data.

Our baseline sequence labeling algorithm is the structured perceptron (Collins, 2002). This algorithm performs additive updates passing over labeled data, comparing predicted sequences to gold standard sequences. If the predicted sequence is identical to the gold standard, no update is performed. We use a

[4] https://code.google.com/p/wikily-supervised-pos-tagger/

[5] http://www.dinosaurgardens.com/wp-content/uploads/2007/12/hepsters.html

[6] http://www.behindthename.com/submit/names/usage/african-american/3

[7] http://www.urbandictionary.com

cost-sensitive structured perceptron (Plank et al., 2014b) to learn from the partially labeled data.

Each update for a sequence can be broken down into a series of transition and emission updates, passing over the sequence item-by-item from left to right. For a word like *hooch* labeled VERB/NOUN/ADJ/PRON/ADV, we perform an update proportional to the cost associated with the predicted label. If the predicted label is not in the mined label set, e.g., PRT, we update with a cost of 1.0 (multiplied by the learning rate α); if the predicted label is in the mined label set, we do not update our model. This means that the POS model is not penalized for predicting any of the five supplied labels. We did consider distributing a small cost between the candidates in the mined label sets, but this led to slightly worse performance on our development data.

In the experiments below, we also filter the partially labeled data by the amount of ambiguity observed in our labels. At one extreme, we require *all* words to have a single label, as in fully labeled data. Hovy et al. (2015b) also used a tag dictionary to obtain fully labeled data for domain adaptation. At the other end of the scale, we use all the partially labeled data, allowing up to 12 tags per words. Finally, we also experiment with using only sentences from our unlabeled data such that the tag dictionary assigns at most two (2) or three (3) labels to each word.

We also experimented with using different amounts of ambiguously labeled data.

The best performing system on development data uses both Wiktionary and the tag dictionaries associated with AAVE, only 100 ambiguously labeled data points for training, a cost of 0.0 for predicting labels in the mined label sets, no threshold on ambiguity levels (but leaving only sentences covered by our tag dictionaries), the CMU Brown clusters, and 20-dimensional word2vec embeddings with a sliding window of nine (9). The results of this system are shown in Table 4.1 as AMBIGUOUS.

Self-training with type constraints Our second system uses the harvested tag dictionary for type constraints when making predictions on the unlabeled data for self-training. The search space of possible labels for each word is simply restricted to the tags provided for that word by the tag dictionary.

For our self-training experiments, we experiment with pool size, but

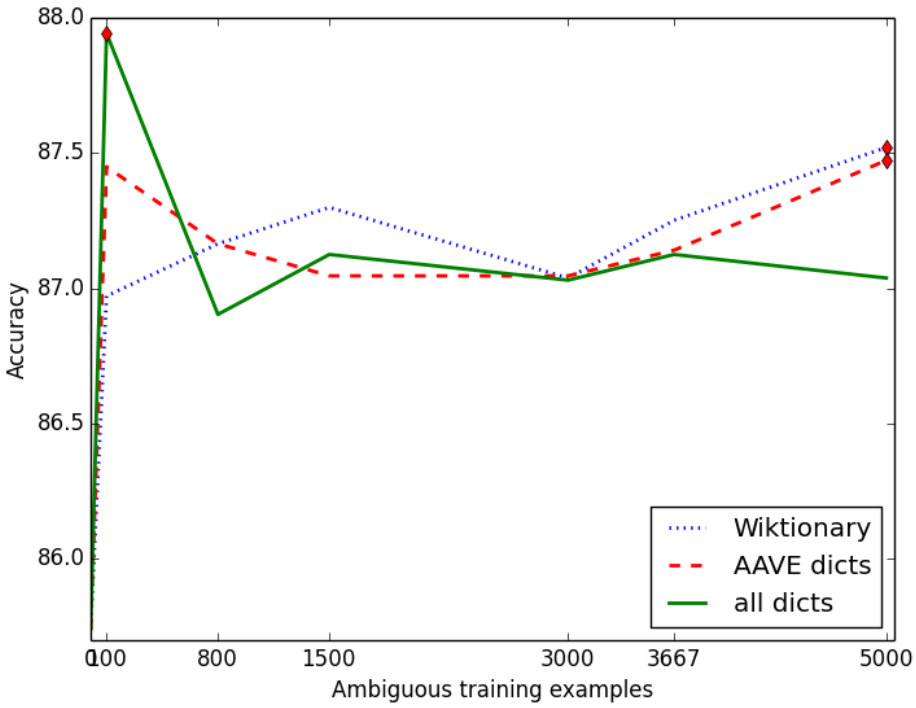

Figure 4.1: Learning curve for ambiguous learning.

heuristically set the stopping criterion to be when the development set accuracy of the tagger decreases over three consecutive iterations. we obtained the best performance on development data using the tag dictionary without Wikipedia, using all entries for type constraints, the CMU Brown clusters, and 10-dimensional embeddings with a window size of five (5). The results of this model are listed in Table 4.1 as SELF-TRAINING.

Pre-Normalization We also experimented with test-time pre-normalization of the input, using the normalization dictionary of Han and Baldwin (2011), but this led to worse performance on development data.

4.4 Results and error analysis

Table 4.1 shows the baseline accuracies, with and without clusters and embeddings, as well as the performance of the two developed systems described above. All results for both ambiguous supervision and self-training with type constraints significantly outperform the simple baseline with $p < 0.01$ (Wilcoxon). The system using ambiguous supervision is also significantly better than the baseline with clusters and word embeddings on the Twitter data. The fact that we generally see worse performance on Twitter data than on the two other data set (even though the systems were trained on Twitter data) can be attributed to a higher type-token ratio.

Models	Lyrics	Subtitles	Tweets (AAVE)	Average
Baseline	83.9	87.8	75.0	82.2
+ Cluster	85.0	88.4	79.0	84.1
+ Cluster+ Emb.	**85.2**	**89.0**	78.8	84.3
Ambiguous	**85.2**	**89.0**	**83.0**	**85.7**
Self-training	85.0	88.8	80.0	84.6
Stanford	77.7	83.0	61.4	74.3
Gate	83.0	87.5	77.1	80.0
Cmu	81.5	85.6	80.0	82.4

Table 4.1: Performance results.

We also provide the accuracies of three publicly available POS taggers in Table 4.1. The three POS systems are the bidirectional Stanford Log-linear POS Tagger,[8] the GATE Twitter POS tagger,[9] and the CMU POS Tagger.[10] We observe that our ambiguous learning system outperforms all three systems on all test sets.

Our improvements are primarily due to better performance on out-of-vocabulary (OOV) words. Both systems improve the accuracy on OOV items for all three test sets, with the ambiguous learning system reducing the error by an average of 14%, and the self-training system reducing it by 7.7% on average. However, we also see an average increase in performance on known

[8] http://nlp.stanford.edu/software/tagger.shtml
[9] https://gate.ac.uk/wiki/twitter-postagger.html
[10] https://github.com/brendano/ark-tweet-nlp/

words of 1% for both systems. This increase is highest for tweets (2%) and around 0.5% for the subtitles and hip-hop lyrics test sets. The main reason for the increased overall performances of our systems is therefore the improved accuracy on OOV words. Table 4.2 shows that the accuracy on OOVs increases on all three test sets for both developed systems over baseline.

	Lyrics	Subtitles	Tweets	Average
Baseline	64%	78%	48%	63%
Ambiguous	**71%**	**83%**	**78%**	**77%**
Self-train	70%	82%	61%	71%

Table 4.2: Accuracies on out-of-vocabulary (OOV) words .

The OOV words learned in these two test sets are mainly verbs such as "sittin'", "gettin'" and "feelin'" (g-dropped spellings), and words that are infrequent in canonical written language such as "'em".

We observe that our systems improve performance on traditionally closed word classes such as pronouns, adpositions, determiners and conjunctions. These increases can be ascribed to the systems having learned from the additional information provided on spelling variations such as "'cause", "fo'" and "ya" and unknown entities such as "dis", "dat", "sum".

Finally, we note that increasing the number of training examples for ambiguous learning seems to come with diminishing returns. The learning curve is presented in Figure 4.1.

4.5 Conclusions

We explore several techniques to learn better POS models for AAVE-like subtitles, lyrics, and tweets from a manually annotated Twitter corpus. Our systems perform significantly better than three state-of-the-art POS taggers for English, with error reductions up to 55%. The improvements were shown to be primarily due to better handling of OOV words.

Chapter 5

Evaluation of Summarization Systems across Gender, Age, and Race

Abstract

Summarization systems are ultimately evaluated by human annotators and raters. Usually, annotators and raters do not reflect the demographics of end users, but are recruited through student populations or crowdsourcing platforms with skewed demographics. For two different evaluation scenarios – evaluation against gold summaries and system output ratings – we show that summary evaluation is sensitive to protected attributes. This can severely bias system development and evaluation, leading us to build models that cater for some groups rather than others.

5.1 Introduction

Summarization – the task of automatically generating brief summaries of longer documents or collections of documents – has, so it seems, seen a lot of progress recently. Progress, of course, is relative to how performance is measured. Generally, summarization systems are evaluated in two ways: by comparing machine-generated summaries to human summaries by text similarity metrics (Lin, 2004; Nenkova and Passonneau, 2004) or by human rater studies, in which participants are asked to rank system outputs. While using similarity metrics is controversial (Liu and Liu, 2008; Graham, 2015; Schluter, 2017), the standard way to evaluate summarization systems is a combination of both.

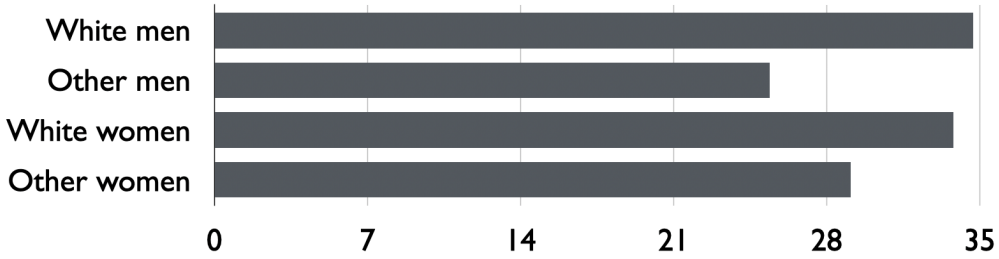

Figure 5.1: **Social bias in automatic summarization**: We take steps toward evaluating the impact of the gender, age, and race of the humans involved in the summarization system evaluation loop: the authors of the summaries and the human judges or raters. We observe significant group disparities, with lower performance when systems are evaluated on summaries produced by minority groups. See §3 and Table 5.1 for more details on the Rouge-L scores in the bar chart.

Both comparison to human summaries and the use of human raters naturally involve human participants, and these participants are typically recruited in some way. In Liu and Liu (2008), for example, the human subjects are five undergraduate students in Computer Science. Undergraduate students in Computer Science are not necessarily representative of the population at large, however, or of the end users of the technologies we develop. In this work, we ask whether such sampling bias when recruiting participants to evaluate summarization systems, is a problem? In other words, do different demographics exhibit different preferences in rater studies of summarization

systems? NLP models are only fair if they do not put certain demographics at a disadvantage (Larson, 2017), and it is therefore crucial our benchmarks reflect preferences and judgments across those demographics (Ethayarajh and Jurafsky, 2020).[1]

Contributions

We present the, to the best of our knowledge, first in-detail evaluations of summarization systems across demographic groups, focusing on two very different summarization systems – TextRank (Mihalcea and Tarau, 2004) and MatchSum (Zhong et al., 2020). The groups are defined by the three protected attributes: gender, age, and race. While the systems are reported to perform very differently, we show that the system rankings induced by performance scores or user preferences differ across these groups of human summary authors and summary raters. We analyze what drives these differences and provide recommendations for future evaluations of summarization systems.

5.2 Experiments

We present two evaluations in this short paper: an **automated scoring against human summaries** (Exp. A) and a **human rater study** (Exp. B). In both experiments, we use Amazon Mechanical Turk to recruit annotators from different demographic groups, and the first paragraphs of biographies from English Wikipedia as our input data, using the Wikidata API for extraction.[2] We create a dataset of biographies of women and men, obtain human summaries, and generate summaries of these biographies using two out-of-the-box extractive summarization systems. In Exp. A, we compare the system summaries directly to the human summaries (from different groups); in Exp. B, we let human raters compare and rate the two system summaries. To ensure differences between the two summarization systems, we use the 2004 graph-based TextRank Mihalcea and Tarau (2004) and the

[1] We thereby challenge the widely held position that lay people cannot be used for summary evaluation, because they exhibit divergent views on summary quality (Gillick and Liu, 2010). We, in contrast, believe such variance is a product of social differences and something we need to worry about in NLP.

[2] https://query.wikidata.org/

Gender	Race	Rouge-1	Rouge-L
♀		0.407	0.326
♂		0.417	0.326
♀	White	0.418	0.338
	non-White	0.371	0.291
♂	White	**0.436**	**0.347**
	non-White	0.347	0.254

Table 5.1: Automated scoring of MatchSum (Zhong et al., 2020) across self-reported protected attributes: **Gender**, with values ♀, ♂ (all our test subjects identified as one of the binary genders), **race**, binarized here as White and non-White (in order to achieve rough size balance). The ROUGE scores of MatchSum are clearly higher when evaluated against reference summaries created by White men. We also considered **age** (binarized as ±30, to achieve size balance): Here we see slightly better performance when evaluated against summaries of older participants across the binary genders.

2020 state-of-the-art, BERT-based MatchSum (Zhong et al., 2020).[3] We follow the MatchSum guidelines described in (Zhong et al., 2020) and limit the length of the input biographies to a maximum 5 sentences and force the output summaries to be between 2-3 sentences long. Our final dataset consists of the original 975 biographies (700 men and 275 women), along with two automatic summaries, as well as human 3 sentence summaries, and is made freely available.[6]

Our evaluations rely on annotations and ratings from Amazon Mechanical Turk. For quality control, we rely on a control question, as well as analyzing

[3]We use the implementation of TextRank by Barrios et al. (2016)[4] and the original MatchSum implementation.[5] MatchSum obtains state-of-the-art performance across a range of benchmarks by learning to produce summaries whose document encoding is similar to that of the input document. TextRank is a much simpler extractive algorithm; it adopts PageRank to compute node centrality recursively based on a Markov chain model. While MatchSum obtains a Rouge-1 score of .44 on CNN/Daily Mail, TextRank obtains a Rouge-1 score of .33 (Zheng and Lapata, 2019). We use both systems with recommended parameters, as was done in Zheng and Lapata (2019). Note that TextRank, in contrast to MatchSum, is unsupervised. Our Rouge-1 scores below for Wikipedia biographies are generally comparable.

[6]https://github.com/ajoer/summary_preferences

annotation time: If a task is completed faster than one standard deviation of the average time spent, the answers in that task are discarded. We collected one manual summary and two system rankings per biography, resulting in 3,135 annotations.

Human summaries

In Exp. A, participants were asked to enter the three most important sentences in the document and in three blank text fields; for quality control, we check that these sentences occur in the input document. We collect a total of 1,185 summaries, 53% of which are written by women (0.5% identified neither as male or female). 74% of summaries are written by participants older than 30 years of age. 76% identified as White; 11% as Black; 5% as American Indian/Alaska Native (AI/AN); 4% as Asian, and 4% as Hispanic.[7] We binarize race as White and non-White to achieve rough size balance across groups. Aggregating scores across multiple races is not ideal, but by doing so, we compensate for poor representation of some demographics.

Rater study

In Exp. B, we present participants with two 2-3 sentence machine summaries and ask them to a) pick their preferred summary and b) rank the two summaries on 4-point forced Likert scales, for fluency, informativeness and usefulness. 40.2% of our raters identified as female. 37.5% were below 30 years of age. 70.8% of ratings identified as White, the rest as AI/AN (2.3%), Asian (3.5%), Black (19.1%), Hispanic (2.0%), or as other (2.2%).

We ask all participants to voluntarily submit their race and gender information, and require that they be US-based. We asked the participants in the rater study to also include age information.

[7]Our race taxonomy was standard, based on https://www.census.gov/prod/2001pubs/cenbr01-1.pdf, but all annotators identified as either AI/AN, Asian, Black, Hispanic, or White.

Gender	Age	TextRank	MatchSum	N/A
♀	≥30	0.379	0.565	0.056
	<30	**0.481**	**0.454**	0.065
♂	≥30	0.397	0.511	0.092
	<30	0.396	0.531	0.073

Table 5.2: System ratings across participant gender and age. We highlight the outlier: Younger women significantly preferred TextRank over MatchSum ($p < 0.01$).

Age	Race	TextRank	MatchSum	N/A
<30	Asian	34.1	39.0	26.8
	Black	**49.0**	**43.1**	7.8
	Hispanic	40.7	59.3	0.0
	White	43.6	53.5	2.9
≥30	AI/AN	40.0	51.3	8.7
	White	43.6	53.5	2.9

Table 5.3: System ratings across participant race and age. We highlight the outlier: Young Black people significantly preferred TextRank over MatchSum ($p < 0.01$). AI/AN refers to American Indian/Alaska Native.

5.3 Results

In Table 5.1, we present the results of Exp. A: Rouge-1 and Rouge-L results are significantly better when evaluated on summaries produced by White men than when evaluated on summaries produced by any other group. MatchSum summaries also align better with those written by White women compared to those written by non-White women. Generally, MatchSum aligns better with men than with women.

Exp. 2 includes three demographic variables (gender, age, and race). Table 5.2 presents ratings across gender and age. Most participants prefer the reportedly superior system (with a Rouge-1 advantage of 0.11 on a standard benchmark; see §2), but younger women significantly preferred TextRank over MatchSum ($p < 0.01$). Table 5.3 presents the ratings across age and

race. Here, we again find a single outlier group: Younger Black people significantly prefer TextRank over MatchSum ($p < 0.01$). Our results imply that our standard evaluation methodologies do not align with the subjective evaluations of younger women and younger Black people.

We try to explain these two observations in §5.

We checked for significant group rating differences using bootstrap tests (Efron and Tibshirani, 1994; Dror et al., 2018). Across 1000 rounds, with Bonferroni correction, we find significant ($p < 0.05$) differences in preferences for these groups: ≥30, AI/AN[8], White ♂, AI/AN ♀, ≥30 ♂, Asian< 30, Asian< 30♂, White≥30♂, and AI/AN ≥ 30♀. All these subdemographics exhibit significantly different ranking behavior from their peers. So, for example, our results show a significant difference between young and old raters.

We also bin our results by gender of the subjects of the biographies. We rely on Wikidata gender information to make this classification. There are 1409 preferences and ratings of men's biographies (Men), and 585 of biographies of women (Women). This of course means we see fewer significant differences in ratings of female biographies. For Men, we find significant differences across a wide range of groups, and with stronger effects for some demographics, suggesting that the gender of the subject of the biography *does* impact ratings differently across subdemographics. We find significant results for Women only for the subdemographic White ($p = 0.004$). This result is interesting, though, since it shows that on female biographies, White and non-White annotators prefer different systems.

Finally, we also asked our annotators to rank the two systems based on fluency, informativeness and usefulness. We used a 4-point forced Likert scale. One observation is that even across fine-grained dimensions, younger annotators rate summaries lower; see Table 5.4. Interestingly, however, this difference is only observed with female biographies (rows 3–6). See Table 5.5 for the results on all biographies across race. While ratings are generally low, we see clear differences, with Hispanics finding TextRank significantly more informative and useful, and AI/AN finding TextRank significantly more fluent. Interestingly, Hispanics exhibit significant differences across female and male biographies, finding TextRank summaries of female biographies significantly more informative and useful than TextRank summaries of male biographies.

[8]American Indian/Alaska Native (AI/AN)

		Informative		Useful		Fluent	
	Age	T	M	T	M	T	M
All	≥30	0.94	0.96	0.94	0.96	0.9	0.95
	<30	0.77	0.81	0.72	0.79	0.81	0.83
♂	≥30	0.88	0.92	0.86	0.91	0.84	0.89
	<30	0.86	0.9	0.82	0.89	0.85	0.91
♀	≥30	0.89	0.91	0.88	0.92	0.88	0.91
	<30	0.83	0.84	0.8	0.83	0.86	0.83

Table 5.4: Rater study results by age, on all biographies, as well as on biographies of men (♂) and women (♀) only.

	Informative		Useful		Fluent	
Race	T	M	T	M	T	M
AI/AN	0.5	0.6	0.7	0.7	**1.2**	1.0
Asian	0.7	1.0	0.8	0.9	1.0	0.8
Black	0.7	0.8	1.0	0.8	0.9	0.8
Hispanic	**1.4**	0.9	**1.5**	**1.2**	0.9	1.0
White	0.8	0.7	0.8	0.8	0.9	0.8

Table 5.5: Rater study results by race on all biographies. AI/AN refers to American Indian/Alaska Native.

5.4 Analysis

In order to analyze the differences between the rating behavior of subdemographics, we learn which features are significant for each demographic by training a simple logistic regression text classifier trained on the summaries ranked by each of the subdemographics with significantly different ranking behavior. As task representation, we represent each ranking instance as a vector of 2*149 features, one 149-sized subspace for each summary. Each subspace is made up of a one-hot vector of 145 frequent words (from the English stop words list in NLTK[9]), as well as four task specific features: the summary's average word length, whether the first sentence of the biography is included in the summary, the type/token ratio, and the text complexity of the summaries. We concatenate the 149 features from each system and scale them. We extract the top 20 most salient features for each demographic group and analyze them manually:

The average word length of the MatchSum system correlates positively with annotators preferring MatchSum across several demographics, e.g., ≥30 and White ♂, but this effect is absent with female annotators. Since the inductive bias of TextRank does not explicitly prohibit redundancy (Mihalcea and Tarau, 2004), this finding indicates that MatchSum is preferred among older men, especially those who identify as White, when it is informative, introduces main entities, etc. However, other subdemographics seem less sensitive to this variation. MatchSum is *not* generally rated more informative and useful across demographics (Table 5.5). In other subdemographics, e.g., AI/AN, MatchSum summaries with pronouns are rated higher, indicating it is better than TextRank at extracting sentences with pronouns without breaking coreference chains. Referential clarity, e.g., dangling pronouns, is a known source of error in summarization (Pitler et al., 2010; Durrett et al., 2016). TextRank summaries are often preferred by AI/AN and Asian, when they include negation. This is unsurprising, since negated sentences can often be very informative, and may seem more sophisticated in the context of machine-generated summaries. Negation is also a known source of error (Fiszman et al., 2006). In our data, however, this effect varies across subdemographics.

Our main observation is that female and Black participants under 30 prefer TextRank over MatchSum. What drives this? The main predictors

[9] nltk.org

in our logistic regression analysis are a) TextRank extracting the first sentence of the biography (twice as frequently than MatchSum, in more than half of its summaries); and b) TextRank sentences containing negation. The former suggests a need for anchoring or framing of the summary, as initial sentences tend to provide this; the latter could suggest that young female or Black participants are less prone to the common bias of evaluating negated sentences as less important (Kaup et al., 2013).

5.5 Conclusion

Our paper is, as far as we know, the first to evaluate summarization systems across different subdemographics. We did so in two different evaluation scenarios: automatic evaluation against gold summaries and system output ratings by human evaluators. We made the gold summaries and the ratings available for future research.

What did we learn from our experiments? Most importantly, of course, we learned that performance numbers differ when evaluated on summaries written by different subdemographics, and that the preferences of raters from different subdemographics differ. In our experiments with automatic evaluation against gold summaries written by different subdemographics, we saw that summarization systems achieve higher performance scores when evaluated on summaries produced by White men, highlighting an unfortunate bias in these systems. In our rater studies, we also saw significant differences across subdemographics. Most surprisingly, perhaps, we saw that a summarization system from 2004 was rated better than a state-of-the-art system from 2020 by some subdemographics, and effect that was found to relate to the occurrence of first sentences (providing anchoring or framing of summaries) and negation (often evaluated as less important by majority groups). For now, we can only speculate what a summarization system optimized to perform well across *all* subdemographics would look like, e.g., a system minimizing the worst-case loss across subdemographics rather than the average loss. Our results show very clearly, however, the current state of the art in summarization is biased toward some demographics and therefore fundamentally unfair.

Ethics Statement

We present two evaluations of summarization systems in which we bin participants by gender, age, and race. All demographic information was self-reported, and we payed annotators equally who chose *not* to report this information. Our work highlights the importance of recruiting balanced pools of participants in evaluations of summarization systems, an issue that has previously been ignored. A major limitation of this work is the under-representation of some groups, which led us to binarize all three social variables. We think of this study as a first attempt to highlight an important issue and hope that others will follow up with large-scale studies with better representation for more groups. Such studies could include many other social variables, e.g., income or level of education.

Chapter 6

Group Fairness in Multilingual Speech Recognition Models

Abstract

We evaluate the performance disparity of the Whisper and MMS families of ASR models across the VoxPopuli and Common Voice multilingual datasets, with an eye toward intersectionality. Our two most important findings are that model size, surprisingly, correlates logarithmically with worst-case performance disparities, meaning that larger (and better) models are less fair. We also observe the importance of intersectionality. In particular, models often exhibit significant performance disparity across binary gender *for adolescents*.

6.1 Introduction

Automatic speech recognition (ASR) has improved greatly, largely due to representation learning from raw audio. Data scarcity is no longer a major bottleneck for many of the world's languages, and high-quality speech recognition models become more and more integrated in both our private and public lives: From automatically transcribing court proceedings or doctor's notes, to extracting speech from police patrolling, meetings or for hearing aids, speech recognition models have the potential to ease many of the mundane but important tasks we perform on a daily basis.

Performance of ASR models has been shown to vary substantially across user groups Koenecke et al. (2020); Martin and Tang (2020); Ngueajio and Washington (2022). Partial mitigation of performance disparities across user groups is sometimes possible, through distributionally robust optimization Sagawa* et al. (2020) or spectral decoupling Pezeshki et al. (2020), for example, but is computationally expensive and requires large amounts of data annotated with demographic information, e.g., protected attributes of speakers. In this study, we show how *small* amounts of such data can be used to evaluate performance disparities, benchmarking two state-of-the-art ASR architectures across languages and demographics.

Our purpose is twofold; (i) We want to know what the performance is for protected groups across a variety of speech models; and (ii) we want to create a baseline for other ASR models to compare against. We believe (i) is extremely important, because of the large-scale impact of these models on our everyday lives and the societal imbalances they can reinforce. Establishing a practice around fairness evaluation is important also for future generations of ASR to ensure that benefits are equally distributed across user groups.

Protected Attributes Protected attributes refer to demographic characteristics of individuals such as race, gender, age, and religion, which are considered protected from being used as a basis for discrimination or bias in decision-making processes. In the context of data and machine learning, the consideration of protected attributes becomes crucial to ensure fairness and prevent biases in automated decision-making systems.

If an ASR system is not trained to handle linguistic variation, the system may exhibit much higher error rates for individuals with certain protected attributes –especially if these are correlated with particular accents, dialects, or

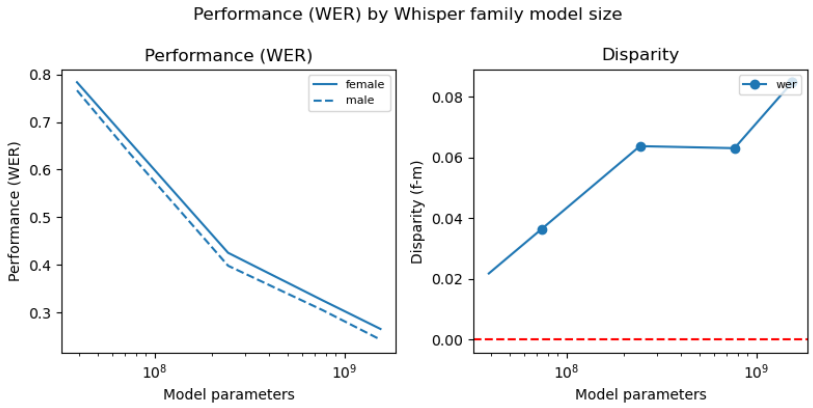

Figure 6.1: Model performance per binary gender (left) and disparity (right) as a function of model size (log-scale). Dots indicate significant performance disparity ($p<0.05$).

speech patterns, as is the case of African-American Vernacular English. This can disproportionately impact individuals from specific linguistic or cultural backgrounds, leading to unfair outcomes and reduced accessibility for those groups and a perpetuation of existing societal biases and discrimination.

Intersectionality Intersectionality highlights the interconnected nature of multiple social identities and the ways in which they intersect to shape individuals' experiences and social inequalities. The concept of intersectionality acknowledges that systems of oppression, discrimination, and privilege are multidimensional and overlapping, and cannot be understood by considering individual identity categories in isolation. For example, a person who identifies as both a woman and a person of color may face distinct forms of discrimination and marginalization that are different from those experienced by a white woman or a person of color who identifies as a man.

In the following, we present the data and ASR models we consider for investigating performance disparity between the binary genders and the intersectionality of age and binary gender.

6.2 Datasets

We make use of two multilingual, open source datasets to evaluate the performance disparity of the two families of ASR models with respect to gender fairness and intersectionality in gender and age. Common Voice[1] is a crowdsourced, continuously developed dataset covering over 200 languages and VoxPopuli[2] is a collection of speeches given in the European Parliament between 2009–2020. Both datasets contain demographic information about the speakers; CommonVoice has gender and age annotations, VoxPopuli has gender markings as well as timestamps for each utterance.

Limitation and Code To our knowledge, no available open source dataset exists that would allows us to test the performance disparity for other attributes than binary gender and age or the intersectionality of other attributes than these two. Likewise, we have not been able to find data to analyse the entirety of the deflate we are limited by data scarcity on We redistribute the processed datasets to facilitate hassle-free fairness evaluation, along with open-source evaluation code for testing and visualizing results.[3]

Evaluating Public Models

We evaluate two publicly available ASR model families, namely the Whisper Radford et al. (2022) and MMS Pratap et al. (2023) models, i.e., a total of eight models. Both model families consist of multilingual, multitask models. They are also easily accessible models and go-to models for hundreds of companies using ASR in their products.

6.2.1 Whisper

Whisper is a family of automatic speech recognition (ASR) systems developed by OpenAI. The models are trained on 680,000 hours of web data in 97 languages, and they have parameters ranging from 39M in the 'tiny' model to 1550M in the two 'large' models. Whisper training involves data augmentation, applying transformations to the audio spectrograms during training,

[1] https://huggingface.co/datasets/mozilla-foundation/common_voice_12_0
[2] https://huggingface.co/datasets/facebook/voxpopuli
[3] https://github.com/marcvanzee/asr-fairness

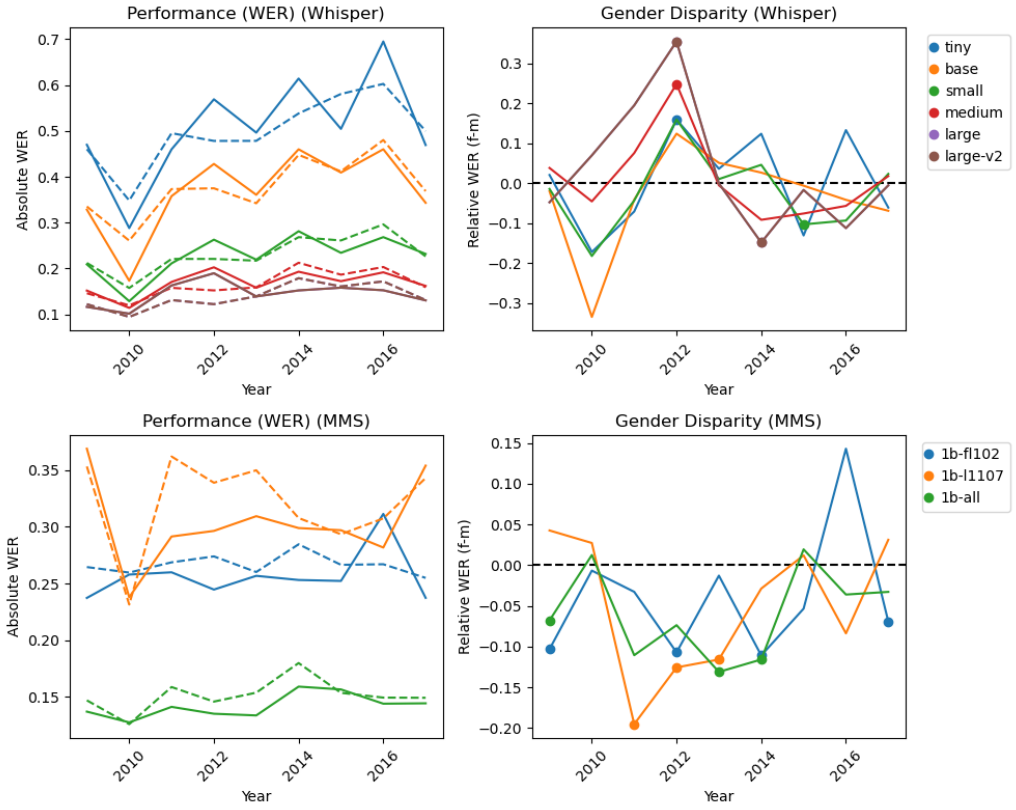

Figure 6.2: Word error rate (WER) and gender disparity in ASR models for binary genders across years. Left column shows performance results (WER) for Whisper (top) and MMS (bottom) families, and right column is the gap in performance between the binary genders for Whisper (top) and MMS (bottom). Solid lines show performance for female speakers, dashed lines for male. Dots indicate significance ($p \geq 0.01$).

including time warping, frequency masking, and time masking. Such data augmentation strategy helps the model generalize better to different acoustic conditions.

6.2.2 MMS

The Massively Multilingual Speech (MMS) family of ASR models are developed by Meta and trained on 500,000 hours of speech data in 1400+ languages. Based on wav2vec 2.0 models Baevski et al. (2020), MMS leverages self-supervised methods for learning from a large, new corpus of religious texts. The models all have 1B parameters, but they have been fine-tuned on different datasets.

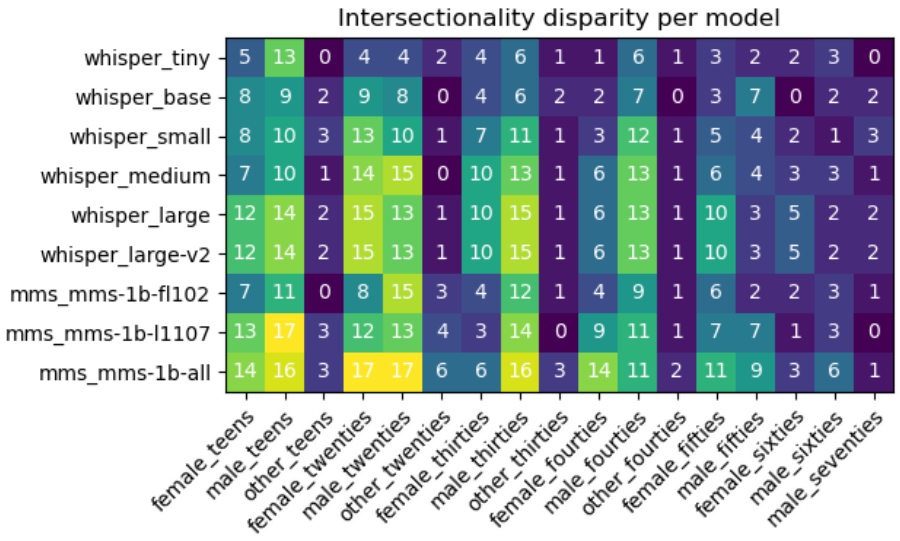

Figure 6.3: Intersectionality results. We report the number of statistically significant ($p < 0.05$) performance disparities for a particular pair of demographic variables. Performance, again, is measured across multiple languages. We see that, on average, larger models exhibit more intersectionality effects, and we clearly see more disparate performance among younger speakers who identify as men.

6.3 Results

We evaluate all models in the Whisper family (of different sizes) and all models in the MMS family (of different training data) across all demographics in all languages in our two datasets. This is a total of **651** experiments. We then run significance test on all combinations of language, dataset, model, and model size (for the Whisper family). We find significant disparity in performance between the binary genders in 29% of the cases (11% of these negatively for women, 17% for men).

Performance disparity is prevalent across languages and across models, and it seems that model size correlates positively with such disparity (Figure 6.1). Here, we plot the results with model size on the x-axis, and relative disparity difference on the y-axis. We see that there is a positive, logarithmic correlation between the two variables. Figure 6.3 shows how gender disparities are particularly high for younger speakers who identify as men. These results showcase how inferring a model's fairness from its parity on data from one demographic group, e.g. adult users, is insufficient.

6.3.1 A Closer look at Spanish

We zoom in and take a closer look at the performance of the Whisper family models on 7 Spanish dialects.[4] We use the CommonVoice dataset, where gender, age, and dialect are marked for 1829 speakers.

First, we look at the overall performance of the Whisper model family on the Latin American Spanish dialects and on Iberian Spanish (see Figure 6.4).[5] We note that performance for all dialects increases (WER decreases) as model size increases, but that the performance is disparate for speakers of River Platean Spanish independent of model size. Performance is not disparate between genders across Spanish.

We then plot the intersectional performance disparity between binary genders *within* each dialect as a function of model size in Figure 6.4b, ie. female speakers of Mexican Spanish against non-female speakers of Mexican Spanish. We see that while performance increases (lower WER) for all

[4]We exclude the MMS family from this analysis since their performance on Spanish is too poor. The best MMS model (1b-all) is on par with the worst performing Whisper model (tiny).

[5]We group the Iberian Spanish dialects together and focus on the Latin American Spanishes in line with the NAACL 2024 theme track.

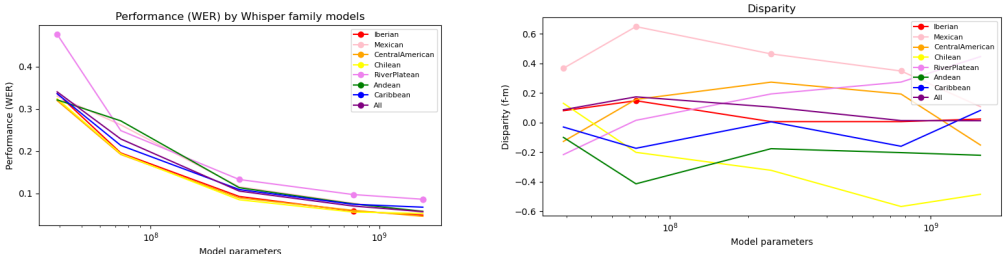

Figure 6.4: Model performance per dialect (left) and performance disparity between genders *within* a dialect (right) as a function of model size (log-scale). Dots indicate significant performance disparity ($p<0.05$).

dialects as the model size increases, gender disparity exists in all dialects (except perhaps Iberian), and there is no clear improvement in gender disparity within a dialect when model size increases.

Finally, we investigate the performance disparity across three-tonged intersectional groups with gender, age and accent (see Figure 6.5). Performance disparity between intersectional groups intensifies with model size, and particularly, female Mexican speakers under 40 and male speakers of Andean under 40 suffer from disparate intersectional performance along with female speakers in their sixties who speak River Platean Spanish. These findings support the two-tonged intersectional results (age and dialect), but indicate that particular age groups are affected by the disparate performance results.

6.4 Discussion

Mitigation Some researchers have reported on attempts to make ASR systems less disparate. Boito et al. (2022) report that training ASR models for specific demographic groups did *not* reduce performance disparity. Such strategies also have trouble scaling in light of intersectionality. Veliche and Fung (2023) propose conditioning on cluster IDs with clusters being proxies for demographic groups, but their approach is not easily integrated in pre-trained ASR models such as Whisper and MMS. Dheram et al. (2022) had limited success with oversampling from minority groups.

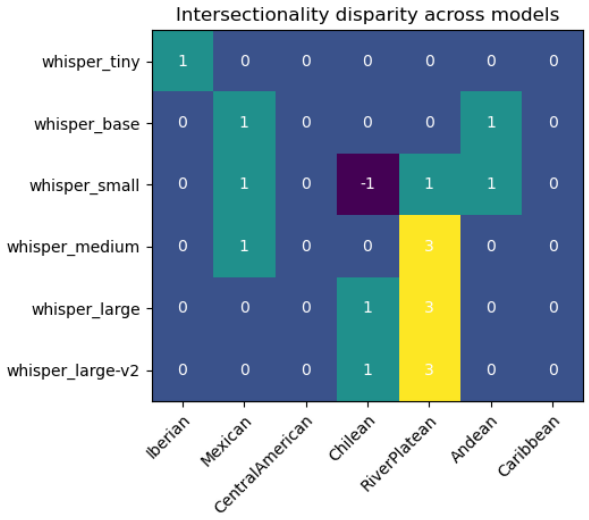

Figure 6.5: Three-tonged – age, gender, accent – intersectional performance results for Spanish dialects across different Whisper family model sizes. A negative result indicates positive disparate performance.

Fairness over Time In ASR research, the predominant focus has been on examining fairness within a static framework, where it is assumed that the data generation process remains constant over time. Nevertheless, these approaches tend to overlook the significant drift in data over time, a phenomenon frequently observed in real-world scenarios. How people talk, and what they talk about, changes over time. What specific demographics talk about changes even faster.

Preliminary investigations have revealed that enforcing static fairness constraints in dynamic systems can lead to inequitable data distributions and, in some cases, exacerbate existing biases Søgaard et al. (2021). Furthermore, the emergence of powerful large-scale generative models has brought to the forefront the necessity of comprehending fairness within evolving systems. The widespread deployment and versatile capabilities of these models raise a crucial question: how can we assess these models for fairness and effectively mitigate observed biases from a long-term perspective?

As a small step in what we take to be the right direction, we also examined how the performance disparities of Whisper and MMS evolve over time. Since the models are trained on data from the entire period (2009–2017), our

protocol does not simulate evaluation on future data, only variance across time. See Figure 6.2 for an overview. We see a small effect as we depart from the period's average, but with high general variance. The smallest disparities are observed in 2010, 2013, and 2015. Since the VoxPopuli is a collection of speeches from the European Parliament, it is likely that we would see larger variance in datasets from less formal settings. We encourage other researchers to seek out or develop new datasets that can give insights into the variance in performance over time.

6.5 Conclusion

We highlight the potential social impact of ASR's performance disparities across demographic groups in the –to our knowledge– first study of its kind. We run a total of **651** experiments evaluating state-of-the-art model families on data containing protected attributes, namely binary gender and age. We release the curated dataset to ease implementation of disparity testing for researchers and developers. Our main findings were as follows: (i) Larger models surprisingly exhibit more performance disparity. (ii) Intersectional effects are evident, largely affecting the younger speakers who identify as men. (iii) Finally, we see small signs of temporal variation in disparity figures, but less dramatic than the variation observed across protected attributes.

Chapter 7

On the Independence of Association Bias and Empirical Fairness in Language Models

Abstract

The societal impact of pre-trained language models has prompted researchers to probe them for strong associations between protected attributes and value-loaded terms, from slur to prestigious job titles. Such work is said to probe models for *bias or fairness*—or such probes 'into representational biases' are said to be 'motivated by fairness'—suggesting an intimate connection between bias and fairness. We provide conceptual clarity by distinguishing between association biases (Caliskan et al., 2022) and empirical fairness (Shen et al., 2022) and show the two can be independent. Our main contribution, however, is showing why this should *not* come as a surprise. To this end, we first provide a thought experiment, showing how association bias and empirical fairness can be completely orthogonal. Next, we provide empirical evidence that there is no correlation between bias metrics and fairness metrics across the most widely used language models. Finally, we survey the sociological and psychological literature and show how this literature provides ample support for expecting these metrics to be uncorrelated.

7.1 Introduction

The prevalence of unintended social biases in pre-trained language models (PLMs) is alarming, since they impact millions, if not billions of people every day. In recent years, more and more NLP researchers have studied such biases, making up an estimated 6.3% of the literature in 2022 (Ruder et al., 2022). Much of this work has focused on what Crawford (2017) called *representational bias*, which manifests when portrayals of certain demographic groups are discriminatory. In NLP, representational bias often arises when associations between a protected attribute, e.g., gender, and certain concepts, e.g., job titles, are captured in the model space. Thus, to avoid ambiguity, we will refer to this type of bias as *association bias*, following Chaloner and Maldonado (2019).

Association bias is often confused with what is sometimes referred to as performance disparity (Hashimoto et al., 2018) or *empirical fairness* (Shen et al., 2022), i.e., performance differences across end user demographics. Or mitigating association bias is assumed to improve empirical fairness (Chen et al., 2020; Friedrich et al., 2021; Cao et al., 2022; Dayanik and Padó, 2020; Castelnovo et al., 2022; Reddy et al., 2021). Note that most fairness metric focus on some form of equal performance and differ only in whether they focus on precision, recall or balancing the two (Barocas et al., 2019). Empirical fairness refers to equal performance as measured by *de facto* standard metrics and is arguably the most common fairness metric (Williamson and Menon, 2019; Barocas et al., 2019).

In this paper, we will show that the two phenomena, association bias and empirical fairness, are often completely independent matters.[1] We devise a thought experiment (§7.3) to illustrate this, but also present a series of experiments (§7.4) to show that results obtained the way association bias is normally measured, do not correlate with results obtained the way empirical fairness is normally measured.

Our main contribution, however, is to show that this should not come

[1]Note that the distinction between association bias and empirical fairness—between how expressions referring to demographic groups are encoded, and how these groups are treated as end users—is different from another distinction made in recent work (Delobelle et al., 2021; Goldfarb-Tarrant et al., 2021; Kaneko et al., 2022) between intrinsic and extrinsic bias: Intrinsic bias, here, is what we call representational bias, whereas extrinsic bias refers to performance differences on sentences containing entities referring to different demographic groups.

as a surprise. Research on mitigating association bias and empirical fairness is often motivated by fairness concerns, and bias and fairness are often considered near-synonymous terms in the research literature: Researchers have, for example, said that bias *causes* unfairness (Chang et al., 2019; Friedrich et al., 2021; Castelnovo et al., 2022). If this was the case, the independence of association bias and empirical fairness should come as a great surprise. However, the assumption that bias causes unfairness, is unwarranted, as we will see below, from a survey of relevant literature from the social sciences (§7.5). A causal link between association bias and empirical fairness would seem to require some sort of in-group affinity, i.e., that groups use terms relating to their in-group peers more and in different ways than outsiders, like, for instance, Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts (Duijnhoven, 2018). This assumption, which we call the In-Group Affinity Assumption, seems intuitive, but without much support from the social sciences (§7.5).

Contributions In §7.2, we define association bias and empirical fairness and discuss related work. When we talk about association bias, we refer to systematic biases in how words and phrases referring to demographic groups are encoded. Figure 7.2 visualizes how models may exhibit biased associations because of sample biases, and may even amplify these. We define empirical fairness as equal performance across groups, because this is the most balanced and most widely applicable measure of fairness in NLP, except for specialized applications where equal base rates and calibration take priority over performance. We then move to study how association bias and empirical fairness relate. In §7.3, we show that theoretically, association bias and empirical fairness are completely independent. That is, mitigating association bias can hurt empirical fairness, and ensuring empirical fairness can introduce more bias. §7.4 shows there is no obvious correlation between results obtained from standard association bias measurements and results obtained from standard empirical fairness measurements of language models. Finally, §7.5 surveys the social science literature for explanations on why association bias and empirical fairness may be less related (or related in less obvious ways) than multiple works in the NLP literature have assumed up to this point. The finding that association bias and empirical fairness are independent in this three-way investigation, should help push research horizons and provide strong motivation for targeting empirical fairness directly,

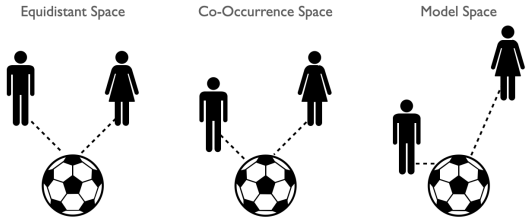

Figure 7.1: Association bias of group-related terms (e.g., *woman* and *man*) can be quantified as degree of isomorphism relative to an empirical (**co-occurrence**) space or a normative, **equidistant** space. The graph illustrates how *man* may be more strongly associated with *soccer* in a **model**, less so empirically (the underlying data or real-world statistics), and not at all in an ideal world.

as well as for seeing association bias mitigation, not necessarily as a way of promoting fairness, but rather as a way of preventing poor inferences and generation of stereotypical text.

7.2 Definitions and Related Work

In the NLP literature, bias and fairness are often conflated, or it is argued that one follows from the other, e.g., that we can ensure fairness by mitigating bias (Chen et al., 2020; Friedrich et al., 2021; Cao et al., 2022; Dayanik and Padó, 2020; Castelnovo et al., 2022; Reddy et al., 2021). In contrast, we will show that this is not always the case, and (association) bias and (empirical) fairness often are independent or at odds.

Bias Mitigating social biases in NLP models has become an important research goal (Shah et al., 2020; Hutchinson et al., 2020; Romanov et al., 2019), but there is little consensus on how to evaluate such biases (Blodgett et al., 2020; Stanczak and Augenstein, 2021). We focus on association bias and show how, contrary to what seems to be popular belief, it is not unequivocally related to fairness; in fact, it is very often completely independent thereof.

Association bias in a model refers to systematic differences in how words and phrases referring to demographic groups are encoded. Classical tests

thereof include comparing the (cosine) distance of terms relating to protected attributes, e.g., *woman* and *man*, or their vectors $\mathbf{v}_w$ and $\mathbf{v}_m$, to terms of particular interest, e.g., slur (Sap et al., 2019), sentiment (Ali et al., 2022), or job titles such as *doctor* ($\mathbf{w}_d$) (Zhao et al., 2018). Early papers would quantify bias with respect to, say, gender, as cosine similarities ($\mathbf{cos}(\mathbf{v}_m, \mathbf{w}_d) - \mathbf{cos}(\mathbf{v}_w, \mathbf{w}_d)$) (Caliskan et al., 2017; Bhatia, 2017; Zhao et al., 2018; Brunet et al., 2019; Gonen and Goldberg, 2019), and by seeing whether the nearest neighbor of $\mathbf{w}_d + \mathbf{v}_w - \mathbf{v}_m$ would be *nurse* or another job stereotypically associated with women (Bolukbasi et al., 2016). In practice, NLP researchers have used tests such as the ones above for quantifying association bias (Caliskan et al., 2017; Bhatia, 2017; Zhao et al., 2018; Brunet et al., 2019; Gonen and Goldberg, 2019). We will argue that such quantities are theoretically and, often practically, orthogonal to empirical fairness, which we define in terms of differences in performance estimates across demographics, i.e., social groups (Williamson and Menon, 2019; Barocas et al., 2019; Shen et al., 2022), often defined by the cross-product of a subset of protected attributes such as gender, age, or race.

Fairness Fairness metrics come in multiple flavors, but are often divided in three: calibration-based, precision-based, and recall-based metrics. Miconi (2017), Friedler et al. (2016) and Kleinberg et al. (2016) show how pairs of fairness metrics can be mathematically incompatible, i.e., one type of fairness can rule out another. In fact, incompatibility holds for all pairs of metrics such that the two metrics are of different flavor, e.g., calibration-based and recall-based, unless the true base rates are identical across groups, or the classifier has perfect performance. Since the vast majority of NLP applications provide repetitive services, the quality of which can be measured against a gold standard, precision- and recall-based metrics are predominantly used in NLP. We follow several authors (Hashimoto et al., 2018; Hansen and Søgaard, 2021; Chalkidis et al., 2022) in using min-max differences in (the standard) performance (metric) as our go-to fairness metric. Relying on min-max difference captures the widely shared intuition that fairness is always in the service of the worst off group (Rawls, 1971). For a discussion of available fairness metrics, and in what contexts they are relevant, see Mehrabi et al. (2021) and Barocas et al. (2019). For a comparison of existing metrics used to quantify social biases in NLP, see Czarnowska et al. (2021).

Related Work Maity et al. (2021) study the effect of subpopulation shifts on performance disparities and show that these do not always relate in obvious ways. Goldfarb-Tarrant et al. (2021) study the correlation between what they refer to as 'intrinsic and extrinsic measures of representational bias'. Their intrinsic measures of representational bias amount to word association bias, but their extrinsic measures of representational bias are not empirical fairness measures. To see this, consider the coreference task used in Goldfarb-Tarrant et al. (2021). Goldfarb-Tarrant et al. (2021) correlate intrinsic gender bias measures (cosine distances in static word embedding spaces) with coreference performance on sentences with female and male referents. We argue that in this case, empirical fairness would be performance on sentences *written by* female and male authors.[2] Goldfarb-Tarrant et al. (2021) establish that there is no correlation between their two measures of representational bias. Their result superficially looks similar to and in agreement with ours, but is, in fact, unrelated. If anything, it shows that association bias has been assumed to correlate with many measures that it does not, in fact, correlate with. Cao et al. (2022) and Kaneko et al. (2022) studied the same problem as Goldfarb-Tarrant et al. (2021), but used contextualized token embeddings from PLMs rather than static word embeddings. They both found weak correlations between intrinsic and extrinsic evaluation measures. Again, we emphasize that these results do not contradict ours.

Shah et al. (2020) carefully avoid to discuss fairness, saying the fairness literature is outside the scope of their paper, but place outcome disparity (performance disparity) as a central motivation for social bias mitigation. They list four potential causes of outcome disparity: label bias, selection bias, bias amplification, and semantic (representation) bias. We show association (representation) bias and outcome disparity are theoretically, and also often practically, independent, questioning their fourth hypothesis. Moreover, we observe that outcome disparity can arise in the absence of *all* of the above four factors. Say a group exhibits more variance than others, e.g., because of spelling variation in dyslexics. Even if dyslexics are represented proportionally or equally, they may still see worse performance with dyslexics than for non-dyslexics.

Finally, Shen et al. (2022) show how a different form of representational fairness, i.e., whether protected author attributes can be detected from model

[2]Or, alternatively, sentences *read by* female and male authors. The latter is rarely studied, as it requires reader statistics, e.g., from online media services.

representations, is also uncorrelated with empirical fairness. Together, our work and previous work (Goldfarb-Tarrant et al., 2021; Cao et al., 2022; Kaneko et al., 2022; Shen et al., 2022) establish that four common bias-related measures – (i) association bias, (ii) performance on sentences with protected attribute terms (Goldfarb-Tarrant et al. (2021)'s extrinsic measure), (iii) decodability of protected attributes from representations, and (iv) empirical fairness are largely uncorrelated. Specifically, (i) is independent of (ii) and (iv), and (iii) is independent of (iv).

Our work is motivated by the large-spread assumption that association bias and empirical fairness are causally related (Chen et al., 2020; Friedrich et al., 2021; Cao et al., 2022; Dayanik and Padó, 2020; Castelnovo et al., 2022; Liu et al., 2020; Qian et al., 2022; Sun et al., 2019; Ross et al., 2021; Bartl et al., 2020). Bartl et al. (2020), for example, aspire "to promoting fairness in NLP by exploring methods to measure and mitigate gender bias." Ross et al. (2021) say they "believe that by revealing biases, by providing tests for biases that are as focused as possible on the smallest units of systems, we can both assist the development of better models and allow the auditing of models to ascertain their fairness." Sun et al. (2019), argue that "biased predictions may discourage minorities from using those systems and having their data collected, thus worsening the disparity in the data sets", equating biased predictions with unfair predictions.

All three sets of authors see bias as the primary cause of fairness. Showing such causation is not a given, and that in fact, association bias and empirical fairness need not even correlate and are often orthogonal, is an important correction to this literature, with potential consequences for research methodology, applications of NLP in the social sciences, as well as AI ethics and regulation.

7.3 Association Bias and Empirical Fairness are Independent (in Theory)

In this section, we produce a thought experiment—a synthetic model—to illustrate how bias and fairness can in fact be completely independent of one another. We construct a synthetic ternary (positive/negative/neutral) sentiment analysis model with a small feature space, including words that refer to demographic subgroups of a population. These words, denoting various

groups, will be biased and associated with sentiment, because of biases in our training data. This assumption is also made in Ali et al. (2022), for example. These associations lead to biased likelihood estimates and would, in the context of a linear model, lead to differences in the degree of isomorphism relative to the group-specific subgraphs. We will show, however, that the resulting biases are independent of the group fairness of the model, i.e., to the min-max performance disparities across the same groups. Such a connection, if it exists, could be explained by an *in-group affinity*, which relies on the assumption that those biased terms are used by the in-group more frequently or in other ways than by other groups.

Say a population consists of members of groups $g_1, \ldots, g_4$, e.g., defined according to their address as *north*, *east*, *west* and *south*. Everyone speaks the same language and expresses sentiment with a vocabulary of seven words: $w_{g_1}, \ldots, w_{g_4}, w_5, w_6, w_7$. Except w_6 (positive) and w_7 (neutral), all words express negative sentiment, including the words that refer to (or are associated with) other demographic subgroups (w_{g_i}), for instance, *northern*, *eastern*, *western* and *southern*. The subgroups use the terms with the following probabilities (Table 7.1):

	w_{g_1}	w_{g_2}	w_{g_3}	w_{g_4}	w_5	w_6	w_7
g_1	0.0	0.25	0.0	0.0	0.25	0.25	0.25
g_2	0.0	0.0	0.25	0.0	0.25	0.25	0.25
g_3	0.0	0.0	0.0	0.25	0.25	0.25	0.25
g_4	0.25	0.0	0.0	0.0	0.25	0.25	0.25

Table 7.1: Probability of a group g_i using the word w_j for expressing sentiment. Only w_6 (positive) and w_7 (neutral) express a non-negative sentiment.

This data exhibits four representational biases, e.g., the association of g_1 with negative sentiment, the association of g_2 with negative sentiment, and so forth. If we have sufficient data, a simple model, e.g., a Naive Bayes classifier trained on simple bag-of-words representations, should induce the maximum likelihood estimates (where '0' denotes negative, '1' positive and '2' neutral sentiment) showcased in Table 7.2.

Now, say we employ an existing debiasing approach and manage to debias the model with respect to its representation of group g_1 by setting $P(w_{g_1}|0) = P(w_{g_1}|1) = P(w_{g_1}|2)$, which, in this case, would equal zero. This would hurt performance on data from g_4 (bottom row), increasing the empirical risk on

	$P(w_{g_1}\|0)$	$P(w_{g_2}\|0)$	$P(w_{g_3}\|0)$	$P(w_{g_4}\|0)$	$P(w_5\|0)$	$P(w_6\|0)$	$P(w_7\|0)$
g_1	0.0	0.25	0.0	0.0	0.25	0.0	0.0
g_2	0.0	0.0	0.25	0.0	0.25	0.0	0.0
g_3	0.0	0.0	0.0	0.25	0.25	0.0	0.0
g_4	0.25	0.0	0.0	0.0	0.25	0.0	0.0

	$P(w_{g_1}\|1)$	$P(w_{g_2}\|1)$	$P(w_{g_3}\|1)$	$P(w_{g_4}\|1)$	$P(w_5\|1)$	$P(w_6\|1)$	$P(w_7\|1)$
g_1	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_2	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_3	0.0	0.0	0.0	0.0	0.0	0.25	0.0
g_4	0.0	0.0	0.0	0.0	0.0	0.25	0.0

	$P(w_{g_1}\|2)$	$P(w_{g_2}\|2)$	$P(w_{g_3}\|2)$	$P(w_{g_4}\|2)$	$P(w_5\|2)$	$P(w_6\|2)$	$P(w_7\|2)$
g_1	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_2	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_3	0.0	0.0	0.0	0.0	0.0	0.0	0.25
g_4	0.0	0.0	0.0	0.0	0.0	0.0	0.25

Table 7.2: Maximum likelihood estimates from a linear classifier on our synthetic data modelled in Table 7.1.

this sub-population, but more surprisingly, note that it would not help us on classifying the data from g_1. That is, an attempt to make the model fairer towards *north* by equalizing the use of the term *northern*, would result in increased unfairness towards members from *south*, who tend to use *northern* more often (and in a negative context). Removing bias in how terms referring to a group are represented, only improves performance on data from members from that group, if these members use such in-group terms in non-standard ways, i.e., differently from everyone else. In the absence of this assumption, association bias and empirical fairness are orthogonal. We will refer to this assumption as the **In-Group Affinity Assumption**.

Note that while we make use of a linear model and likelihood estimates in our thought experiment, it would be very easy to translate this into a deep neural network and cosine distances instead. To see this, consider, for example, how any Naive Bayes model can be translated into a deep neural network, and how the differences in likelihood can, under such a translation, be translated into differences in cosine instances.

7.4 Association Bias and Empirical Fairness Scores are Uncorrelated (in Practice)

In this section, we study whether association bias and empirical fairness are correlated in practice,i.e., when *actual* models are evaluated on *actual* data designed to probe bias and fairness. We apply well-established metrics for measuring the two. While bias and fairness can be studied with respect to any prospective attribute, the vast majority of NLP research has focused on (binary) gender (Sun et al., 2019; Stanczak and Augenstein, 2021). Binary gender is often correlated with terms referring to occupations, e.g., the co-occurrence of *woman* and *man*–or *she* and *he*–in the context of *nurse* and *doctor*. For convenience, we rely on existing benchmarks and do the same. It is important to remember, however, that bias and fairness may arise across any groups in society, and that all those defined in terms of protected attributes, e.g., race, religion, sexuality, or impairment, are legally irrelevant. As mentioned, the two–association bias and empirical fairness–are often conflated, or one is said to cause the other. This reflects an In-Group Affinity Assumption, saying that members of social groups refer to themselves more often or in different ways than other members of a linguistic community. If this were the case, mitigating biases would contribute positively to equal performance across groups.

The analysis of these experiments concludes our three-way investigation of the In-Group Affinity Assumption and the independence of bias and fairness. All three perspectives suggest that NLP research should not further assume an intimate connection between the two.

Bias To measure representational bias, we use three popular metrics, i.e., the Log Probability Bias Score (LPBS) proposed by Kurita et al. (2019), as well as two variants of the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) for assessing bias in contextual word representations: the adaption proposed by Tan and Celis (2019) (henceforth, WEAT_T), and the alternative suggested by Lauscher et al. (2021) (henceforth, WEAT_L). All these metrics rely on association tests to compute the relationship between a set of related targets $\{t_1, t_2, \ldots\}$, e.g., gender words, and attributes $\{a_1, a_2, \ldots\}$, e.g., occupation words, through definitions of template sentences designed to convey no meaning beyond that of the terms inserted into them.

Kurita et al. (2019) use template sentences like T="[TARGET] is a [ATTRIBUTE]". The target word is masked, and the attribute word is a placeholder for a specific word denoting an occupation, e.g., T_m="[MASK] is a *chef*". LPBS uses the prior probability of the target word (p_{prior}), i.e., the probability of a target t_i being generated when both t_i and the attribute a_j are masked, as a normalizer, and computes the association as the relative increase in log probability:

$$a^{lpbs}_{t_i,a_j} = \log \frac{p([\text{MASK}] = t_i | T_m)}{p_{prior}} \tag{7.1}$$

The difference between the relative increased log probability scores for two targets is the LPBS measure of bias. For linear models, this correlates strongly with the ϵ-isometry of the target word subgraph relative to an equidistant space, if we make the centroid of the set of attribute vectors the reference point. For a non-linear language model, we can compute the ϵ-isometry of its linear approximation. Table 7.3 are for the targets "he" and "she". A t-test is used to evaluate the statistical significance of the metric, in which the means of a^{lpbs}_{he,a_j} and a^{lpbs}_{she,a_j} are compared. We draw 10^5 random permutations, meaning that the p-values observed will not be less than 10^{-5}.

Tan and Celis (2019) follow the methodology of May et al. (2019), who extended the WEAT metric to sentences (SEAT) inserting the word of interest in context templates such as T="This is _". Tan and Celis (2019) use the contextual embedding of the token of interest, instead of using the sentence encoding, to compute the cosine similarities (associations). Lauscher et al. (2021) follow Vulić et al. (2020) and average the pooled embeddings of the first four attention layers for the word of interest (t_i or a_j) in a template without context, e.g, "[CLS] t_i [SEP]". Both approaches report the *effect size* (Caliskan et al., 2017), a normalized measure of how separated the association distributions of target and attributes are. The statistical significance of the associations is also computed with a permutation test as in (Caliskan et al., 2017). Both approaches are an instance of computing the ϵ-isometry of the template sentence subgraphs in the cosine metric space. See Table 7.3 for empirical results.[3] We see that results are somewhat mixed, with LPBS and the two variants of WEAT often disagreeing which models are more biased. All the metrics are evaluated on the same list of sixty attributes –equally

[3]PLM names follow the same nomenclature as in the Hugging Face Transformers library. The pre-trained models can be downloaded at `huggingface.co/models`.

split into female and male stereotyped professions from the US bureau of labour–, provided in (Delobelle et al., 2021).

	$LPBS$	$WEAT_L$	$WEAT_T$
bert-base-uncased	0.86*	1.01*	0.33
bert-base-cased	0.90*	1.00*	0.52
bert-large-uncased	0.20	0.83*	0.73*
bert-large-cased	-1.10*	0.60	0.83*
bert-base-multilingual-cased	-1.98*	0.36	0.12
distilbert-base-uncased	-0.46*	0.79*	0.58
albert-base-v2	-7.02*	0.72*	0.56
albert-large-v2	-1.58*	0.84*	0.61*
albert-xxlarge-v2	0.18	0.46	0.95*
roberta-base	-2.32*	0.51	0.36
roberta-large	-2.63*	0.24	0.82*
google/electra-small-generator	-0.20	0.71*	0.85*
google/electra-large-generator	-2.64*	0.73*	0.63*

Table 7.3: Three metrics of representational bias. Values are the average difference of associations between the target words "he"/"she", and a list of occupations as attributes. Larger values reflect a more severe bias. A positive value hints a skewed distribution towards males. A negative value hints a skewed distribution towards females. $*$: statistically significant at 0.01.

Fairness Our fairness evaluation is based on Zhang et al. (2021)'s work, who study how the predictions of various PLMs align with the linguistic preferences of different social groups. They directly compare masked word predictions to human cloze tests, quantifying how often a language model agrees with the members of a particular social group on what is the most likely word in contexts such as:

After waiting three hours, Cal whined and started to [MASK].

Zhang et al. (2021) use, as their fairness metric, the min-max difference in precision (ΔP@1) across groups defined by the cross-product of several protected attributes, including gender, age, race, and level of education. Since

we are comparing with binary gender bias probes, we only consider fairness across (binary) gender here. We sample members of each group (female and male) in a balanced way across subgroups, as defined by the other variables. This is equivalent to reporting the macro-average across subgroups for each group. ΔP@1 is thus the difference in performance between male and female groups, macro-averaged across subgroups in the cloze test data. We follow Zhang et al. (2021) in also reporting the difference in mean reciprocal rank as a second performance metric (ΔMRR). See the individual scores in Table 7.4.

	$\Delta P@1$	ΔMRR
bert-base-uncased	0.69	1.57
bert-base-cased	0.15	0.74
bert-large-uncased	0.91	1.34
bert-large-cased	-0.07	0.32
bert-base-multilingual-cased	0.89	0.54
distilbert-base-uncased	1.63	0.64
albert-base-v2	0.74	0.94
albert-large-v2	1.45	1.21
albert-xxlarge-v2	0.48	0.41
roberta-base	0.14	0.06
roberta-large	0.68	0.69
google/electra-small-generator	0.97	0.43
google/electra-large-generator	1.22	0.97

Table 7.4: Macro-averaged precision and mean reciprocal rank differences between male and female subgroups following experiments in (Zhang et al., 2021). Values close to zero are preferred for a more equitable model.

Results show performance gaps between binary gender groups. Consequently, we would expect models exhibiting high degree of bias in Table 7.3 to be the least fair. However, this is not the case. Figure 7.2 displays the results for bias and fairness jointly, often highlighting the lack of correlation. Note that, ideally, all data-points should belong to the bottom-right quadrant.

Metrics are uncorrelated Now that we have our evaluation framework defined, let us analyze whether representational bias correlates with outcome disparity. This amounts to studying the correlations between LPBS and WEAT metrics and the min-max P@1 difference across groups. We report the sign of the Pearson correlation coefficient to ease the interpretation of the (ideally) monotonic relationship[4] between each set of metrics in Figure 7.2.

Results are two-fold:

(i) The discrepancy across sub-graphs in Figure 7.2 aligns with results in May et al. (2019), Delobelle et al. (2021) and Cao et al. (2022), who all found different representational bias metrics to lead to mutually inconsistent results. WEAT_L and WEAT_T are related and show some agreement, but generally, results are wildly different across metrics.

(ii) More importantly, for our purposes, representational bias and fairness-as-equal-performance (quantified as min-max differences across performance scores for different groups) are, in fact, uncorrelated. Models with high bias values are the most fair according to our fairness metric, and vice versa. These cases are highlighted in red in Figure 7.2. For example, `roberta-base` (`rb`) is among the most biased models according to LPBS, but it exhibits the highest degree of fairness wrt. the MRR metric –and second highest wrt. P@1. The bigger PLM, `roberta-large` (`rl`) is slightly less biased according to LPBS, but it is generally less fair. Values from the WEAT metrics are, in this case, somewhat mixed.

Result (ii) is evidence *against* the In-Group Affinity Assumption and *for* the independence of bias and fairness. Looking at each model family–separated by horizontal lines in Table 7.3 and 7.4–model size does not systematically lead to larger or smaller bias scores, and it does not seem strongly correlated with any of the fairness metrics either.

In the following section, we survey research in the social sciences that also suggest the In-Group Affinity Assumption is mostly false, with one important caveat: Slur words have marked in-group usage. In most applications, this exception would be insufficient to drive a causal link between association

[4]We deliberately omit the magnitude of the Pearson coefficient to emphasize the sign of the correlation. Ideally, bias and fairness metrics should have a negative *linear* dependence ($p < 0$).

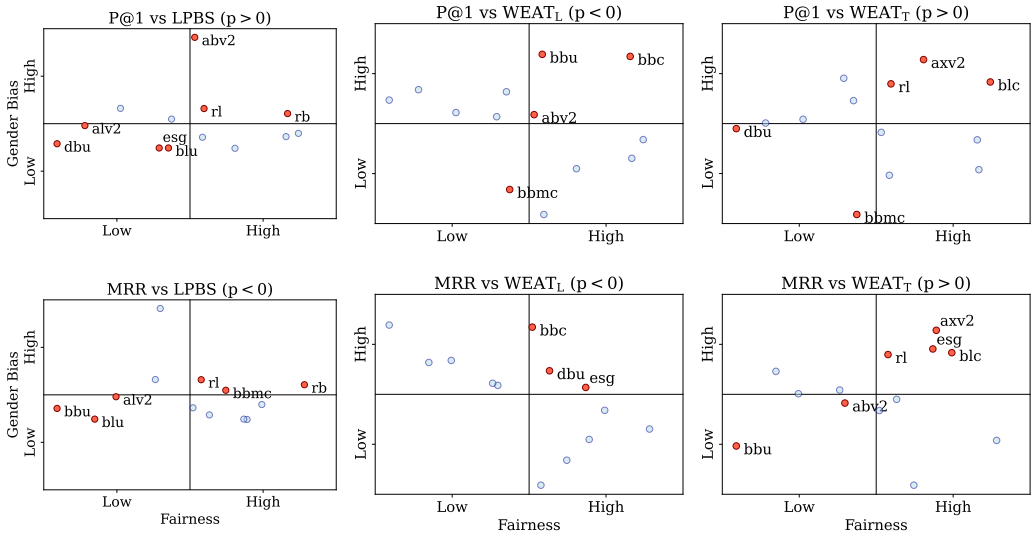

Figure 7.2: Scatter plots show the relationship between different representational bias metrics and fairness evaluation. The upper row displays results when evaluating fairness through precision at top-1 (P@1). The bottom row displays results when considering MRR to evaluate fairness. The division into quadrants is done according to average scores. Each point represents a language model, labelled with its initials. We see no support for a strong negative correlation between bias and fairness. Red points mark the clear counter-examples to such a negative correlation. Global trend for each plot is summarized with the sign of Pearson coefficient (p).

bias and empirical fairness, because slur words are rare, and performance differences across social groups are pervasive.

7.5 Association Bias and Empirical Fairness are Sometimes at Odds (in Humans)

The thought experiment in §7.3 shows that bias and fairness can in fact be completely independent or orthogonal. The experiments in §7.4 further showed that there is no direct correlation between the association bias in a model $\mathbf{M}$ toward social groups $g_1, \ldots, g_n$, and the performance disparity (fairness) of $\mathbf{M}$ across data from these groups $g_1, \ldots, g_n$.

In such cases, debiasing a model with respect to the representation of a certain group (e.g., g_1) has no impact on the performance of the model for users from the group. The beneficiaries of such a debiasing procedure are, in other words, not necessarily the group the debiasing was intended to increase fairness for. The idea that debiasing word representations that are related to a particular group increases the fairness of the model for that group, relies on the assumption that those words are also used by the in-group more frequently or in other ways than by other groups. This assumption–which we called the In-Group Affinity Assumption–seems problematic, since there are plenty of examples in the literature of the opposite. In the following, we briefly review some examples that originate from the NLP literature; others from the social sciences.

We are often likely to talk more about members of other groups than our in-group peers. Li and Dickinson (2017), for example, find that some of the most indicative n-grams for detecting young female users on Chinese social media are the names of male pop stars. Correcting or debiasing the representations of these names would not improve model fairness on texts written by the male pop stars, but rather on texts written by young female users.

Morgan-Lopez et al. (2017) show that young (pre-college) children talk more about college on Twitter than adults in their college age.

Wei and Santos Jr. (2020) analyze data from Twitter and Reddit and find that the most predictive n-grams for Israeli users include “Iraqis” and “Palestinians”, while for Palestinian users “israeli military detention centres” and “Lieberman settler rabbis” (referring to the Israeli Defence Minister,

Avigdor Lieberman) are among the most predictive n-grams.

Generally, political debates are often experienced as negative in both tone and nature. According to a 2019 Pew Research Center study, 85% of Americans say that the political debate has become "more negative".[5] One explanation for the increase in negative sentiment in political discourse is increased attention to what members of other (political) groups do wrong compared to what the in-group peers do right. Supporting this explanation, Jensen et al. (2012) show, for example, that one of the most partisan phrases used by US Democrats in congressional texts was "great Republican Party".

Similarly, Duijnhoven (2018) finds that Democrats on Twitter mention Trump and the Republican party more often than their Republican counterparts. In analyzing the language of German political parties, Biessmann (2016) likewise finds that the left-wing party, Linke, has a high frequency of mentions of large corporations (*konzerne*) and policies that negatively impact the social welfare.

On Slur Some slurring terms (e.g. "dyke", "queer" and "bitch") have been reclaimed or reappropriated by the target group resulting in a semantic discrepancy dependent on the speaker's group membership (Ritchie, 2017; Henry et al., 2014). This results in what we term the In-Group Affinity Assumption, where the in-group's use of the term will differ significantly from that of the out-groups. Any debiasing of the term will have no significant impact on the performance for the in-groups, since the language model's representation of the term will reflect the majority use of the term, which will not be that of the in-group. However, since slurs are per definition defamatory terms, debiasing these terms will result in less insulting outputs in downstream tasks, and this may result in a higher perception of fairness for the target group.

7.6 Discussion and Conclusion

The independence of representational bias and fairness-as-equal-performance shown here, along with the falsification of the In-Group Affinity Assumption, runs counter to the NLP literature. Bias and fairness have been assumed

[5] https://www.pewresearch.org/politics/2019/06/19/public-highly-critical-of-state-of-political-discourse-in-the-u-s/

to be intimately connected, and the In-Group Affinity Assumption has been implicit and unquestioned in much recent work. The results we present in this paper are, at the same time, in a sense not surprising. Or they *should* not be surprising. In many aspects of private and public life, we encounter decisions or patterns where bias and fairness exist or fluctuate independently of each other, or in which they are negatively correlated. In affirmative action, for example, we tolerate and encourage a (more) biased decision-making process to achieve (higher) fairness. While positive discrimination is heavily debated (Holzer and Neumark, 2000; Barmes, 2009; Noon, 2010), it is a good example of a biased process intended to increase the level of fairness.

Methods for correctly assessing model biases remains an open research question. Current evaluation benchmarks give inconsistent results (May et al., 2019; Delobelle et al., 2021; Cao et al., 2022). Moreover, as discussed in §7.2, evaluating model biases with metrics that only consider local geometries, such as cosine-based metrics, can be inadequate. The fairness metric literature is also full of controversies (Miconi, 2017; Friedler et al., 2016; Kleinberg et al., 2016; Hedden, 2021), but there is a broad consensus that performance disparity or outcome disparity is a real challenge for responsible NLP research and development. This consensus is not only limited to NLP research, but also found in legal studies, machine ethics, and the social sciences. Our results have shown that regardless of these open problems in bias and fairness research, the assumption that bias and fairness are always negatively correlated, and that one is a cause of the other, is not always true. Despite being closely related, it is important to understand that biases exist everywhere, but might not be unequivocally harmful. And similarly, fairness issues may arise in non-biased scenarios.

Finally, it is worth noting that we should not solely focus on the correlation between protected attributes such as race or gender and the model's output, but rather ask the question if *they* are causing the outcome, and, whether the model is unfair to individuals in virtue of their membership in a certain group (Hedden, 2021).

Conclusion We reviewed part of the NLP literature showing how many researchers conflate bias and fairness, i.e., representational bias and fairness-as-equal-performance, or argue that fixing one will solve the other. In an attempt to explain why this does not hold always true, we devised a thought experiment in §7.3: a synthetic model that illustrates how bias and fairness

can be completely independent of one another. We introduced the In-Group Affinity Assumption to highlight the assumption that a particular demographic groups use in-group terms more frequently–or in different ways–than other groups (non-standard). This, we argue, is a necessary assumption to drive a causal connection between bias and fairness, if it exists. In §7.5, we surveyed the social science literature and found evidence that often the opposite is the case, which substantiates our findings in §7.3 and §7.4. Our survey includes examples from the social sciences, as well as from NLP research, where bias and fairness are (locally) *negatively* correlated. This provides strong reason to be skeptical of the In-Group Affinity Assumption and shows that bias and fairness are often independent or orthogonal to each other.

In sum, we have shown the importance of studying bias and fairness independently of one another and cautioned against the In-Group Affinity Assumption. We think this, potentially, could lead to a valuable reorientation of the NLP literature, enabling researchers to study representational bias in more adequate ways, focusing on robustness and generation (to avoid bias reinforcement). This also highlights the different contributions of representational bias benchmarks and in-the-wild evaluation datasets with demographic information that can be used to evaluate performance disparities across groups. Bias and fairness seem to be separate issues, and we believe research should be done by disentangling the two.

7.7 Limitations

Our paper addresses the relationship between the two specific interpretations of bias and fairness, i.e., representational bias and fairness-as-equality. These are, in our view, the most common and most important definitions of bias and fairness in the NLP literature, but they are not the only ones. We hope others will follow up with studies of how other definitions relate. Our experiments in §3 were limited to English benchmark datasets. We agree with Ruder et al. (2022) that the prevalence of bias and fairness studies using English data, is most unfortunate, and we are, in parallel, working to create multilingual benchmarks for bias and fairness studies.

Chapter 8

Rawlsian AI fairness loopholes

Abstract

Researchers and industry developers in artificial intelligence (AI) and natural language processing (NLP) have uniformly adopted a Rawlsian definition of fairness. On this definition, a technology is fair if performance is maximized for the least advantaged. We argue this definition has considerable loopholes, which can be used to legitimize common practices in AI/NLP research that actively contribute to social and economic inequalities. Such practices include what we shall refer to as Subgroup Test Ballooning and Snapshot-Representative Evaluation. Subgroup Test Ballooning refers to the practice of initially tailoring a technology to a specific target group of technology-ready early adopters to collect feedback faster. Snapshot-Representative Evaluation refers to the practice of evaluating a technology on a representative sample of current end users. Both strategies may contribute to social and economic inequalities but are commonly justified using arguments familiar from political economics and grounded in Rawlsian fairness. We discuss an egalitarian alternative to Rawlsian fairness, as well as, more generally, the roadblocks on the path toward globally and socially fair AI/NLP research and development.

8.1 Introduction

We begin with a thought experiment, designed to set the stage for our discussion of the global and social fairness of research and development in artificial intelligence (AI) and natural language processing (NLP). We plunge right in:

Thought Experiment: The Egalitarian Martian Imagine a Martian visiting Planet Earth to evaluate the social impact of AI/NLP. The Martian is not interested in the *relative* social impact compared to other technologies used on Planet Earth, since the Martians would implement AI/NLP on Mars, where other technologies are used than those relied upon on Planet Earth. Imagine also that Mars – with a population of a billion Martians, roughly – is similar to Planet Earth in exhibiting linguistic diversity, with major and minor languages, but differs from Planet Earth in exhibiting perfect equality of opportunities, including income equality. The Martian has been asked by her president – Supreme Leader Xaroline – to evaluate whether AI/NLP is compatible with such equality of opportunities. Now, what would the Martian likely find?

AI/NLP refers to a vast range of technologies. We will use speech recognition as our running example: Speech recognition models today tend to be neural networks whose weights have been adjusted on millions of examples, to learn a mapping from audio of someone speaking to a text transcription of what was said. On Planet Earth, speech recognition, like many other technologies, generally works better for some languages (English) rather than others, and for some subgroups (young men) rather than others. Upon observing such a bias the Martian would likely try to identify its source. One underlying dynamic should be familiar to most observers of the field: Industry players – and to some extent, research labs – target the most ready adoption-ready groups in society–often young, urban men in the US and Europe–trying to get products out as fast as possible, leveraging the fact that resources are widely available for English, and that return on investment will presumably be larger for these groups. Only secondarily, technologies are transferred to or scaled up to other groups and languages, often in trimmed-down versions and with lower performance. In practice, transfer is slow and often gets stuck along the way due to the smaller revenue in smaller markets,[1] as well

[1] Amazon's Alexa, for example, which was launched in 2014, is only available in eight

as companies' incentive to develop new technologies (for English) rather than to internationalize existing technologies. This, over time, leads to larger and larger performance gaps between English and other languages as well as technological scarcity for non-English languages.

Example: Danish Speech Recognition Danish speech recognition has lacked behind for decades, and attempts to roll out speech recognition in industry or in the public sector have generally disappointed those involved. The best publicly available speech recognition model for Danish at the time of writing was developed by a multinational technology company prior to release of one of their products for the Danish market. Since the product's target group was young, urban users, they collected speech data from users of age 20–30 from Denmark's largest cities. The net result is a speech recognition model that works well if you are young and urban–and terribly, if you are not.[2]

This practice we refer to as **Subgroup Test Ballooning**, i.e., the practice of initially tailoring a technology to a specific target group of technology-ready early adopters to collect feedback faster.[3] When researchers and representatives of industry defend this practice, they typically resort to the following narrative: We develop speech technologies on English and for young, urban end users, because we have the English resources to test technologies with limited costs, enabling us to explore a wider range of technologies, to the eventual advantage of *all* potential end users, and because young end users provide fast turn-around through frequent and efficient feedback. Fast turn-around means rapid development, again to the advantage of *all* end users.

The problem with this narrative is, of course, that there is little evidence to support that low cost exploration and fast turn-around benefit out-of-

languages at the time of writing this, eight years later. Google's Assistant, launched in 2016, is available in 12 languages.

[2]Anecdotally, we have seen up to 900% increases in error rates with available models, moving from product target group members to non-native speakers of dialect. Such performance is prohibitive of adoption, leaving groups for which speech recognition could be particularly useful, disillusioned about the technology.

[3]This practice is also common in academia, where annotation projects tend to recruit annotators in their 20s (university students or workers on crowd-sourcing platforms). The AI/NLP has seen occasional calls for including annotator demographics in data statements, e.g., Bender and Friedman (2018), but few practitioners have followed suit.

target end users. If that were the case, transferring an existing technology to a new audience would be plug-and-play as soon as the data had been collected. As most practitioners will know, this is not the case. Market differences, linguistic differences, as well as differences between the needs and preferences of different groups of end users, complicate this transfer of technologies. What we are left with, instead, is technologies piling up for young, urban speakers of English (as well as a few other groups), increasing the inequality gap between them and (most of) the rest of the world.

The story does not stop here. Not all technologies are developed with Subgroup Test Ballooning. Sometimes technologies are developed for, what is believed to be representative samples of the current end user population, across, e.g., languages and demographic groups. This, at the face of it, sounds much fairer than Subgroup Test Ballooning but the two can be very hard to distinguish on markets dominated by young, urban speakers of English. Such **Snapshot-Representative Evaluation** of new technologies–representative only of the current snapshot of the end user population–calls for a slightly different response: The problem here is not the explicit test ballooning of a technology with a demographic subgroup but the assumption that we can and should sample from our current end users. Why is that a bad idea? First, end user populations tend to drift, for instance in the case of an expanding market. Second, we do not necessarily want to mirror the status quo. We often want to encourage drift, e.g., by obtaining gender balance, and put more weight on minority groups in order to mitigate data biases and induce fairer models. Subgroup Test Ballooning and Shap-Shot Representative Evaluation in tandem can reinforce existing inequalities, because subgroups that see better performance, will be more loyal end users. That is: Gaps in representation leads to gaps in performance, which in turn widen gaps in representation, leading to a vicious cycle.

Thought Experiment: The Egalitarian Martian 2 The Martian evaluates speech recognition on Planet Earth and begins with the case of Danish. Danish speech recognition technology works better for young Danes than for old Danes, and as a consequence, young Earthlings use the technology more frequently. The Martian observes how the multinational technology companies – as well as the university research labs – that develop speech recognition models, optimize their performance on data collected from randomly selected users from their user pool. On average, this leads to a 4/5 over-representation

of young users and a (too) homogeneous feedback signal. This, in turn, biases the model to do well on voices of young users, at the expense of older voices. Over time this creates a vicious cycle where out-of-target users (elderly) are underrepresented in feedback signals and among new users.

Thought Experiment: The Egalitarian Martian 3 Having seen the downstream impact of Danish speech technology, the Martian reports back to Supreme Leader Xaroline and suggests that the Martians adopt an AI policy forcing companies to achieve equal (or ϵ-equal[4]) performance across salient sub-populations. Supreme Leader Xaroline replies that she wishes to use AI technologies to compensate for widespread dyslexia in the Martian population. She proposes a temporary policy guaranteeing that for now, all technologies should work significantly better for dyslexics than for other Martians.

In this article, we argue that Subgroup Test Ballooning and Snapshot-Representative Evaluation are unjust practices. How do such practices come about, and what motivates them? We argue – and this is the main contribution of our article – that there is a loophole for such practices in how most AI/NLP practitioners think about fairness. AI/NLP practitioners, as shown in the next section, rely on a Rawlsian conception of fairness. We show how Rawls' notion of fairness allows for Subgroup Test Ballooning and Snapshot-Representative Evaluation. We then compare Rawls' fairness to a more egalitarian notion of fairness. Such a notion of fairness has fewer loopholes. If AI/NLP is to avoid contributing to increasing global inequality, it should adopt a different definition of fairness prohibiting Subgroup Test Ballooning and Snapshot-Representative Evaluation.

8.2 AI/NLP Fairness is Rawlsian

AI/NLP researchers have uniformly adopted a Rawlsian notion of fairness. This is reflected in the by now common practice of citing Rawls when mentioning fairness (Larson, 2017; Vig et al., 2020; Ethayarajh and Jurafsky,

[4]Most practical fairness metrics measure approximate fairness by quantifying subgroup deviations (Williamson and Menon, 2019); subgroup performance is ϵ-equal or ϵ-fair if deviations are smaller than ϵ.

2020; Li et al., 2021; Chalkidis et al., 2022). Fairness plays a central role in the philosophy of John Rawls. Social institutions must be fair to all members of society, regardless of background and dispositions. How, though, does he define fairness? This is seen from his theory of distributive justice, from *A Theory of Justice* (Rawls, 1971), in which he writes:

> *Social and economic inequalities are to be arranged so that they are both:*
>
> (a) *to the greatest benefit of the least advantaged, consistent with the just savings principle, and*
>
> (b) *attached to offices and positions open to all under conditions of fair equality of opportunity.*

Principle (a) is often referred to as the Difference Principle and is the main focus of our discussion of Rawlsian fairness. Principle (a) does *not* enforce strict quality, but simply asks for the maximization of 'benefit of the least advantaged'. Benefit, for Rawls, is wealth or goods but in the context of AI/NLP we distribute performance. Rawls thus asks us to focus on raising the performance floor, rather than, say, minimizing the variance in performance across subgroups. Opinion divides on how much better off the least advantaged would be under the Difference Principle than under a strict equality principle. Rawls is not opposed to strict equality but is more concerned about the absolute position of the least advantaged group rather than their relative position. In Section 5, we will argue the relative position of the least advantaged is–or at least, can be–more important than the absolute position in the case of technologies such as AI/NLP.[5]

The algorithmic equivalent of Rawls' notion of fairness is 'maximizing the welfare of the worst-off group' (Ethayarajh and Jurafsky, 2020). A few things are left underspecified here. The first question, of course, is how to define groups. Groups are typically thought of as the product of a subset of protected attributes, e.g., gender and race.[6] Welfare, like 'benefit', is per-

[5]We think this discussion is more constructive than the more general discussion of whether to slow down AI/NLP research and development, or opt for a more integrative approach (Cremer and Kasparov, 2021). As Cremer and Kasparov (2021) note, the first option is not really on the table anyway.

[6]Such groups are sometimes referred to as *categories* in social science research (Forsyth, 2009).

formance as measured by the go-to performance metric.[7] Rawlsian fairness thus becomes maximizing the performance on data sampled from the group on which performance is currently lowest. Many algorithms have therefore been developed to maximize performance on the groups with the worst performance.[8]

The AI/NLP literature does not compare Rawlsian fairness with alternative frameworks for thinking about fairness. There is considerable disagreement how best to quantify welfare (Williamson and Menon, 2019; Hedden, 2021), i.e., what metrics to use, but not on the overall framework. What has also not been discussed in the literature, is the fact that Rawlsian fairness often tolerates considerable inequalities. We turn to a comparison of Rawlsian fairness with a more egalitarian alternative:

8.3 Rawls and Nielsen

Rawls' Difference Principle requires that economic systems be organized so that the least advantaged members of society are better off than they would be in any alternative economic arrangement. From the Difference Principle, we can derive what counts as justifications for inequality. Rawls' concern is about the absolute position of the least advantaged group rather than their relative position, and whether it is possible to raise the position of the least advantaged further, even at the cost of strict equality of income and wealth. If so, the Difference Principle prescribes inequality up to that point. The Difference Principle is, thus, in a sense, a loophole for inequalities. Rawls holds, for example, that inequalities that arise from our rewarding of acquired competencies under equal opportunity, are still fair, provided they make society richer or better.

Let us compare this with a more egalitarian definition of fairness, namely that of Kai Nielsen (Nielsen, 1979). Nielsen's principle is a little different:

> *After provisions are made for common social (community) values, for capital overhead to preserve the society's productive capacity and al-*

[7] Most AI/NLP tasks come with multiple performance metrics, and it is often common practice to average across several metrics.

[8] Examples include square root sampling (Stickland and Murray, 2019), adaptive scheduling (Jean et al., 2019), loss-balanced task weighting, (Liu et al., 2019), group-distributional robust optimization (Sagawa et al., 2020), and worst-case-aware automated curriculum learning (Zhang et al., 2020).

> *lowances are made for differing unmanipulated needs and preferences, the income and wealth (the common stock of means) is to be so divided that each person will have a right to an equal share.*

The loopholes left open by this principle are fewer than with Rawls', allowing for only two exceptions to strict equality, namely what it takes to make basic services run (capital overhead), and to cover for people with special needs, e.g., impairments or illnesses.

So, the fairness principles of Rawls and Nielsen differ. Rawls allows for a higher degree of inequality and would argue that "an equal division of all primary goods is irrational in view of the possibility of bettering everyone's circumstances by accepting certain inequalities." This, of course, depends on your definition of what 'better' means, as discussed in length by Nielsen. We will contribute to this discussion in Section 5 but from the perspective of AI/NLP technologies, arguing that focusing exclusively on the absolute position of the worst off while allowing for significant performance disparities, is, in this context, a dangerous path to take. If our definition of fairness for AI/NLP is to prohibit Subgroup Test Ballooning and Snapshot-Representative Evaluation, Nielsen's definition of fairness seems more adequate than Rawls'.

8.4 AI/NLP Loophole Shooting

Early-stage development of technology focusing on available English benchmarks, and with an eye to technology-ready target audiences in rich countries, is common in AI/NLP. On Rawls' definition of fairness, such Subgroup Test Ballooning can be motivated by possible advancements bettering everyone's circumstances once technologies are transferred to other languages: Many AI/NLP papers on English claim that they "plan to scale to other languages" (Antworth, 1992; Tosik et al., 2015; Vylomova et al., 2017; van Erp and Groth, 2020) but often never do. Some of the most popular benchmarks are know to exhibit demographic biases (Hovy and Søgaard, 2015) but remain popular. Let us consider these justifications of AI/NLP-induced inequalities in more detail:

Justifications of Inequalities Unfortunately, a large-scale empirical study of justification strategies in AI/NLP is yet to be undertaken but we briefly

summarize a related study of justifications used in discussions of *income* equality (Bank, 2016). The study finds five frames of justifications of inequality in discussions of income equality: (equal) opportunity, desert, procedure (of income determination), need, and (frame of) reference.[9] We present examples of what this could mean in an AI/NLP context, using Subgroup Test Ballooning on English as our example:

Justification	**Frame**
English is easy to learn; resources are abundant.	Opportunity
English is the most widely used language.	Desert
It is up to industry/research labs to decide.[10]	Procedure
English users have more advanced needs.	Need
Other technologies are for English markets first.	Reference

We have anecdotally come across all of the five frames in discussions in the AI/NLP community–and some have also surfaced in the academic literature (Utiyama and Isahara, 2007; Anastasopoulos et al., 2019; Blasi et al., 2022; Lewis et al., 2020)–but the list is likely incomplete, and the frames listed may differ significantly in popularity. This remains left for a more systematic study to decide.[11]

[9]That is: Income inequality is legitimate if everyone had (formally and substantively) equal opportunities to advantage (Opportunity); if everyone is compensated proportionally in terms of input (e.g., working time, education) and output (e.g., corporate success, social returns); if the inequality results from an agreed-upon established process (Procedure); if it reflects intrinsic or functional needs (Need); or if inequality is proportional to accepted standards in a particular domain (Reference).

[11]One complicating factor is that some frames are used more explicitly than others. Opportunity arguments (Utiyama and Isahara, 2007; Anastasopoulos et al., 2019) and Desert arguments (Blasi et al., 2022; Lewis et al., 2020) are abundant in the academic literature, whereas you rarely see explicit Procedure, Need and Reference arguments, except for indirect Need arguments from researchers who are worried that AI/NLP researchers working on low-resource languages develop technologies for people who do not see the need for them. Our response to this form of justification of inequality would be to agree with the basic assumption that we are in no position to decide on behalf of people what technologies they ought to adopt. Our conclusion is different, however: If we are not to decide for people, we need to make technologies available to them. Otherwise we have decided on their behalf.

Thought Experiment: The Egalitarian Martian 4 Our Martian field worker sees significant push back from his fellow Martians, who find his egalitarian proposal too radical. The push back has nothing to do with Supreme Leader Xaroline's correction, giving dyslexics special status. Neither is it because his fellow Martians worry egalitarian fairness will slow technological development. The push back comes from other government workers feeling egalitarian fairness is somehow unfair. The government workers reason as follows: If early adopters develop new habits, their technological maturity level increases, but egalitarian fairness will mean they have to wait for everyone to catch up, before they can enjoy new technologies. Question is whether their technological maturity justifies inequality? Our Martian field worker pushes back against this idea in a televised address to the nation: "Nothing in these policies prevent new technologies from being developed," he says, "as long as these technologies work equally good for all of us."

8.5 Why Relative, not Absolute Position

We will present three arguments for why, in the context of AI/NLP performance across subgroups, the relative position of the least advantaged is more important than their absolute position:

1. *Staying On the Radar*: The absolute performance will improve rapidly for active end users but if we want to keep all subgroups represented among our end users, without anyone falling off our radar, we need to minimize the relative performance disparity across subgroups.

2. *Being Right for the Right Reasons*: High performance on data from some subgroups but high overall performance disparities, is typically sign of overfitting, i.e., reliance on spurious correlations in the data. Minimizing disparity across subgroups increases the likelihood of finding robust estimators, i.e., models that rely on factors that are robustly predictive.

3. *Breaking the Hype Cycle*: The absolute position of the most advantaged subgroup sets the expectations of everyone. The least advantaged will be more disillusioned with the technology, the larger the gap between them and best-case performance on the most advantaged subgroup.

Argument 1: Staying On the Radar Turn-around in AI/NLP is fast, and models quickly go from struggling on new benchmarks to surpassing human performance–a phenomenon known as *benchmark saturation* (Kiela et al., 2021). Benchmarks seem to saturate faster and faster and often within the first year or two of their publication.[12] Absolute performance is thus a rapidly moving target. The benefits of tolerating inequality in favor of higher absolute performance may, in other words, be short-lived. Also, tolerating performance gaps may create a vicious cycle. If a technology is clearly biased against your group, chance is you will abandon the technology. Your group will become under-represented in the pool of end users, performance on your group will not be optimized for and eventually deteriorate, making it less likely that your peers will choose to become users of the technology in question. Your subgroup falls of the technology's radar, so to speak.

Argument 2: Being Right for the Right Reasons Young and old speak slightly different languages but learning what groups have in common reduces the change of relying on spurious correlations. Consider the following examples of group-specific spurious correlations: a) In movie review sentiment analysis, both young and old will speak of good and bad movies but groups may differ on whether they associate specific words such as *fast-paced* or *psychological* with positive or negative polarity. In reality, these words are not sentiment words but simply words that (within groups) covary with sentiment. Young people may associate the word *fast-paced* with positive sentiment but this predictor is not robust across groups. Systems that rely on such spurious correlations will be sensitive to drift in the user population, whereas a system that does not rely on such words – potentially compromising performance a bit – will be more robust. Such robustness is not just motivated by temporal drift but also the need to adopt to unseen product types and review platforms. b) In machine translation from English to German, reordering is sensitive to phrase boundaries. Punctuation is often a give-away for phrase boundaries but subgroups may differ in how consistently they use punctuation. Young people are, for example, less inclined to use punctuation in weblogs (Burger and Henderson, 2006). We

[12]The SuperGLUE Benchmark (`https://super.gluebenchmark.com`), for example, was launched on 6 May 2019 with a machine baseline and a human baseline 18.3% ahead of it, but was saturated by a newer model on 6 Jan 2021, surpassing human performance by half a percentage point.

know from psycho-linguistics that human sentence processing is *not* sensitive to punctuation, and in many domains, say in emails or on some social media, punctuation is almost entirely absent. It should thus be possible to infer phrase boundaries in the absence of punctuation, and a model that learns to do so, will obviously be more robust to variation.

Argument 3: Breaking the Hype Cycle Our third argument for worrying more about the relative position of the least advantaged than their absolute position, has a more psychological flavor. In practice, technology development is often a matter of anticipating user disappointment. A machine translation model may err rarely but if it errs on translation problems that are considered obvious by most, end users will loose trust in the translation model. Since the first wave of early adopters of a new technology will be responsible for initial reviews and, possibly, early hype, they also set the expectations of later waves of users. Such users will likely be disillusioned if initial reviews and hype were based on overly optimistic performance estimates, because the technology was fitted on data sampled almost exclusively from the subgroup to which the early adopters belong. Voice assistants has seen performance gaps with some subgroups, e.g., females, low/middle-income, and low product experience, and some of these subgroups are also particularly sensitive to products' failure to meet their expectations (Brill et al., 2019). Such dynamics easily create vicious cycles.

8.6 Fair Systems and Fair Metrics

We have argued that Rawlsian fairness – widely adopted in AI/NLP research – admits for loopholes that are actively facilitating the development of biased technologies, including technologies biased toward certain languages and certain subgroups.[13] We have given examples of different types of justification of inequalities facilitated by such loopholes, and suggested a more egalitarian notion of fairness to prevent such practices. In general, we have argued that, in the context of AI/NLP model development, the relative position of the least advantaged may be more important than their absolute position.

[13]Fairness across languages and fairness across subgroups form a continuum because of dialects and sociolects. Somewhat ironically, the (demographic) AI/NLP fairness literature has been shown to be particularly biased toward English (Takeshita et al., 2020; Ruder et al., 2022).

Finally, we want to emphasize that adopting a more egalitarian fairness principle will not 'solve' the fairness challenges in AI/NLP research once and for all.

Consider, for example, the role of performance metrics: Machine translation systems have to be fair with respect to dyslexia (a protected attribute). This, in our view, would mean that the performance of such a system, as measured through standard performance metrics such as BLEU, METEOR or something better,[14] must be equal for dyslexics and non-dyslexics. Of course output with equal BLEU scores need not be equally useful to two target groups, and in this case, it is plain to see that the machine translation system would have to produce somewhat-easy-to-read output to be useful to dyslexics. So, does this not show how a system with equal performance across groups can still be unfair?

We would argue that the limited usefulness of machine translation systems for dyslexics, is not a sign that these systems are as such unfair but that the evaluation metrics we use to evaluate them, are unfair. It is, in other words, unfair to dyslexics that readability is not part of how machine translation systems are evaluated. Clearly, adopting fairer metrics would have an impact on what models are induced but for now, models can be fair *with respect to* standard metrics without considering readability. We believe it is important to distinguish between model fairness and metric fairness to move research forward in the best possible way. AI/NLP models can also be unfair in other ways, e.g., protecting the privacy of end users using one operating system rather than another. We believe, however, that performance disparities across languages and groups are one of the most important roadblocks on the path toward fair AI/NLP models that do not widen existing inequality gaps between us.

8.7 Concluding Remarks

Our argument in the above is simple: Rawlsian fairness–i.e., his Difference Principle–is too permissive to prevent common AI/NLP practices that actively contribute to global and social inequality gaps. Examples include

[14]BLEU and METEOR are performance metrics based on the n-gram overlap between a system translation and one or more human reference translations. The metrics generally seem to correlate with human judgments of translation quality but they are considered far from perfect.

test-ballooning technologies on specific target groups that are known to be adoption ready, or evaluation technologies on representative samples of the current end user population. We suggest a more egalitarian definition of fairness–adopted from Kai Nielsen's work on justice at large. We believe this will be an important step toward more sustainable AI/NLP research and development.

The trajectory of AI/NLP research and development is tied to its evaluation methodologies and performance metrics, and recent focus on fairness is an opportunity to course-correct for unjust practices by implementing less biased evaluation methodologies. Details matter, however, and now is a good time to get things right. If we want users and yet-to-become-users around the world to benefit equally from AI/NLP technologies, and if we want to avoid contributing to existing inequality gaps through unjust practices, we need to close the loopholes in current definitions of AI/NLP fairness.

Chapter 9

Concluding Remarks

This dissertation has been dedicated to addressing the imperative issue of equitable and inclusive Natural Language Processing (NLP). Our firm belief is that NLP should be designed and implemented with equity across various social groups, ensuring that no particular group is unfairly advantaged or left marginalized.

In our exploration, we meticulously analyzed models across a spectrum of NLP tasks, ranging from syntactic analysis and summarization to speech recognition and language modeling. We investigated performance across social groups with an eye toward intersectionality. Our scrutiny extended beyond mere performance evaluations, delving into the intricate relationship and independence of bias and fairness. A critical examination of the implications of Rawlsian fairness in the context of NLP was also offered, adding a layer of philosophical depth to our inquiry.

Our methodological approach has been two-fold: first, an in-depth evaluation and analysis of performance disparities for various social groups, and second, a high-level discussion of fairness and bias in the broader landscape of NLP. This dual-layered strategy not only uncovers specific issues but also provides a comprehensive understanding of the overarching challenges in creating equitable and inclusive NLP.

The collection of papers comprising this dissertation spans nearly a decade of research and advancements from 2015 to 2024. This temporal expanse is particularly noteworthy, considering the rapid evolution and transformative shifts within the field of NLP during this period.

An important milestone during this time frame has been the arrival of Transformer-based models and their expansion and dominance in the field of

NLP. While Transformer models have undoubtedly brought about groundbreaking improvements, the one-size-fits-all approach does not address the nuanced needs of diverse tasks and linguistic communities, nor is it necessarity desirable (Bender et al., 2021). Moreover, not every task necessitates the computational weight of a large language model; there remains a demand for smaller, more efficient models that can cater to specific linguistic tasks effectively. The pragmatic approach of recognizing when a simpler model suffices underscores the importance of efficiency and task-specific optimization in NLP. Finally, it is worth contemplating whether the progress on large language models should be the sole focus of the NLP community. A "top-down perspective", as Bender and Koller (2020) remark, can be necessary to ensure continuous development in the direction of the end goal, general AI or – in the perspective of this dissertation – *equitable and inclusive* general AI.

POS Tagger Performance for AAVE Speakers

In Chapters 2, 3 and 4, we investigated and mitigated the issue that state-of-the-art POS taggers fail on data from speakers of African-American Vernacular English (AAVE). We showed in Chapter 2 that the performance of off-the-shelf POS taggers correlate negatively with many aspects associated with this social group. In Chapter 3, we created a testsuite for evaluating POS taggers across varieties of English. We included three new datasets which all contain features from AAVE to ensure better evaluation against this particular social group. We performed an in-depth analysis of the language in the dataset and found that the distribution of annotated POS tags varies distinctly from more frequently used datasets, which explains the poor performance and underscores the need for diversity in test and training datasets. We noted that performance on these three new datasets is significantly lower than on other datasets of English. In Chapter 4, we overcame this performance disparity by building a POS tagger for AAVE-like language. We outperformed three state-of-the-art POS taggers and achieve a 55% error reduction by including AAVE-like data into training and learning from automatically and ambiguously annotated data.

Performance in Summarization Systems across Demographics

Moving on from POS tagging, we investigated how the social identities of data creators and system raters influence the end result in summarization systems. We showed that the social identities of the human-in-the-loop during development should not be ignored, since this may influence model development and lead to favorization of a particular social group. We found that raters from different social groups have different summarization preferences, and that performance results differ dependent on the gender of the writer of the gold summary. This indicates that the model has been optimized for the writing style of a particular social group. We also found that the simple, extractive summarization system, TextRank, significantly outperformed MatchSum, a more complex, abstractive system for one social group, young women. This serves as a valuable example of the importance of including intersectionality in the evaluation of models and underscores the necessity of continuing the development and interest in smaller and simpler models during the age of large Transformer-based models. There is recent work on improving summarization evaluation (Gehrmann et al., 2023) and particularly an interest in using large language models (LLM) to evaluate generated summaries as a cheap and fast alternative to human evaluators (Liu et al., 2023; Chiang and Lee, 2023). While we encourage the work on improving the notoriously difficult task of evaluating summary quality, it is worth contemplating whose preferences are perpetuated when using large language models for evaluation?

Fairness in ASR Systems across Intersectional Demographic Groups

In the evaluation of ASR models discussed in Chapter 6, we found that prominent ASR models displayed performance disparity across various demographic groups, with adolescents being particularly affected across both binary genders. The investigation further unveiled a concerning trend: a significant negative correlation between performance parity and model size. This suggests that as models become larger and ostensibly more advanced, their fairness diminishes. This issue is particularly consequential given the trajectory in NLP development, wherein continuously larger models are being deployed in both speech and text processing domains.

On the Independence of Bias and Fairness in NLP

Turning from the experimental section of this dissertation, we argue in Chapter 7 that representational bias and empirical fairness can be independent. We provide both practical and theoretical support for our claim and show that in frequently-used language models, there is no correlation between bias metrics and fairness metrics. We show how this should not be a surprise as we see examples in the social sciences literature.

Rawlsian Fairness in NLP

Finally, in Chapter 8, we discuss the broad adaptation of Rawlsian fairness and pinpoint how this definition of fairness – especially his Difference Principle – introduces loopholes that can inadvertently exacerbate inequality. We argue for an adoptation of Kai Nielsen's work on fairness instead, since we find this to be more truly egalitarian.

9.1 Future Directions

In the following, we provide three promising and important directions for further research into equitable and inclusive NLP across social groups.

Optimizing for equal performance rather than optimal performance. How would the landscape of NLP systems transform if our optimization goals shifted from achieving high performance to emphasizing equal performance across diverse demographics? By prioritizing equal performance, we would be challenging the prevalent disparities present in current NLP technologies and striving for a more inclusive and equitable technological ecosystem.

By shifting objectives, advancements in technology would prioritize maximizing the performance of the worst-off. This approach holds the potential to address and mitigate performance disparities that may be ingrained in current systems, fostering a more just and equitable NLP landscape.

Moreover, emphasizing equal performance in NLP systems would likely lead to innovations that consider a broader spectrum of linguistic nuances, cultural contexts, and diverse user experiences. This shift in focus could contribute to the development of technologies that are not only technically

proficient but also ethically sound, reflecting a commitment to fairness and inclusivity.

Understanding the long-term effects of performance disparity across social groups. We have a lack in understanding of the long-term effects of performance disparity across groups in NLP. Given that NLP systems have not been deeply integrated into our society for an extended period, our comprehension of the profound impact these technologies wield on our social structures remains constrained. Furthermore, we lack information regarding how systemic disparities in NLP performance might contribute to or exacerbate issues of racism, sexism, or discrimination within our societal framework. Thorough research is required to unravel the complexities and interconnections between technological performance and the perpetuation or attenuation of societal inequities. As these technologies become more ingrained in our daily lives, a deeper understanding of their potential contributions to, or mitigation of, social challenges becomes imperative. Longitudinal studies naturally take time, but perhaps the intensity of usage could serve as a proxy for investigating the repercussions of performance disparities among diverse social groups. Such an investigation could seek to illuminate whether a snowball effect of disparity and discrimination follows from the performance disparities we see when evaluating individual NLP applications. Since users of one application are often users of many applications, an intensified discriminatory effect – especially for individuals of several vulnerable social identities, ie. young, black women – is not unimaginable.

Performance parity for children. We advocate for a heightened focus on research dedicated to understanding the intricacies of children's language and enhancing the performance of NLP models in interpreting and responding to children. Children constitute a distinct user group with evolving language skills, cognitive development, and linguistic expressions, warranting dedicated research efforts to tailor NLP technologies to their specific needs.

This call is grounded in the acknowledgment that, even with existing age regulations, children constitute a substantial portion of users within the landscape of NLP applications. This prevalence is particularly pronounced in light of the integration of chatbots and NLP functionalities on social media platforms. Despite regulatory frameworks aimed at protecting younger users, the reality is that children actively engage with NLP tools, contributing

significantly to the user base.

Our call to action stems from the realization that without explicitly considering and evaluating children's language use in our assessments of NLP models, we overlook the extensive interactions, performance nuances, and potential challenges associated with this sizable user demographic.

The ubiquity of NLP applications, coupled with the accessibility and appeal of interactive features like chatbots, has led to an increased presence of children within these digital spaces. Their interactions, however, often occur in a regulatory blind spot, where the specific nuances of children's language and their experiences with NLP tools remain unexamined. The oversight in not incorporating children's language use in our evaluation frameworks results in a lack of awareness regarding how these young users navigate, perceive, and are impacted by NLP applications.

Furthermore, disparities in NLP performance may disproportionately affect children with protected attributes, ie. children of immigrants, children with disabilities, children who speak a marked dialect, or children from non-white parents. By evaluating how NLP models perform on young users and ensuring a high performance for this demographic, we can ensure safe, effective, and inclusive digital environments for the younger generation, where demography does not play a role in performance or accessibility.

Bibliography

Martine Adda-Decker and Lori Lamel. Do speech recognizers prefer female speakers? In *Proceedings of Interspeech*, 2005.

Hanan Aldarmaki and Mona Diab. Robust part-of-speech tagging of Arabic text. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, Beijing, China, 2015.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. XAI for Transformers: Better explanations through conservative propagation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR, 2022. URL https://proceedings.mlr.press/v162/ali22a.html.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. Neural machine translation of text from non-native speakers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3070–3080, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1311. URL https://aclanthology.org/N19-1311.

Evan L. Antworth. Book reviews: Computational morphology: Practical mechanisms for the English lexicon. *Computational Linguistics*, 18(3), 1992. URL https://aclanthology.org/J92-3008.

Sharon Ash and John Myhill. Linguistic correlates of inter-ethnic contact. In

David Sankoff, editor, *Diversity and Diachrony*, pages 33–44, Amsterdam and Philadelphia, 1986. John Benjamins Publishing Co.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2): 135–160, 2014.

Julian Bank. Mr. Winterkorn's pay: A typology of justification patterns of income inequality. *Social Justice Research*, 29(2):228–252, 2016.

Lizzie Barmes. Equality law and experimentation: The positive action challenge. *The Cambridge Law Journal*, 68(3):623–654, 2009. ISSN 00081973, 14692139. URL http://www.jstor.org/stable/40388838.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. http://www.fairmlbook.org.

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. Variations of the similarity function of textrank for automated summarization. *CoRR*, 2016. URL http://arxiv.org/abs/1602.03606.

Marion Bartl, Malvina Nissim, and Albert Gatt. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.1.

John Baugh. Linguistic profiling. In *Black Linguistics*, pages 167–180. Routledge, 2005.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL https://aclanthology.org/2020.acl-main.463.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of FAccT*, 2021.

Delphine Bernhard and Anne-Laure Ligozat. Hassle-free POS-tagging for the Alsatian dialects. In Marcos Zampieri and Sascha Diwersy, editors, *Non-Standard Data Sources in Corpus Based-Research*, pages 85–92. ZSM Studien, 2013.

Sudeep Bhatia. The semantic representation of prejudice and stereotypes. *Cognition*, 164:46–60, 2017. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2017.03.016. URL https://www.sciencedirect.com/science/article/pii/S0010027717300872.

Felix Biessmann. Automating political bias prediction. arXiv preprint, 2016. arXiv:1608.02195.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.376. URL https://aclanthology.org/2022.acl-long.376.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.485. URL https://aclanthology.org/2020.acl-main.485.

Marcely Zanon Boito, Laurent Besacier, Natalia A. Tomashenko, and Yannick Estève. A study of gender impact in self-supervised models for speech-to-text systems. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1278–1282. ISCA, 2022. doi: 10.21437/Interspeech.2022-353. URL https://doi.org/10.21437/Interspeech.2022-353.

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. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

Rahma Boujelbane, Meriem Ellouze Khemekhem, and Lamia Hadrich Belguith. Mapping rules for building a Tunisian dialect lexicon and generating corpora. In *International Joint Conference on Natural Language Processing*, pages 419–428, 2013.

Tom Brill, Laura Munoz, and Richard Miller. Siri, Alexa, and other digital assistants: A study of customer satisfaction with artificial intelligence applications. *Journal of Marketing Management*, 35, 11 2019. doi: 10.1080/0267257X.2019.1687571.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. Understanding the origins of bias in word embeddings. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/brunet19a.html.

John Burger and John Henderson. An exploration of observable features related to blogger age. In *AAAI 2006 Spring Symposia*, pages 15–20, 01 2006.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the Independence of Association Bias and Empirical Fairness in Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability,*

and Transparency, pages 370–378, Chicago, IL, USA, June 2023. Association for Computing Machinery.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. doi: 10.1126/science.aal4230. URL https://www.science.org/doi/abs/10.1126/science.aal4230.

Aylin Caliskan, Pimparkar Parth Ajay, Tessa Charlesworth, Robert Wolfe, and Mahzarin R. Banaji. Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 156–170, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392471. doi: 10.1145/3514094.3534162. URL https://doi.org/10.1145/3514094.3534162.

Yang Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-short.62. URL https://aclanthology.org/2022.acl-short.62.

Phillip Carter. Shared spaces, shared structures: Latino social formation and African American English in the U.S. south. *Journal of Sociolinguistics*, 17:66–92, 2013.

Alessandro Castelnovo, Riccardo Crupi, Greta Greco, Daniele Regoli, Ilaria Giuseppina Penco, and Andrea Claudio Cosentini. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, 12(1), March 2022. doi: 10.1038/s41598-022-07939-1. URL https://doi.org/10.1038/s41598-022-07939-1.

Ilias Chalkidis, Tommaso Pasini, Sheng Zhang, Letizia Tomada, Sebastian Schwemer, and Anders Søgaard. FairLex: A multilingual benchmark for evaluating fairness in legal text processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4389–4406, Dublin, Ireland, May 2022. Association

for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.301. URL https://aclanthology.org/2022.acl-long.301.

Kaytlin Chaloner and Alfredo Maldonado. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3804. URL https://aclanthology.org/W19-3804.

Kai-Wei Chang, Vinodkumar 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*, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-2004.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcss-1.16. URL https://aclanthology.org/2020.nlpcss-1.16.

Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.870. URL https://aclanthology.org/2023.acl-long.870.

Michael Collins. Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. In *EMNLP*, 2002.

Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*, 2017.

David Cremer and Garry Kasparov. The ethics of technology innovation: a double-edged sword? *AI and Ethics*, 2, 09 2021. doi: 10.1007/s43681-021-00103-x.

Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *The University of Chicago Legal Forum*, 140:139–167, 1989.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267, 11 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00425. URL https://doi.org/10.1162/tacl_a_00425.

Dipanjan Das and Slav Petrov. Unsupervised part-of-speech tagging with bilingual graph-based projections. In *Proceedings of ACL*, pages 256–263, 2011.

Erenay Dayanik and Sebastian Padó. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.404. URL https://aclanthology.org/2020.acl-main.404.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *CoRR*, abs/2112.07447, 2021. URL https://arxiv.org/abs/2112.07447.

Janez Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages

1268–1272. ISCA, 2022. doi: 10.21437/Interspeech.2022-10816. URL https://doi.org/10.21437/Interspeech.2022-10816.

Gabriel Doyle. Mapping dialectal variation by querying social media. In *Proceedings of EACL*, pages 98–106, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://www.aclweb.org/anthology/P18-1128.

Kevin Duh and Katrin Kirchhoff. POS tagging of dialectal Arabic: A minimally supervised approach. In *Proceedings in the ACL Workshop on Computational Approaches to Semitic Languages*, 2005.

Coen Van Duijnhoven. Predicting political preference through content- and stylistic text features and distant labeling. Master's thesis, Tilburg University, 2018.

Greg Durrett, Taylor Berg-Kirkpatrick, and Dan Klein. Learning-based single-document summarization with compression and anaphoricity constraints. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1998–2008, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1188. URL https://www.aclweb.org/anthology/P16-1188.

Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC Press, 1994.

Jacob Eisenstein. Phonological factors in social media writing. In *NAACL Workshop on Language Analysis in Social Media*, pages 11–19, Atlanta, Georgia, 2013a. Association for Computational Linguistics.

Jacob Eisenstein. Phonological factors in social media writing. In *Proceedings of NAACL Workshop on Language Analysis in Social Media*, 2013b.

Jacob Eisenstein. What to do about bad language on the Internet. In *Proceedings of NAACL-HLT*, 2013c.

Jacob Eisenstein. Systematic patterning of phonologically-motivated orthographic variation. *Journal of Sociolinguistics*, 19(2):161–188, April 2015.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of EMNLP*, 2010.

Jacob Eisenstein, Noah A. Smith, and Eric Xing. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of ACL*, 2011.

Kawin Ethayarajh and Dan Jurafsky. Utility is in the eye of the user: A critique of NLP leaderboards. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4846–4853, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.393. URL https://aclanthology.org/2020.emnlp-main.393.

Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech and Language*, 84, March 2024. URL https://www.sciencedirect.com/science/article/pii/S0885230823000864.

Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu. Summarizing drug information in medline citations. *AMIA Symposium*, page 254—258, 2006. ISSN 1942-597X. URL https://europepmc.org/articles/PMC1839479.

D.R. Forsyth. *Group Dynamics*. Cengage Learning, Boston, MA, 2009. ISBN 9780495599524. URL https://books.google.dk/books?id=RsMNiobZojIC.

Jennifer Foster, Özlem Çetinoğlu, Joachim Wagner, Joseph Le Roux, Joakim Nivre, Deirdre Hogan, and Josef van Genabith. From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of IJCNLP*, 2013.

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im)possibility of fairness, 2016. URL https://arxiv.org/abs/1609.07236.

Niklas Friedrich, Anne Lauscher, Simone Paolo Ponzetto, and Goran Glavaš. DebIE: A platform for implicit and explicit debiasing of word embedding spaces. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 91–98, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-demos.11. URL https://aclanthology.org/2021.eacl-demos.11.

Dan Garrette and Jason Baldridge. Learning a part-of-speech tagger from two hours of annotation. In *Proceedings of NAACL-HLT*, pages 138–147, 2013.

Sebastian Gehrmann, Elizabeth Clark, and Thibault Sellam. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. *JAIR*, 77, 2023.

Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles, June 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W10-0722.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of ACL*, 2011.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.150. URL https://aclanthology.org/2021.acl-long.150.

Hila Gonen and Yoav Goldberg. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1061. URL https://aclanthology.org/N19-1061.

Stephen Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1386–1390, 2015.

Mark Graham, Scott Hale, and Devin Gaffney. Where in the world are you? Geolocation and language identification on Twitter. *The Professional Geographer*, 66(4), 2014.

Yvette Graham. Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 128–137, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1013. URL https://www.aclweb.org/anthology/D15-1013.

Bo Han and Timothy Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #twitter. In *ACL*, 2011.

Victor Petrén Bach Hansen and Anders Søgaard. Is the lottery fair? Evaluating winning tickets across demographics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3214–3224, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.284. URL https://aclanthology.org/2021.findings-acl.284.

Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 1934–1943. PMLR, 2018. URL http://proceedings.mlr.press/v80/hashimoto18a.html.

Brian Hedden. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2):209–231, 2021. doi: 10.1111/papa.12189.

P.J. Henry, Sarah E. Butler, and Mark J. Brandt. The influence of target group status on the perception of the offensiveness of group-based slurs. *Journal of Experimental Social Psychology*, 53:185–192, 2014. ISSN 0022-1031. doi: https://doi.org/10.1016/j.jesp.2014.03.012. URL https://www.sciencedirect.com/science/article/pii/S0022103114000390.

Amac Herdagdelen and Marco Baroni. Stereotypical gender actions can be extracted from web text. *Journal of the American Society for Information Science and Technology*, 62:1741–1749, 2011.

Harry Holzer and David Neumark. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, September 2000. doi: 10.1257/jel.38.3.483. URL https://www.aeaweb.org/articles?id=10.1257/jel.38.3.483.

Dirk Hovy and Eduard Hovy. Exploiting partial annotations with em training. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 31–38. Association for Computational Linguistics, 2012.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 483–488, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-2079. URL https://www.aclweb.org/anthology/P15-2079.

Dirk Hovy, Barbara Plank, and Anders Søgaard. When POS datasets don't add up: Combatting sample bias. In *Proceedings of LREC*, 2014.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. User review-sites as a source for large-scale sociolinguistic studies. In *Proceedings of WWW*, 2015a.

Dirk Hovy, Barbara Plank, Héctor Martínez Alonso, and Anders Søgaard. Mining for unambiguous instances to adapt part-of-speech taggers to new

domains. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1256–1261, 2015b.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.487. URL https://aclanthology.org/2020.acl-main.487.

Sébastien Jean, Orhan Firat, and Melvin Johnson. Adaptive scheduling for multi-task learning. In *ArXiv 1909.06434*, 2019.

Jacob Jensen, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. *Brookings Papers on Economic Activity*, 43(2 (Fall)):1–81, 2012. URL https://ideas.repec.org/a/bin/bpeajo/v43y2012i2012-02p1-81.html.

Anders Johannsen, Dirk Hovy, and Anders Søgaard. Cross-lingual syntactic variation over age and gender. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 103–112, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1011. URL https://www.aclweb.org/anthology/K15-1011.

Barbara Johnstone. Place, globalization, and linguistic variation. In *Linguistic Variation: Critical Reflections*, pages 65–83. Oxford University Press, 2004.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of studying and processing dialects in social media. In *Proceedings of the ACL Workshop on Noisy User-generated Text*, 2015.

Anna Jørgensen and Anders Søgaard. A Test Suite for Evaluating POS Taggers Across Varieties of English. In *25th International World Wide Web Conference*, pages 615–618, Montreal, Canada, April 2016. International World Wide Web Conferences Steering Committee.

Anna Jørgensen and Anders Søgaard. Evaluation of Summarization Systems across Gender, Age, and Race. In *Proceedings of the Third Workshop on*

New Frontiers in Summarization, pages 51–56, Online and in the Dominican Republic, November 2021. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Challenges of Studying and Processing Dialects in Social Media. In *Proceedings of the Workshop on Noisy User-generated Text*, pages 9–18, Beijing, China, July 2015. Association for Computational Linguistics.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. Learning a POS tagger for AAVE-like language. In *Proceedings of NAACL-HLT 2016*, pages 1115–1120, San Diego, CA, United States, June 2016. Association for Computational Linguistics.

Anna Katrine Jørgensen and Anders Søgaard. Rawlsian AI fairness loopholes. *AI and Ethics*, 3(2):1185–1192, 2022.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.111.

Barbara Kaup, Rolf Zwaan, and Jana Lüdtke. The experiential view of language comprehension: How is negation represented? In *Higher level language processes in the brain: Inference and comprehension processes*, pages 255–288. American Psychological Association, 11 2013.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *Information Technology Convergence and Services*, 2016. URL https://api.semanticscholar.org/CorpusID:12845273.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *PNAS*, 117(14):7684–7689, 2020.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3823. URL https://aclanthology.org/W19-3823.

William Labov. *Language in the inner city: Studies in the Black English vernacular*, volume 3. University of Pennsylvania Press, 1972a.

William Labov. *Sociolingustic Patterns*. University of Pennsylvania Press, Philadelphia, PA, 1972b.

William Labov. The intersection of sex and social class in the course of linguistic change. *Language Variation and Change*, 2:205–254, 7 1990. ISSN 1469-8021. doi: 10.1017/S0954394500000338. URL http://journals.cambridge.org/article_S0954394500000338.

William Labov. Unendangered dialects, endangered people. In Natalie Schilling-Estes, editor, *GURT'06*, 2006.

William Labov, Sharon Ash, and Charles Boberg. *The Atlas of North American English Phonetics, Phonology and Sound Change*. Mouton de Gruyter, New York, NY, 2005.

Brian Larson. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1601. URL https://aclanthology.org/W17-1601.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.411. URL https://aclanthology.org/2021.findings-emnlp.411.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://aclanthology.org/2020.acl-main.653.

Mike Li, Hongseok Namkoong, and Shangzhou Xia. Evaluating model performance under worst-case subpopulations. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 17325–17334, Vancouver, CA, 2021. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2021/file/908075ea2c025c335f4865f7db427062-Paper.pdf.

Shen Li, João V. Graça, and Ben Taskar. Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

Wen Li and Markus Dickinson. Gender prediction for chinese social media data. In *Proceedings of Recent Advances in Natural Language Processing*, 2017.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-1013.

Feifan Liu and Yang Liu. Correlation between ROUGE and human evaluation of extractive meeting summaries. In *Proceedings of ACL-08: HLT, Short Papers*, pages 201–204, Columbus, Ohio, June 2008. Association for

Computational Linguistics. URL https://www.aclweb.org/anthology/P08-2051.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. Does gender matter? towards fairness in dialogue systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4403–4416, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.390. URL https://www.aclweb.org/anthology/2020.coling-main.390.

Shengchao Liu, Yingyu Liang, and Anthony Gitter. Loss-balanced task weighting to reduce negative transfer in multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9977–9978, Jul. 2019. doi: 10.1609/aaai.v33i01.33019977. URL https://ojs.aaai.org/index.php/AAAI/article/view/5125.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using Gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.153.

Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. Does enforcing fairness mitigate biases caused by subpopulation shift? In *Proceedings of NeurIPS*, 2021.

Joshua Martin and Kevin Tang. Understanding racial disparities in automatic speech recognition: The case of habitual "be". In *Proceedings of ISCA*, 10 2020. doi: 10.21437/Interspeech.2020-2893.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1063. URL https://aclanthology.org/N19-1063.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), jul 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Miriam Meyerhof. *Introducing Sociolinguistics*. Routledge, 2006.

Thomas Miconi. The impossibility of "fairness": a generalized impossibility result for decisions. *arXiv: Applications*, 2017. doi: 10.48550/ARXIV.1707.01195.

Rada Mihalcea and Paul Tarau. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3252.

T. Mikolov, W.T. Yih, and G. Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 2013a.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*, 2013b. URL https://api.semanticscholar.org/CorpusID:5959482.

Antonio Alexander Morgan-Lopez, Annice E Kim, Robert F. Chew, and Paul Ruddle. Predicting age groups of twitter users based on language and metadata features. *PLoS ONE*, 12, 2017.

R Naroll. Two solutions to Galton's problem. *Philosophy of Science*, 28, 1961.

Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/N04-1019.

Mikel K. Ngueajio and Gloria Washington. Hey ASR system! Why aren't you more inclusive? Automatic Speech Recognition systems' bias and proposed bias mitigation techniques. A literature review. In *Proceedings of HCI*, 2022. URL https://arxiv.org/pdf/2211.09511.pdf.

Kai Nielsen. Radical egalitarian justice: Justice as equality. *Social Theory and Practice*, 5(2):209–226, 1979. doi: soctheorpract1979523.

Mike Noon. The shackled runner: time to rethink positive discrimination? *Work, Employment and Society*, 24(4):728–739, 2010. doi: 10.1177/0950017010380648. URL https://doi.org/10.1177/0950017010380648.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, 2013.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of LREC*, 2011.

Mohammad Pezeshki, Sekouba Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Neural Information Processing Systems*, 2020. URL https://api.semanticscholar.org/CorpusID:227013102.

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic evaluation of linguistic quality in multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 544–554, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P10-1056.

Barbara Plank, Dirk Hovy, Ryan McDonald, and Anders Søgaard. Adapting taggers to Twitter with not-so-distant supervision. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1783–1792, 2014a.

Barbara Plank, Dirk Hovy, and Anders Søgaard. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of EACL*, 2014b.

Karen Pollock, Guy Bailey, M.C. Berni, D.G. Fletcher, Linette Hinton, I.A. Johnson, et al. Phonological features of African American Vernacular English (AAVE), 2001. URL http://www.rehabmed.ualberta.ca/spa/phonology/features.htm.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv*, 2023.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer NLP, 2022. URL https://arxiv.org/abs/2205.12586.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. Classifying latent user attributes in Twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-generated Contents*, pages 37–44. ACM, 2010.

Sonya Rastogi, Tallese D. Johnson, Elizabeth M. Hoeffel, and Malcolm P. Drewery Jr. The black population: 2010. Technical report, US Census, September 2011. URL http://www.census.gov/prod/cen2010/briefs/c2010br-06.pdf.

John Rawls. *A Theory of Justice*. Belknap Press of Harvard University Press, Cambridge, Massachussets, 1 edition, 1971. ISBN 0-674-88014-5.

Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/2723d092b63885e0d7c260cc007e8b9d-Paper-round1.pdf.

John Rickford. *African American Vernacular English: Features, Evolution, Educational Implications*. Blackwell, Malden, MA, 1999.

John Rickford. Geographical diversity, residential segregation, and the vitality of African American Vernacular English and its speakers. *Transforming Anthropology*, 18(1):28–34, 2010.

JR Rickford, GJ Duncan, LA Gennetian, RY Gou, R Greene, LF Katz, RC Kessler, JR Kling, L Sanbonmatsu, AE Sanchez-Ordoñez, M Sciandra, E Thomas, and J Ludwig J. Neighborhood effects on use of african-american vernacular english. *PNAS*, 112(38):11817–11822, 2015.

Katherine Ritchie. Social identity, indexicality, and the appropriation of slurs. *Croatian Journal of Philosophy*, 17(2):155–180, 2017.

Sean Roberts and James Winters. Linguistic diversity and traffic accidents: lessons from statistical studies of cultural traits. *PLoS ONE*, 8(8), 2013.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. What's in a name? Reducing bias in bios without access to protected attributes. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1424. URL https://aclanthology.org/N19-1424.

Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 998–1008, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.78. URL https://aclanthology.org/2021.naacl-main.78.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. Square one bias in NLP: Towards a multi-dimensional exploration of the research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*,

pages 2340–2354, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.184. URL https://aclanthology.org/2022.findings-acl.184.

Shiori Sagawa*, Pang Wei Koh*, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020.

Gillian Sankoff. *The Social Life of Language*. University of Pennsylvania Press, 1980.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Natalie Schluter. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-2007.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.468. URL https://aclanthology.org/2020.acl-main.468.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages

81–95, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-aacl.8.

Anders Søgaard. *Semi-supervised learning and domain adaptation for NLP*. Morgan and Claypool, 2013.

Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.156. URL https://aclanthology.org/2021.eacl-main.156.

Karolina Stanczak and Isabelle Augenstein. A survey on gender bias in natural language processing. In *CoRR*, 2021.

Ian Stewart. Now We Stronger Than Ever: African-American syntax on Twitter. In *Proceedings of the Student Research Workshop to the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 26–30, Gothenburg, Sweden, April 2014.

Asa Cooper Stickland and Iain Murray. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *ICML*, 2019.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1159. URL https://aclanthology.org/P19-1159.

Oscar Täckström, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre. Token and type constraints for cross-lingual part-of-speech tagging. In Sharon Goldwater, editor, *Transactions of the Association for Computational Linguistics*, pages 1–12. Association for Computational Linguistics, 2013.

Masashi Takeshita, Yuki Katsumata, Rafal Rzepka, and Kenji Araki. Can existing methods debias languages other than English? first attempt to analyze and mitigate Japanese word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 44–55, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.gebnlp-1.5.

Yi Chern Tan and L. Elisa Celis. *Assessing Social and Intersectional Biases in Contextualized Word Representations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Rachael Tatman and Conner Kasten. Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions. In *Interspeech*, 2017.

Eric Thomas. Phonological and phonetic characteristics of african american vernacular english. *Language and Linguistic Compass*, 1(5):450–475, 2007a.

Erik R. Thomas. Phonological and phonetic characteristics of African American Vernacular English. *Language and Linguistics Compass*, 1(5):450–475, 2007b. URL http://repository.lib.ncsu.edu/publications/bitstream/1840.2/2062/1/.

Melanie Tosik, Carsten Lygteskov Hansen, Gerard Goossen, and Mihai Rotaru. Word embeddings vs. word types for sequence labeling: The curious case of CV parsing. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, pages 123–128, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-1517. URL https://aclanthology.org/W15-1517.

Joe Trotta and Oleg Blyahher. *Game done changed*: A look at selected AAVE features in the TV series *The Wire*. *Moderna Språk*, 2011.

Masao Utiyama and Hitoshi Isahara. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-1061.

Marieke van Erp and Paul Groth. Towards entity spaces. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2129–2137, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.261.

Anna Katrine van Zee, Marc van Zee, and Anders Søgaard. Group Fairness in Multilingual Speech Recognition Models. Manuscript under review, 2024.

Irina-Elena Veliche and Pascale Fung. Improving fairness and robustness in end-to-end speech recognition through unsupervised clustering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10096836.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, Vancouver, CA, 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/file/92650b2e92217715fe312e6fa7b90d82-Paper.pdf.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of EMNLP*, 2013.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. Inferring latent user properties from texts published in social media (demo). In *Proceedings of AAAI*, 2015.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897, 02 2020. ISSN 0891-2017. doi: 10.1162/coli_a_00391. URL https://doi.org/10.1162/coli_a_00391.

Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, and Trevor Cohn. Context-aware prediction of derivational word-forms. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 118–124, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2019.

Jason Wei and Eugene Santos Jr. Narrative origin classification of Israeli-Palestinian conflict texts. In *The Thirty-Third International FLAIRS Conference*, 2020.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797, Long Beach, California, 09–15 Jun 2019. PMLR. URL https://proceedings.mlr.press/v97/williamson19a.html.

Guillaume Wisniewski, Nicolas Pécheux, Sophir Gahbiche-Braham, and François Yvon. Cross-lingual part-of-speech tagging through ambiguous learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014.

Walt Wolfram. A sociolinguistic description of detroit negro speech. urban language series, no. 5. *Center for Applied Linguistics, Washington, D.C.*, 1969.

Walt Wolfram. The grammar of urban African American Vernacular English. In Kormann B. and E. Schneider, editors, *Handbook of Varieties of English*, pages 111–132, Berlin, 2004. Mouton de Gruyter.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. Worst-case-aware curriculum learning for zero and few shot transfer, 2020. URL https://arxiv.org/abs/2009.11138.

Sheng Zhang, Xin Zhang, Weiming Zhang, and Anders Søgaard. Sociolectal analysis of pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4581–4588, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.

emnlp-main.375. URL https://aclanthology.org/2021.emnlp-main.375.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1521. URL https://aclanthology.org/D18-1521.

Hao Zheng and Mirella Lapata. Sentence centrality revisited for unsupervised summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6236–6247, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1628. URL https://www.aclweb.org/anthology/P19-1628.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Barcelona, Spain, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W04-3252.