BEYOND ENGLISH: OFFENSIVE LANGUAGE DETECTION IN LOW-RESOURCE NIGERIAN LANGUAGES

Saminu Mohammad Aliyu
Department of Computer Science
Bayero University Kano
smaliyu.cs@buk.edu.ng

Gregory Maksha Wajiga & Murtala Muhammad
Department of Computer Science
Modibbo Adama University, Yola
Adamawa, Nigeria
{gwajiga,mmurtala1875}@mautech.edu.ng

Lukman Jibril Aliyu
HausaNLP Research Community
Kaduna state
lukman.j.aliyu@gmail.com

ABSTRACT

The proliferation of online offensive language necessitates the development of effective detection mechanisms, especially in multilingual contexts. This study addresses the challenge by developing and introducing novel datasets for hate speech detection in three major Nigerian languages: Hausa, Yoruba, and Igbo. We collected data from Twitter and manually annotated it to create datasets for each of the three languages, using native speakers. We used pre-trained language models to evaluate their efficacy in detecting offensive language in our datasets. The best-performing model achieved an accuracy of 90%. To further support research in offensive language detection, we plan to make the dataset and our models publicly available.

1 INTRODUCTION

Social networking sites has become one of the most powerful media for sourcing information, expressing opinions and feelings as well as posting of multimedia contents. People communicate freely and anonymously with one another through these platforms irrespective of geographical locations (Alsafari et al., 2020). However, the sharing of hateful content and online harassment is becoming a menace in these online communities. Therefore, these social networks sites developed some regulations that prohibit the posting of offensive and hateful contents, but the lack of clear distinction between freedom of speech and offensive or hate speech has rendered these regulations not very effective (Chetty & Alathur, 2018; Alkiviadou, 2019). Furthermore, these platforms employ some automatic and semi-automatic approaches using Artificial Intelligence to send warnings to users against posting offensive and hateful contents, in addition to detecting and removing offensive and hateful posts and comments Schneider & Rizoiu (2023).

The huge amount of data generated from these social networking sites has made the traditional method of manually identifying and removing offensive and hateful contents almost impractical. Hence, automatic approaches using Natural Language Processing (NLP) and using Artificial Intelligence are employed to develop models that can be used not only to detect but remove offensive and hateful contents from user posts and comments in social media.

According to Husain & Uzuner (2021), what constitutes offensive language depends on the contextual meaning and the intention of the author as well as the society. Offensive language has been defined as any language that belittles, attacks, disparages, mocks, or insults and individual or a group of people (Díaz-Torres et al., 2020; Fieri & Suhartono, 2023). Offensive language can be expressed as hate speech, cyberbully, sexism, abusive language and many other forms with hate speech as one of the most research alongside offensive language (Caselli et al., 2020; Davidson et al., 2017; Dorris et al., 2020).

Hate speech has no universally accepted definition. Different authors have given similar opinions on what constitutes hate speech (MacAvaney et al., 2019). Among the most common definition of hate speech is that it is any form of communication, verbal or written that attacks an individuals or groups based on characteristics like race, religion, gender, ethnicity, nationality, disability, political affiliations, and more (Aliyu et al., 2022; Patil et al., 2023; Fortuna & Nunes, 2018; Alkomah & Ma, 2022). Hate speech has been reported to negatively affect the victim's psychology and has translated into physical crises in some cases (Saha et al., 2019; Bilewicz & Soral, 2020).

Nigeria is a multi-cultural country located in West Africa. There are more than 522 languages are spoken in the country with Hausa, Yoruba and Igbo languages as the most dominants. Hausa language is predominantly spoken in the North, Yoruba in the West and Igbo in the Eastern part of the country (Orekan, 2010; Burns, 2023). The official language of the country is English, but these three languages alongside the Nigerian Pidgin English have dominated conversions especially, on the social networking sites. The country has recorded a number of communal, tribal and religious crises which are believed to be fueled through the spread of hate speech on social media (Pate & Ibrahim, 2020). Twitter is one of the most used social media networking site by Nigerians. According to a 2023 report by Statista[1], there are over 4.9 million twitter users in Nigeria. Most Nigerian communicate on twitter in native languages. These communications are full of hateful and offensive comments which are detrimental to victim's health and also can lead to physical confrontations.

The existing studies on automatic offensive and hate speech detection are mostly in high resource languages (Davidson et al., 2017; Mollas et al., 2022; Mathew et al., 2021; De Gibert et al., 2018) with little studies covering low resource languages especially, African languages (Demilie & Salau, 2022). To the best of our knowledge, there is no study on offensive and hate speech detection in the three major Nigerian languages. Consequently, we intend to create a novel dataset that can help in the automatic detection of hateful and offensive contents in tweets that are written in Hausa, Yoruba and Igbo languages.

The main contributions of the study are:

- We created the first manually annotated data for hate and offensive speech detection in Hausa, Yoruba and Igbo languages.
- We conducted a baseline experiment for the detection of hate and offensive language in Hausa, Yoruba and Igbo social media text.

2 Literature review

The exponential growth of user generated data on social media has rendered the manual approach of content moderation ineffective. Hateful and offensive contents sharing are in the rise on social media partly because of the lack of clear definition of what constitute hate or offense and the user anonymity. These Social Network Sites (SNS) like Facebook, YouTube and X (Twitter) have drawn a line between offensive speech and freedom of speech as well as using different approaches to detect and remove offensive and hateful contents. However, these measures by the SNS are not adequate, especially for low resource languages. Consequently, there are remarkable number of studies in the academia that have proposed solutions for the automatic detection of hate in social media contents as contained in (Poletto et al., 2021; Fortuna & Nunes, 2018). Most of these research are on high resource languages with English taking the lead (Swamy et al., 2019; Waseem & Hovy, 2016; Davidson et al., 2017). Recently, there has been a significant rise in the research on offensive and hate speech detection in low resource languages like Arabic (Husain & Uzuner, 2021), Indonesia (Ibrohim & Budi, 2018) and India (Bohra et al., 2018). Some authors treated the problem as a binary classification task Risch et al. (2020); Pelicon et al. (2021); Mozafari et al. (2022), multi-class Djandji et al. (2020); Plaza-del Arco et al. (2022) and multi-label Ibrohim & Budi (2019); Omar et al. (2021); Azzi & Zribi (2023). In terms of approach, many have experimented with classical machine learning algorithms (Pitenis et al., 2020; HaCohen-Kerner & Uzan, 2021; De Souza & Da Costa-Abreu, 2020; Swain et al., 2022), deep learning models (Wei et al., 2021; Roy et al., 2022; Mahibha et al., 2021) and the state-of-the-art transformer models (Molero et al., 2023; Elmadany et al., 2020; Ranasinghe & Zampieri, 2021; Subramanian et al., 2022).

[1]https://www.statista.com/statistics/1325514/number-of-potential-twitter-advertising-audience-in-nigeria/

Ali et al. (2021) used a combination of keywords and lexicons to collect tweets in Urdu. The tweets were pre-processed and a final corpus of 16,000 tweets was obtained. They used Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) to classify the tweets as either offensive or not-offensive. Essefar et al. (2023) used machine learning and deep learning algorithms to classify social media comments written in Moroccon Arabic Dialect as offensive or not. They observed that emojis are mostly used to express offensiveness. In a related study, Pookpanich & Siribor-vornratanakul (2024) explored the performance of five different transformer model in the task of detecting offensive language in Thai sports comments. The authors observed that the models performances are almost similar with XLM-ROBERTa outperforming the rest. Mazari & Kheddar (2023) developed a dataset of 14150 Algerian Arabic comments from various online social media platforms. The explored word2vec and FastText embedding with some classical and deep learning models to detect offensive, hateful and cyberbullying comments and achieved the best performance with an average of over 75% F1-score. Aliyu et al. (2022) created a dataset of English, Hausa and Nigerian-Pidgin language for the detection of hate speech against the Fulani herdsmen in Nigeria. The dataset comprises of about 6000 manually annotated tweets. They used mBERT, XLM-T and AfriBERTa pretrained models to create baseline models and XLM-T performed best with and f1-score of 99.3%. A related study was conducted by Ndabula et al. (2023), where they collected code-mixed tweets in English, Hausa, Yoruba, Igbo and pidgin language posted during the EndSARS protest and 2023 general elections. They used bags-of-words and tf-idf features with Support Vector Machine (SVM) and Random Forest (RF) to classify the tweets as positive (hateful) or negative (not-hateful). The SVM obtained the best f1 score of 93.57%.

3 Methodology

This section discusses the method of tweets collection, annotation, and exploratory data analyses.

3.1 Data source

Different Authors have used various data sources for data collection. A research by Jahan & Oussalah (2023) has shown that Twitter is the most used source of data for offensive and hate speech detection task. This has been attributed to the huge amount of public data available from Twitter and its free access. In Nigeria, Twitter has become one of the most used social media platforms were people of different culture, religion and political affiliations interacts. Hence, our choice of Twitter as the data source.

3.2 Search strategy

We employed the keyword approach to search for tweets in the three languages. Offensive and hate keywords were collected through crowd-sourcing and validated by language experts. Queries were developed using the keywords and the twitter academic API was used to crawl 20,000 tweets per language. Keywords approach are used to increase the chances of collecting offensive and hateful tweets (Warner & Hirschberg, 2012; Davidson et al., 2017). All the collected tweets were in the original languages, that is, we collected tweets written in Hausa, Yoruba and Igbo language.

3.3 Pre-processing and annotation

The tweets were pre-processed by removing duplicates and tweets that are written in other languages or unintelligible. We replaced all mentions of usernames with *"@USER"*, emails with *"@email"* and urls with *"@URL"*. Three native speakers per language were employed and trained for the annotation task. We drafted an annotation guideline similar to that of Sigurbergsson & Derczynski (2019) with some modifications. The annotation task consists of two different levels:

Level 1: Tweet category At this level, each tweet is labeled as offensive, hate, indeterminate or normal. A tweet is offensive if it contains any form of bad language against an individual or group. A tweet is labeled as hate if it offensive and based on characteristics like religion, race, etc. A tweet is labeled indeterminate if it is completely in a different language or unintelligible. A tweet is labeled normal if it intelligible and no use of any bad language.

Level 2: Hate Target identification Tweets labeled as hateful are further annotated to identify the target of the hate. These targets include : religion, ethnicity, gender, disability, Politics, others. If an annotator selects the ”others” category, he/she will be prompted with an input box to write down the category.

After the first round of the training, the annotators were given a set of 100 tweets each to annotate and the Inter-annotator agreement (IAA) was computed using Fleiss' kappa (Fleiss, 1971). We accepted an IAA of 60% and above. Where the IAA score is less than our threshold, the annotators were re-trained and given another set of training sample to annotated. This process was repeated until the annotators score an IAA above 60

A quick analysis of the manually annotated samples shows that most of the tweets were labeled as normal across all the languages. We therefore, conducted another pre-annotation selection were we sampled some potentially harmful tweets before the main annotation. This was done to increase the possibility of having more offensive and hate classes. Our final datasets after dropping the tweets labeled as 'Indeterminate' contained a total of 6476, 4926 and 2974 tweets for Hausa, Yoruba and Igbo languages respectively. Each of these datasets was experimented independently. Table 1 shows the distribution of the datasets classes per language.

3.4 METHODS

3.4.1 DATA PREPARATION AND TRAIN/TEST SPLIT

In our study, we focused on analyzing tweet datasets in three major Nigerian languages: Hausa, Igbo, and Yoruba. We adopted a systematic approach to manage these language-specific datasets. A custom dataset class was developed, tailoring to the unique text characteristics of each language. This class handled specific preprocessing requirements, such as normalization of text and handling of unique language constructs, ensuring efficient feature extraction and embedding.

For each language dataset, we implemented a train/test split of 80/20. This means that 80% of the tweets were used for training our models, while the remaining 20% formed the test sets. This split ensured a comprehensive evaluation of the models' performance on unseen data, reflecting their real-world applicability.

3.4.2 MODELS AND FEATURE EXTRACTION

We trained four distinct models, each obtained from the Huggingface model repository, Huggingface model repository:known for its extensive collection of advanced NLP models. The models used were:

1. XLM-Roberta-Base: Served as a baseline for comparison. It provided a broad understanding of multilingual context.
2. BERT-Base-Multilingual-Cased: Chosen for its enhanced language context capabilities, offering a more nuanced understanding of multilingual nuances.
3. Morit/XLM-T-Full-XNLI: Selected for its expanded language context, having been trained for hate speech detection outside our target languages.
4. Davlan/Naija-Twitter-Sentiment-Afriberta-Large: Originally trained on Nigerian Twitter sentiment, we fine-tuned this model to focus specifically on offensive speech classification in our target languages.

Each model, primarily encoder-based and akin to BERT architecture, underwent a fine-tuning process on our specific datasets. This involved adapting the pre-existing knowledge of these pretrained models to the linguistic contexts of Hausa, Igbo, and Yoruba tweets.

The feature extraction process was significantly enhanced by the use of AutoModel and AutoTokenizer classes from Huggingface. AutoModel dynamically adapted to each chosen model's architecture, while AutoTokenizer ensured accurate and consistent tokenization and encoding of the multilingual text data. This was especially crucial given the linguistic peculiarities of our target languages.

Table 1: Datasets label distribution

LABEL	HAUSA	YORUBA	IGBO
Hate	75	221	231
Normal	4008	2221	715
Offensive	2384	2484	2028

Table 2: Models results with accuracy scores

MODEL	HAUSA	YORUBA	IGBO
XLM-RoBERTa-base (Conneau et al., 2019)	0.79	0.82	0.69
Bert-based-multilingual-cased (Devlin et al., 2018)	0.83	0.83	0.87
morit/XLM-T-full-xnli (Barbieri et al., 2021)	0.81	0.85	0.90
Davlan/Naija-Twitter-Sentiment-Afriberta-Large (Muhammad et al., 2022)	0.85	0.85	0.88

3.4.3 Evaluation and Analysis

The performance of each model was evaluated on the separate test sets for Hausa, Igbo, and Yoruba tweets. The primary metric for evaluation was model accuracy, which provided crucial insights into each model's effectiveness in accurately classifying language-specific tweets. This comprehensive evaluation allowed us to ascertain the relative strengths and areas for improvement in our multilingual classification approach. Table 2 shows the models and accuracy obtained on the tests sets. The the morit/XLM-T-full-xnli model achieved the highest result of 0.85 and 0.90 for Yoruba and Igbo language. The Devlan/Naija-Twitter-Sentiment-Afriberta-Large acheived a competitive results with 0.85 for Hausa, Yoruba and 0.88 for Igbo language. On average, this model gives the best accuracy across all the three dataset. This may be because our data source is also twitter. The other two models also achieved a reasonable accuracy scores. These shows the adaptability of these models in detecting offensive comments in African language.

4 Conclusions and Future Work

This paper presents datasets for offensive language detection in the three major Nigerian languages. We used a crowd-sourcing approach to collect keywords which were used to collect tweets in these languages. We developed guidelines that were used to manually annotated these data into offensive, hate, normal and indeterminate. The indeterminate class was drop and the final datasets contain three class. Using pre-trained language models, we developed baselines for each of the language evaluated their performances using accuracy scores. Some of these models achieved a very good results on all the three languages while others perform better on one language only. These has shown the significance of taking into account linguistics diversity in creating and evaluating multilingual models. As future work, we intend to use annotate more comments from YouTube and Instagram to have larger datasets and also to detect the categories and targets of offensive and hateful tweet

Acknowledgments

We express our profound gratitude to Data Science Africa (DSA) for their significant financial assistance during the course of this research. Their dedication to promoting data science, especially in Africa has greatly enhanced the achievements of our endeavours. We look forward to further collaborate with DSA to continue to explore more solutions to our local problems.

References

Muhammad Z Ali, Sahar Rauf, Kashif Javed, Sarmad Hussain, et al. Improving hate speech detection of urdu tweets using sentiment analysis. *IEEE Access*, 9:84296–84305, 2021.

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. *arXiv preprint arXiv:2211.15262*, 2022.

Natalie Alkiviadou. Hate speech on social media networks: towards a regulatory framework? *Information & Communications Technology Law*, 28(1):19–35, 2019.

Fatimah Alkomah and Xiaogang Ma. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273, 2022.

Safa Alsafari, Samira Sadaoui, and Malek Mouhoub. Hate and offensive speech detection on arabic social media. *Online Social Networks and Media*, 19:100096, 2020.

Salma Abid Azzi and Chiraz Ben Othmane Zribi. A new classifier chain method of bert models for multi-label classification of arabic abusive language on social media. *Procedia Computer Science*, 225:476–485, 2023.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. *arXiv preprint arXiv:2104.12250*, 2021.

Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pp. 36–41, 2018.

Alan Burns. *History of Nigeria*, volume 30. Taylor & Francis, 2023.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*, 2020.

Naganna Chetty and Sreejith Alathur. Hate speech review in the context of online social networks. *Aggression and violent behavior*, 40:108–118, 2018.

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. *arXiv preprint arXiv:1911.02116*, 2019.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pp. 512–515, 2017.

Ona De Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*, 2018.

Gabriel Araújo De Souza and Márjory Da Costa-Abreu. Automatic offensive language detection from twitter data using machine learning and feature selection of metadata. In *2020 international joint conference on neural networks (IJCNN)*, pp. 1–6. IEEE, 2020.

Wubetu Barud Demilie and Ayodeji Olalekan Salau. Detection of fake news and hate speech for ethiopian languages: a systematic review of the approaches. *Journal of big Data*, 9(1):66, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

María José Díaz-Torres, Paulina Alejandra Morán-Méndez, Luis Villasenor-Pineda, Manuel Montes, Juan Aguilera, and Luis Meneses-Lerín. Automatic detection of offensive language in social media: Defining linguistic criteria to build a mexican spanish dataset. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 132–136, 2020.

Marc Djandji, Fady Baly, Wissam Antoun, and Hazem Hajj. Multi-task learning using arabert for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp. 97–101, 2020.

Wyatt Dorris, Ruijia Hu, Nishant Vishwamitra, Feng Luo, and Matthew Costello. Towards automatic detection and explanation of hate speech and offensive language. In *Proceedings of the sixth international workshop on security and privacy analytics*, pp. 23–29, 2020.

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. Leveraging affective bidirectional transformers for offensive language detection. *arXiv preprint arXiv:2006.01266*, 2020.

Kabil Essefar, Hassan Ait Baha, Abdelkader El Mahdaouy, Abdellah El Mekki, and Ismail Berrada. Omcd: Offensive moroccan comments dataset. *Language Resources and Evaluation*, pp. 1–21, 2023.

Brillian Fieri and Derwin Suhartono. Offensive language detection using soft voting ensemble model. In *MENDEL*, volume 29, pp. 1–6, 2023.

Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

Yaakov HaCohen-Kerner and Moshe Uzan. Detecting offensive language in english, hindi, and marathi using classical supervised machine learning methods and word/char n-grams. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*, 2021.

Fatemah Husain and Ozlem Uzuner. A survey of offensive language detection for the arabic language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 20(1):1–44, 2021.

Muhammad Okky Ibrohim and Indra Budi. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229, 2018.

Muhammad Okky Ibrohim and Indra Budi. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the third workshop on abusive language online*, pp. 46–57, 2019.

Md Saroar Jahan and Mourad Oussalah. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, pp. 126232, 2023.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152, 2019.

C Jerin Mahibha, Sampath Kayalvizhi, Durairaj Thenmozhi, and Sundar Arunima. Offensive language identification using machine learning and deep learning techniques. 2021.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Ahmed Cherif Mazari and Hamza Kheddar. Deep learning-based analysis of algerian dialect dataset targeted hate speech, offensive language and cyberbullying. *International Journal of Computing and Digital Systems*, 2023.

José María Molero, Jorge Pérez-Martín, Alvaro Rodrigo, and Anselmo Peñas. Offensive language detection in spanish social media: Testing from bag-of-words to transformers models. *IEEE Access*, 2023.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896, 2022.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Anuoluwapo Aremu, Saheed Abdul, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *arXiv preprint arXiv:2201.08277*, 2022.

Joseph Nda Ndabula, Oyenike Mary Olanrewaju, and Faith O Echobu. Detection of hate speech code mix involving english and other nigerian languages. *Journal of Information Systems and Informatics*, 5(4):1416–1431, 2023.

Ahmed Omar, Tarek M Mahmoud, Tarek Abd-El-Hafeez, and Ahmed Mahfouz. Multi-label arabic text classification in online social networks. *Information Systems*, 100:101785, 2021.

George Orekan. Language policy and educational development in africa: The case of nigeria. *Scottish Languages Review*, 21:17–26, 2010.

Umaru A Pate and Adamkolo Mohammed Ibrahim. Fake news, hate speech and nigeria's struggle for democratic consolidation: A conceptual review. *Handbook of research on politics in the computer age*, pp. 89–112, 2020.

Prachi Patil, Sakshi Raul, Dhanisha Raut, and Tatwadashi Nagarhalli. Hate speech detection using deep learning and text analysis. In *2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pp. 322–330. IEEE, 2023.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559, 2021.

Zeses Pitenis, Marcos Zampieri, and Tharindu Ranasinghe. Offensive language identification in greek. *arXiv preprint arXiv:2003.07459*, 2020.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and María-Teresa Martín-Valdivia. Integrating implicit and explicit linguistic phenomena via multi-task learning for offensive language detection. *Knowledge-Based Systems*, 258:109965, 2022.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523, 2021.

Peerat Pookpanich and Thitirat Siriborvornratanakul. Offensive language and hate speech detection using deep learning in football news live streaming chat on youtube in thailand. *Social Network Analysis and Mining*, 14(1):18, 2024.

Tharindu Ranasinghe and Marcos Zampieri. An evaluation of multilingual offensive language identification methods for the languages of india. *Information*, 12(8):306, 2021.

Julian Risch, Robin Ruff, and Ralf Krestel. Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pp. 137–143, 2020.

Pradeep Kumar Roy, Snehaan Bhawal, and Chinnaudayar Navaneethakrishnan Subalalitha. Hate speech and offensive language detection in dravidian languages using deep ensemble framework. *Computer Speech & Language*, 75:101386, 2022.

Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, pp. 255–264, 2019.

Philipp J Schneider and Marian-Andrei Rizoiu. The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences*, 120(34):e2307360120, 2023.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. Offensive language and hate speech detection for danish. *arXiv preprint arXiv:1908.04531*, 2019.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. Offensive language detection in tamil youtube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404, 2022.

Manswini Swain, Manish Biswal, Priya Raj, Abhinav Kumar, and Debahuti Mishra. Hate and offensive language identification from social media: A machine learning approach. In *Electronic Systems and Intelligent Computing: Proceedings of ESIC 2021*, pp. 335–342. Springer, 2022.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*, pp. 940–950, 2019.

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

Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pp. 88–93, 2016.

Bencheng Wei, Jason Li, Ajay Gupta, Hafiza Umair, Atsu Vovor, and Natalie Durzynski. Offensive language and hate speech detection with deep learning and transfer learning. *arXiv preprint arXiv:2108.03305*, 2021.