International Journal of Scientific Research and Modern Technology (IJSRMT) *ijsrmt.com*
Volume 4, Issue 12, 2025
DOI: https://doi.org/10.38124/ijsrmt.v4i12.1075

Bias Amplification in AI Models Trained on Low-Resource Data: A Sociolinguistic Audit

Nneoma Udeze[1]

Publication Date: 2025/12/30

Abstract

This paper explored how the rapid advancement of artificial intelligence (AI) has shaped global communication while simultaneously creating linguistic inequalities. AI systems that had been trained on high-resource languages showed high performance compared to those that had been trained on low-resource languages with low digital presence or datasets. The study revealed that data scarcity not only decreased the accuracy of the models but also exacerbated the existing social and linguistic inequalities. Amplification of bias was made possible by processes like statistical over-fitting, representational gaps, and normalizing standardized forms of language in training corpora. Therefore, speakers of marginalized or minoritized languages had increased chances of misclassification, exclusion, and digital discrimination. This study introduced a sociolinguistic audit framework that included community involvement, documentation of data sets, and modeling methods that were based on fairness. Overall all the review highlighted the necessity of fair AI development that would be based on collaborative, justice-oriented approaches that would consider the linguistic rights and agency of the underrepresented language communities.

Keywords: *Bias Amplification, Low-Resource Languages, Sociolinguistics, AI Ethics, Language Equity, Natural Language Processing (NLP).*

I. INTRODUCTION

The systems of artificial intelligence (AI) and natural language processing (NLP) have already defined the way individuals interact, retrieve information, and engage in online living. Nevertheless, these technologies mostly serve high-resource languages that have rich digital text, standardized orthographies, and institutionalization (Joshi et al., 2020). By comparison, thousands of languages around the world, and particularly indigenous languages, minoritized and economically marginalized languages are grossly underrepresented in the data used to drive modern AI. Thus, AI systems often fail when presented with these languages or dialects or even multilingualism (Blodgett et al., 2020).

This difference is not a technical matter only. The lack of data is in contact with social histories of colonization, socioeconomic marginalization, and linguistic hierarchy to generate and reproduce what sociolinguists define as language inequality (Ruiz, 1984; Lippi-Green, 2012). When the model of AI is trained on the data sets that favor standardized, elite, or institutional forms, it recreates them. More importantly, they tend to boost the biased tendencies within small datasets, which is called bias amplification (Zhao et al., 2017). These distortions are even more severe in low-resource settings, where datasets are small, skewed, or based on restricted domains.

For already marginalized communities speaking different languages, it is a matter of life and death. Automated speech systems misrecognize the speaker, dialect classification systems misclassify dialects, content moderation systems over-flag users, and digital services discriminate against them all uphold structural inequalities (Sap et al., 2019; Koenecke et al., 2020). It is a sociotechnical problem and can only be solved by not only enhancing model architectures but also being aware of the sociolinguistic realities that influence data availability and utilization.

The language-and-society approach, adopted by this paper, aims at analyzing the amplification of bias in AI systems under the influence of low-resource conditions. This paper argues that the key solutions have to be based on sociolinguistic knowledge, community involvement, and ethical data management. The subsequent sections (1) conceptualize the low-resource languages outside of the technical definition, (2) examine the mechanisms of amplifying bias, (3) overview real-world case studies of harmful outcomes, and (4) introduce a sociolinguistic

Udeze, N. (2025). Bias Amplification in AI Models Trained on Low- Resource Data: A Sociolinguistic Audit. *International Journal of Scientific Research and Modern Technology*, *4*(12), 109–114. https://doi.org/10.38124/ijsrmt.v4i12.1075

audit system of developing AI fairly. Finally, we argue that language justice needs to be propagated by re-conceptualizing AI as a technology that is neither neutral nor neutral, but a field of power, representation, and social inequality.

II. THE LOW-RESOURCE LANGUAGE PROBLEM

Low-resource languages are frequently represented in a strictly technical manner, languages with no large annotated corpora, digital texts, or parallel translation data, or NLP tools (Joshi et al., 2020). Nevertheless, this definition loses the insight into the more profound social, political, and historical forces that precondition the emergence of the languages as "resource-rich" and the digital sidelines of the languages. The sociolinguistic aspect of low-resource conditions not being intrinsic to languages per se is indicative of historical disparities of power, representation, and technological investment.

➢ *Defining Low-Resource Languages Beyond Data Availability*

Although the low-resource languages are frequently characterized by the community of NLP according to the size of the dataset, the quality of the corpus, or the performance metrics of the model, these parameters may be viewed as surface symptoms of the underlying structural problems. Numerous low-resource languages, such as Indigenous languages, minority ethnolects, Creoles, and local dialects, are historically marginalized in institutions that create and maintain written documents (Ruiz, 1984). These languages have also been limited in the online space by colonial language policies, socioeconomic disadvantage, and poor access to digital infrastructure (Piller, 2016). Therefore, the state of being a "low-resource" is not as much concerned with linguistic complexity, but rather:

- Past oppression (e.g., language oppression during colonialism)
- Economic disparities (literacy, access to the internet, and creation of digital content)
- Neglect at the level of the institution (absence of educational and technological aid).
- Standardized varieties are privileged in monolingual design ideologies.

Thus, the "low-resource" situation is a result of the sociopolitical dynamics, rather than the language shortage.

➢ *Sociolinguistic Implications of Data Scarcity*

Data scarcity has a severe sociolinguistic impact since it strengthens language hierarchies, both online and offline. The datasets that are used to train AI systems are filled with Standard Language Ideology, the notion that standardized, elite, or institutionally supported varieties are naturally superior (Lippi-Green, 2012). Consequently, the models that are trained with an assumption of these varieties usually misunderstand, degrade, or omit non-standard forms. For example.

- The AI systems that have been trained on Standard American English label the African American Vernacular English as either grammatically "incorrect" or emotionally "aggressive" (Sap et al., 2019).
- Multilingual societies that practice frequent code-switching are punished by the monolingual models that perceive such activity as a mistake (Dogruoz et al., 2021).
- Corpora are unaware of youth slang, oral storytelling traditions, and informal registers, and thus, a model is insensitive to culturally significant language practices.

Data scarcity also produces a digital feedback loop which could manifest as such:

- Poor digitization translates to poor online linguistic presence.
- The AI models fare badly on such languages.
- Low performance decreases the levels of trust and adoption of digital tools.
- The decreased usage leads to the reduction of available data.

This marginalization cycle is a reflection of societal inequalities in general, and therefore, technological marginalization is another aspect of linguistic injustice.

➢ *The Political Economy of Low-Resource Data*

The development of linguistic data is influenced by the unequal distribution of power. Most of the existing corpora in African, Indigenous, or minority languages are based on missionary work, political texts, or scholarly ethnographies, all of which represent a very limited worldview and cannot be used to reflect the sociolinguistic complexity of everyday life (Leonard, 2017). These sources incorporate colonial ideologies and formal registers and result in datasets that bias models to elite language forms. In this way, low-resource language as a problem can be best interpreted as a sociotechnical issue that is the result of historical injustices, data asymmetry, and power relations that control the visibility of which languages can be seen in the digital world.

III. MECHANISMS OF BIAS AMPLIFICATION

The amplification of bias is the process by which machine learning models not only acquire the biased tendencies in training data but also magnify them in their predictions (Zhao et al., 2017). This amplification is worse in low-resource settings, where the data is sparse, skewed, or unrepresentative. These distortions are caused by three fundamental mechanisms: statistical amplification, sampling bias, and representational gaps.

➢ *Statistical Amplification in Low-Data Regimes*

When resources are limited, the models use few examples to acquire linguistic patterns. Small datasets are an incentive to over fit to large amounts of data, when that data reflects large amounts of harmful stereotypes, or is non-generalizable to the wider speech community.

- *Key Mechanisms Include:*

✓ Overgeneralization: Models based on limited examples. In situations where a linguistic group has only a few examples, these limited patterns are used as general.
✓ Sparse counterexamples: When the data has biased relationships (e.g., gendered occupations, negative feelings associated with particular dialects), there should be no counterexamples, and therefore the model is free to enhance such relationships.
✓ Embedding distortion: Word embeddings with small corpora generate inflated semantic associations, which support stereotypes.

For instance, Zhao et al. (2017) posit that gender associations in word embeddings become more polarized through training, although the initial text may have moderate bias. This is further enhanced in the case of low-resource languages in which the model possesses fewer signals to anchor balanced linguistic representations.

➢ *Sampling Bias and Representation Gaps*

Language datasets of low resources can be of a limited or non-representative source, religious literature, government documents, educational resources, or online resources with a restricted audience. These datasets are structurally biased regarding:

- Register (formal > informal)
- Class (literate elites > rural speakers)
- Domain (written texts > oral traditions)
- Age (elderly people have more language practices than young)

Models that have been trained on these sources recreate their hierarchies. An example would be that a model trained on the missionary translations of an Indigenous language will be likely to do the following:

- Misunderstands the youth lingo or the new language.
- Consider non-standard spellings to be mistakes.
- Do not appreciate dialect continuums in the language community.
- Favor colonial orthographies to decolonised writing systems.

Sampling bias thus limits the types of language sampled and strengthens the hegemonic language norms.

➢ *Ideological Bias in Model Training*

In addition to the statistical and sample-related mechanisms, the bias is further increased by the ideological assumptions that are coded into training pipelines. Standard language ideology can be implicit in the processes of constructing datasets, making tokenization choices, annotation policies, and metrics of model evaluation. In the process of optimizing models to standardized forms, linguistic diversity is implicitly penalized, and dialects, creoles, and mixed codes are treated as noise.

- *This Leads to:*

✓ Greater error rates among dialect speakers.
✓ Misclassification of culturally based expressions.
✓ Systematic lack of protection by safety- or access-critical technologies.
✓ A constriction of what is considered to be legitimate language in the online realm.

This kind of ideological bias makes technological development compatible with the already established hierarchies of race, classification, and colonialism (Rosa & Flores, 2017).

➢ *Interaction Effects: Scarcity + Bias*

Low-resource conditions not only allow bias to persist; they create an environment where models *actively magnify* unequal patterns.
When:

- Data is scarce,
- The degree of linguistic variation is high.
- There is already a sociopolitical marginalization,

AI systems transform into amplifiers of linguistic discrimination of the high-risk type.

IV. CASE STUDIES IN BIAS AMPLIFICATION

The practical applications of AI systems demonstrate how the situation of low resources interacts with social inequities to generate detrimental results. The following section discusses three examples, hate speech detection, automated speech recognition, and machine translation, which show how bias amplification is disproportionately applied to marginalized linguistic communities.

➢ *Hate Speech Detection and Dialect Misclassification*

The detection systems of hate speech often incorrectly label dialectal speech as hate speech since the training sets do not reflect all possible variations in sociolinguistic diversity. Sap et al. (2019) demonstrated that tweets written in African American Vernacular English (AAVE) had a higher chance of being considered abusive or hateful by a margin of 1.5x than semantically equivalent tweets written in Standard American English.

This difference manifests itself in the form of underrepresentation of AAVE in training corpora, semantic relationships between dialect markers and negative sentiment, and annotation bias, with many crowd workers treating dialect markers as signs of aggression.

There are material implications of these misclassifications. There is an unequal content moderation, algorithmic silencing, and visibility of AAVE speakers on social platforms. When other mechanisms are used in the Global South, such as with Nigerian Pidgin, Sheng (Kenya), or South African township varieties, the danger of digital suppression is even greater, as these varieties already face sociopolitical stigmatization offline.

➢ *Automated Speech Recognition Failures*

There is a pronounced performance difference between the automated speech recognition (ASR) systems for the racial and linguistic groups. According to Koenecke et al. (2020), the error rates on words dictated by major ASR systems (including the systems of the largest technology companies) were significantly higher with the African American speakers than with the white ones, even when both groups were speaking Standard American English.

The factors that cause ASR failures in low-resource settings include the lack of acoustic data that represent multiple accents, training corpora that are biased towards the middle-class and urban, standardized speech, the lack of community-specific phonological characteristics, and inadequate modeling of multilingual speakers with frequent code-switching. ASR systems in multilingual African settings have issues with tonal languages, non-standardized orthographies, and the mixed influence of colonial and indigenous languages.

The outcome is that they become technologically excluded in the usage of voice-based authentication, virtual assistants, disabled user accessibility tools, and emergency communications.

➢ *Machine Translation Bias in African and Indigenous Languages*

Machine translators (MT) systems often increase the bias of low-resource languages because either the corpora is incomplete or limited to the domain. African and Indigenous languages have a variety of datasets of MT based on religious translations, governmental documents, missionary writings, or historical ethnographies.

These few sources bias models with old vocabulary, official languages, and colonialism. Admittedly, gender bias is the most widespread: the translation usually falls into masculine pronouns, professional roles, and terms of authority (Zhao et al., 2018).

For instance, Translations of either Yoruba or Swahili to English tend to give gender roles incorrectly. Local meaning is flattened, idioms, proverbs, and culturally based phrases are not handled by the MT systems, Code-switched or urban (e.g., Sheng, Camfranglais) are mistranslated or not recognized at all. Such distortions increase inequities by portraying cultures in false ways, misleading the global audiences, and decreasing the linguistic agency in intercultural communication.

V. A SOCIOLINGUISTIC AUDIT FRAMEWORK

To reduce the amplification of bias in the low-resource languages, AI systems have to be tested on the basis of the technical parameters only, but also within the frameworks that consider sociolinguistic diversity, power structures, and communal demands. This section outlines a sociolinguistic audit model that can guide researchers, practitioners, and institutions to create equitable NLP systems.

➢ *Community-Centered Assessment*

An effective audit will start with the communities whose language is modeled or falsely modeled by AI systems. Community-based assessment contains:

- Participatory Design Processes: for engaging speakers in co-design ensures that dataset boundaries reflect real linguistic practices, labeling decisions honor local meaning, and model goals align with community priorities, not only technological convenience.
- Language Ideology Assessment: This step analyzes the beliefs embedded in datasets and models, such as: privileging standard varieties over vernaculars, prioritizing written forms over oral traditions, assuming monolingualism in multilingual communities.
- Power Dynamics Analysis: The framework asks: Who gains after system deployment? Who bears risk or harm? Whom are they systematically excluded? This brings social effects into focus, instead of being seen as abstract measures of fairness.
- Cultural Evaluation Valuation: The cultural norms, politeness strategies, humor, metaphor, and indirect communication and styles should be adequately described. In the absence of this, AI systems introduce external standards and wipe out local communicative practices.

➢ *Dataset Documentation and Critique*

Based on the principles of Data Statements provided by Bender and Friedman (2018), the audit presupposes the sound dataset documentation along multiple axes:

- Data Provenance: Some questions are, who created the data? Under what conditions? What influence do power relations have on the form, content, and representation of the data? This prevents the sources from being dominated by colonial, missionary, or elite sources.
- Representation Analysis: It entails mapping, i.e., dialect variation, social class variation, registers (formal, informal, youth, rural, online), age distribution, gender diversity, and multilingual practices. Representation analysis determines systematic underrepresentation and overrepresentation.
- Historical Context: Each dataset is a socio-historical event. The audit asks the teams to take into account the legacies of colonial language policy, patterns of literacy and digital access, and social stigma on vernaculars, language shift, or endangerment processes. The contextualization of this type of model does not allow models to reproduce linguistic hierarchies under the guise of being “neutral” or “objective.”

- *Technical Complement: Integrating Fairness with Sociolinguistics*

Sociolinguistic audit is not a substitute of technical evaluation, it is complementary. There should also be models that combine the quantification of uncertainty on low-confidence predictions, loss functions that are mindful of fairness, dialect-aware tokenization and embeddings, multi-dialect data augmentation and continuous community review cycles. The framework combines sociolinguistic understanding with technical enhancement, which guarantees that justice is not introduced later but rather a design principle.

VI. CONCLUSION AND RECOMMENDATIONS

The sociotechnical solution to the problem of bias amplification in AI systems on low-resource languages involves a concerted effort. To start with, community-driven data practices should be put at the center of the developer. This involves establishing long-term relationships with speaker communities, gathering various and culturally pertinent data, and making sure that communities have control over the use of their language information. To avoid extractive or colonial data practices, ethical data governance with its focus on transparency, consent, and respect towards the cultural meaning is necessary.

Second, fairness and linguistic diversity should be explicitly included in technical design processes. The developers ought to record dataset composition, tokenize dialect-sensitive, apply fairness-based training purposes, and test systems on various language varieties and not just on aggregate measures. Some of the strategies, like domain adaptation, uncertainty estimation, and targeted augmentation, can be used to decrease the overfitting and representational gaps that intensify bias amplification in low-resource environments.

Third, progress requires institutional support and policy commitments in order to remain. Governments, universities, and technology firms need to invest in the digital infrastructure of marginalized languages and invest in community-driven language documentation. Digital policies on linguistic rights could be used to guarantee that the speakers of minoritized languages should not be locked out of such fundamental technologies as speech recognition, translation systems, and automated online services.

Finally, the amplification of bias is not only a technical but also a very social problem. The histories, inequalities, and language ideologies of the data used to train AI systems are reflected in them. When the low-resource languages are not represented or misrepresented, these systems reproduce and reproduce existing hierarchies, determining who can be part of digital life in a full manner. To establish language equity in AI, then, it is necessary to move away from thinking of these languages as a barrier and towards the centrality of these languages to ethical and inclusive technological design. Through community collaboration, language knowledge, and equity-based engineering, AI systems can start benefiting a greater number of language communities in a more equitable and precise manner in an ever-connected world.

REFERENCES

[1]. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics, 6*, 587–604. https://doi.org/10.1162/tacl_a_00041

[2]. Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

[3]. Doğruöz, A. S., Sitaram, S., Bullock, B. E., & Toribio, A. J. (2021). A survey of code-switching: Linguistic and social perspectives for language technologies. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *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)* (pp. 1654–1666). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.131

[4]. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.560

[5]. Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., & Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America, 117*(14), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[6]. Leonard, W. (2017). Producing language reclamation by decolonising 'language'. *Language Documentation and Description, 14*, 15–36. https://doi.org/10.25894/ldd146

[7]. Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the United States*. Routledge. https://doi.org/10.4324/9780203348802

[8]. Piller, I. (2016). *Linguistic diversity and social justice: An introduction to applied sociolinguistics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199937240.001.0001

[9]. Rosa, J., & Flores, N. (2017). Unsettling race and language: Toward a raciolinguistic perspective. *Language in Society*, *46*(5), 621–647. https://doi.org/10.1017/S0047404517000562

[10]. Ruíz, R. (1984). Orientations in language planning. *NABE: The Journal for the National Association for Bilingual Education*, *8*(2), 15–34.

[11]. Sap, M., Card, D., Gabriel, S., Choi, Y., & Smith, N. A. (2019). The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1668–1678). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1163

[12]. Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P., Durand, J.-L., Elliott, J., Ewert, F., Janssens, I. A., Li, T., Lin, E., Liu, Q., Martre, P., Müller, C., Peng, S., Asseng, S. (2017). Temperature increase reduces global yields of major crops in four independent estimates. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(35), 9326–9331. https://doi.org/10.1073/pnas.1701762114