Language Choice in Nigerian Social Media Hate Speech

Nneoma C. Udeze
Northwestern University
nneomaudeze2027@u.northwestern.edu

Rob Voigt
University of California, Davis
robvoigt@ucdavis.edu

Abstract

Language choice in multilingual societies is rarely arbitrary. In Nigerian, English, Nigerian Pidgin (NP) and indigenous languages are strategically deployed in online discourse, yet little is known about how they function in hostile contexts. Here we conduct the first systematic analysis of NP in online hate speech on two platforms, Twitter and Instagram. Using a linguistically enriched annotation scheme, we label each post for class, targeted group, language variety, and hate type. Our results show that NP is disproportionately used in offensive and hateful discourse, particularly against Hausa, women, and LGBTQ+ groups, and that insults are the dominant hate strategy. Cross-domain evaluation further reveals that classifiers trained on Twitter systematically overpredict hate on Instagram, highlighting challenges of domain transfer. These findings underscore NP's role as a linguistic resource for hostility and its sociolinguistic salience in amplifying stereotypes and affect. For NLP, the work demonstrates the need for NP-specific resources, sensitivity to figurative strategies, and domain adaptation across platforms. By bridging sociolinguistics and computational modeling, this study contributes new evidence on how language choice shapes online hate speech in a multilingual African context.

1 Introduction

The choice of language in multilingual contexts such as Nigeria is rarely arbitrary. It reflects social alignments, ideological positions, and strategic rhetorical intent. Among the languages spoken in Nigeria, Nigerian Pidgin (NP), an English-based pidgin/creole, is the most widely spoken language despite lacking an official political government status (Faraclas, 2021; Soneye, 2019). Although NP has long been studied for its roles in informality, solidarity, and identity expression, especially in digital and youth culture (Osoba, 2014; Agantiem and Alagbe, 2023; Usoro and Nsit, 2024; Nweke et al., 2024; Adegbija, 2004; Oluyinka Adebayo, 2023), its role in conflict-driven discourse remains underexplored. It is deeply embedded in everyday life and plays a role as a unifying force in Nigeria's multicultural society, functioning as both a street and home language, valued for humor and storytelling. Hate speech adds a layer of complexity to its sociolinguistic analysis. Existing research often focuses on hateful content or political context, but pays limited attention to the specific language varieties through which hate is articulated. In Nigeria's highly multilingual digital landscape, users appear to alternate between English, NP, and indigenous languages in ways that can be interpreted as framing hostility, intensifying emotion, or encoding in-group messages. Previous datasets (Muhammad et al., 2025) indicate that multilingual hate expression is common and moderation of large-scale and targeted hate speech remains limited due to scarcity of high-quality data in local local languages and the exclusion of local communities from data collection, annotation, and moderation efforts.

This study therefore examines the highly charged context of online hate speech to ask whether language choice matters, particularly the use of Nigerian Pidgin. This question is significant because language choice in a multilingual society like Nigeria is never neutral; it reflects social identities, power relations, and ideological alignments. Nigerian Pidgin, in particular, functions as both a unifying lingua franca and a marker of authenticity, solidarity, and informality. Its use in online hate speech therefore offers a lens into how speakers draw on linguistic resources to frame hostility, express stance, and index in-group belonging.

Although computational models have begun to incorporate NP in hate detection systems (Adegoke et al., 2024), these approaches may overlook the discursive and socio-ideological functions of NP in expressing toxicity. As a result, the role of NP

in framing hostility, reinforcing stereotypes, and drawing in-group boundaries remains poorly understood. To address this gap, we present these research questions:

1. Is Nigerian Pidgin used more frequently in offensive and hate content than in neutral content within the multilingual Nigerian social media context?

2. Are there specific social groups that are more frequently targeted with Nigerian Pidgin in online discourse?

3. What are the lexical and multi-word patterns associated with various hate types (insults, dehumanization, demonization, incitement to violence) and which groups are most frequently targeted with specific hate types?

To achieve this, we examine two datasets of online discourse: a re-annotated sample from the NaijaHate Twitter Corpus (Tonneau et al., 2024) and a novel dataset of Instagram comments scrapped from Instablog9ja (from July-Sept. 2024). We combine computational classification with manual linguistic annotation to identify languages used, isolate languages expressing toxicity, and categorize hate speech types. Our contributions are as follows:

1. A cross-platform analysis of Nigerian Pidgin in online hate speech: We provide the first systematic investigation of NP across Twitter and Instagram, examining how language choice varies by class (neutral, offensive, hateful) and targeted group.

2. A linguistically enriched annotation framework for hate speech: We develop and apply an annotation scheme that captures not only class labels and targeted groups, but also hate types (insults, dehumanization, demonization, and incitement to violence) and language choice, enabling both quantitative modeling and qualitative analysis.

3. Empirical and computational insights into multilingual hate detection: Through cross-domain error analysis and regression modeling, we show how NP disproportionately encodes hostility, and we highlight the implications for building culturally informed hate-speech detection systems.

Systematically analyzing the role of Nigerian Pidgin in constructing hostility across platforms is critical because hate speech classifiers trained on a single platform or on majority-language data (e.g. English) may fail when applied to other platforms or to local language varieties. This paper addresses this gap by providing a linguistically enriched, cross-platform analysis of NP in hate speech, combining computational modeling with fine-grained annotation of language choice, hate type, and targeted groups. By doing so, we contribute new empirical evidence to sociolinguistics and NLP, demonstrating how language choice functions as a strategic mechanism in online hostility.

2 Background of Study

Research on language choice in multilingual societies shows that speakers switch among codes to express identity, solidarity, authority, or informality (Ifukor, 2011; Oduma and Gomwalk, 1986; Igboanusi, 2008a). Language switching in digital forums and commercial discourse often serve rhetorical goals, reflecting the strategic use of the language that best resonates with target audiences (Ifukor, 2011; Dalamu, 2017; Doğruöz et al., 2021). English often indexes authority, while Nigerian Pidgin indexes informality and solidarity, particularly in informal interactions (Oluyinka Adebayo, 2023; Balogun, 2013; Taiwo, 2010). It is often preferred for its neutrality and broad intelligibility in diverse multilingual public settings such as markets, prisons, or interethnic gatherings. On social media, NP appears frequently for strong opinions, identity marking, and rhetorical effects (Taiwo, 2010; Oluyinka Adebayo, 2023).

Despite this rich body of research, little is known about how NP functions in hostile discourse. Most sociolinguistic studies focus on uses of NP in positive or neutral contexts (Akande and Salami, 2010; Osoba, 2014; Balogun, 2013), while computational research has only recently begun to include NP in hate speech detection (Adegoke et al., 2024). Hate speech introduces an added layer of complexity to the study of language choice. While much of the existing literature concentrates on the semantic content or ideological thrust of hateful messages (AYENI, 2018, 2024; Adepoju and Kalu, 2022; Ononye and Nwachukwu, 2019), less attention has been paid to the linguistic strategies (like language choice and hate types) through which such content is framed and delivered.

Hate speech detection and characterization has been investigated extensively in high-resource contexts (Zannettou et al., 2020; ElSherief et al., 2018; Warner and Hirschberg, 2012) and, increasingly, in multilingual and low-resource contexts with specialized datasets and models (Vargas et al., 2022; Geleta et al., 2023; Muhammad et al., 2025; Ayele et al., 2023). For this study, we adopt the definition of hate speech proposed by the United Nations and used by Tonneau et al. (2024), as "any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor" (United Nations, 2019). Hate speech as noted by Waseem et al. (2017) can manifest in both explicit and implicit forms (Talat et al., 2017). Explicit hate speech typically includes clearly offensive language, slurs, or threats, making it more easily recognizable through automated detection methods (Talat et al., 2017). Implicit hate speech, on the other hand, is often expressed through sarcasm, irony, or coded language, and requires greater contextual awareness to identify, posing significant challenges for both human annotators and machine learning models. Papcunová et al. (2023) further contribute to the operationalizations of hate speech by proposing a structured set of indicators (Papcunová et al., 2023). These include expressions that promote violence, deny human rights, invoke negative stereotypes, employ ad hominem attacks, or manipulate historical facts. Such indicators offer a measurable framework for annotation and detection.

Within Nigerian-focused work, several datasets and computational systems have emerged (Ilevbare et al., 2024; Adegoke et al., 2024; Aliyu et al., 2022; Tonneau et al., 2024; Asogwa et al., 2022; Nkemdilim and Somtochukwu, 2024), but most prioritize model evaluation and performance rather than linguistic profiling of hate expressions. Recent studies show that abusive content is often more prevalent in Yoruba, Nigerian Pidgin, and code-switched messages than in Standard English in political contexts (Ilevbare et al., 2024). However, these studies may often neglect the nuanced interplay between linguistic form, hate type, and language variety. Therefore, our work bridges sociolinguistics and NLP by directly testing the prevalence of NP in hate contexts, profiling hate expressions at the lexical and phrasal level, and mapping their group-specific targets, offering the contributions mentioned above.

3 Data Collection and Annotation

3.1 Datasets

This study is based on two social media datasets: Twitter and Instagram. The primary dataset is the NaijaHate Twitter Corpus (Tonneau et al., 2024), a curated collection of approximately 36,000 tweets collected between July 2021 and July 2023, annotated for tweet class (neutral, offensive, hateful) and target groups. From the NaijaHate Twitter corpus, we drew a stratified sample of 6,000 tweets (2,000 per class) to support a balanced comparative analysis. To compare cross-domain behavior, we scraped Instablog9ja, a high-traffic Nigerian Instagram account, using Instaloader (Instaloader Developers, 2024) and collected 35,000 public comments between July and September 2024. The NaijaXLM-T classifier, an existing XLM-R pretrained model finetuned on the NaijaHate Twitter corpus, initially classified the scraped Instagram comments.[1] From the classification set, we selected a balanced sample of 1,500 comments (500 per class) for manual re-annotation and analysis.[2] We used the NaijaHate-trained classifier to propose language labels and an initial class label for Instagram. All further analysis reported here rely on manual re-annotation to ensure consistency of labels across platforms.

3.2 Annotation scheme

To capture linguistic and ideological nuances, we used an expanded annotation schema based on Bahador (2020) (Bahador, 2020) (figure 1). Rather than a binary distinction, we annotate class (neutral, offensive, hateful), language variety (English, Nigerian Pidgin, Yoruba, Igbo, Hausa), target groups (Women, Igbo, Hausa, Yoruba, Christians, Muslims, LGBTQ+, Northerners, Southerners, Fulani, Herdsmen, Other), and hate type. Hate types include:

- **Insults:** Negative group characterizations (e.g. 'stupid', 'lazy').
- **Dehumanization:** Equating a group to subhuman entities (e.g. 'rats', 'pigs').

[1]The NaijaHate dataset and the NaijaXLM-T classifier can be accessed here https://github.com/worldbank/NaijaHate

[2]The annotated Instagram dataset can be accessed here https://github.com/Nneoma-Udeze/Language-Choice-In-Nigerian-Social-Media-Hate-Speech

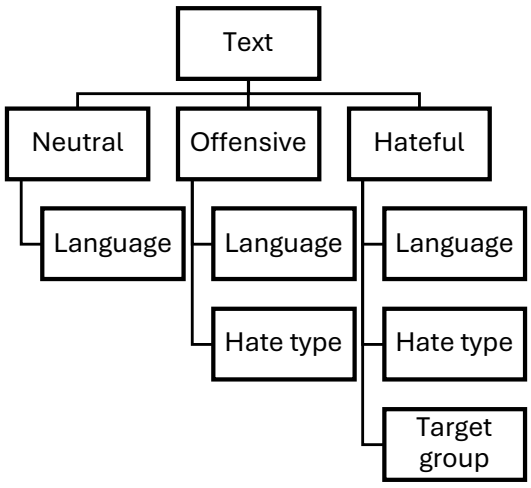

Figure 1: Hierarchical Annotation Framework for Classes, Languages, hate Types, and Target Groups

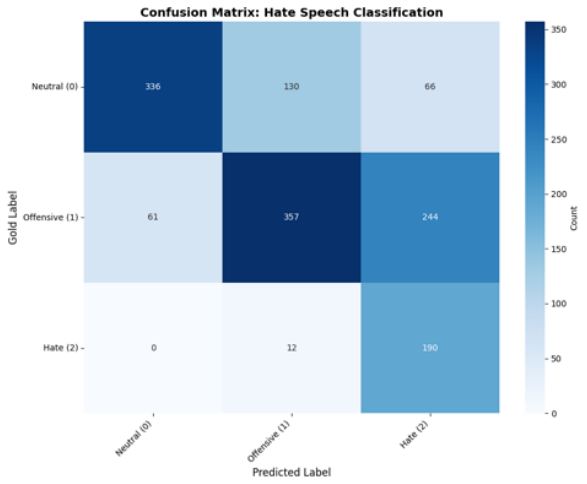

Figure 2: Confusion matrix evaluating NaijaHate model performance against human annotations

- **Demonization:** Portraying group as an existential threat (e.g. 'monsters', 'diseases').
- **Incitement to violence:** Calls for harm against a group.

Both target group and hate type annotations are multi-label; a single comment can target many groups and use multiple hate strategies.

3.3 Annotators

The annotations were conducted by the lead author and a graduate student in linguistics. The NaijaHate sample was re-annotated for speech class, targeted group, language(s) used, and hate type. A 20% overlapping sample (1,200 tweets) was used to compute inter-annotator agreement for the speech class alone. The agreement was high (Cohen's Kappa = 0.88), and most disagreements occurred between offensive and hateful classes, reflecting their inherent ambiguity.

4 NaijaHate Cross-Domain Performance

We evaluated the NaijaHate Twitter-trained classifier on the manually re-annotated Instagram dataset to test domain adaptability. This revealed an imbalanced class distribution in the Instagram sample, underscoring the discrepancy between model predictions and human annotations. The model achieved an overall accuracy of 63.3% when evaluated against human annotations (N = 1,396). Figure 2 shows a confusion matrix that compares model predictions and human labels.

We observe a clear bias toward hate classification, and the model exhibits exceptionally high recall for the hate class (94.1%) at the cost of very low precision (38.0%). This bias is problematic for automated moderation systems as it risks inflating rates of false positives, potentially leading to unjustified censorship of neutral or offensive content. In contrast, the model shows more conservative behavior for neutral content, achieving high precision (84.6%) but moderate recall (63.2%), indicating that it correctly identifies neutral content when predicted, but misses a substantial portion of neutral instances.

We initially hypothesized that the poor performance of the NaijaHate model on Instagram data could be attributed to the prevalence of Nigerian Pidgin, a language not seen in large proportions on the Twitter-based training corpus but representing a substantially larger proportion of the Instagram discourse.

However, contrary to this expectation, the classifications in the Nigerian Pidgin comments demonstrated a better classification performance across all metrics. Comments containing only NP achieved the highest hate class F1 score (0.67), substantially outperforming English-only content (F1 = 0.57) despite English being the primary training language. The English-NP code-mix yielded an F1 score of 0.61.

Our findings point to fundamental differences between platforms in the manifestation of hate speech rather than language-specific classification challenges. The poor performance of the model on English content suggests that Twitter English and Instagram English could represent different communicative domains with different rhetorical

strategies for expressing toxicity. The superior performance on the NP content implies that Nigerian Pidgin employs more direct rhetorical strategies that align better with the model's learned patterns, while the English discourse on Instagram may utilize more sophisticated linguistic strategies that diverge from the Twitter training patterns. This reveals the need for domain-sensitive calibration or strategies to prevent harmful overgeneralization.

5 Analysis and Discussion of Language Choice in Hate Speech

Analysis of language use within Instagram and Twitter datasets revealed distinct distributions across the three classes (neutral, offensive, and hateful). English and Nigerian Pidgin dominate both neutral and harmful content; indigenous languages (Yoruba, Igbo and Hausa) occur less frequently. Figure 3 & 4 shows the distribution of language by comment and tweet class in the Instagram and twitter datasets.

Among indigenous languages, Yoruba appears the most frequently, and many affective terms in NP are borrowed from Yoruba. When such loanwords occur embedded within a Pidgin syntactic frame, they are labeled as Nigerian Pidgin; when they appear in isolation, they are annotated as Yoruba, illustrating fluid boundaries and annotation challenges. This overlap introduces annotation ambiguity that has direct implications for downstream modeling. In supervised classification, such hybrid usage effectively introduces label noise, as lexical items associated with Yoruba may appear in instances labeled as Nigerian Pidgin and vice versa. As a result, models relying on surface-level lexical features or language identification assumptions may conflate language membership with affective intensity, potentially inflating or obscuring language-specific effects. These challenges suggest that computational models applied to Nigerian social media should move beyond strictly discrete language categories and instead accommodate mixed-language representations, for example, through subword modeling, contextual embeddings, or multi-label language tagging. Explicitly acknowledging and modeling language overlap is therefore crucial for both predictive performance and the interpretability of sociolinguistic findings.

Given the data, we structure our analysis around three central research questions that aim to address key gaps in existing studies on hate speech in Nigeria. The findings of these analyses offer a nuanced understanding of how NP functions within online hate speech and contributes new insights into the sociolinguistic dynamics of digital discourse in Nigeria.

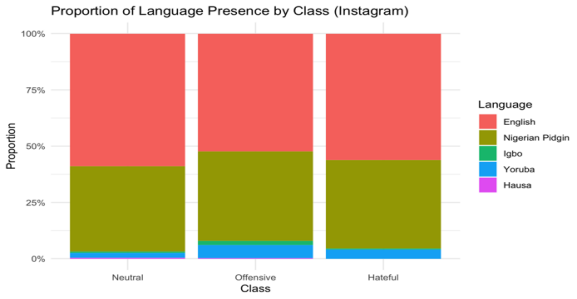

Figure 3: Language Distribution by Comment Class in Instagram Data

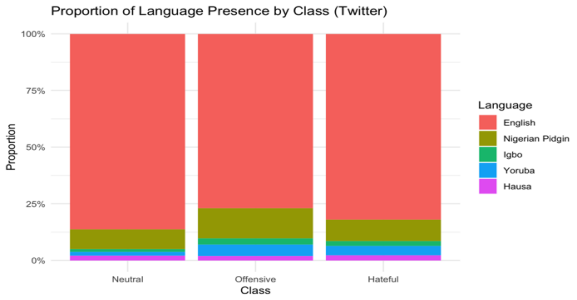

Figure 4: Language Distribution by Comment Class in Twitter Data

5.1 RQ1: Is Nigerian Pidgin more frequent in offensive and hateful comments than in neutral content?

In figures 5 & 6, we notice that Nigerian Pidgin (both used alone and mixed) is used more in the Instagram data than in the Twitter data. Two regression models were fit on 1) the full dataset to determine the relationship of NP with toxic comments and tweets, and 2) a subset of the full dataset to text if an effect was mostly driven by single language use only (NP) or code-mixed language (NP mixed with other languages)

To answer RQ1, a regression model was fitted using a generalized linear model (GLM) framework in R to predict the probability that a comment is written in Nigerian Pidgin as a function of class, controlling for platform and code-mixing status. Two models were estimated: 1) on the full dataset to assess language-class relationships; 2) on a subset examining single-language NP vs NP in code-mixed contexts.

Statistical testing revealed a significant effect

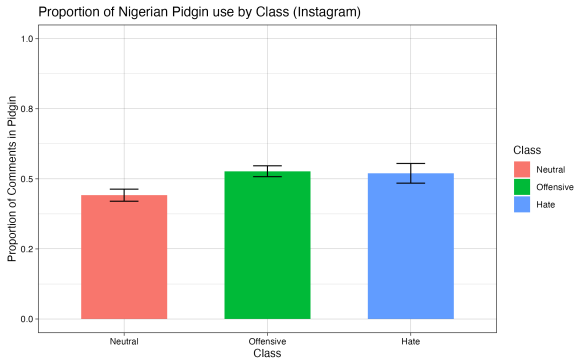

Figure 5: Distribution of Nigerian Pidgin Use across Neutral, Offensive, and Hateful Comments in Instagram Data

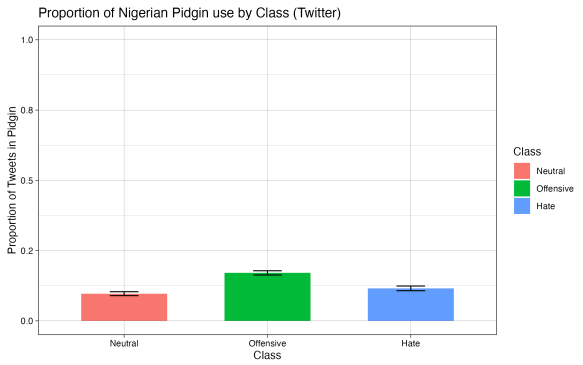

Figure 6: Distribution of Nigerian Pidgin Use across Neutral, Offensive, and Hateful Comments in Twitter Data

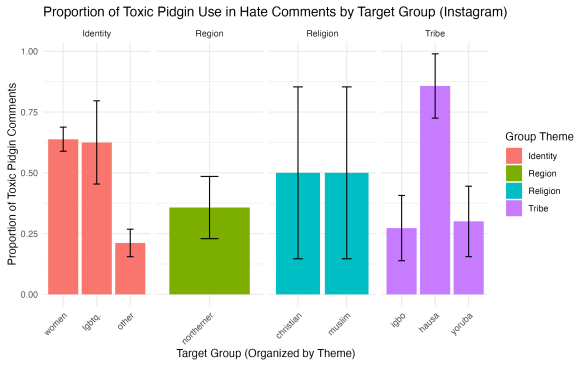

Figure 7: Distribution of Nigerian Pidgin in Hate Speech by Target Group in Instagram Data

for the offensive class compared to neutral ($p < 0.005$) in both Instagram and Twitter, indicating offensive content is more likely to be written in NP than neutral content. For the hateful class, effects were marginal on Instagram (b = -0.234, p = 0.059) and non-significant on Twitter (b = -2.234, p = 0.083), suggesting trends but limited power given fewer hateful instances and potential moderation effects prior to data collection. These patterns should be interpreted in light of pronounced class imbalance, as hateful content constituted a substantially smaller proportion of the dataset relative to offensive and neutral comments. This imbalance reduces statistical power by inflating standard errors for class-specific estimates, making it more difficult to detect statistically significant effects even when underlying trends are present. As a result, the coefficients associated with the hateful class are inherently less stable and more sensitive to sampling variability. For content expressed solely in NP versus code-mixed with English or other languages, neither offensive nor hateful classes showed significant effects driving the association of NP in these classes for both Instagram and Twitter ($p > 0.05$). This suggests that the observed association between NP and harmful speech may not be driven by one subset of NP usage alone. Instead, the effect appears to emerge from the combined influence of both monolingual NP comments and code-mixed comments.

From a sociolinguistic perspective, this is consistent with previous work showing that NP functions as an informal and affective code used to express humor, critique, and solidarity (Oluyinka Adebayo, 2023). However, in hostile contexts, these same pragmatic features make NP a natural vehicle for insults and ridicule. The fact that NP is more predictive of offensive than neutral discourse highlights its strategic use in emotionally charged interactions. The trend observed in the hateful class warrants further investigation with a more balanced dataset, as it may reveal whether NP also systematically functions as a language of hate speech, beyond offensive expression.

5.2 RQ2: Which groups are disproportionately targeted with Nigerian Pidgin?

Figure 7 & 8 shows the proportion of hateful comments and tweets (which is the percentage of comments and tweets in NP compared to other languages seen in the dataset) expressed in Nigerian Pidgin across targeted groups, organized by thematic categories (identity, region, religion, and tribe).

NP is used disproportionately when targeting certain groups, especially Hausa, women, and LGBTQ+ people on Instagram. By contrast, non-Nigerians ('other') received the least hate expressed in NP. This suggests that NP is strategically

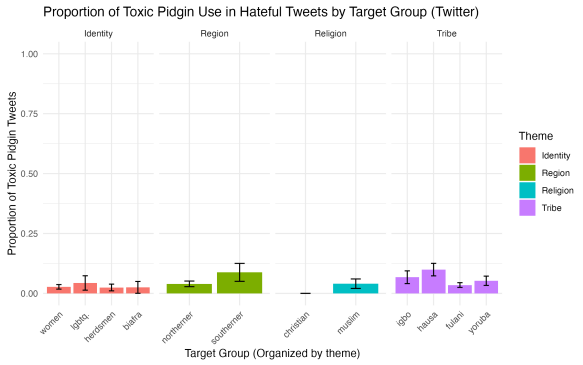

Figure 8: Distribution of Nigerian Pidgin in Hate Speech by Target Group in Twitter Data

mobilized in online hostility where the speaker seeks resonance within an in-group audience. On Twitter, the overall proportions are smaller for all groups. Across the two platforms, the Hausa and LGBTQ+ groups are the dominant NP targets. Thus, the selective use of NP demonstrates that language choice in hate speech is sensitive to the audience.

The Hausa case is particularly notable. As the group most frequently targeted on both social media platforms in NP, their prominence suggests that the language is strategically mobilized in relation to the Hausa group, potentially reinforcing their marginalization. Similarly, NP's recurrent use in hate directed at women and LGBTQ+ groups highlights its affective force in amplifying ridicule and hostility within these group audiences. Nigerian Pidgin was least frequently employed in hostile comments directed at non-Nigerians. Instead, English predominated in these cases. This suggests that NP is primarily reserved for intra-national targeting, where it achieves stronger affective resonance within the local speech community. This supports Osoba's (2015: 138) claim that NP can be regarded as a highly sensitive informal language where "nastier, sharper, more basal, and more naturally unobtrusive conceptions and inclinations towards a brutally lower level of emotion laden acquaintance can be observed to feature more prominently" than any other language captured in the dataset. Thus, the selective use of NP demonstrates that language choice in hate speech is sensitive to the audience. Speakers appear to choose NP when addressing co-nationals, amplifying hostility through a language that maximizes shared cultural meaning, while turning to English in addressing outsiders. These findings show that hate detection cannot be based solely on generic NP lexicons. Models must account for target-specific usage. Bias-aware training and group-sensitive annotation schemes are needed to mitigate misclassification and under-detection of harm.

5.3 RQ3: What rhetorical strategies appear in NP hate speech?

Insults are the dominant hate strategy across platforms and languages. The frequency analysis of the hateful class revealed that Nigerian Pidgin (NP) terms such as "mumu" (fool) and "ashawo" (prostitute) were among the 20 most frequently used words. These elements, documented in online resources such as Naija Lingo, exemplify how NP functions as a lexical resource for emotionally charged expressions. Their prominence underscores the role of NP as a key medium through which hostility and contempt are linguistically encoded in Nigerian digital spaces. In contrast, a frequency analysis of the hateful class in the Twitter data showed that English terms such as "stupid" and "useless" were among the 35 most frequent words. This pattern is not surprising, given that only 8% of the Twitter dataset is written in Nigerian Pidgin. Notably, across both datasets, women emerged as the group most frequently targeted with insults.

The dependence on NP for insults must be understood within Nigeria's multilingual ecology. As an English-based creole shaped by indigenous languages, NP combines broad intelligibility with an affective, informal register (Obi, 2014; Osoba et al., 2015). Previous studies show that speakers often choose NP to express humor, solidarity, or critique (Affia, 2025; Oluyinka Adebayo, 2023). However, in hostile contexts, this same versatility allows NP to be strategically mobilized to intensify insult and ridicule, particularly against marginalized groups. Negative attitudes toward NP, considering it an inferior language and associated with low education, vulgarity, and lack of prestige (Osoba et al., 2018; Akande and Salami, 2010; Igboanusi, 2008b; Mann, 1996; Nwoda, 2023) further reinforce its role as a linguistic resource to position targets as socially inferior.

Beyond insults, dehumanizing language appeared mostly against the Muslim group on Instagram and women on Twitter. Mendelsohn & Budak (2025) highlight how metaphors such as natural disasters (e.g. *'floods of immigrants'* or *'infestations'*, operate at the level of discourse to re-

inforce exclusionary ideologies and justify hostile policies toward migrants. In the Nigerian context, while the specific metaphors differ, the logic of dehumanization remains consistent. These framings are not merely rhetorical; they have material consequences, as dehumanizing discourse lowers thresholds for violence and legitimizes social or physical harm (Haslam, 2006; Haslam and Stratemeyer, 2016). By showing that Nigerian online discourse deploys dehumanization in ways like how western immigration discourse deploys metaphors, we provide further evidence that metaphor is a mechanism through which hostility is linguistically constructed. It also emphasizes the urgent need for hate speech detection models to account for metaphorical and figurative language, which may encode hostility more subtly than direct insults but can be equally potent in legitimizing violence.

Demonizing language was used against several groups, with women being the most frequent targets on Instagram, reflecting broader discourses that portray women as the source of societal decline. However, it was used the most to target the Fulani group on Twitter. Notably, incitement to violence was directed solely against women on Instagram, where examples explicitly encourage harm, attack, and even death. These patterns indicate that women occupy a uniquely vulnerable position in Nigerian online hate discourse, targeted not only with contempt but also with explicit calls for violence. On Twitter, it was used across targets but in very low proportions.

6 Discussion

Taken together, our results show NP plays a systematic rhetorical role in hostile online discourse. NP functions as a resource for expressing affective meaning in ways English often does not. Its strategic deployment reflects broader sociolinguistic patterns of language choice in multilingual societies, where codes carry ideological and affective meanings (Igboanusi, 2008a; Ifukor, 2011). NP, often dismissed as "low status," is mobilized to reinforce social hierarchies and intensify hostility against vulnerable groups. This is evident in the dataset, where NP is disproportionately used in hateful expressions targeting specific groups. For NLP, these findings underscore the necessity of incorporating sociolinguistic insight into computational models. This is because hate speech expression in Nigerian social media is deeply shaped by local linguistic practices and sociocultural meanings that current NLP models fail to capture. The cross-domain overprediction by the NaijaHate classifier reveals practical consequences: models trained on platform-specific data (Twitter) can misclassify content in other environments (Instagram), especially where language use and conversational norms differ. This has implications for moderation systems and for dataset curation: domain adaptation and platform-aware annotations are necessary. Computationally, the use of NP (and NP's overlap with Yoruba loanwords) suggests model architectures need to be sensitive to code-mixed tokens, loanword handling, and figurative devices. Multilingual embeddings, metaphor detection modules, and bias-aware metrics can help, as can human-in-the-loop annotation practices that incorporates local linguistic expertise.

7 Conclusion

This paper provided the first systematic corpus analysis of Nigerian pidgin in online hate speech across Twitter and Instagram. By combining manual annotation with error analysis of a classifier trained on the NaijaHate Twitter corpus, we show that NP is disproportionately used in offensive contexts and is an important medium for insults and other figurative hate strategies, particularly targeting Hausa, women and LGBTQ+ groups. Our findings suggest that NP is not merely an informal code, but a linguistic resource strategically deployed to reinforce stereotypes and intensify hostility. The sociolinguistic associations of NP with low prestige make it especially potent in positioning groups as inferior, while its broad intelligibility ensures affective resonance in multiethnic online settings. Cross-domain evaluation also reveals that Twitter-trained classifiers over-predict hate on Instagram, underscoring the need for domain adaptation.

For NLP, these results underscore three key implications. First, hate detection in Nigerian and other multilingual contexts requires NP-specific resources, including lexicons, embeddings, and annotated corpora. Second, models must be sensitive to figurative and metaphorical strategies, which encode hostility more subtly than direct slurs. Third, domain adaptation across platforms is essential, as language use varies significantly between Twitter and Instagram. By foregrounding both sociolinguistic and computational perspectives, this work bridges a critical gap and provides a foundation

for building more robust, culturally informed hate detection systems.

8 Future Work

Future research should expand the size and balance of annotated datasets, particularly for the hateful class, to improve statistical power and classifier robustness. They should also expand demographic annotation to contextualize language choice. Extending analysis beyond Instagram and Twitter to platforms such as TikTok or WhatsApp would test the portability of models across even more diverse digital environments. Incorporating multilingual embeddings, metaphor detection modules, and bias-aware evaluation metrics will be crucial for capturing the subtle, context-dependent ways in which hostility is encoded. Finally, collaboration with linguists, sociologists, and local communities can help ensure that computational approaches remain sensitive to the cultural and political dynamics of language choice in Nigeria and other multilingual societies.

Limitations

This work has limitations. Datasets lack user demographics, limiting sociolinguistic contextualization. It remains unclear whether NP's emotional impact is carried by specific words or by full-language usage in code-mixed vs single-language contexts; experimental work is needed because the findings here are largely descriptive. Resource constraints meant annotators were not expert specialists, which may affect label quality. Distinguishing offensive vs hateful categories remains challenging.

References

Efurosibina E Adegbija. 2004. *Multilingualism: A Nigerian case study*. Africa World Press.

Folake Oluwatoyin Adegoke, Bashir Tenuche, and Eneh Agozie. 2024. Development of pidgin english hate speech classification system for social media. *American Journal of Information Science and Technology*, 8(2):34–44.

B Adepoju and T Kalu. 2022. An ideological analysis of hate speech comments on selected online new reports.

Precious Isaac Affia. 2025. Nigerian pidgin: The identity of a nigerian away from home.

AA Agantiem and Joseph O Alagbe. 2023. A comparative study of nigerian english pidgin usage among students in federal university of lafia. *Ansu Journal of Language and Literary Studies*, 2(3).

Akinmade T Akande and L Oladipo Salami. 2010. Use and attitudes towards nigerian pidgin english among nigerian university students. In *Marginal Dialects: Scotland, Ireland and Beyond. Aberdeen: Forum for Research on the Languages of Scotland and Ireland*, pages 1–79.

Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. Herdphobia: A dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification using svm and naive bayes. *arXiv preprint arXiv:2204.07057*.

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023. Multilingual racial hate speech detection using transfer learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 41–48.

ABIODUN COMFORT AYENI. 2018. Socio-pragmatic analysis of online hate speeches in nigeria.

ABIODUN COMFORT AYENI. 2024. *Pragmatic Analysis of Hate Rhetoric on Selected Social Media Platforms in Nigeria*. Ph.D. thesis, Doctoral dissertation, Federal University Lokoja.

Babak Bahador. 2020. Classifying and identifying the intensity of hate speech. *Social Science Research Council. https://items. ssrc. org/disinformation-democracy-and-conflictprevention/classifying-and-identifying-the-intensity-of-hate-speech*.

Temitope Abiodun Balogun. 2013. In defense of nigerian pidgin. *Journal of languages and culture*, 4(5):90–98.

Taofeek Olaiwola Dalamu. 2017. *A discourse analysis of language choice in MTN® and Etisalat® advertisements in Nigeria*. University of Lagos (Nigeria).

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. 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 1654–1666.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Nicholas Faraclas. 2021. Naija: A language of the future. *Current trends in Nigerian Pidgin English: A sociolinguistic perspective*, pages 9–38.

Raisa Romanov Geleta, Klaus Eckelt, Emilia Parada-Cabaleiro, and Markus Schedl. 2023. Exploring intensities of hate speech on social media: A case study on explaining multilingual models with xai. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 532–537.

Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.

Nick Haslam and Michelle Stratemeyer. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29.

Presley Ifukor. 2011. Linguistic marketing in "... a marketplace of ideas": language choice and intertextuality in a nigerian virtual community. *Pragmatics and Society*, 2(1):110–147.

Herbert Igboanusi. 2008a. Changing trends in language choice in nigeria. *Sociolinguistic Studies*, 2(2):251–269.

Herbert Igboanusi. 2008b. Empowering nigerian pidgin: A challenge for status planning? *World Englishes*, 27(1):68–82.

Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola, and Oluwaseyi Adeyemo. 2024. Ekohate: Abusive language and hate speech detection for code-switched political discussions on nigerian twitter. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 28–37.

Instaloader Developers. 2024. Instaloader: Download instagram photos and metadata. urlhttps://instaloader.github.io/.

Charles C Mann. 1996. 10 anglo-nigerian pidgin in nigerian education: A survey of policy. *Language, Education, and Society in a Changing World*, page 93.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Sa'id Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, and 1 others. 2025. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871.

Mbeledogu Njideka Nkemdilim and Ike-Okonkwo Mishael Somtochukwu. 2024. Navigating the dark web of hate: Supervised machine learning paradigm and nlp for detecting online hate speeches. *International Journal of Advanced Engineering Research and Science*, 11:3.

Ogochukwu C Nweke, Koryoe Anim-Wright, and Bukola Towuru. 2024. The influence of pidgin english in nigerian media and its societal impact: A systematic literature review. *Journal of Humanities, Arts and Social Science*, 8(12).

Chinazor Nwoda. 2023. Attitudes regarding the use of nigerian pidgin english among nigerian students at coventry university. *SCHOOL OF LANGUAGES, CULTURES AND LINGUISTICS SOAS UNIVERSITY OF LONDON*, page 46.

Edith Ifeyinwa Obi. 2014. Language attitude and nigerian pidgin. *AFRREV IJAH: An International Journal of Arts and Humanities*, 3(4):34–46.

A Oduma and V Gomwalk. 1986. Towards a typology of variation in nigerian english: A critique of some existing frameworks of analysis. In *17th NESA Conference. University of Lagos*.

Mosunmola Oluyinka Adebayo. 2023. Exploring the meaning of pidgin english on social media: A sociolinguistic analysis of nigerian pidgin hashtags as adapted speech. , 4(2):68–87.

Chuka Fred Ononye and Nkechinyere Juliana Nwachukwu. 2019. Metalinguistic evaluators and pragmatic strategies in selected hate-inducing speeches in nigeria. *Indonesian Journal of Applied Linguistics*, 9(1):48–57.

Joseph Babasola Osoba. 2014. The use of nigerian pidgin in media adverts. *International Journal of English Linguistics*, 4(2):26.

Joseph Babasola Osoba and 1 others. 2015. Analysis of discourse in nigerian pidgin. *Journal of universal language*, 16(1):131–159.

Joseph Babasola Osoba and 1 others. 2018. Power in nigerian pidgin (np) discourse. *Journal of Universal Language*, 19(1):1–32.

Jana Papcunová, Marcel Martončik, Denisa Fedáková, Michal Kentoš, Miroslava Bozogáňová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems*, 9(3):2827–2842.

Taiwo Soneye. 2019. David jowitt: Nigerian english.

Rotimi Taiwo. 2010. The dynamics of language mixing in nigerian digital communication. In *Handbook of research on discourse behavior and digital communication: Language structures and social interaction*, pages 180–190. IGI Global Scientific Publishing.

Zeerak Talat, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online*, pages 78–84.

Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040.

United Nations. 2019. United Nations Strategy and Plan of Action on Hate Speech. https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf. Accessed on [Date you accessed the document].

Donatus George Usoro and Afaha Nsit. 2024. The status of nigerian pidgin in nigeria's linguistic landscape: Another examination. *Erudite Compendiums in Education*, pages 89–100.

Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.

Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM conference on web science*, pages 125–134.

Appendix: Annotated Hate Speech Examples

Platform	Targeted group	Insults	Dehumanization	Demonization	Incitement to violence
Twitter	Women	women are actually the greatest hypocrites of all time, all these trolls coz it was not the other way round? A creature that always wants to be on the receiving end... Enjoy your family...	Women are Scum	na only God go save us from all this girls hand, they are all blood suckers (bed bug)	
	Lgbtq+	play good football and stay off of that madness,homosexuals are insane people animals don't gay not to talk of humans	Gay baboon doesn't know about getting consent?	Gay Terrorist	
	Herdsmen	You all really thought you had the monopoly of crime. The only reason I think herdsmen and some other northerners are doing all this crap is because they are uneducated and brainwashed, but "uneducated" comes first for a reason.	We demand aâ€¦ @USER Yet one foolish talkative will come here 2 bark like a curse dog dat they don't suppose 2 chase or killed Fulani cows, but when killer herdsmen were massacring, maiming and raping our people she lost her voice.	Killer herdsmen are heartless terrorists, they killed and they are killing people in North. The same herder-terrorists (might have also) disguised as UGM to kill the police. LINK	
	Biafra	The moment you ask questions, they resort to insults, abuse and name-calling. @USER Swears nothing theyâ€™d will make f00lish tribe who thinks everyone is against them like theyâ€™re the best thing in the world win. Biafraud biafools now theyâ€™re Nigerians suddenly confused souls These Biafrans just want to waste a part of Nigeria	Crowd funding a terrorist organization, u are a magnetic fool Igbos are smart nd industrious people , but IPOB are the pigs and idiotic morons among them	Accept it ! You ipobians , are terrorist . How many people did members of IPOB kill before President Buhari ordered an armed invasion of the South East?	
	Northerner	Northern Nigeria and their lies. They formed BH and they are Reaping the result. My heart bleeds for the innocent ppl loosing their lives.	I stand with southern governors. No to open grazing! Northern extremists have continuously destroyed investments in alcoholic beverage belonging to southerners with impunity but want the South to consult them before Open Grazing can be banned in the South. Away with parasites! LINK	Northerners are the biggest, greediest, erratically selfish and nepotistic thieves & corrupt people in Nigeria...	They are terrorists supported by northern leadersâ€¦. The truth is the north is ok with themâ€¦.. let them burn! But should not cross down to the south. They should kill their own people.
	Southerner	Should worry his southern Christian morons.	Some people are dogs They are blaming	You are very stupid, an undesirable idiot, Fulani	

Appendix: Annotated Hate Speech Examples

			northerners for the deed of SA people Northerners are innocent but southerners are looking for trouble look at how heartless southerners people are blaming us	are not the course of any conflict the southerners are because of their greedy and selfishness. You want dominate the country so you start hating the Fulani after that the northern, we know all I'm with @USER Nigeria will never change, cause of these Narcissistic Southerners especially Igbo people they hate us, they are happy with how this #COVID__19 spreading in kano.	
	Christian	If there are two christians and they do not understand each other,then one of them must be a bastard in christ jesus.		I'm an atheist born of Christians but at the same time I'm in love with a Muslim. I actually see Christians as the biggest problem but muslim fanatics the biggest mistake the world has	
	Muslim	Some of you Muslims are THE most judgemental people ever. You say the most disgusting things to your fellow brothers and sister	If this is what Islam teaches, then, Mohammad and Allah are scum and shit	Islamic terrorists would have overrun Igbo land if not the presence of ESN and the so called IPOB. Those animals are afraid of these two - IPOB and ESN. Don't ever attribute evil deed to these two, but the government mercenaries	useless fool say islamic terrorism, the world has to kill islam b4 islam kills the world. we know you added ojukwu to ur name to look an Igbo but u r just a pig.
	Igbo	Igbo people are lazy,jobless, deceitful? Nah sorry we dont fit into that category…	Igbos are cannibals	idiot, don't change the rhetoric.These are simply Igbo kidnappers.Most of your people hide under the guise of Fulani herdsmen to kidnap people for a living, and fools like you would crucify and persecute the Fulani herders for your evils	STOP SPREADING HATE MSGS! By the way I am IGBO! @USER Igbos re d scum of d Earth. Jeopardizing d future of Nigerians..kill them all"
	Hausa	A Hausa man is an ungrateful human. Hypocrites, fanatics,pretenders and evil doers	Na Hausa you be I can't blame you Come to southern you will see what's happening	I agree with you on this, we need to find a way of repatriating these abokis in the south back to their	Shoot me 4 saying dis bt I tink Hausas especially d muslim ones shld b exterminated "Le

Appendix: Annotated Hate Speech Examples

			Mumu man You go be the next person that sars go kill even your family #animal #EndSarsNow	north. The ploy to take over Nigeria is so real.	strugle to trend. Bitch have a seat on thorns
	Fulani	Idiots! All of you are idiots! All this talk is to kill more Igbo people on the east without investigating anything. Buhari has never sounded this way about Terrorists ravaging Nigeria, murderous fulani herdsmen God will wash you and	Leave the deaf Fulani baboon	Are you happy that fulani herdsmen are killing, raping women and destroying people's farms and property? Do fulani have any land in the south west or south east? Can an ibo/yoruba man come to the north with pigs or goats to destroy the farmland	Sometimes, Fayose is my guy. Other times I just cringe at his actions.But on this murderous Fulani bastards, I support him! KILL THEM ALL!
	Yoruba	God!!! Yoruba's are so daft! I never see	They are mostly yoruba pigs	Yorubas are the problem of this country, you can never see hausa man to treat his brother like this coz of a peaceful protest... Animal in uniform, you guys will soon get sense by force Yorubas are mugus nah, we've sold our right to the Hausa people so they are the one controlling us. Why can't they just share the materials in Lagos and distribute to every other state? Wicked souls ðŸ˜□ðŸ˜□ðŸ˜□ LINK RT @USER: Yorobber criminals & thieves ve been defeated. Many of them are having heart attack right now as I am writing.	Yoruba dirty goat like u should slaughtered by boko haram. Idiat fool
Instagram	Women	women na real wa for una anything for money some of you would do totally ashamed	everything about this gender na cheating cheating as if we have no life aside cheating men doesnt own women any form or loyalty i brought you into my life	women are more deadly when it comes to power this your analysis is not a reason	girls too many for this world sef let them continue make feminist reduce
	Lgbtq+	thats why you decided to be a gay you are very use less		na gay people cause em	
	Northerner	northern nigeria has to be the shame of africa		anything wey no be religious violence no concern northerners	
	Christian	sebi bible says if dem slap u u turn the other cheek look at christians embarrassing themselves las las everybody just dey play christ like kor			
	Muslim	mumu muslims under bridge na muslims full there dey suffer	you see this when a man or group of men start forcefully controlling women they themselves are weak very weak thats why they look for the weak beings to		

Appendix: Annotated Hate Speech Examples

			control and treat any how animals in veils always causing problems in the name of islam		
	Igbo	some brainless ibo sha the price of rice in ibadan must be different from imo state abi always the victim		later now they will say igbo are not tribalistic igbos are the must tribalistic and the haters of Nigeria we dont even want the protest so allow people to complain as united nigerians but the protest is not a good thing but its the last option	
	Hausa	this hausa people for this video so and sense are far apart cause wtf are you singing for someone who doesnt even care about your existence someone who dislikes the black		hausa na the problem of this country sense dey always pain them	
	Yoruba	yoruba president the worst president ever		yorubas people are the one destroying nigeria and yoruba leaders are all control by usa and nato countries	