# EKOHATE: OFFENSIVE 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 (LSV), 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 utilized 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 curate *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 and hate speech on political discussion. We develop 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 provide 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 two publicly available offensive datasets (OLID and HateUS2020) with at least 62.7 F1 points.

## 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 & 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; Ron et al., 2023; Piot et al., 2024).

However, the majority of these datasets were created for elections in the US (Suryawanshi et al., 2020; Grimminger & Klinger, 2021; Zahrah et al., 2022) and other non-African countries (Alfina et al., 2017; Febriana & 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 and verification, and even 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

[*] Equal contribution.

| 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, their labels under two labelling schemes. Where N is Normal, O is offensive (i.e. Abusive & Hateful), A is abusive, and C is contempt.

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 are 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 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) and HateUS2020 (Grimminger & Klinger, 2021) which achieved 71.1 F1 and 58.6 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 and 62.7 F1 on HateUS2020, 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. For reproducibility, we release our code, data and models on GitHub[1]

## 2 EKOHATE DATASET

### 2.1 LAGOS GUBERNATORIAL ELECTIONS

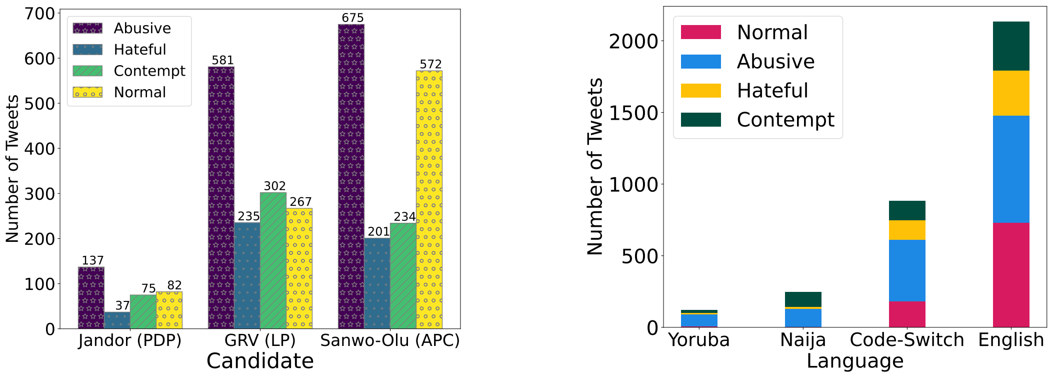
Lagos (also known as Èkó) is the commercial nerve centre of Nigeria, the former federal capital of Nigeria, and the most populous city in Nigeria and Africa with over 15 million residents according to Sasu (2023). In the 2023 Nigerian election, Lagos is probably the most strategic state because of its 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 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..).

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

**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.



(a) The distribution of the classes per candidate.

(b) The label distribution according to languages.

Figure 1: **EkoHate**: The dataset in Summary.

**EkoHate data statistics** Figure 1a 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. Across these candidates, tweets categorized as *abusive* accounted for 40% of the total, while *hateful* tweets were the least common category across the board.

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 1b 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).

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 two widely known datasets which are offensive language identification dataset (OLID) (Zampieri et al., 2019), and a corpus of offensive speech and stance detection from the 2020 US elections (HateUS2020) (Grimminger & Klinger, 2021). These are both datasets collected from Twitter and manually annotated. While

| | Number of tweets | | |
|---|---|---|---|
| **Data** | **train** | **dev** | **test** |
| ***Binary*** | | | |
| OLID (N-O) | 11916 | 1324 | 860 |
| HateUS2020 (N-H) | 2160 | 240 | 600 |
| EkoHate (N-O) | 1950 | 278 | 559 |
| EkoHate (N-H) | 976 | 139 | 280 |
| ***Multi class*** | | | |
| EkoHate (N-A-H) | 1950 | 278 | 559 |
| EkoHate (N-A-H-C) | 2377 | 339 | 682 |

Table 2: The split of the different datasets.

| schema | normal | offensive | abusive | hateful | contempt | F1 |
|---|---|---|---|---|---|---|
| N-O | $93.4_{\pm 0.4}$ | $96.8_{\pm 0.2}$ | - | - | - | $95.1_{\pm 0.3}$ |
| N-H | $94.6_{\pm 0.3}$ | - | | $89.2_{\pm 0.7}$ | - | $91.9_{\pm 0.5}$ |
| N-A-H | $93.4_{\pm 0.5}$ | - | $85.9_{\pm 1.3}$ | $55.4_{\pm 4.7}$ | - | $78.2_{\pm 2.2}$ |
| N-A-H-C | $90.5_{\pm 0.6}$ | - | $78.6_{\pm 0.8}$ | $51.1_{\pm 2.2}$ | $61.1_{\pm 1.7}$ | $70.3_{\pm 1.3}$ |

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

OLID used offensive and non-offensive schema, HateUS2020 used hateful and non-hateful schema. These datasets had no validation set, therefore, we sampled $10\%$ of their training splits as the dev set. See Table 2 for the splits and sizes of data.

**Models and Training** Using the respective datasets, we fine-tune Twitter-RoBERTa-base (Barbieri et al., 2020). [2] 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 Transformers framework (Wolf et al., 2020). We reported label-wise F1 score as well as the average F1 of 5 runs for the different models.

## 4 Results

**EkoHate baseline** We fine-tune 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 with high F1, however, multi-class configurations are difficult where classes are not predicted equally well. Lastly, we observed that in all settings, hateful class was the most difficult.

**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 corpora are evaluated on their corresponding test data, we obtained a high F1 score with the lowest being HateUS2020, where we obtained 77.8 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.

**Effect of model ensembling** Instead of reporting the average F1 score, we opted to assess the impact of ensembling the EkoHate baseline models. Table 5 shows a $+2.3$ improvement in the N-A-H-C scheme with ensembling, while other schemes showed only marginal improvement.

[2] 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

| dataset | normal | offensive | hateful | F1 |
|---|---|---|---|---|
| OLID | $88.3_{\pm 0.2}$ | $69.5_{\pm 1.0}$ | - | $78.9_{\pm 0.6}$ |
| → EkoHate | $69.2_{\pm 0.2}$ | $73.1_{\pm 0.4}$ | - | $71.1_{\pm 0.3}$ |
| EkoHate | $93.4_{\pm 0.4}$ | $96.8_{\pm 0.2}$ | - | $95.1_{\pm 0.3}$ |
| → OLID | $80.4_{\pm 0.7}$ | $63.2_{\pm 0.8}$ | - | $71.8_{\pm 0.7}$ |
| HateUS2020 | $95.2_{\pm 0.5}$ | - | $60.7_{\pm 2.5}$ | $77.8_{\pm 1.5}$ |
| → EkoHate | $83.1_{\pm 0.6}$ | - | $34.1_{\pm 4.7}$ | $58.6_{\pm 2.6}$ |
| EkoHate | $94.6_{\pm 0.3}$ | - | $89.2_{\pm 0.7}$ | $91.9_{\pm 0.5}$ |
| → HateUS2020 | $87.2_{\pm 1.2}$ | - | $38.3_{\pm 1.6}$ | $62.7_{\pm 1.4}$ |

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

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

Table 5: Result of model ensembling in different settings.

## 5 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 focus on racial hate. In this work, we focus on election-related hate speech, which includes racial elements.

## 6 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.

### 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.

## References

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.382.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pp. 233–238, 2017. doi: 10.1109/ICACSIS.2017.8355039.

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

Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. Multilingual racial hate speech detection using transfer learning. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 41–48, Varna, Bulgaria, September 2023a. INCOMA Ltd., Shoumen, Bulgaria. URL `https://aclanthology.org/2023.ranlp-1.5`.

Abinew Ali Ayele, Seid Muhie Yimam, Tadesse Destaw Belay, Tesfa Asfaw, and Chris Biemann. Exploring Amharic hate speech data collection and classification approaches. In Ruslan Mitkov and Galia Angelova (eds.), *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pp. 49–59, Varna, Bulgaria, September 2023b. INCOMA Ltd., Shoumen, Bulgaria. URL `https://aclanthology.org/2023.ranlp-1.6`.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.148. URL `https://aclanthology.org/2020.findings-emnlp.148`.

Matt Carlson. 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, 2020. doi: 10.1080/1369118X.2018.1505934. URL `https://doi.org/10.1080/1369118X.2018.1505934`.

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

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

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 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, pp. 512–515, 2017.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. Detox: A comprehensive dataset for German offensive language and conversation analysis. In Kanika Narang, Aida Mostafazadeh Davani, Lambert Mathias, Bertie Vidgen, and Zeerak Talat (eds.), *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pp. 143–153, Seattle, Washington (Hybrid), July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.woah-1.14. URL `https://aclanthology.org/2022.woah-1.14`.

Trisna Febriana and Arif Budiarto. Twitter dataset for hate speech and cyberbullying detection in indonesian language. In *2019 International Conference on Information Management and Technology (ICIMTech)*, volume 1, pp. 379–382, 2019. doi: 10.1109/ICIMTech.2019.8843722.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1), Jun. 2018. doi: 10.1609/icwsm.v12i1.14991. URL `https://ojs.aaai.org/index.php/ICWSM/article/view/14991`.

Thomas Fujiwara, Karsten Müller, and Carlo Schwarz. The effect of social media on elections: Evidence from the united states. Working Paper 28849, National Bureau of Economic Research, May 2021. URL `http://www.nber.org/papers/w28849`.

Lara Grimminger and Roman Klinger. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In Orphee De Clercq, Alexandra Balahur, Joao Sedoc, Valentin Barriere, Shabnam Tafreshi, Sven Buechel, and Veronique Hoste (eds.), *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pp. 171–180, Online, April 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wassa-1.18`.

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 u.s. presidential election. *Science*, 363(6425):374–378, 2019. doi: 10.1126/science.aau2706. URL `https://www.science.org/doi/abs/10.1126/science.aau2706`.

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

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, May 2021. doi: 10.1609/aaai.v35i17.17745. URL `https://ojs.aaai.org/index.php/AAAI/article/view/17745`.

Agaptus Nwozor, Olanrewaju OP Ajakaiye, Onjefu Okidu, Alex Olanrewaju, and Oladiran Afolabi. 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, 2022.

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

Gal Ron, Effi Levi, Odelia Oshri, and Shaul Shenhav. Factoring hate speech: A new annotation framework to study hate speech in social media. In Yi-ling Chung, Paul R\"ottger, Debora Nozza, Zeerak Talat, and Aida Mostafazadeh Davani (eds.), *The 7th Workshop on Online Abuse and Harms (WOAH)*, pp. 215–220, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.woah-1.21. URL `https://aclanthology.org/2023.woah-1.21`.

Doris Dokua Sasu. Population of lagos, nigeria 2000-2035. *statista*, Dec. 2023. URL `https://www.statista.com/statistics/1308467/population-of-lagos-nigeria/`.

Alexandra A. Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua A. Tucker. 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, 2021. ISSN 1554-0626. doi: 10.1561/100.00019045. URL `http://dx.doi.org/10.1561/100.00019045`.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In Ritesh Kumar, Atul Kr. Ojha, Bornini Lahiri, Marcos Zampieri, Shervin Malmasi, Vanessa Murdock, and Daniel Kadar (eds.), *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 32–41, Marseille, France, May 2020. European Language Resources Association (ELRA). ISBN 979-10-95546-56-6. URL `https://aclanthology.org/2020.trac-1.6`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

Turgay Yerlikaya and Seca Toker. Social media and fake news in the post-truth era: The manipulation of politics in the election process. *Insight Turkey*, 22:177–196, 2020. URL `https://api.semanticscholar.org/CorpusID:225728790`.

Fatima Zahrah, Jason R. C. Nurse, and Michael Goldsmith. 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, pp. 1797–1800, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387132. doi: 10.1145/3477314.3507226. URL https://doi.org/10.1145/3477314.3507226.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 1415–1420, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1144. URL https://aclanthology.org/N19-1144.

## 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.

| **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 6: Comparing variants of RoBERTa on EkoHate N-A-H-C. We report the average Macro F1 after 5 runs.