https://dx.doi.org/10.4314/dujopas.v10i4c.10
ISSN (Print): 2476-8316
ISSN (Online): 2635-3490
Dutse Journal of Pure and Applied Sciences (DUJOPAS), *Vol. 10 No. 4c December 2024*

VocalTweets: Investigating Social Media Offensive Language Among Nigerian Musicians

[1]Sunday Anthony Oluyele, [1]Juwon Akingbade, [1]Victor Akinode

[1]Department of Computer Engineering,
Federal University Oye Ekiti,
Oye,
Ekiti State,
Nigeria

Email: Sunday.oluyele.2826@fuoye.edu.ng

Abstract
Musicians frequently use social media to express their opinions, but they often convey different messages in their music compared to their posts online. Some utilize these platforms to abuse their colleagues, while others use it to show support for political candidates or engage in activism, as seen during the #EndSars protest. There is extensive research done on offensive language detection on social media, though, the usage of offensive language by musicians has received limited attention. In this study, we introduce VocalTweets, a code-switched and multilingual dataset comprising tweets from 12 prominent Nigerian musicians, labeled with a binary classification method as Normal or Offensive. We trained a model using HuggingFace's base-Twitter-RoBERTa, achieving an F1 score of 74.5. Additionally, we conducted cross-corpus experiments with the OLID dataset to evaluate the generalizability of our dataset.

Keywords: Offensive Language, Natural Language Processing, Multilingual, Social Media

INTRODUCTION
Social media platforms like Twitter (X), Facebook, and others have significantly influenced how individuals, including public figures, interact and communicate with their audience. These platforms have created opportunities for people worldwide to share their thoughts instantly. Public figures, such as musicians with large follower bases, play a crucial role in shaping public opinion and discourse. Their messages often reach a wide audience and influence public attitudes, particularly when engaging in social or political issues (Jin & Phua, 2014; Marshall, 2014). However, the increasing prevalence of offensive language on social media, especially on platforms like Twitter (X), poses a serious threat to healthy online interactions (Mutanga et al., 2020).

Offensive language is characterized by sharp, rude content, including vulgarity, insults, or attacks on individuals or groups (Touahri & Mazroui, 2022; Mubarak et al., 2020). It can have severe psychological impacts, contributing to stress, anxiety, depression, and even PTSD. The constant exposure to derogatory or harmful content online also leads to emotional desensitization and social isolation (Wachs et al., 2020). In regions with diverse linguistic and social structures, such as Africa, the challenge of detecting hate speech is compounded by the

Author for Correspondence

variety of languages and dialects, along with distinct cultural contexts. Researchers face difficulties in creating comprehensive datasets that account for evolving communication trends, including the use of emoticons, emojis, hashtags, slang, and linguistic contractions (Mody et al., 2023).

Offensive language detection traditionally relied on rule-based systems and classical machine learning techniques, such as keyword detection and lexicon-based approaches. These methods, while simple, struggled with contextual nuances, sarcasm, and evolving slang. For example, keyword-based systems would often flag specific derogatory terms, resulting in false positives in non-offensive contexts (Warner & Hirschberg, 2012). Lexicon-based methods, relying on predefined hate-related word dictionaries, were similarly ineffective, as they couldn't adapt to the diversity of languages on social media platforms (Davidson et al., 2017). With the advent of machine learning algorithms, such as Support Vector Machines (SVM) and Naive Bayes classifiers, there were improvements in accuracy. However, detecting indirect or implicit hate speech remained a challenge, given their reliance on manual feature extraction and lack of deep contextual understanding (Davidson et al., 2017).

The challenge of detecting offensive language becomes particularly pronounced in multilingual settings like Nigeria, where social media communication often involves code-switching. In these contexts, offensive language detection models face difficulties due to the blending of languages and evolving cultural expressions. African musicians, for instance, use multilingual and code-switched language in their online interactions, making it difficult to analyze using existing Natural Language Processing (NLP) models. This paper focuses on Nigerian musicians because they frequently engage in controversial topics and interpersonal disputes on social media platforms. Some, like Wizkid and Davido, have had public quarrels, while others use their platforms for political commentary and activism. The spread of offensive language online has increased significantly, which raises concerns about the social and political risks associated with it (Citron & Norton, 2011).

Previous studies have contributed valuable insights into hate speech detection on social media. For instance, Ilevbare et al. (2024) introduced EkoHate, a dataset focused on code-switched political discussions in Nigeria, highlighting the challenges of detecting hate speech in code-switched contexts. Yuan and Rizoiu (2024) proposed a Multi-Task Learning framework to improve cross-dataset generalization in hate speech detection, showing that learning generalized representations can enhance performance across diverse datasets. Similarly, Egode et al. (2023) explored deep learning methods like LSTM for detecting hate speech, revealing the potential of advanced models to handle class imbalance and limited data. Other works by Badjatiya et al. (2017) and Gambäck and Sikdar (2017) showed the effectiveness of deep learning models like CNN and LSTM in outperforming traditional methods like TF-IDF in detecting hate speech.

Despite the progress made in offensive language detection, there remain challenges in handling the complex and evolving nature of online language. Recent studies, such as those by Ibrahim et al. (2024), highlight the limitations of traditional rule-based systems in dealing with the contextual and dynamic nature of online speech. Their work shows the importance of integrating machine learning and NLP techniques to improve detection accuracy. In the context of Nigerian musicians, our study aims to bridge the gap by focusing on how offensive language is used in the social media interactions of these influential figures. This study contributes to the development of NLP resources for offensive language detection in African languages and addresses a critical gap in the research by focusing on Nigerian musicians. The

rest of this paper outlines the methodology, presents the results, and concludes with recommendations for future work.

METHODOLOGY

The VocalTweet dataset

VocalTweets (https://github.com/Tonycrux/VocalTweets) is a dataset comprising tweets from 12 prominent Nigerian musicians: Falz, Wizkid, Davido, Vector, Erigga, Teni, Tiwa Savage, Peter of P-Square, Buju, Burna Boy, Don Jazzy, and Tems. These artists are influential in the Nigerian musical scene and actively engage with their fan base through social media interactions via X (formerly Twitter). VocalTweets captures their online expressions, which goes beyond their music lyrics but provides a sneak peek into their daily lives, personal opinions, and social activism. Note that the figures presented after this section (Figures 1–6) were generated based on our own data analysis of the VocalTweets dataset.

Data Collection

The dataset comprises 4,677 tweets collected from the musicians' official X accounts. Figure 1 shows the distribution of tweets per musician; VocalTweets has 775 tweets from Davido, who has the highest number, followed by Don Jazzy with 661 tweets, Falz and Tiwa have the lowest number of tweets with 187 and 145, respectively.

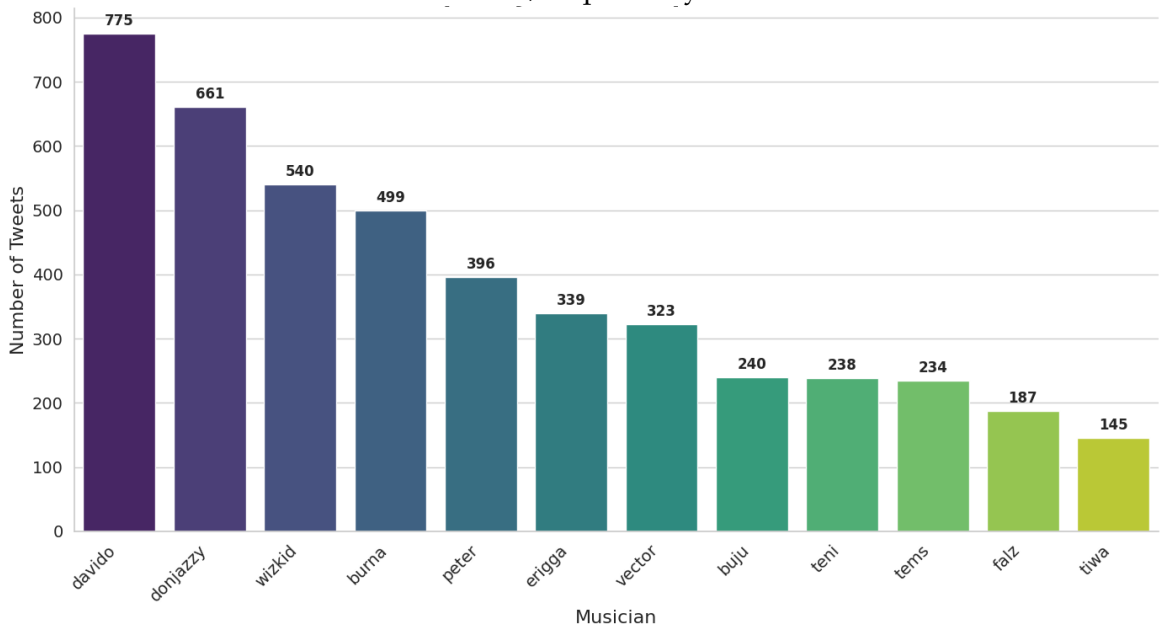

Figure 1: Frequency of tweets per musician

These tweets were manually collected by copying and pasting directly into a Google Sheet. To ensure the dataset's relevance and focus, we intentionally omitted promotional tweets—primarily advertising music releases and concerts—during the collection process.

Data Annotation

The annotation of the dataset was done using LabelStudio, which is a data labeling tool. We engaged two undergraduate annotators to categorize each tweet into one of the two classes:

- Normal (N): Tweets that do not contain any offensive language.
- Offensive (O): Tweets with offensive sentiments.

Instructions were given to the annotators to ensure unbiased classification to maintain objectivity and consistency. The inter-annotator agreement was measured using Fleiss'

Kappa, with a score of 0.76, indicating a good level of agreement between the annotators. In cases with inconsistencies, an objective resolution was resolved, and a final sentiment was allocated to such tweets. After annotation, we observe that many of the annotated tweets are in the Normal category, while others are offensive. Also, more of the tweets are in English, while others are in other languages. An overview of this is visualized in figure 2.

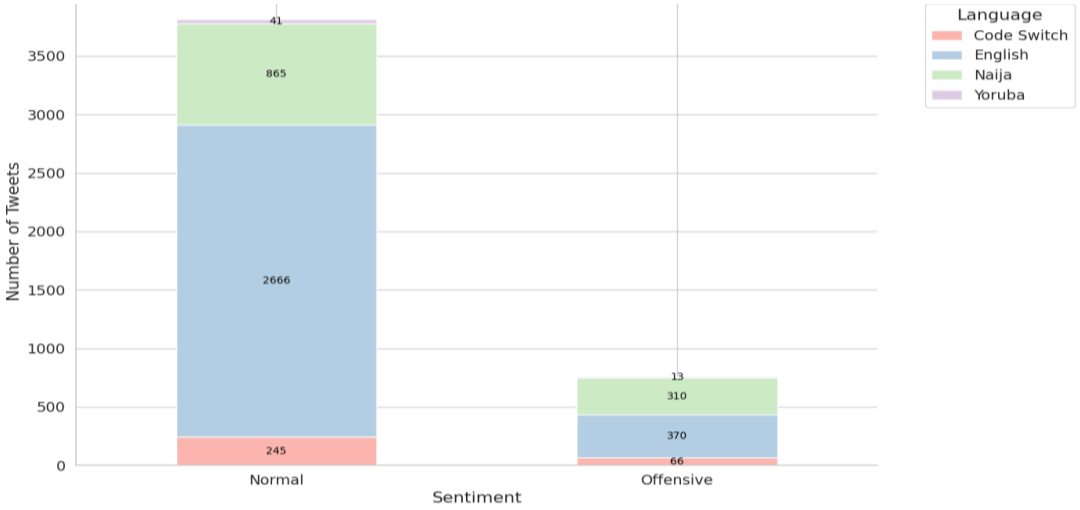

Figure 2: Sentiment and language distribution in VocalTweets

Class distribution

The dataset exhibits a notable imbalance in class distribution as there are 3817 (83.4%) normal tweets and 759 (16.6%) offensive tweets, as shown in Figure 3. A closer look at the dataset in figure 4 reveals that specific musicians exhibit higher tendencies of offensive language in their tweets; for instance, 43% of Vector's tweets are offensive; taking a closer look at these tweets, most of them are posted during the Nigerian #EndSARS protest - this was a mass mobilization against the police organization, seeking greater accountability in Nigeria (Ojedokun et al., 2021). In contrast, Don Jazzy has the highest number of normal tweets, with 96.8% of his tweets in the normal category, while only 3.2% are offensive.

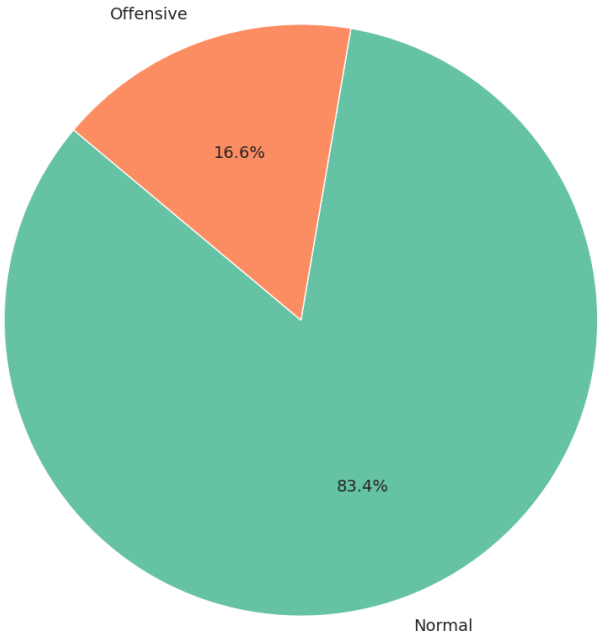

Figure 3: Overall sentiment distribution

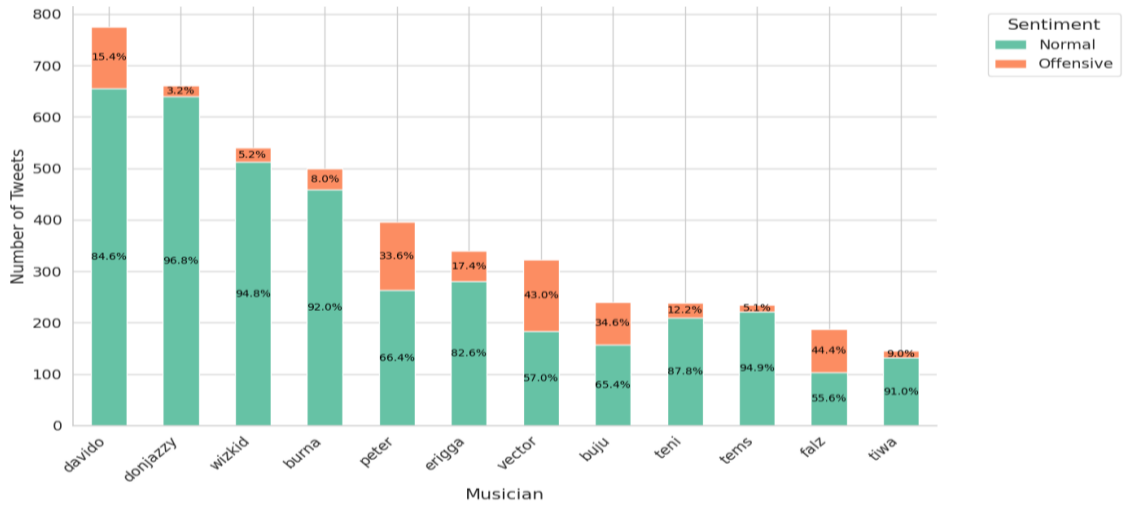

Figure 4: Sentiment distribution of tweets per musician

Language Distribution

VocalTweets is characterized by its Multilingual and code-switching nature, which reflects the diverse language usage in Nigeria. Figure 5 reveals that 3,036 (66.3%) tweets are in English, 1,175 (25.7%) are in Nigerian Pidgin (or Naija), 311 (6.8%) are code-switched, and only 54(1.2%) tweets are in Yoruba.

Looking more closely, we observed, as seen in Figure 6, that one of the musicians, named Erigga, predominantly tweets in Nigerian Pidgin, with approximately 60% of his tweets in the language, and Vector also has 42.9% of his tweets in Nigerian Pidgin. Davido has the highest usage of Code-Switching with 12.4% of his tweets switches between Yoruba and English. An example of such a tweet is: "*Orie ti daru that yo bitch ass was crying screaming bout they intimidating ur family u a real bitch for life!!!*". Teni has the highest usage of Yoruba with about 5% of her tweets in Yoruba.

The dataset possesses noisy characteristics, a common characteristic of social media communications. This is compounded by the fact that these musicians tweet with alphanumeric characters and many emojis.

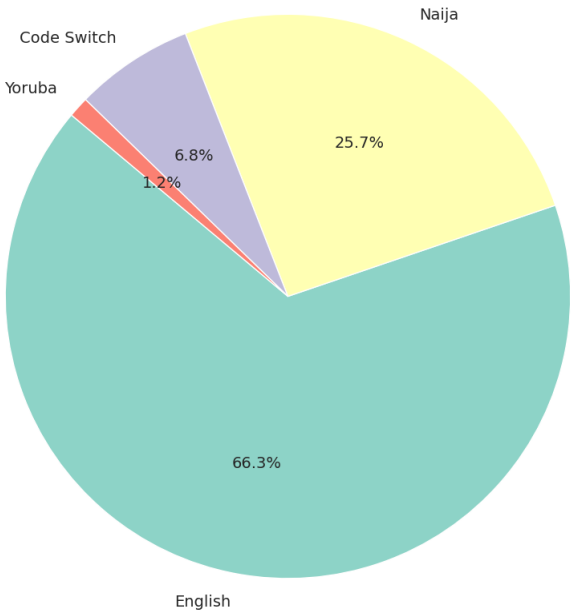

Figure 5: Overall language distribution

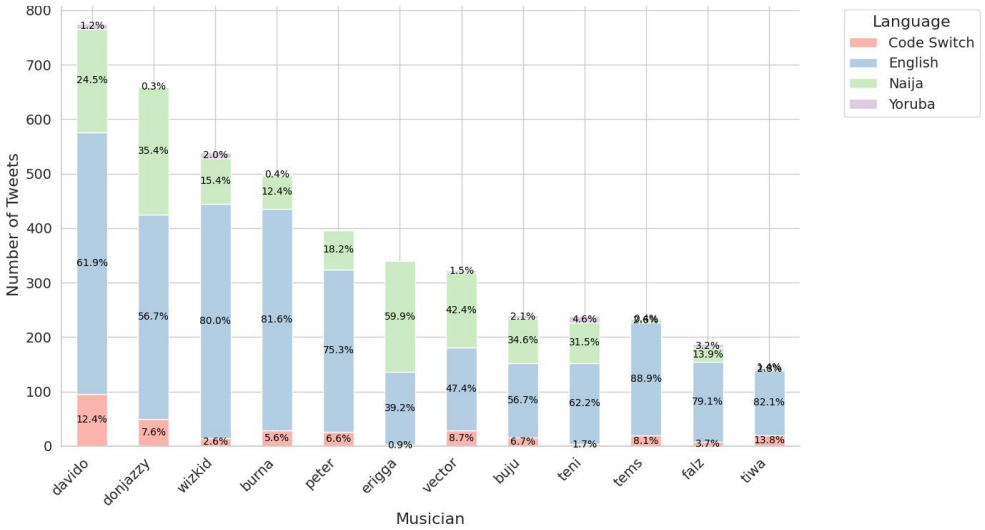

Figure 6: Language distribution of tweets per musician

Data Splitting

To facilitate effective training and evaluation in the creation of the model, the dataset was split into training and validation subsets using an 80-20 ratio, with 80%(approximately 3,742 tweets) of the data being the training set and 20% (approximately 934 tweets) of the data being the validation set. This split is necessary to ensure that models trained on the VocalTweets dataset can be evaluated for performance and generalization.

Experimental Design

This work uses a binary classification framework to categorize tweets into two classes: Normal and Offensive. The objective is to create a model that can accurately detect non-offensive and offensive language used on social media by Nigerian musicians.

To evaluate the generalizability of our dataset, we conducted cross-corpus transfer experiments using VocalTweets and the Offensive Language Identification Dataset (OLID) (Zimmerman et al, 2018), which comprises annotated tweets categorized into normal and offensive. We adapted OLID to fit our binary classification framework and conducted cross-corpus transfer experiments on both datasets. For each classification task, in-domain (VocalTweets on VocalTweets and OLID on OLID) and cross-domain (VocalTweets on OLID and vice versa), we conducted five independent runs with different random seeds to ensure the reliability of our results.

Model and Training

We used the Twitter-roberta-base model - a transformer-based architecture optimized for Twitter data by HuggingFace. This model, based on RoBERTa, an improved version of BERT, has been pre-trained on large-scale Twitter data, making it well-suited for handling the informal, noisy, and often code-switched language found on social media platforms like Twitter (X). The models are trained under the following configurations:

- **Learning Rate (2e-5)**: A low learning rate was selected to ensure stable training, allowing the model to learn progressively without overshooting optimal solutions.
- **Batch Size (16)**: 16 was chosen to balance memory constraints and training efficiency.
- **Epochs (10)**: Training was performed over 10 epochs to allow the model sufficient time to adapt to our dataset.
- **Weight Decay (0.01)**: This regularization technique helps prevent overfitting by penalizing large weights.

- **Logging Steps (50)**: Training progress was logged every 50 steps to track performance and ensure convergence.
- **Metric for Best Model (eval_Macro F1)**: The Macro F1 score was chosen as the primary evaluation metric, ensuring balanced performance across both the Normal and Offensive classes, especially given the class imbalance in the dataset.

Evaluation Metrics

To assess the model's performance, we reported both label-wise F1 scores and the overall Macro F1 Score:

- **Label-wise F1 Scores**: These measure the model's accuracy in predicting each class (Normal and Offensive).
- **Macro F1 Score**: This aggregates the F1 scores of both classes, giving equal weight to both the majority and minority classes, and offering a fair assessment of the model's performance in class-imbalanced datasets.

Results and Discussion

We fine-tuned the Twitter-RoBERTa-base model on our VocalTweets dataset for binary classification, categorizing tweets as Normal or Offensive. Due to the imbalanced nature of the dataset, we observed relatively low performance in the Offensive class, which is the least occurring category. Models trained on imbalanced datasets often struggle to detect minority classes, leading to lower F1 scores for the Offensive class. In our dataset, only 17% of the tweets are labeled as Offensive, leading to a naturally skewed distribution. To address this, future work could explore techniques such as oversampling, undersampling, or using class weights. We trained the model across five runs, and the label-wise F1 scores are summarized in Table 1. Note that the tables presented after this section (Tables 1–6) were populated based on the information and training results of the VocalTweets dataset.

Table 1: Label-wise F1 score after training

Metric	F1 Score
Normal F1	91.2 ± 0.4
Offensive F1	57.8 ± 3.5
Macro F1	74.5 ± 1.9

In addition to the overall performance, we examined the model's performance across different language categories. Table 2 below presents the F1 scores for each language:

Table 2: Model performance across the language categories

Language	Normal F1	Offensive F1	Macro F1
English	94.1 ± 0.3	59.7 ± 0.8	76.9 ± 0.5
Code-Switched	93.8 ± 0.8	76.1 ± 3.5	85.0 ± 2.1
Naija	85.7 ± 1.9	63.9 ± 1.3	74.8 ± 1.4
Yoruba	85.5 ± 9.2	67.3 ± 24.1	76.4 ± 16.6

The results indicate that the model performs best for the Code-Switched language class, with an Offensive F1 score of 85.0 ± 2.1. This suggests that the model performs well when handling tweets containing a mix of languages, which is common in social media discourse, especially among Nigerian users. The high performance in the Code-Switched category may be attributed to the model's ability to leverage the diversity of linguistic features present in mixed-language contexts. This finding aligns with previous work by Ilevbare et al. (2024), where their work shows a high F1-Score for the Code-Switched class. Also other classes followed closely after the Code-Switched class, with minor differences while English, on the other hand, has a lower Offensive F1 score due to the fact that our dataset has higher Offensive representations in English for Offensive sentiments, this we suggest leads to a lower overall Macro F1 score of 76.9 for English. Table 3 below shows the language distribution of the test set.

Table 3: Language distribution in the test set

Language	Count	Percentage (%)
English	585	63.9
Naija	249	27.2
Code-Switched	71	7.8
Yoruba	11	1.2

We did a cross-corpus experiment, by training the Twitter-RoBERTa-base model on the OLID dataset and evaluated its performance on both the VocalTweets and OLID datasets. The results of this zero-shot transfer experiment are shown in Table 4. As expected, the model performed well when trained and evaluated on the same dataset, yielding high F1 scores on the OLID test set. However, when we tested the model across different corpora, the performance dropped significantly. For instance, transferring from OLID to VocalTweets resulted in an F1 score of 69.2. This performance gap can be attributed to several factors, including differences in class distribution between the two datasets. The OLID dataset has a much higher proportion of Offensive tweets (approximately 50%), while the VocalTweets dataset only contains 17% Offensive tweets. This class imbalance likely led to a lower F1 score for the Offensive class, as the model was trained on a dataset with a greater representation of Offensive content. To address these issues, further work could focus on augmenting the VocalTweets dataset with more Offensive tweets to balance the class distribution.

Table 4: Cross-corpus experimental result across each dataset

Experiment	**Normal F1**	**Offensive F1**	**Macro F1**
OLID → OLID	89.5 ± 1.1	79.6 ± 2.6	84.5 ± 1.9
VocalTweets → OLID	72.5 ± 3.7	61.1 ± 0.9	66.8 ± 2.0
VocalTweets → VocalTweets	97.2 ± 1.5	87.8 ± 6.5	92.5 ± 4.0
OLID → VocalTweets	91.2 ± 0.1	47.2 ± 2.3	69.2 ± 1.2

We further evaluated the model's performance by performing random classifications for 10 samples of the tweets, as shown in Table 5. The results of this classification reveal that the model struggles a bit with accurately predicting Offensive tweets. Out of the three Offensive tweets in the sample, the model correctly classified two, while misclassifying one Normal

tweet as Offensive. This shows the model's tendency to incorrectly label non-offensive tweets as offensive, an issue often observed in imbalanced datasets. We also generated a confusion matrix in Table 6, the model is confident in classifying Normal tweets, with 697 correct predictions (true positives). However, there are 51 instances where Normal tweets are misclassified as Offensive (false positives). This suggests that the model may occasionally identify certain neutral expressions as offensive. For the Offensive Class, the model correctly classifies 102 Offensive tweets (true positives), it struggles with 66 instances of Offensive tweets being misclassified as Normal (false negatives), this indicates that the model has difficulty detecting Offensive language. As mentioned earlier, this issue can be addressed by augmenting the Offensive class with more diverse examples, particularly those in Naija, Yoruba and Code-Switched. Techniques such as synonym replacement or back translation could also help introduce more variation in the Offensive tweets.

Table 5: 10 Random classifications performed by the model

	Tweet	Language	True Label	Predicted Label
0	😄 😄 it's the best accent ever!	English	Normal	Normal
1	Let the youth embarrass the elders who don't respect themselves.	English	Offensive	Offensive
2	U don dey post cold, If them send transport now U go use am fill gas Thief!	Naija	Offensive	Offensive
3	Otilo	Yoruba	Normal	Normal
4	Live on @fox5ny now tune in!!	English	Normal	Normal
5	I lied to you sorry	English	Normal	Normal
6	That's why they must protest with us.	English	Normal	Normal
7	Congratulations on your win	English	Normal	Normal
8	Try make money so you fit comot from relationship when you no like	Naija	Normal	Offensive
9	I never enter gear but shift Abeg. If your soap guy no get Drip, switch shop hehe	English	Offensive	Normal
10	Hmmm 😧	English	Normal	Normal

Table 6: Confusion matrix

	Normal	**Offensive**
Normal	697	51
Offensive	66	102

Conclusion

We present the VocalTweets dataset for offensive language detection, comprising tweets from 12 famous Nigerian musicians. This dataset is characterized by its code-switched and multilingual nature. We conduct an empirical evaluation using the binary classification method to classify tweets as Normal or Offensive; this results in a Macro F1-Score of 74.5, which shows that the dataset faces a challenge with the Offensive class. To assess the generalization ability of our model, we conducted cross-corpus experiments between the VocalTweets and OLID datasets, which indicates that our dataset falls slightly short when compared to OLID due to its imbalance nature, the difference is minimal, and it suggests that despite the nature of the dataset we are still able to achieve some level of accuracy compared to OLID which is well balanced. VocalTweets is a valuable resource for advancing research in offensive language detection, particularly for low-resource languages like Yoruba and Nigerian Pidgin.

Future work could focus on augmenting the VocalTweets dataset with additional Offensive tweets to ensure more balanced classes between Normal and Offensive courses to improve the model's performance and robustness. Additionally, cross-corpus experiments could be expanded beyond the OLID dataset to provide more understanding of the model's generalization ability. The dataset can also be expanded to include more Nigerian musicians who are not included in this study and other famous African musicians; this will help create a more diverse dataset that extends beyond Nigerian languages and also other African languages. The dataset can still be expanded to include tweets from famous influencers and public personalities to get a sense of offensive language usage by popular people that influences people's opinions on social media.

References

Badjatiya, P., Gupta, S., Gupta, M. and Varma, V., (2017). Deep Learning for Hate Speech Detection in Tweets. In Proceedings of the 26th International Conference on World Wide Web Companion. Perth, Australia: International World Wide Web Conferences Steering Committee, pp.759-760. https://doi.org/10.48550/arXiv.1706.00188 .

Citron, D.K. and Norton, H., 2011. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, 91(4), pp.1435-1484.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), pp.512–515. doi:https://doi.org/10.1609/icwsm.v11i1.14955.

Egode, K.O., Oraegbunam, L., Oyatunji, A.S. and Akwue, O.S., (2023). Hate Speech Detection in Twitter: Natural Language Processing Exploration. *Global Advanced Research Journal of Educational Research and Review,* 11(8), pp.325-336.

Gambäck, B. and Sikdar, U.K., (2017). Using Convolutional Neural Networks to Classify Hate-Speech. *In Proceedings of the First Workshop on Abusive Language Online. Vancouver, BC,*

Canada: Association for Computational Linguistics, pp.85-90. doi:https://doi.org/10.18653/v1/W17-3013 .

Ibrahim, U., Guma, U.L., Lawal, I.A. (2024). Social sensing with big data: Detecting hate speech in social media. *International Journal of Science and Research Archive*, 11(2), pp.1146–1152. https://doi.org/10.30574/ijsra.2024.11.2.0540 .

Ilevbare, C.E., Alabi, J.O., Adelani, D.I., Bakare, F.D., Abiola, O.B. and Adeyemo, O.A. (2024). EkoHate: Abusive Language and Hate Speech Detection for Code-switched Political Discussions on Nigerian Twitter. doi:https://doi.org/10.18653/v1/2024.woah-1.3 .

Hamdy Mubarak, Kareem Darwish, Walid Magdy, Tamer Elsayed, and Hend Al-Khalifa. (2020). Overview of OSACT4 Arabic Offensive Language Detection Shared Task. In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, pp. 48–52, Marseille, France. European Language Resource Association.

Jahan, M.S. and Oussalah, M., (2023). A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546, pp.126232. https://doi.org/10.1016/j.neucom.2023.126232 .

Jin, S.-A.A. and Phua, J. (2014). Following Celebrities' Tweets about Brands: the Impact of Twitter-Based Electronic Word-of-Mouth on Consumers' Source Credibility Perception, Buying Intention, and Social Identification with Celebrities. *Journal of Advertising*, 43(2), pp.181–195. https://doi.org/10.1080/00913367.2013.827606 .

Mody, D., Huang, Y. and Alves de Oliveira, T.E. (2023). A curated dataset for hate speech detection on social media text. *Data in Brief*, 46, p.108832. doi:https://doi.org/10.1016/j.dib.2022.108832 .

Mutanga, R.T., Naicker, N. and Olugbara, O.O. (2020). Hate Speech Detection in Twitter using Transformer Methods. *International Journal of Advanced Computer Science and Applications*, 11(9). doi:https://doi.org/10.14569/ijacsa.2020.0110972 .

P David Marshall (2014). *Celebrity and power : fame in contemporary culture ; with a new introduction*. Minneapolis: University Of Minnesota Press.

Touahri, I., Mazroui, A. (2022). Offensive Language and Hate Speech Detection Based on Transfer Learning. In: Kacprzyk, J., Balas, V.E., Ezziyyani, M. (eds) Advanced Intelligent Systems for Sustainable Development (AI2SD'2020). AI2SD 2020. Advances in Intelligent Systems and Computing, vol 1418. Springer, Cham. https://doi.org/10.1007/978-3-030-90639-9_24

Warner, W. and Hirschberg, J. (2012) . Detecting hate speech on the World Wide Web , *in Proceedings of the Second Workshop on Language in Social Media,* Montréal, Canada, pp. 19–26. Association for Computational Linguistics.

Yuan, L. and Marian-Andrei Rizoiu (2024). Generalizing Hate Speech Detection Using Multi-Task Learning: A Case Study of Political Public Figures. *Computer Speech & Language*, 89, pp.101690–101690. https://doi.org/10.1016/j.csl.2024.101690 .

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)