Wétin dey with these comments? Modeling Sociolinguistic Factors Affecting Code-switching Behavior in Nigerian Online Discussions

Innocent Ndubuisi-Obi*
School of Information
University of Michigan
innoobi@umich.edu

Sayan Ghosh*
Department of EECS
University of Michigan
sayghosh@umich.edu

David Jurgens
School of Information
University of Michigan
jurgens@umich.edu

Abstract

Multilingual individuals code switch between languages as a part of a complex communication process. However, most computational studies have examined only one or a handful of contextual factors predictive of switching. Here, we examine Naijá-English code switching in a rich contextual environment to understand the social and topical factors eliciting a switch. We introduce a new corpus of 330K articles and accompanying 389K comments labeled for code switching behavior. In modeling whether a comment will switch, we show that topic-driven variation, tribal affiliation, emotional valence, and audience design all play complementary roles in behavior.

1 Introduction

Multilingual individuals frequently switch between different languages throughout a discourse, a process known as code switching (Heller, 2010; Gambäck and Das, 2016). This switching process is thought to be driven from a variety of factors, including grammatical constraints (Pfaff, 1979; Poplack, 1980), audience design (Gumperz, 1977; Bell, 1984), or even to evoke a specific perception of the speaker's identity (Niedzielski, 1999; Schmid, 2001). In common social situations, many of these factors are in play, yet we often do not have an idea of how they interact. Here, we present a large scale study of code switching in Nigeria between English and Naijá, the widely-spoken Nigerian creole, to quantify which factors predict switching.

Computational studies of code switching have largely focused on linguistic aspects of switching (Solorio and Liu, 2008; Adel et al., 2013; Vyas et al., 2014; Hartmann et al., 2018). However, several recent works have begun to examine the contextual factors that influence switching behavior, finding that the topic driving a discussion spurs on language variation (Shoemark et al., 2017; Stewart et al., 2018) and that individuals are sensitive to the scope of their audience when choosing a language (Papalexakis et al., 2014; Pavalanathan and Eisenstein, 2015). Given that the social context is known to be strongly influential on code switching (Gumperz, 1977; Thomason and Kaufman, 2001; Gardner-Chloros and Edwards, 2004), our work builds on these recent advancements to quantify the impact of social and contextual factors influencing code switching.

Here, we examine the social and contextual factors predictive of English-Naijá code switching in online discussions across five major Nigerian newspapers. Our work makes three contributions towards computational sociolinguistics. First, we introduce a massive new corpus of Naijá and English text that presents code switching behavior in context, using 330K articles and 389K comments from nine years of longitudinal data. Second, we develop a new classifier for distinguishing Naijá and English, identifying over 24K cases of code switching. Third, we show that although topic-driven variation drives much of code switching behavior, tribal affiliation, emotional valence, and audience design play important roles in which language is used.

2 Identifying Naijá and English

Naijá is an English creole spoken by approximately 80 million people throughout Nigeria, with 3 to 5 million speaking it as a first language (Uchechukwu Ihemere, 2006), leading to many popular services generating content in Naijá, e.g., BBC Pidgin. While official business is frequently conducted in English, Naijá is considered the main language of social interaction in Nigeria (Ifeanyi Onyeche, 2004). Although spo-

*Authors contributed equally.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 6204–6214
Florence, Italy, July 28 - August 2, 2019.

Source	Articles	Tokens	Comments
The Nation	150,724	80,596,156	6,232
The Guardian	73,894	39,411,837	59,232
The Punch	39,576	19,453,935	152,928
Vanguard	30,279	29,315,637	178,734
Daily Trust	29,019	14,481,549	723
BBC (Naijá)	6,999	1,114,844	n/a

Table 1: Corpus of Nigerian news in English and Naijá

ken widely, no language detection systems support recognizing the creole, in part due to the lack of existing corpora with examples.[1] Therefore, to support our ultimate goal of modeling the social factors influencing code switching, we first introduce a new corpus of Naijá and English texts and then develop a classifier to distinguish them.

Data A longitudinal sample of Nigerian news was collected from six major news sources; five of these are in Nigerian Standard English, while one is in Naijá. Table 1 summarizes the datasets. Articles span from 2010 to present day and all but the BBC Pidgin site allow users to comment on the article, with activity rates ranging significantly. Notably, all sites share a common commenting framework through Disqus, which allows consistent extraction and identification of individuals and observing commenter's global statistics.

As news media, all six datasets use a formal register in their style, which does not necessarily match that of the comments. Therefore, to supplement the news data, two annotators labeled a sample of 2,500 comments across all sites. As Naijá is less frequent, the sample was bootstrapped to potentially contain more Naijá by first training our classifier (described next) from the news data and then sampling comments uniformly across its posterior distribution. A held out set of 682 randomly sampled comments (not bootstrapped) was additionally doubly annotated (Krippendorff α=0.511) as a test set, 9.5% of which were Naijá; note that due to class imbalance, α represents a highly-conservative estimate of agreement.

Method and Experimental Setup Our goal is to create a classifier that identifies whether a sentence contains Naijá. English is significantly more frequent in our news dataset and therefore we downsample English to a 9:1 ratio following the observed frequency in test data, using 461K English and 51K Naijá sentences from our news corpora, in addition to 1,887 English and 613 Naijá manually-annotated comment sentences.

As a primarily spoken language, Naijá has significant orthographic variation in its spelling (Deuber and Hinrichs, 2007). Therefore, we follow insights from language detection approaches (Lui and Baldwin, 2012; Jauhiainen et al., 2018; Zhang et al., 2018) and adopt character-based features, which are more robust to such variation. Here, character sequences of length 3 to 7 are used as features with a logistic regression with L2 loss. The resulting model is evaluated using AUC in two ways: using 5-fold cross validation within the training data and the held-out comment test set.

Conf.	Example
0.99	See dem people as dem dey steal our money.
0.89	Your brain don sour...Tufiakwa!
0.84	If you no like Kemi go bring Iweala.

Table 2: High confidence Naijá classification examples

Results The classifier was highly accurate at learning to distinguish Naijá and English in the mostly-news training data, achieving a cross-validation AUC of 0.996, compared with the random baseline of 0.5. The model performed less accurately on the comments, which have a more informal register, achieving an AUC of 0.724.

3 Social Factors Influencing Switching

People code switch in part to signal a part of their identity (Nguyen, 2014) and online discussion provides an intersectional context that combines social and topic features that could each elicit the use of Naijá (Myers-Scotton, 1995). Here, we outline the social and contextual factors that could affect whether Naijá is used and identify outline specific research hypotheses to test.

Article Topic The content of a discussion has the potential to elicit a response in a particular language, especially if content, language, and identity interrelate. For example, in online discussions of independence referendums, Shoemark et al. (2017) and Stewart et al. (2018) show evidence of topic-based language variation, with additional modulation based on expected audience. These results point to hypothesis **H1** that we should observe topic-induced variation in which Naijá would be more frequent for certain topics.

[1] Nigerian Standard English is different from Naijá, with each having its own syntax and separate lexicon—to the point that individuals code switch between them (Akande, 2010).

Social Setting The audience imagined by an author leads to differing code switching behavior, where computational studies have found that messages intended for broader audiences typically use the major language (Papalexakis et al., 2014; Shoemark et al., 2017). Similarly, Nguyen et al. (2015) notes that individuals switch to a minority language during a conversation with other individuals. We operationalize audience design in three ways: (1) the number of prior comments to an article, which signals general its potential audience size, (2) the depth of the comment in the discussion thread, which is often a signal of more interpersonal discussion (Aragón et al., 2017), and (3) the time of day the comment is made, as an expectation of future audience size. These three factors lead to hypothesis **H2a** that initial comments will be *less likely* to be in Naijá as they would have a wider audience and **H2b** comments made to a smaller audience are more likely to be made in Naijá.

Tribal affiliation Nigeria is home to individuals identifying with over a hundred different tribal identities which are concentrated in different regions. These tribal affiliations are the strongest aspect of self identity in present day Nigeria (Mustapha, 2006) and have also historically served as sources of conflict due to social stratification along tribal and geographic lines (Akiwowo, 1964; Himmelstrand, 1969). Tribal identity and salience is closely linked with language in Nigeria (Bamiro, 2006), with individuals alternating between English, Naijá, and local languages to emphasize identity. Language choice is driven in part by these cultural identities (Gudykunst and Schmidt, 1987; Myers-Scotton, 1991; Moreno et al., 1998). We test hypothesis **H3** that tribal affiliation will be predictive of code-switching.

As our dataset does not initially come with tribal affiliation, we follow previous work (Rao et al., 2011; Fink et al., 2012) and train a classifier (described in Appendix A) to automatically label all article authors as Igbo, Hausa-Falani, Yoruba, or other. These three tribes constitute over 71% of the population. Similar to prior work, our method attains an 81.0 F1 on author names, with slightly lower performance (67.7 F1) on the noisier commenter names.

Social Status Code switching behavior is connected to perceived notions of status, especially along the perceived status of each language in context (Genesee, 1982). Kim et al. (2014) notes that higher status individuals tend to speak in the majority language. Here, we operationalize status through users' meta-data from Disqus that provides their number of followers, which acts as a proxy for their reputation on the platforms. In hypothesis **H4**, individuals with higher status are more likely to use the majority language, English.

Emotion The language spoken by a bilingual individual is intimately connected to emotion (Rajagopalan, 2004). Indeed, individuals are more likely to swear in their native language (Dewaele, 2004; Rudra et al., 2016) or code switch when being impolite (Hartmann et al., 2018), underscoring a unconscious connection during emotional moments. Odebunmi (2012) notes that Naijá is used in the more formal setting of doctor-patient interactions to express emotions. These results suggest hypothesis **H5** that in high-emotion settings, individuals are more likely to code-switch into Naijá.

4 When is Naijá Used?

What sociocultural factors influence a person's choice of communicating in Naijá or English? Here, we analyze the comments from data in Table 1 to test the hypotheses from Section 3.

Experimental Setup The Naijá-English classifier was run on all comments made to the 330K articles in the dataset, classifying each sentence within the comment separately. If any one sentence is classified as Naijá, we consider the comment to have code-switched, noting that we are not making a distinction about what level the switch is occurring, e.g., word, phrase, or sentence (Gambäck and Das, 2016). Ultimately, this process resulted in 365,420 English and 24,232 Naijá-containing comments.

User-based statistics were extracted for each commenter from their Disqus profile. As only 15K individuals use Disqus accounts (4%), we include an additional binary indicator variable for whether the individual has an account. To test for the effect of content, a 20-topic LDA model (Blei et al., 2003) was run on the article text and included as variables (due to collinearity, topic 20 is excluded). We model tribal affiliation in four ways: (i) the commenter, (ii) the article author,

and, where possible, (iii) the affiliation of the parent being replied to, and (iv) whether the parent explicitly mentions a tribe. For the first, three the "Other" category is the reference coding. Emotion is measured using VADER (Hutto and Gilbert, 2014), a lexicon designed for sentiment analysis in social media on a scale of [-1,1]. We incorporate sentiment in four ways: (1-2) the sentiment scores of the post and its parent, using 0 for the parent's sentiment if the current comment has no parent and (3-4) the absolute value of the sentiment and parent's sentiment. The latter two variables enable us to separately test whether any emotionality (positive of negative) influence using Naijá, rather than the particular direction. Each platform is included as a fixed effect to control for differences in baseline rates of Naijá. After testing for collinearity, all features had VIF<3.1 indicating the model's features are largely independent. As each hypothesis uses different regression variables, this low VIF also indicates that any results are likely not confounded by correlations within the data.

Results A logistic regression model is fit using all the features, and the resulting coefficients, shown in Figure 1, provide support for all five hypotheses. However, the effect sizes of each hypotheses variables differed substantially, pointing to the complexity of code switching behavior.

The strongest effects of Naijá usage in the comment section came from the topic of the article, supporting H1. Topics related to business, social issues, and tribal and electoral politics were more likely to see code switching into Naijá. However, topics related to more general, legislative politics and individual sectors of the economy do not promote Naijá usage. Further, this trend is seen in the newspapers' relative rates: being more oriented towards business topics and targeting an educated audience, *The Guardian* features less code-switching in its comment sections compared to *The Punch*, a tabloid with a wider audience (Marcus, 1999). In particular, the code switching effect is strongest for topics that relate to societal tensions (e.g., political, socioeconomic, and tribal). While prior work on topic-induced variation (Shoemark et al., 2017; Stewart et al., 2018) identified behaviors for political identity-based content (national referendums on independence), in contrast, here, we also observe that individuals are sensitive to audience for more domestic topics like education and health care.

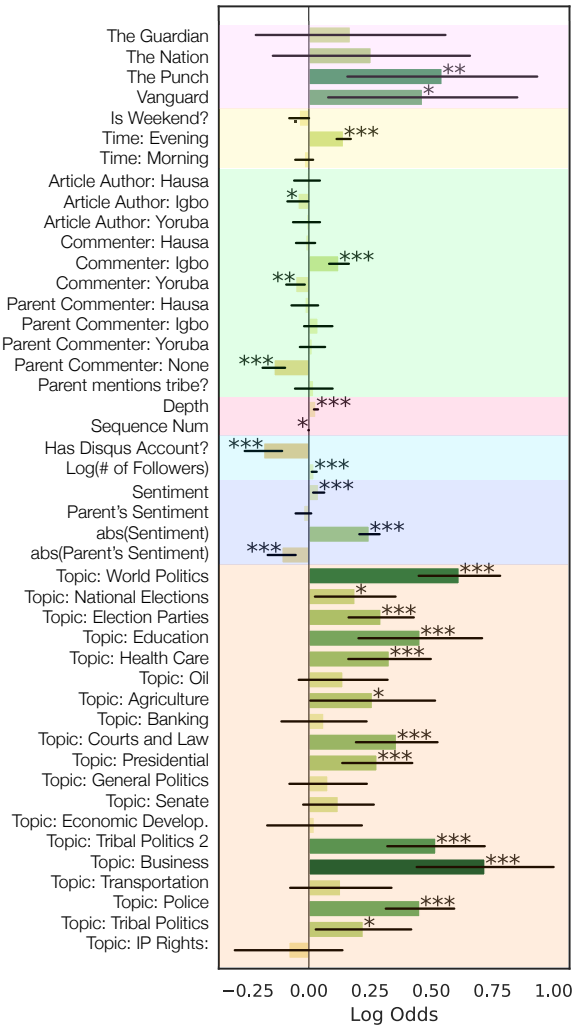

Figure 1: Regression results for whether a comment will have Naijá in it. Error bars show standard error, with *** denoting p<0.001, ** p<0.01, and * p<0.05. Shaded regions group similar variables. Full results are detailed in Appendix Table 7.

The use of Naijá did vary by audience, with strongest support for H2b. Comments deeper in a reply thread are more likely to be Naijá as well as those made in the evening when much of the discussion has taken place and when replies are more likely to be conversational with a particular person, rather than commentary on the article. The total effect is seen by considering both the depth and when "Parent Commenter: None" (i.e., the comment is at the top level). Such initial comments are much more likely to be in English, after which as the discuss turns more conversational, more Naijá is used. Our results agree with those of Nguyen et al. (2015) who found more minority language using in interpersonal communication.

The initial comments to an article (low sequence number) were less likely to be in Naijá (H2a; $p<0.05$), though the effect is relatively weaker.

Tribal affiliation only had limited association with use of Naijá (H3), where Igbo commenters are more likely and Yoruba commenters are less likely to use Naijá. A subsequent model tested for interaction effects between author and parent tribe, which revealed only one significant trend that individuals from all tribes are more likely to reply to Yoruba commenters in Naijá. As Naijá is widely spoken throughout the country, compared with Standard English, which is spoken more frequently at higher socioeconomic levels (Faraclas, 2002), our results suggest its use is not to emphasize tribal affiliation.

The expectation of H4 was observed: higher status (as measured by number of followers) was as predictive of use of the higher prestige language (English), though the effect is relatively small and the effect is estimated only from those users with Disqus accounts. As a complementary analysis, we performed a second test where we replace the number of followers with the number of total upvotes as a proxy of status, with the rationale that users who generate content that is well-received by the community might aquire a positive reputation. The regression results using total upvotes also found a similar weak effect of higher status users writing more in English (and highly similar coefficients for all other features). However, we note that this second analysis has a potential confound, as an English comment could be read by a wider audience and therefore receive more upvotes simply due to audience size rather than status. As all newspapers in our study are primarily read by a Nigerian national audience who is likely bilingual in English and Naijá, this potential effect is expected to be small. Nevertheless, given the limitations of both operationalizations of status, we view their similar results as tentative evidence of the effects of status on Naijá code switching in social discussions (H4).

The effects associated with H5 were strongly shown: when expressing any kind of sentiment, authors were much more likely to do it in Naijá, with a positive effect for using Naijá in positive sentiment comments. Surprisingly, a parent's use of sentiment was negatively associated with Naijá indicating a reaction to emotional language does not elicit a code switch. Given that our model controls for topics that may be more likely to elicit certain emotions, this result suggests that emotion is a driving factor code switching behavior.

5 Conclusion

This work provides the first computational examination of code switching behavior in Naijá through introducing a large corpora of articles in Naijá and Nigerian Standard English, along with comments to these articles. We develop new methods for distinguishing these two languages and identify over 24K instances of code switching in the comments. Through examining code switching in an intersectional social context, our analysis provides evidence of complementary social factors influencing switching. Notably, we find that topical modulation has the largest effect on switching to Naijá, with use of emotion surpassing the effect for a few topics. However, as no one factor was sufficient for predicting code switching, our results point to the need for holistically modeling the social context when examining factors influence code-switching behavior. All data and code are made available at https://blablablab.si.umich.edu/projects/naija/.

Acknowledgments

We thank the three reviewers for their helpful comments.

References

Heike Adel, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 206–211.

Akinmade T Akande. 2010. Is nigerian pidgin english english? *Dialectologia et Geolinguistica*, 18(1):3–22.

Akinsola A Akiwowo. 1964. The sociology of nigerian tribalism? *Phylon (1960-)*, 25(2):155–163.

Pablo Aragón, Vicenç Gómez, and Andreaks Kaltenbrunner. 2017. To thread or not to thread: The impact of conversation threading on online discussion. In *Eleventh International AAAI Conference on Web and Social Media*.

Edmund O Bamiro. 2006. The politics of code-switching: English vs. nigerian languages. *World Englishes*, 25(1):23–35.

Allan Bell. 1984. Language style as audience design. *Language in society*, 13(2):145–204.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Dagmar Deuber and Lars Hinrichs. 2007. Dynamics of orthographic standardization in jamaican creole and nigerian pidgin. *World Englishes*, 26(1):22–47.

Jean-Marc Dewaele. 2004. Blistering barnacles! what language do multilinguals swear in? *Estudios de sociolingstica: Linguas, sociedades e culturas (Issue dedicated to: Bilingualism and emotions)*, 5(1).

Nick Faraclas. 2002. *Nigerian pidgin*. Routledge.

Clayton Fink, Jonathon Kopecky, Nathan Bos, and Max Thomas. 2012. Mapping the twitterverse in the developing world: An analysis of social media use in nigeria. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 164–171. Springer.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *LREC*.

Penelope Gardner-Chloros and Malcolm Edwards. 2004. Assumptions behind grammatical approaches to code-switching: when the blueprint is a red herring. *Transactions of the Philological Society*, 102(1):103–129.

Fred Genesee. 1982. The social psychological significance of code switching in cross-cultural communication. *Journal of language and social psychology*, 1(1):1–27.

William B Gudykunst and Karen L Schmidt. 1987. Language and ethnic identity: An overview and prologue. *Journal of Language and Social Psychology*, 6(3-4):157–170.

John J Gumperz. 1977. The sociolinguistic significance of conversational code-switching. *RELC journal*, 8(2):1–34.

Silvana Hartmann, Monojit Choudhury, and Kalika Bali. 2018. An integrated representation of linguistic and social functions of code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Monica Heller. 2010. *Codeswitching: Anthropological and sociolinguistic perspectives*, volume 48. Walter de Gruyter.

Ulf Himmelstrand. 1969. Tribalism, nationalism, rank-equilibration, and social structure: A theoretical interpretation of some socio-political processes in southern nigeria. *Journal of Peace Research*, 6(2):81–102.

Clayton J Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*.

Joseph Ifeanyi Onyeche. 2004. As naija pipo dey tok: a preliminary analysis of the role of nigerian pidgin in the nigerian community in sweden. *Africa & Asia*, 4:48–56.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2018. Automatic language identification in texts: A survey. *arXiv preprint arXiv:1804.08186*.

Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248. ACM.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Judith Marcus. 1999. *Surviving the Twentieth Century: Social Philosophy From the Frankfurt School to the Columbia Faculty Seminars*. Transaction Publishers.

Luis Moreno, Ana Arriba, and Araceli Serrano. 1998. Multiple identities in decentralized spain: The case of catalonia. *Regional & Federal Studies*, 8(3):65–88.

Abdul Raufu Mustapha. 2006. *Ethnic structure, inequality and governance of the public sector in Nigeria*. United Nations Research Institute for Social Development Geneva, Switzerland.

Carol Myers-Scotton. 1991. Making ethnicity salient in codeswitching. *Language and ethnicity*, 2:95–109.

Carol Myers-Scotton. 1995. *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.

Dong-Phuong Nguyen, Rudolf Berend Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on twitter. In *Proceedings of the Ninth International AAAI Conference on Web and Social Media, ICWSM 2015*, pages 666–669. AAAI Press. Eemcs-eprint-26675.

Thuy Nguyen. 2014. *Code Switching: A sociolinguistic perspective*. Anchor Academic Publishing (aap_verlag).

Nancy Niedzielski. 1999. The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology*, 18(1):62–85.

Akin Odebunmi. 2012. the baby dey chuk chuk: Language and emotions in doctor–client interaction. *Pragmatics and Society*, 3(1):120–148.

Evangelos Papalexakis, Dong Nguyen, and A Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of the first workshop on computational approaches to code switching*, pages 42–50.

Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Audience-Modulated Variation in Online Social Media. *American Speech*, 90(2):187–213.

Carol W Pfaff. 1979. Constraints on language mixing: intrasentential code-switching and borrowing in spanish/english. *Language*, pages 291–318.

Shana Poplack. 1980. Sometimes ill start a sentence in spanish y termino en espanol: toward a typology of code-switching1. *Linguistics*, 18(7-8):581–618.

Kanavillil Rajagopalan. 2004. Emotion and language politics: The brazilian case. *Journal of multilingual and multicultural development*, 25(2-3):105–123.

Delip Rao, Michael Paul, Clay Fink, David Yarowsky, Timothy Oates, and Glen Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social media. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Monojit Choudhury, Kalika Bali, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do hindi-english speakers do on twitter? In *Proceedings of EMNLP 2016*. Association for Computational Linguistics.

Carol L Schmid. 2001. *The politics of language: Conflict, identity and cultural pluralism in comparative perspective*. Oxford University Press.

Philippa Shoemark, Debnil Sur, Luke Shrimpton, Iain Murray, and Sharon Goldwater. 2017. Aye or naw, whit dae ye hink? scottish independence and linguistic identity on social media. In *Proceedings of EACL*, volume 1, pages 1239–1248.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

Ian Stewart, Yuval Pinter, and Jacob Eisenstein. 2018. Sío no, qué penses? catalonian independence and linguistic identity on social media. In *Proceedings of NAACL*.

Sarah Grey Thomason and Terrence Kaufman. 2001. *Language contact*. Edinburgh University Press.

Kelechukwu Uchechukwu Ihemere. 2006. A basic description and analytic treatment of noun clauses in nigerian pidgin. *Nordic Journal of African Studies*, 15(3):296–313.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of EMNLP*.

A A Classifier for Tribal Affiliation

As our dataset does not come with tribal affiliations to start with, we first create a classifier to identify affiliations on the basis of name. Due to cultural norms in Nigeria, individual's names often reveal their tribal affiliation (Rao et al., 2011; Fink et al., 2012), which lends itself to developing computational methods for distinguishing between the affiliations. Here, we develop a classifier for distinguishing between the three largest tribal affiliations: Hausa-Falani (29%), Yoruba (21%), and Igbo (21%), which together account for over 71% of the population thereby providing solid coverage of online users. Data for the tribal affiliation classifier was compiled using online databases and annotated names extracted from a held-out set of article authors and commenter names from the dataset of articles. The final training dataset included 493 Hausa-Falani names, 500 Yoruba names, 351 Igbo names, and 511 "other" names, which encompassed Nigerian names not fulling under the aforementioned three categories as well as non-Nigerian names (e.g., "The Editorial Board" or "flexingbenny"). Table 4 shows examples of names used in training. We note that some tribes' names have similar cultural origins and therefore our data could result in systematic misclassifications for some tribes; for example, both the Hausa and the Kanuri (an ethnic group comprising roughly 3-4% of the Nigerian population) share names that are Arabic in origin. Our model would likely label all such names as Hausa, though due to population size differences, the impact of such errors are likely to be small.

A logistic regression classifier was trained using L2 regularization with character n-grams ranging from 2 to 5 in length. To evaluate perfor-

Model	Article Author	Commenter
Our method	0.81	0.68
majority class	0.12	0.17
random	0.24	0.21

Table 3: Tribal affiliation classifier Macro F1

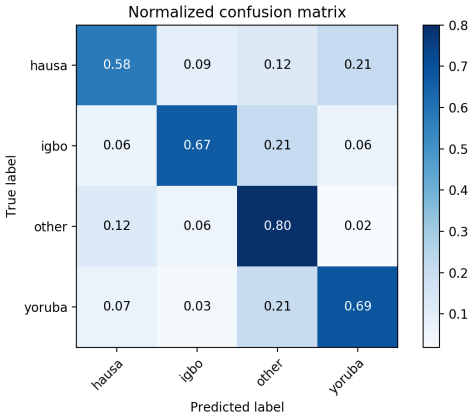

Figure 2: Normalized confusion matrix of tribal affiliation classifier

mance, two trained annotators labeled 200 held out names of article authors and 200 commenter names; Krippendorff α agreement was 0.516, with disagreements resolved through adjudication.

Performance of our model is shown in Table 3. While absolute performance on article authors is on par with similar approaches to classifying tribal affiliation (Rao et al., 2011; Fink et al., 2012), which applied their classifiers to clean name data. Performance on commenter names is slightly lower due noise from lexical variation, misspellings, and web extraction. Table 5 shows examples of names with tribal affiliation in the test data. The confusion matrix of the tribal affiliations, shown in Figure 2, reveals no systematic misclassification bias, suggesting that any errors will only increase variance in the downstream results without biasing findings towards one particular affiliation.

Category.	Example
Hausa	Murtala Mohammed, Saheed Ahmad Rufai, abubakar umar, ismail mudashir, mamman usman
Yoruba	Olajide Olatundun, Yetunde Arebi, Ayo Olododo, Ahmad Olawale, Aderonke Adeyeri
Igbo	Kelechi Akunna, Davies Iheamnachor, Uche Okeke, Chukwudi Enekwechi, Bartholomew Madukwe
Other	John Marks, Aaron Frost, Charles Frederick, Bush Jenkins, Victor Jonah

Table 4: Tribal affiliation training data examples

Category.	Example
Hausa	Muhammad Hassanto, AK Mohammed, Suleiman Alatise, Alalere Tajudeen, Zahraddeen Yakub
Yoruba	Olatunji Omirin, Adenike Grace, Anthony Akinola, Tayo Aiyetoro, Vincent Ikuoola
Igbo	Ochuko Akuoph, Nwanchor Friday, John Megbechi, Adache Ene, Cynthia Onana
Other	Leon Willems, Michael Johnbull, Pamela John, Roses Moses, Tim Daiss

Table 5: Tribal affiliation test data examples

B Additional Naijá Classification Examples

Table 8 shows a sample of instances classified by the final trained language-distinguishing model. Instances are sampled uniformly across the posterior to show the variety of confidence scores.

C Additional Regression Details

Table 7 shows the full regression coefficients for the model depicted in Figure 1. We additionally show the most probable words for each topic in Table 6. Note that the final topic ("Security") was intentionally omitted from the regression to remove the effects of collinearity between topic probabilities.

Category.	**Example**
World Politics	africa president african countries world trump united south country international
National Elections	election inec elections electoral commission anambra party governor political national
Election Parties	party pdp apc governor national election political chairman congress candidate
Education	university school education students schools nigeria teachers universities lagos prof
Health Care	god health children women church life family medical hospital child
Oil	oil power gas petroleum nigeria company electricity nnpc government crude
Agriculture	nigeria food farmers products production agriculture rice government country agricultural
Banking	cent bank billion market cbn nigeria exchange million banks capital
Courts and Law	court justice efcc law accused judge appeal federal trial judgment
Presidential	president nigeria buhari country nigerians jonathan government national political nation
General Politics	people time nigeria don political country money nigerians power government
Senate	national senate president assembly government house committee budget federal public
Economic Politics	nigeria government development country economic sector economy people national support
Tribal Politics 2	governor delta government rivers people edo niger bayelsa local chief
Business	usiness bank customers nigeria company services mobile technology service brand
Transportation	road lagos government roads federal airport project air aviation safety
Police	police arrested incident command told suspects security officer lagos killed
Tribal Politics	governor lagos ekiti government people osun fayose ondo chief ogun
IP Rights	punch government workers rights email written protected website published broadcast
Security	security government boko haram military people army kaduna nigeria nigerian

Table 6: Key words corresponding to topic

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.5180	0.208	-16.922	0.000	-3.925	-3.111
The Guardian	0.1704	0.198	0.862	0.389	-0.217	0.558
The Nation	0.2553	0.205	1.243	0.214	-0.147	0.658
The Punch	0.5449	0.198	2.754	0.006	0.157	0.933
Vanguard	0.4649	0.197	2.359	0.018	0.079	0.851
Is Weekend?	-0.0398	0.021	-1.926	0.054	-0.080	0.001
Time: Evening	0.1422	0.015	9.558	0.000	0.113	0.171
Time: Morning	-0.0193	0.019	-1.031	0.302	-0.056	0.017
Article Author: Hausa	-0.0077	0.027	-0.285	0.776	-0.060	0.045
Article Author: Igbo	-0.0452	0.022	-2.080	0.038	-0.088	-0.003
Article Author: Yoruba	-0.0098	0.028	-0.347	0.728	-0.065	0.045
Commenter: Hausa	-0.0136	0.020	-0.682	0.496	-0.053	0.026
Commenter: Igbo	0.1229	0.021	5.969	0.000	0.083	0.163
Commenter: Yoruba	-0.0547	0.019	-2.826	0.005	-0.093	-0.017
Parent Commenter: Hausa	-0.0167	0.028	-0.602	0.547	-0.071	0.038
Parent Commenter: Igbo	0.0382	0.030	1.291	0.197	-0.020	0.096
Parent Commenter: Yoruba	0.0147	0.026	0.557	0.577	-0.037	0.066
No parent (top-level comment)	-0.1432	0.023	-6.097	0.000	-0.189	-0.097
Parent mentions tribe?	0.0200	0.039	0.511	0.609	-0.057	0.097
Comment Depth	0.0290	0.004	6.781	0.000	0.021	0.037
Sequence Number	-0.0013	0.001	-2.257	0.024	-0.002	-0.000
Has Disqus Account?	-0.1858	0.039	-4.761	0.000	-0.262	-0.109
log(Number of Followers)	0.0217	0.005	4.171	0.000	0.011	0.032
Sentiment	0.0400	0.012	3.434	0.001	0.017	0.063
Parent's Sentiment	-0.0224	0.016	-1.372	0.170	-0.054	0.010
abs(Sentiment)	0.2474	0.021	11.713	0.000	0.206	0.289
abs(Parent's sentiment)	-0.1112	0.029	-3.807	0.000	-0.168	-0.054
Topic: World Politics	0.6148	0.085	7.240	0.000	0.448	0.781
Topic: National Elections	0.1893	0.084	2.252	0.024	0.025	0.354
Topic: Election Parties	0.2953	0.068	4.344	0.000	0.162	0.428
Topic: Education	0.4549	0.129	3.530	0.000	0.202	0.708
Topic: Health Care	0.3294	0.086	3.830	0.000	0.161	0.498
Topic: Oil	0.1399	0.093	1.511	0.131	-0.042	0.321
Topic: Agriculture	0.2604	0.130	2.001	0.045	0.005	0.515
Topic: Banking	0.0618	0.089	0.695	0.487	-0.112	0.236
Topic: Courts and Law	0.3587	0.085	4.216	0.000	0.192	0.525
Topic: Presidential	0.2792	0.073	3.823	0.000	0.136	0.422
Topic: General Politics	0.0785	0.081	0.973	0.331	-0.080	0.237
Topic: Senate	0.1208	0.074	1.641	0.101	-0.023	0.265
Topic: Economic Develop.	0.0230	0.099	0.233	0.816	-0.170	0.216
Topic: Tribal Politics 2	0.5192	0.102	5.113	0.000	0.320	0.718
Topic: Business	0.7199	0.143	5.044	0.000	0.440	1.000
Topic: Transportation	0.1304	0.106	1.235	0.217	-0.077	0.337
Topic: Police	0.4538	0.071	6.362	0.000	0.314	0.594
Topic: Tribal Politics	0.2230	0.099	2.242	0.025	0.028	0.418
Topic: IP Rights	-0.0828	0.112	-0.738	0.461	-0.303	0.137

Table 7: Logistic regression results for predicting the use of Naijá in a comment (cf. Figure 1 in Main Paper)

p(Naijá)	Sentence
0.966469	Me, I don taya for awa piple oo!
0.962135	You fit correct o because na only Igbos be the major tribe for Nigeria wey no get tribal fellow as citizens of neighboring West African countries.
0.927030	I don't blame you.
0.906231	APC na Edo, Edo na APC.
0.863062	Abeg make I go collect small brandy from terrydgreat.
0.824909	Watch for August 14
0.812487	Guess your bet don cast by now.
0.798906	Abeg make we hear word.
0.798014	I tire for you!
0.793962	No spillage go affect my life.
0.792577	Make I come, joor!
0.783273	If e break or e crack, all na spoil.
0.782503	London.
0.752294	But you be "entourage" abi "High commissioner" dat one na another chapter.
0.727051	Abeg, Make we hia word.
0.696996	#NO2Buhari
0.691851	aspirant for mouth.
0.690936	im done.
0.670130	The guy no get money, make him no get something to press after the whole stress again?
0.659617	So please don't refer me to it.
0.649868	Uba no case.
0.637665	Is it by land mass...abegii na population.
0.620910	I am done with you for ever!
0.613897	I weep for my country
0.606911	When am supposed to be charged 100Naira for bus fare, am charged 150Naira because of some party men.
0.530184	Happy New Year !!
0.530156	Like father like son.
0.497918	How come Saraki suddenly forgot Ekwe??
0.489173	Thanks dear.
0.421127	A year from now?
0.404989	I got N4.6b from Dasuki for spiritual purposes - Bafarawa 6.
0.373616	DG, Immigration ..... Northern Muslim Hausa-Fulani 18.
0.356982	Solomon Grundy, Born on a Monday, Christened on Tuesday, Married on Wednesday, Took ill on Thursday, Grew worse on Friday, Died on Saturday, Buried on Sunday.
0.316233	India to come and help run government refinery?.
0.302539	Good morning in this hot afternoon Dr.Buhari, you just behave like say you don't understand what you are doing?
0.293588	WHY CAN'T ONE NIGERIA DIVIDE - OSINBAJO ?
0.256922	He is crawling inside a 50 bedroom mansion on top a hill at minna.
0.232036	Lolz.
0.154062	Shehu Sani, may God bless you.
0.143232	Well stated .I don't even know as much , as this of him.
0.132101	Vanguard please can you do a research on how much each zone or state contribute yearly to federal government coffer and how much each zone or state get from federal government coffer yearly?
0.103502	But madam your contradiction defeats your standpoint.
0.064607	Some Igbo then came out to claim NRI Kingdom.
0.049916	Now my Thursday is wasted.
0.016798	"... when have we started practicing state...government."
0.013228	They had been issued with bullets but I was unarmed.
0.005393	Yom Kippur war even is mild, the US and Taliban war in Afghanistan is better suited.
0.000042	A lot of the numerous Federal Ministries and agencies should be scrapped, and the funds given to the states to fund what is important to them.

Table 8: A random sample of comment sentences and their classification probabilities