EkoHate: Abusive Language and Hate Speech Detection for Code-switched Political Discussions on Nigerian Twitter

Comfort Eseohen Ilevbare[1*], Jesujoba Oluwadara Alabi[2*], David Ifeoluwa Adelani[3], Firdous Damilola Bakare[1], Oluwatoyin Bunmi Abiola[1] and Oluwaseyi Adesina Adeyemo[1]

[1] Department of Computer Science, Afe Babalola University, Ado-Ekiti, Nigeria

[2] Spoken Language Systems, Saarland University, Saarland Informatics Campus, Germany

[3] University College London

jalabi@lsv.uni-saarland.de, d.adelani@ucl.ac.uk

{abiolaob,adeyemo}@abuad.edu.ng

Abstract

Nigerians have a notable online presence and actively discuss political and topical matters. This was particularly evident throughout the 2023 general election, where Twitter was used for campaigning, fact-checking and verification, and even positive and negative discourse. However, little or none has been done in the detection of abusive language and hate speech in Nigeria. In this paper, we curated *code-switched* Twitter data directed at three musketeers of the governorship election on the most populous and economically vibrant state in Nigeria; Lagos state, with the view to detect offensive speech in political discussions. We developed EkoHate—an abusive language and hate speech dataset for political discussions between the three candidates and their followers using a binary (normal vs offensive) and fine-grained four-label annotation scheme. We analysed our dataset and provided an empirical evaluation of state-of-the-art methods across both supervised and cross-lingual transfer learning settings. In the supervised setting, our evaluation results in both binary and four-label annotation schemes show that we can achieve 95.1 and 70.3 F1 points respectively. Furthermore, we show that our dataset adequately transfers very well to three publicly available offensive datasets (OLID, HateUS2020, and FountaHate), generalizing to political discussions in other regions like the US.

1 Introduction

The internet, with various social media platforms, has interconnected our world, facilitating real-time communication. One area that has benefited from the use of social media platforms is elections at various levels. Research has shown that these platforms have an impact on the outcome of elections in different countries (Fujiwara et al., 2021; Carney, 2022), but not without the spread of false information (Grinberg et al., 2019; Carlson, 2020; Yerlikaya and Toker, 2020), dissemination of hate speech (Siegel et al., 2021; Nwozor et al., 2022), and various other forms of attacks. Therefore, efforts have been made to automatically identify hateful and divisive comments (Davidson et al., 2017). They include supervised methods, that focus on curating hate speech datasets (Mathew et al., 2021; Demus et al., 2022; Piot et al., 2024).

However, the majority of these datasets were created for elections in the US (Suryawanshi et al., 2020; Grimminger and Klinger, 2021; Zahrah et al., 2022) and other non-African countries (Alfina et al., 2017; Febriana and Budiarto, 2019). In this work, we focus on Nigerian elections. Nigerians have a notable online presence and actively discuss political and topical matters. This was particularly evident throughout the 2023 general election, where Twitter was used for campaigning, fact-checking, verification, and positive and negative discourse. However, little or none has been done in the detection of offensive and hate speech in Nigeria.

In this paper, we create EkoHate—a new code-switched abusive language and hate speech detection dataset containing 3,398 annotated tweets gathered from the posts and replies of three leading political candidates in Lagos, annotated using a binary ("normal" vs "offensive" i.e abusive & hateful) and fine-grained four-label annotation scheme. The four-label annotation scheme categorizes tweets into "normal", "abusive", "hateful", and "contempt". The last category was added based on the difficulty to classify some tweets that do not properly fit into "normal" or "abusive" but express strong disliking in a neural tone, suggested by (Ron et al., 2023). Table 1 shows some examples of tweets and their categorization. The last example "You will still be voted out of office sir." does not fit the categorization of "offensive" but can be "contemptuous" to a sitting Governor, implying that despite his campaign, he would still be voted out.

Our evaluation shows that we can identify the

[*]Equal contribution.

Proceedings of the The 8th Workshop on Online Abuse and Harms (WOAH), pages 28–37
June 20, 2024

Tweet	N-O	N-A-H-C
Bro, go to the field and gather momentum. Social media can only do so much	N	N
LOL. This guy na mumu honestly	O	A
A bl00dy immigrant calling another person immigrant...	O	H
You will still be voted out of office sir.	-	C

Table 1: Examples of tweets and their labels under two labelling schemes. In the second example "na mumu" can mean "is a fool" . N is Normal, O is offensive (i.e. Abusive & Hateful), A is abusive, and C is contempt.

offensive tweets with the high performance of 95.1 F1 by fine-tuning a domain-specific Twitter BERT model (Barbieri et al., 2020). However, on a four-label annotation scheme, the F1-score drops to 70.3 F1 showing the difficulty of the fine-grained labeling scheme. Furthermore, we conduct cross-corpus transfer learning experiments using OLID (Zampieri et al., 2019), HateUS2020 (Grimminger and Klinger, 2021), and FountaHate (Founta et al., 2018) which achieved 71.1 F1, 58.6 F1, and 43.9 F1 points respectively on EKOHATE test set. Interestingly, we find that our dataset achieves a good transfer performance to the existing datasets reaching an F1-score of 71.8 on OLID, 62.7 F1 on HateUS2020 and 53.6 on FountaHate, which shows that our annotated dataset generalizes to political discussions in other regions like the US despite the cultural specificity and code-switched nature of our dataset. We hope our dataset encourages the evaluation of hate speech detection methods in diverse countries. The data and code are available on GitHub[1]

2 EKOHATE dataset

2.1 Lagos Gubernatorial Elections

Lagos (also known as Èkó) is the commercial nerve centre of Nigeria, the former federal capital of Nigeria, and the most populous city in Nigeria and Africa with over 15 million residents according to Sasu (2023). In the 2023 Nigerian election, Lagos is probably the most strategic state because of its voting power, and most importantly because the leading candidate for the presidential election is from Lagos. There were three leading candidates from the major political parties: All Progressives Congress (APC), Peoples Democratic Party (PDP), and Labour Party (LP). The latter was particularly popular on social media and especially among the youths because Nigerians saw it as a third force. Therefore, there was a lot of controversial and offensive tweets on social media during the election

[1] https://github.com/befittingcrown/EkoHate

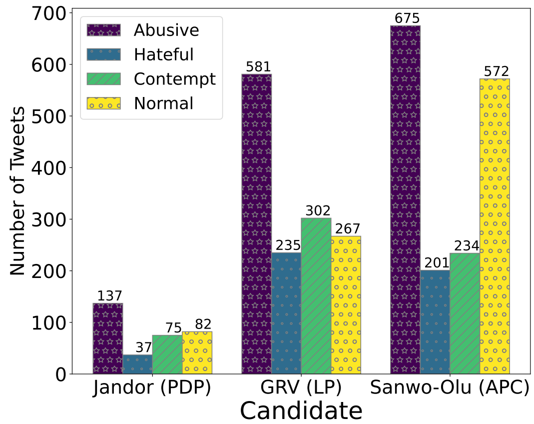

Figure 1: **EkoHate**: The distribution of the classes per candidate.

of Lagos. Thus, we focus on analyzing the political tweets during the last Lagos election.

2.2 Labelling Scheme

There are different labeling scheme for offensive and hate-speech on social media. The simplest approach is to categorize the tweets as either *offensive* or *non-offensive* (Zampieri et al., 2019). In the literature (Davidson et al., 2017; Founta et al., 2018), it is popular to distinguish between the type of **offensive** content as either *abusive* or *hateful*. Here, we adopted the labelling scheme of **normal** (or non-offensive), **abusive**, **hateful**, and **contempt**. The last one was added based on the difficultly of accurately classifying some political tweets showing a strong disliking to someone but expressed using a neutral tone, following the categorization of Ron et al. (2023). Examples of such tweets are: "Just dey play oooo" and "The sheer effrontery! (..to be contesting)", "As if we were sitting before" (a response to—Èkó E dìde (stand up Lagos)!! GRV..).

Anotators The annotators consist of two female individuals: one undergraduate and one postgraduate student in computer science. Neither annotator is from Lagos state nor affiliated with any of the political parties. They underwent a training session for the task, which involved introducing them to

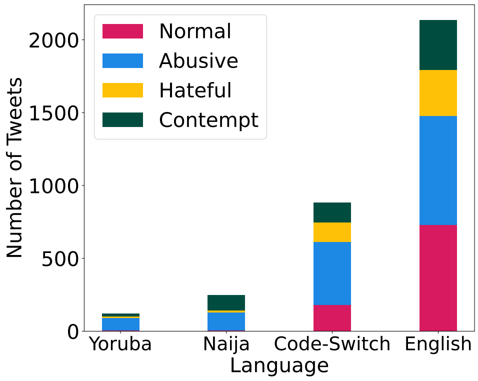

Figure 2: The label distribution according to languages.

Data	Number of tweets train	dev	test
Binary			
OLID (N-O)	11,916	1,324	860
HateUS2020 (N-H)	2,160	240	600
EkoHate (N-O)	1,950	278	559
EkoHate (N-H)	976	139	280
Multi class			
EkoHate (N-A-H)	1,950	278	559
FountaHate (N-A-H)	79,625	2,042	4,299
EkoHate (N-A-H-C)	2,377	339	682

Table 2: The split of the different datasets

the task and Label Studio[2] annotation platform.

Data collection and Annotation Tweets were manually extracted from twitter platform over a period of ten weeks and about 3,398 tweets were collected and annotated. For the purpose of this study, only tweets and replies from three candidates—Babajide Olusola Sanwo-Olu representing APC, Gbadebo Chinedu Patrick Rhodes-Vivour popularly known as GRV representing LP, and Abdul-Azeez Olajide Adediran, popularly known as Jandor representing PDP, were utilized due to the substantial traffic and reactions on their pages, providing ample data for this research. The corpus was annotated by two volunteers for the following five different label categories, *normal*, *contempt*, *abusive*, and *hateful* and *indeterminate*. None of the tweets were classified as indeterminate. The inter-agreement score of the annotation in terms of **Fleiss Kappa** score is **0.43** signifying a moderate agreement. Since, we only have two annotators, we could not use majority voting. To determine the final annotation, we ask the two to meet in-person, discuss and resolve the conflicting annotations. Finally, one of the authors of the paper did a review of the annotation to check for consistency.

EKOHATE data statistics Figure 1 shows the annotated data distribution for the three political candidates: Jandor, GRV, and Sanwo-Olu, with 332, 1385, and 1682 tweets respectively. The incumbent governor, representing APC, garnered the highest engagement, resulting in more tweets. Among the candidates, the proportion of *abusive* tweets is similar at 41%. In contrast, *hateful* tweets associated with the GRV account exceed those from other candidates by more than 4%. Additionally, tweets with the *contempt* are approximately 8% more frequent for Jandor and GRV compared to Sanwo-Olu.

[2] https://labelstud.io/

The dataset exhibits three primary characteristics: it is multilingual, features code-switching, and is inherently noisy due to its social media origin. It has tweets in English, Yoruba, and Nigerian Pidgin (or Naija), which are commonly used languages in Nigeria. Moreover, it includes instances of code-switching between these languages. Figure 2 shows the distribution of tweets across Yoruba, Naija, Code-Switch and English, with 120 (3.5%), 247 (7.3%), 884 (26.0%), and 2,144 (63.2%) tweets respectively. The *abusive* tweets outnumber *normal* tweets across all languages, with Yoruba, Code-Switch, and Naija tweets having a higher proportion of abusive content compared to other categories within each language.

We split the data per label into 70%, 10% and 20% to create the training, development and test.

3 Experiment Setup

Dataset For our study, we opted for both binary and multi-class settings. For binary settings with EkoHate, we consider **binary** label configurations: "normal vs. offensive" (N-O), and "normal vs. hateful" (N-H). For the multi-class, we consider: "normal vs. abusive vs. hateful" (N-A-H), and "normal vs. abusive vs. hateful vs. contempt" (N-A-H-C). And in the multi-class setup, we remove the instances of the excluded classes in the train, development and test splits.

To assess the quality and consistency of our annotations relative to previous work, we conducted cross-corpus transfer experiments. For this task, we opted for three widely known datasets which are offensive language identification dataset (OLID) (Zampieri et al., 2019), a corpus of offensive speech and stance detection from the 2020 US elections (HateUS2020) (Grimminger and Klinger, 2021), and a large hatespeech dataset (FountaHate) (Founta et al., 2018). These are datasets collected from Twitter and manually annotated. While

schema	normal	offensive	abusive	hateful	contempt	F1
N-O	$93.4_{\pm0.4}$	$96.8_{\pm0.2}$	-	-	-	$95.1_{\pm0.3}$
N-H	$94.6_{\pm0.3}$	-		$89.2_{\pm0.7}$	-	$91.9_{\pm0.5}$
N-A-H	$93.4_{\pm0.5}$	-	$85.9_{\pm1.3}$	$55.4_{\pm4.7}$	-	$78.2_{\pm2.2}$
N-A-H-C	$90.5_{\pm0.6}$	-	$78.6_{\pm0.8}$	$51.1_{\pm2.2}$	$61.1_{\pm1.7}$	$70.3_{\pm1.3}$

Table 3: Result of hateful and offensive language detection on EkoHate dataset.

dataset	normal	offensive	abusive	hateful	F1
OLID	$88.3_{\pm0.2}$	$69.5_{\pm1.0}$	-	-	$78.9_{\pm0.6}$
→ EkoHate	$69.2_{\pm0.2}$	$73.1_{\pm0.4}$	-	-	$71.1_{\pm0.3}$
EkoHate	$93.4_{\pm0.4}$	$96.8_{\pm0.2}$	-	-	$95.1_{\pm0.3}$
→ OLID	$80.4_{\pm0.7}$	$63.2_{\pm0.8}$	-	-	$71.8_{\pm0.7}$
HateUS2020	$95.2_{\pm0.5}$	-	-	$60.7_{\pm2.5}$	$77.8_{\pm1.5}$
→ EkoHate	$83.1_{\pm0.6}$	-	-	$34.1_{\pm4.7}$	$58.6_{\pm2.6}$
EkoHate	$94.6_{\pm0.3}$	-	-	$89.2_{\pm0.7}$	$91.9_{\pm0.5}$
→ HateUS2020	$87.2_{\pm1.2}$	-	-	$38.3_{\pm1.6}$	$62.7_{\pm1.4}$
FountaHate	$95.2_{\pm0.1}$	-	$89.0_{\pm0.1}$	$41.1_{\pm1.4}$	$75.1_{\pm0.5}$
→ EkoHate	$63.5_{\pm0.7}$	-	$34.9_{\pm2.7}$	$33.3_{\pm2.3}$	$43.9_{\pm0.7}$
EkoHate	$93.4_{\pm0.5}$	-	$85.9_{\pm1.3}$	$55.4_{\pm4.7}$	$78.2_{\pm2.2}$
→ FountaHate	$82.8_{\pm0.7}$	-	$61.2_{\pm3.4}$	$16.8_{\pm1.5}$	$53.6_{\pm0.9}$

Table 4: Cross-corpus transfer results between EkoHate and other datasets.

OLID used *offensive* and *non-offensive* schema, HateUS2020 used *hateful* and *non-hateful* schema, and FountaHate used four classes which are, *normal*, *abusive*, *hateful*, and *spam*. However, for this work, instances labeled as *spam* were removed.

OLID and HateUS2020 had no validation set, therefore, we sampled 10% of their training splits as the development set. However, due to the large size of FountaHate and the absence of dedicated development and test sets, unlike OLID and HateUS2020, we split the data using the proportions of 92.5%, 2.5%, and 5% for training, development, and test sets, respectively. See Table 2 for the splits and sizes of data.

Models and Training Using the respective datasets, we fine-tuned Twitter-RoBERTa-base (Barbieri et al., 2020). [3] Each model was trained for 10 epochs with a maximal input length of 256, batch size of 16, a learning rate of $2 \cdot 10^{-5}$ using the Huggingface framework. We reported label-wise F1 score as well as macro F1 of 5 runs for the different models for the different classes and also Macro-F1.

Furthermore, given that the baseline model was trained using 5 runs, we explored the effect of model ensembling on the EkoHate dataset. The use of model ensembling has been shown to achieve better results than individual models(Zimmerman et al., 2018; Rajendran et al., 2019; Saha et al., 2021; Singhal and Bedi, 2024). Therefore, we also evaluated hard ensembling, which involved majority voting on five model predictions.

4 Results

EkoHate baseline We fine-tuned Twitter-RoBERTa-base on the EkoHate dataset in both binary and multi-class settings and present the results in Table 3. We observed that binary configurations are easy tasks, achieving high F1 scores of 95.1 and 91.9 for *normal versus offensive and hateful* categories, respectively. However, multi-class configurations are difficult, as classes are not predicted equally well. Lastly, we observed that in all settings, the *hateful* class was the most challenging. We attribute this to the *hateful* class being the least occurring in the EkoHate dataset and the language model's inability to correctly model the class, despite being trained as few-shot learners. Due to class imbalance in the data, we explored models ensembling using majority voting. Our results indicate potential improvements of up to +2.3 for multi-class setups, with relative improvements observed in the binary setups. More details are provided in Appendix D.

Effect of code-switching Going further, we examine the in-language performance of the baseline models, focusing on the F1 scores for the languages present in the test sets (English, Code-switch, Naija and Yoruba). Appendix B shows the distribution of these languages in the test sets, while Table 6 shows the corresponding results. The results indi-

[3] While our data is multilingual and code-switched, we find that English-only model performed better than multilingual model from our early analysis. Result is in Appendix A

schema	Tweet	Lang.	Gold	Pred.
N-A-H	Leave Lagos and return to Anambra omo werey	CDW	hateful	abusive
	Ogun kill you! By the time we're done with you, you'll tell us the real truth behind 20-10-2020. Murderer!	CDW	hateful	abusive
N-A-H-C	The way pitobi failed you will also failed woefully	CDW	hateful	abusive

Table 5: Examples of correct and incorrect predictions.

	Language			
Data	English	Code-Switch	Naija	Yoruba
N-O	$94.7_{\pm 0.3}$	$95.4_{\pm 0.6}$	$82.3_{\pm 0.0}$	$100.0_{\pm 0.0}$
N-H	$91.7_{\pm 0.4}$	$92.6_{\pm 0.8}$	$73.3_{\pm 0.0}$	$100.0_{\pm 0.2}$
N-A-H	$77.5_{\pm 0.6}$	$78.0_{\pm 2.9}$	$57.5_{\pm 7.0}$	$91.4_{\pm 7.4}$
N-A-H-C	$68.9_{\pm 1.0}$	$64.2_{\pm 2.7}$	$60.4_{\pm 1.2}$	$86.2_{\pm 12.7}$

Table 6: In-language performance for English, Code-Switch, Naija, and Yoruba on EkoHate test set.

cate that the models struggle more with Naija, as shown by consistently lower average in-language performance compared to the overall test performance in Table 3. We attribute this primarily to the small size of the Naija examples. In contrast, we observed higher F1 scores for Yoruba. However, considering both Yoruba and Naija have the fewest number of examples, we cautiously attribute their performances to chance and leave this for future work to explore.

Cross-corpus Transfer setting For this experiment, we trained Twitter-RoBERTa-base on existing datasets and evaluated its performance on the EkoHate dataset and vice versa. Table 4 shows the result of our zero-shot cross-corpus transfer result. As expected, when models trained on any of the datasets are evaluated on their corresponding test sets, we obtained a high F1 score with the lowest being FountaHate, where we obtained 75.1 F1 score. However, when these models are evaluated on a different corpora, we observed significantly low performance, for example, HateUS2020→EkoHate gave 58.6 points. Surprisingly, transferring from our newly created data, EkoHate performs slightly better than OLID (+1%) & HateUS2020 (+4%), which shows our dataset generalizes more, possibly due to the fact that EkoHate has a majority of English tweets.

5 Error Analysis

Results from Tables 3 and 4 show that the *hateful* is a difficult class to correctly predict. Hence, we examined the predictions of one of the baseline models for the N-A-H and N-A-H-C. In Appendix C, we showed that *hateful* tweets were often misclassified as *abusive*. Table 5 highlights some misclassified *hateful* tweets. For example, the first N-A-H example expressed hatred toward someone who perhaps is non-Lagosian, asking them to return to their place of origin (*Anambra*) after referring to them as a *mad person* (*omo werey*). The second example is a wish for the recipient to be killed by *Ogun*[4], while the third example shows the recipient being wished failure just like Pitobi (Peter Obi[5]). However, the models failed to capture these tweets as *hateful*. See Table 13 for more examples.

6 Related Work

Several works have been conducted to create hate speech datasets, but the majority have focused on English and other high-resource languages, often within the context of specific countries (Mathew et al., 2021; Demus et al., 2022; Ron et al., 2023; Ayele et al., 2023a; Piot et al., 2024). However, in the context of Africa, only a few hate speech datasets exist to the best of our knowledge. For example, Ayele et al. (2023b) created a hate speech dataset for Amharic tweets using a hate and non-hate speech schema, while Aliyu et al. (2022) created a dataset for detecting hate speech against Fulani herders using hate/non-hate/indeterminate schema. These works, however, primarily focused on racial hate. In this work, we focused on election-related hate speech, which includes racial elements.

7 Conclusion

In this paper, we present **EkoHate** dataset for offensive and hate speech detection. Our dataset is code-switched and focused on political discussion in the last 2023 Lagos elections. We conducted empirical evaluations in fully supervised settings, covering both binary and multi-class tasks, finding multi-class to be more challenging. However, ensemble methods slightly improved multi-class performance. Additionally, cross-corpus experiments between EkoHate and existing datasets confirmed our annotations' alignment and our dataset's usefulness.

[4]Yoruba god of iron and war.

[5]Nigeria's LP presidential candidate in the 2023 elections.

Acknowledgments

Jesujoba Alabi was partially supported by the BMBF's (German Federal Ministry of Education and Research) SLIK project under the grant 01IS22015C. David Adelani acknowledges the support of DeepMind Academic Fellowship programme. Oluwaseyi Adeyemo acknowledges the support of the Founder, Afe Babalola University, Ado-Ekiti, Nigeria. Lastly, we thank Feyisayo Olalere, Nicholas Howell, the anonymous reviewers of AfricaNLP 2024 workshop and WOAH 2024 for their helpful feedback.

References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

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.

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023a. 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, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. 2023b. Exploring Amharic hate speech data collection and classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 49–59, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Matt Carlson. 2020. Fake news as an informational moral panic: the symbolic deviancy of social media during the 2016 us presidential election. *Information, Communication & Society*, 23(3):374–388.

Kevin Carney. 2022. The effect of social media on voters: experimental evidence from an indian election. *Job Market Paper*, pages 1–44.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022. Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Trisna Febriana and Arif Budiarto. 2019. Twitter dataset for hate speech and cyberbullying detection in indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)*, volume 1, pages 379–382.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Thomas Fujiwara, Karsten Müller, and Carlo Schwarz. 2021. The effect of social media on elections: Evidence from the united states. Working Paper 28849, National Bureau of Economic Research.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Agaptus Nwozor, Olanrewaju OP Ajakaiye, Onjefu Okidu, Alex Olanrewaju, and Oladiran Afolabi. 2022. Social media in politics: Interrogating electorate-driven hate speech in nigeria's 2019 presidential campaigns. *JeDEM-eJournal of eDemocracy and Open Government*, 14(1):104–129.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. 2024. Metahate: A dataset for unifying efforts on hate speech detection.

Arun Rajendran, Chiyu Zhang, and Muhammad Abdul-Mageed. 2019. UBC-NLP at SemEval-2019 task 6: Ensemble learning of offensive content with enhanced training data. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 775–781, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. 2023. Factoring hate speech: A new annotation framework to study hate speech in social media. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 215–220, Toronto, Canada. Association for Computational Linguistics.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@DravidianLangTech-EACL2021: Ensembling strategies for transformer-based offensive language detection. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 270–276, Kyiv. Association for Computational Linguistics.

Doris Dokua Sasu. 2023. Population of lagos, nigeria 2000-2035. *statista*.

Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 2021. Trumping hate on twitter? online hate speech in the 2016 u.s. election campaign and its aftermath. *Quarterly Journal of Political Science*, 16(1):71–104.

Kriti Singhal and Jatin Bedi. 2024. Transformers@LT-EDI-EACL2024: Caste and migration hate speech detection in Tamil using ensembling on transformers. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 249–253, St. Julian's, Malta. Association for Computational Linguistics.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Turgay Yerlikaya and Seca Toker. 2020. Social media and fake news in the post-truth era: The manipulation of politics in the election process. *Insight Turkey*, 22:177–196.

Fatima Zahrah, Jason R. C. Nurse, and Michael Goldsmith. 2022. A comparison of online hate on reddit and 4chan: a case study of the 2020 us election. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 1797–1800, New York, NY, USA. Association for Computing Machinery.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Steven Zimmerman, Udo Kruschwitz, and Chris Fox. 2018. Improving hate speech detection with deep learning ensembles. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A Performance using different pre-trained language models

We compared the performance of RoBERTa (Liu et al., 2019) (English PLM model), XLM-RoBERTa (Conneau et al., 2019) (multilingual PLM trained on 100 languages excluding Nigerian-Pidgin and Yoruba), Twitter-RoBERTa (Barbieri et al., 2020) (trained on English tweets) and AfroX-LMR (Alabi et al., 2022) (an African-centric PLM that cover English, Nigerian-Pidgin, and Yoruba in it's pre-training). Our results show that the English models have better performance than the multilingual variants, and the Twitter domain PLM have a similar performance as the RoBERTa model trained on the general domain. We have decided to use the Twitter domain-specific model for the remaining experiments.

B Languages in the test sets of EkoHate

EkoHate contains tweets in English, Yoruba, Naija, and their code-switched versions. While Figure 2 provides a plot comparing the distribution of these

Models	F1
RoBERTa-base (Liu et al., 2019)	$70.4_{\pm 1.2}$
XLM-RoBERTa-base (Conneau et al., 2019)	$66.5_{\pm 1.5}$
Twitter-RoBERTa-base (Barbieri et al., 2020)	$70.3_{\pm 1.1}$
AfroXLM-RoBERTa-base (Alabi et al., 2022)	$69.9_{\pm 1.0}$

Table 7: Comparing variants of RoBERTa on EkoHate N-A-H-C. We report the average Macro F1 after 5 runs.

languages in the whole dataset, Table 8 shows the distribution of these languages within the test split of each EkoHate schema. Yoruba and Naija have the smallest proportion in the test sets.

Data	Number of tweets English	Code-Switch	Naija	Yoruba
N-O	364	150	25	20
N-H	212	62	4	2
N-A-H	364	150	25	20
N-A-H-C	437	170	49	26

Table 8: Language distribution in the EkoHate test sets for English, Code-Switch, Naija, and Yoruba.

C Error analysis with confusion matrix

Tables 3 and 4 shows that the different models struggle with correctly classifying the hateful class. Hence, we examined the predictions of the baseline models in the multi-class setup by computing the confusion matrices for the N-A-H and N-A-H-C, as presented in Tables 9 and 10, respectively, comparing the counts of correct and incorrect predictions given the ground truth and the predictions.

		Prediction normal	abusive	hateful	Total
Gold	normal	173	5	7	185
	abusive	5	236	38	279
	hateful	8	38	49	95
	Total	186	279	94	559

Table 9: Confusion Matrix of one of the models trained and evaluated on EkoHate N-A-H.

Table 9 shows that the baseline model struggle with classifying between abusive and *hateful* tweets in the N-A-H setup, where 40% of *hateful* tweets were misclassified as *abusive*, while 13.5% of *abusive* tweets were predicted as *hateful*. With the inclusion of *contempt* in the label schema, as we have in N-A-H-C, Table 10 shows that more *abusive* tweets were classified as *contempt* than as hateful, with 12.9% and 7.5%, respectively. However,

		Prediction normal	abusive	hateful	contempt	Total
Gold	normal	166	4	2	13	185
	abusive	2	220	21	36	279
	hateful	5	35	42	13	95
	contempt	11	30	6	76	123
	Total	184	289	71	138	682

Table 10: Confusion Matrix of one of the models trained and evaluated on EkoHate N-A-H-C.

schema	F1
N-O	95.3
N-H	92.0
N-A-H	78.8
N-A-H-C	72.3

Table 11: Model ensembling results on EkoHate dataset.

36.8% of hateful tweets were misclassified as abusive, showing how difficult it is for the models to correctly classify hateful tweets which forms the smallest portion of EkoHate.

D Effect of model ensembling

Given the result of the baseline model, we investigate the use of model ensembling, which has been shown to improve model performance by leveraging the different strengths of various underlying models in class imbalance setups like ours. Therefore, instead of reporting the average F1 score, we opted to assess the impact of ensembling the 5 runs of the EkoHate baseline models. Table 11 shows a +0.6 improvement in the N-A-H and +2.3 improvement in the N-A-H-C scheme with ensembling, while binary schemes showed only marginal improvement, perhaps due to their initially good performance. We leave further analysis with model ensembling for future work.

E Annotation guidelines for EkoHate

Introduction This document presents guidelines on how to annotate potentially harmful tweets that can cause emotional distress to individuals, incite violence, or discriminate against, and exclude social groups.

As an annotator, it is important to approach this task with objectivity (as much as possible). We welcome your feedback on how we can update the guidelines based on the peculiarity of the language you are annotating, your background, or any sociolinguistic knowledge that we may have overlooked. Consider the following when performing the task:

Always use the guidelines and you should be objective and consistent in your annotation.

- Focus on the message conveyed in the tweets and try not to focus on your personal opinion on the topic.
- Do not rush to finish the task and always reach out to your language coordinator with questions when in doubt.

Mental health risk and well-being Annotating harmful content can be psychologically distressing. We advise any annotator who feels anxious or uncomfortable during the process to take a break or stop the task and seek help. Early intervention is the best way to cope.

Definitions

- **Hate speech** is language content that expresses hatred towards a particular **group or individual** based on their political affiliation, race, ethnicity, religion, gender, sexual orientation, or other characteristics. **It also includes threats of violence**.
- **Abusive language** is any form of bad language expressions including rude, impolite, insulting or belittling utterance intended to offend or harm an individual.
- **Contempt** is any form of language that **conveys a strong disliking of, or negative attitudes** towards a targeted individual or group, and does so in a **neutral tone** or form of expression.
- **Indeterminate** is any tweet that is not **readable** or is **completely** written in another language other than your language of annotation.
- **Normal** is any form of expression that does not contain any bad language belonging to any of the above classifications.

Task Given a tweet, select the option that best describes it. Table 12 show examples of tweets classified as hate, offensive, contempt, intermediate, and normal.

Label	Tweet
Hateful	We will kill the hoodlums disrupting this election process! it time to take law into our hands. Women belong to the kitchen and not in politics. We hate small boys, you are a small boy with no experience, you can't rule us. Leave that one to ur family members, nobody need ur bitter ass You are Igbo, you can't rule us in Lagos.
Abusive	You are very stupid! Olodo, oloriburuku U be mumu , see gbadego ur mumu never do abi eke nparo funro. Mumu your principal is using Eko o ni baje ...u r using Eko edide..oloshi ..Ori yi ti o pe ye ma pe laipe.
Contempt	Joker Dide Go Where Just dey play oooo U go school so? Vapour abi wetin be ur name?
Normal	I will vote for you. My Incoming Governor. Godbless you May his soul rest in peace
Indeterminate	Tweets that are completely written in languages other than English and Nigerian language of annotation. Tweets that make no sense or do not have any meaning

Table 12: Examples of tweets classified as hateful, abusive, contempt, intermediate, and normal.

schema	Tweet	Lang.	Gold	Pred.
N-A-H	Leave Lagos and return to Anambra omo werey	CDW	hateful	abusive
	Ogun kill you! By the time we're done with you, you'll tell us the real truth behind 20-10-2020. Murderer!	CDW	hateful	abusive
	There's bomb in your brain.	Eng.	hateful	abusive
N-A-H-C	Your tribunal case is being prepared. Enjoy the office while it lasts.	Eng.	hateful	contempt
	The actual election result is loading. Your and your boss will be retired.			
	The way pitobi failed you will also failed woefully	CDW	hateful	abusive
	Bro, go to the field and gather momentum. Social media can only do so much	Eng.	normal	contempt
	Thumb to the working Governor!	Eng.	normal	abusive

Table 13: Examples of correct and incorrect predictions.