Detection and Analysis of Offensive Online Content in Hausa Language

Fatima Muhammad Adam[1*], **Abubakar Yakubu Zandam**[2], **Isa Inuwa-Dutse**[3]
[1]Federal University Dutse, Nigeria
[2] Federal University of Technology Babura, Nigeria
[3]University of Huddersfield, UK
fatima.m@fud.edu.ng, ayzandam95@gmail.com,i.inuwa-dutse@hud.ac.uk

Abstract

Hausa, a major Chadic language spoken by over 100 million people mostly in West Africa is considered a low-resource language from a computational linguistic perspective. This classification indicates a scarcity of linguistic resources and tools necessary for handling various natural language processing (NLP) tasks, including the detection of offensive content. To address this gap, we conducted two set of studies (1) a user study ($n = 101$) to explore cyberbullying in Hausa and (2) an empirical study that led to the creation of the first dataset of offensive terms in the Hausa language. We developed detection systems trained on this dataset and compared their performance against relevant multilingual models, including Google Translate. Our detection system successfully identified over 70% of offensive, whereas baseline models frequently mistranslated such terms. We attribute this discrepancy to the nuanced nature of the Hausa language and the reliance of baseline models on direct or literal translation due to limited data to build purposive detection systems. These findings highlight the importance of incorporating cultural context and linguistic nuances when developing NLP models for low-resource languages such as Hausa. A post hoc analysis further revealed that offensive language is particularly prevalent in discussions related to religion and politics. To foster a safer online environment, we recommend involving diverse stakeholders with expertise in local contexts and demographics. Their insights will be crucial in developing more accurate detection systems and targeted moderation strategies that align with cultural sensitivities.

Trigger Warning: Readers may find some of the terms in this study distressing or disturbing; all examples are for illustration only.

1 Introduction

The modern world is highly interconnected, with millions of people and devices linked through the Internet. This interconnectivity has facilitated research and communication, driving transformative changes across various domains [Kennedy, 2006; Warf, 2018]. One significant outcome of this connectivity is the rise of online social networks, which enable user interactions and generate vast datasets for research purposes [Miller, 2010; Kumar *et al.*, 2014; Haffner, 2018]. Social media platforms such as Facebook[1] and Twitter (now X[2]) allow users to stay connected, express opinions, organize civil movements, and more [Kumar *et al.*, 2014; Rane and Salem, 2012; Olteanu *et al.*, 2015; Freelon *et al.*, 2016; Puspitasari, 2017; Nemes and Kiss, 2021; Dambo *et al.*, 2022; Bello *et al.*, 2023]. These platforms have a profound impact on politics, governance, the economy, business, and other societal issues [Dollarhide, 2021].

However, online participation often involves the use of offensive, racist, and sexist language, contributing to the growing problem of cyberbullying [Caselli *et al.*, 2020b; Caselli *et al.*, 2020a]. Cyberbullying has severe social and psychological consequences, including an increased risk of suicide, particularly among young people [Bully Facts, 2023; Magee *et al.*, 2013]. In response to this issue, various strategies have been proposed to detect, mitigate, and combat cyberbullying [Zhong *et al.*, 2016; Altay and Alatas, 2018; Zois *et al.*, 2018; Rafiq *et al.*, 2018; Yao *et al.*, 2019; Amin *et al.*, 2020; Plaza-Del-Arco *et al.*, 2021]. However, these detection methods and resources are primarily designed for high-resource languages, rendering them ineffective for low-resource languages such as Hausa.

This study aims to support efforts to combat the proliferation of offensive content in low-resource languages, with a specific focus on Hausa. To achieve this, we outline the following objectives:

- Conduct a comprehensive review of existing research on offensive content detection in low-resource languages, with an emphasis on Hausa.
- Engage native speakers through a user study to develop a robust dataset of offensive content in the Hausa language.

*Contact Author

[1]https://facebook.com
[2]https://twitter.com

- Develop an effective detection system tailored to identifying offensive content in Hausa.
- Provide actionable recommendations and strategies to mitigate the spread of offensive content in Hausa.

Beyond these core objectives, our study will also (1) investigate the prevalence of offensive language in Hausa and differentiate between banter and more harmful offensive terms (2) examine the interplay between idiomatic expressions with subtle abusive undertones, and (3) identify the topical issues most likely to attract offensive content. We believe this multifaceted approach will provide a comprehensive understanding of offensive content in Hausa, leading to more effective detection and intervention strategies. As one of the first studies on the detection of offensive online content in the Hausa language, our contributions include:

- First annotated dataset on offensive content in Hausa: We introduce the first collection of annotated datasets[3] on offensive content in Hausa (HOC), enriching resources for future research and the development of online safety tools.
- Hausa offensive content detection model: We develop a robust detection model specifically designed to identify offensive language in Hausa online discourse.
- Insights into Hausa cyberbullying: Through user studies, we provide insights into the prevalence and nature of cyberbullying in Hausa, along with mitigation strategies to foster more civil online interactions.

The remainder of this paper is structured as follows. Section 2 reviews relevant literature and background studies. Section 3 describes our approach, including data collection methods. Section 4 presents the implementation details, while Section 5 discusses the results and key findings. Finally, Section 6 concludes the study and outlines future research directions.

2 Background and Related Work

In this section, we review relevant studies on hate speech, abusive content, and downstream tasks in the Hausa language.

2.1 Hate Speech Detection

According to the General Policy Recommendation of the European Commission against Racism and Intolerance (ECRI), hate speech encompasses any form of advocacy, promotion, or incitement, in any form, of disparagement, abhorrence, criticism, harassment, negative labelling, stigmatization, threats, insults, or other harmful expressions directed at individuals or groups [ECRI Policy, 2023]. Offensive language is closely related to various linguistic and societal issues, including abusive and violent tone, cyberbullying, racism, extremism, radicalization, toxicity, profanity, flaming, discrimination, and hateful speech [Caselli *et al.*, 2020b; Founta *et al.*, 2018]. With the massive increase in social media interactions, offensive content and other forms of cyberbullying have also risen. In May 2016, the EU Commission reached an agreement with major technology companies[4] on a Code of Conduct to counter illegal hate speech online [EU Code, 2023]. These measures have proven most effective in high-resource languages such as English and French. However, detecting cyberbullying-related content in low-resource languages like Hausa remains challenging due to the lack of annotated datasets and relevant linguistic resources. This study aims to address these challenges.

Several studies based on tweets from X have been employed to detect hateful and abusive content across various social issues [Sanguinetti *et al.*, 2018; Basile *et al.*, 2019; Mathew *et al.*, 2021]. For instance, tweets related to hate speech against immigrants [Sanguinetti *et al.*, 2018] and misogynistic content in Spanish and English [Basile *et al.*, 2019] have been utilized to develop multilingual detection systems. Additionally, memes have been incorporated into multimodal classification systems to enhance hate speech detection by leveraging contextual understanding [Kiela *et al.*, 2020]. A benchmark dataset, HateXplain, consisting of hate, offensive, and neutral data points for detecting hate speech, has been introduced by [Mathew *et al.*, 2021]. Inherent racial biases in abusive language detection systems remain a significant concern, as discrepancies in accent and writing style can lead to discriminatory outcomes. Studies have highlighted that such biases disproportionately affect African-American users, undermining detection efforts intended to protect communities that are frequently targeted [Davidson *et al.*, 2019; Davidson *et al.*, 2019].

2.2 Offensive Content Detection

The Oxford English Dictionary defines offensive content as *highly unpleasant and insulting; characterized by persistent violence and cruelty* [Simpson. and Weiner, 2019]. As noted earlier, cyberbullying manifests in various forms, necessitating diverse detection strategies [Zhong *et al.*, 2016; Al-Garadi *et al.*, 2016; Altay and Alatas, 2018; Van Hee *et al.*, 2018; Zois *et al.*, 2018; Rafiq *et al.*, 2018; Yao *et al.*, 2019]. For instance, [Zhong *et al.*, 2016] focused on identifying forms of cyberbullying associated with image data. Another study applied a model that categorizes posts written by bullies, victims, and bystanders to detect online bullying [Van Hee *et al.*, 2018]. Other notable detection strategies emphasize two main aspects: scalability and timeliness of detection systems. Timeliness is a crucial factor in the detection pipeline, ensuring efficient support for cyberbullying victims. A multistage cyberbullying detection system incorporating a scheduling mechanism has been proposed to enhance detection efficiency [Rafiq *et al.*, 2018]. Likewise, a sequential hypothesis testing approach has been suggested as a solution to scalability and timeliness challenges [Zois *et al.*, 2018]. [Yao *et al.*, 2019] explored ways to reduce false positives while improving scalability and timeliness, using data from Instagram. Given that offensive content detection and sentiment analysis are closely related [Schmidt and Wiegand, 2017;

[3] https://github.com/FatimaMuhammadAdam/Detection-and-Analysis-of-Offensive-Online-Content-in-Hausa-Language-Dataset/blob/main/README.md

[4] Facebook, Microsoft, Twitter, and YouTube

Oriola and Kotzé, 2020], sentiment analysis techniques have been employed to identify offensive and inappropriate online content [Plaza-Del-Arco *et al.*, 2021]. Notably, the studies mentioned above primarily focus on cyberbullying detection in high-resource languages. However, there is growing interest in addressing offensive content and hate speech in low-resource languages such as Tamil [Rajalakshmi *et al.*, 2023; Balakrishnan *et al.*, 2023], Pashto [Khan *et al.*, 2023], Urdu [Saeed *et al.*, 2023], and Persian [Kebriaei *et al.*, 2023]. Efforts have also been made to enhance resources for tackling offensive and hateful content detection in these languages [Kovács *et al.*, 2022; Sinyangwe *et al.*, 2023; Miao *et al.*, 2023; Markov *et al.*, 2023].

2.3 Downstream Tasks in the Hausa Language

Datasets from online social media platforms have been leveraged for various computational linguistic tasks, including sentiment analysis [Zishumba, 2019; Sani *et al.*,], topic modelling [Kusumawardani and Basri, 2017; Antypas *et al.*, 2022], sexist language detection [Aliyu *et al.*, 2023], hate speech recognition [Machuve *et al.*, 2022], text classification [Chaure *et al.*, 2019], and named entity recognition [Nie *et al.*, 2020]. Previous studies have also compiled relevant corpora for NLP-related tasks in the Hausa language [Suleiman *et al.*, 2019; Oyewusi *et al.*, 2021; Abubakar *et al.*, 2021; Inuwa-Dutse, 2021; Ibrahim *et al.*, 2022; Muhammad *et al.*, 2022; Zandam *et al.*, 2023]. For instance, a large collection of tweets in Hausa, Igbo, Yoruba, and Nigerian Pidgin has been compiled to improve sentiment lexicons in low-resource languages [Abubakar *et al.*, 2021]. Additionally, a dataset of multilingual tweets annotated with sentiment labels in major Nigerian languages has been developed [Muhammad *et al.*, 2022]. Despite the growing interest in low-resource languages, studies focused on offensive content detection in Hausa remain scarce. Furthermore, many languages lack sufficient linguistic resources for NLP-related tasks [Tsvetkov, 2017]. For Hausa, this scarcity is primarily due to the absence of annotated datasets and lexicons. This study proposes an approach to detect offensive online content in Hausa, one of the most widely spoken low-resource languages, thereby contributing to the broader field of computational linguistics and online safety.

3 Methodology

To achieve our primary objective, we adopt the following strategy (1) engage native Hausa speakers through a user study (2) collect and annotate data comprising offensive terms in the Hausa language (3) develop and evaluate a detection system for offensive content in Hausa. Figure 1 provides an overview of our approach.

3.1 User Study

We conducted two user studies to gain insights into public perceptions of offensive and abusive online content. The study was administered via QuestionPro[5], with 180 volunteers participating. The survey link was distributed on Facebook and X (formerly Twitter) to reach social media users. Informed consent was obtained from all participants, and no personally identifiable information was collected. The study was conducted in accordance with institutional ethical standards. Table 1 presents the demographic details of the participants. The study focused on (1) the role of the Hausa language in information sharing on social media (2) the use and impact of social media as a medium for disseminating offensive and abusive content (3) public perceptions regarding the distinction between abusive terms—whether used offensively or otherwise, and (4) the different meanings and contextual interpretations of offensive and abusive content in Hausa.

3.2 Dataset

Recognising that LRLs like Hausa lack sufficient data for many NLP tasks, we collected posts from Twitter (now X) and Facebook that were likely to contain offensive or abusive content. Table 2 provides examples of the data sources. For data collection, we employed both an automated approach—using X's API—and a manual process for Facebook data. Algorithm 1 outlines the steps taken to search and retrieve relevant posts via the API.

Algorithm 1 *Text Collection*: Using keywords to collect relevant data for the study

Initialisation: $S = [S_i]_{i=1}^{n}$, list of search keywords, n : number of search keywords, M : number of search iterations, N : number of required tweets
for each iteration $m = 1$ to M **do**
    for each query in S above $i = 1$ to n **do** search tweets data with keyword S_i from $D = [D_j]_{j=1}^{k}$ where D is a list of tweets
        if S_i matches D_j **then** // get relevant fields
            get posting date
            get tweet ID
            get retweet count
            get favourite count
            get full text
            get screen-name
            get urls
        else
            continue
        end if
        append D_j to T
        check for N, if satisfies maximum break
    end for
end for
create a dataframe based on T
Output: a dataframe consisting of a list of tweets $[T_t]_{t=1}^{N}$

To maximise the chances of obtaining offensive or abusive content, we focused on topics such as politics, sports, banter, relationships, abusive language, news reports, trending discussions, and accounts known for controversial engagements. A subset of the keywords and accounts used is listed in Table 2. To further refine our dataset, we applied the following strategies for defining offensive content in Hausa:

- *Kamus*[6] (Dictionary) definition: We aligned our dataset with the standard definition of abusive language (*zagi*) as provided in *Kamus*. Additionally, we labelled posts

[5]https://www.questionpro.com/

[6]Hausa dictionary https://kamus.com.ng/

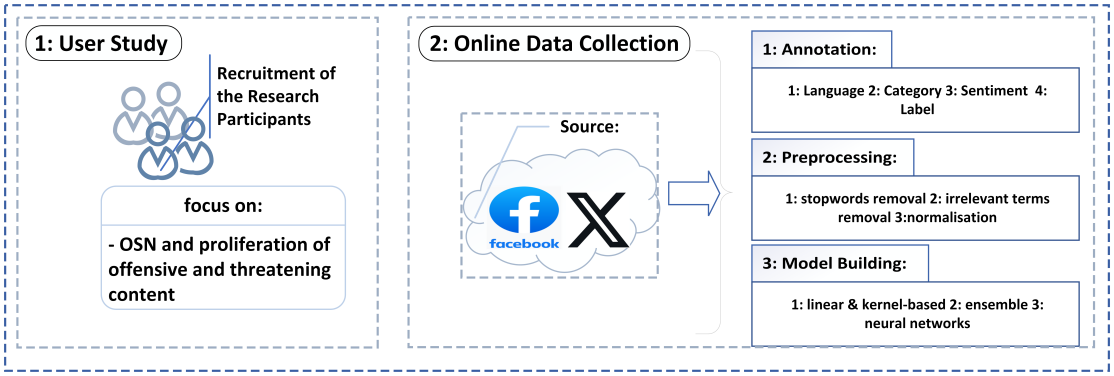

Figure 1: Overview of our approach, which involves (1) conducting user studies and (2) collecting and curating datasets to build detection systems for offensive content.

User study ($n = 180$) **involving offensive content (HOC)**			
Gender	**Age**	**Education**	**Category**
Female **25.3%**	18-25 yrs. **28%**	Sec. Edu. **0%**	Student **46%**
Male **74.7%**	26-35 yrs. **63%**	Undergraduate. **39%**	Self-employed **23%**
	36-45 yrs. **9%**	Postgraduate **46%**	Civil Servant **23%**
		Graduate **15%**	Politician **1%**
			others **19%**

Table 1: Demographic information about the research participants.

based on contextual use of offensive terms, considering anecdotal evidence.

- Past Data and User Study: We leveraged previous datasets [Inuwa-Dutse, 2021] to extract abusive words, informing our collection and annotation approach. We also incorporated offensive examples identified by participants in our user study.

Data Cleaning

The collected dataset, especially from X, contained noise in the form of duplicates, retweets, hyperlinks, emoticons, numbers, punctuation, non-Hausa posts, and other irrelevant symbols. To ensure high-quality data for analysis, we applied the following cleaning strategies:

- **Stopwords removal:** We filtered out function words that add little meaning to texts. In addition to standard stopwords, we defined a custom list for Hausa, including terms like *'a', 'ni', 'to', 'su'* and non-alphabetic characters such as *'.', '?', '/'*.

  Removal of irrelevant terms: We eliminated URLs, HTML tags, numbers, punctuation, and special symbols using regular expressions. Hashtags (#) and mentions (@) were also removed. To filter out non-informative tweets, we calculated the number of unique words per post and discarded those with fewer than eight distinct words.

- **Duplicate removal and normalisation:** Duplicate posts were identified and removed to prevent bias. We normalized text by converting accented vowels, replacing emojis, and correcting misspellings using a predefined Hausa dictionary. Additionally, all tokens were converted to lowercase, and stemming was applied to standardise words.

Data Annotation

To build an effective detection system, we manually annotated the dataset, assigning the following labels to each post:

- *language:* to denote the language of the post with a focus on posts in Hausa language or *Engausa*[7], see Figure 2(a) for distribution.

- *sentiment:* to indicate whether a post is positive, neutral, or negative (see Figure 2(b) for a summary).

- *category:* to denote the theme to which the post belongs (e.g. social, political, religious, security, education, law, sport, health, agriculture, security, and business). Figure 2(c) shows the distribution of topics with the most common posts.

- *offensive:* a column to indicate whether a post is offensive or not; abusive words are considered offensive in this regard.

We utilise a subset of the annotated data and incorporated into a user study, wherein feedback from online participants was used to validate the annotations, especially in instances where abusive content appeared in both playful and non-playful contexts. We compare the survey responses with the annotations to ensure consistency and reliability. While formal inter-rater agreement techniques are valuable for ensuring data quality, we found that incorporating survey-based

[7] a mix of Hausa and English terms

Source	**X (formerly Twitter)**
Topics	politics, strike action, and social events
Keywords	siyasa (politics), ASUU strike, Maulud -Nabiyi , sallah celebration, love relationships, and zagi
Source	**Facebook**
Topics	politics, sport, banter, and news reports
Keywords	Dandalin Siyasar Jigawa, Hakika mabudin faira, legit.ng hausa, Ayi raha asha dariya, Alfijir Hausa, Rariya, Northern hibiscus, Dokin karfe TV

Table 2: Relevant keywords and sources used for data collection.

ratings offered a practical means of validation under the current circumstances. We acknowledge that exposing annotators to large amounts of offensive or sensitive material raises ethical concerns. Consequently, our present ethical approval does not extend to engaging annotators with extensive offensive content without additional measures, such as mental health support and clear handling guidelines in place. Future work will integrate both formal inter-rater agreement methods and enhanced validation processes to further strengthen the work.

4 Experiment

This section describes the development and evaluation of the prediction models used in detection systems.

4.1 Preliminary Analysis

Figure 2 provides relevant statistics about the language, sentiment and major themes in HOC collection. There are many posts written by combining Hausa and English (Engausa). The HOC attracts more diverse discussion topics (2(c)) and the low proportion of negative sentiment (2(b)) suggests that many of the abusive terms have been used in a less offensive context.

We could not find a standard benchmark to compare how effective the detection of offensive terms would be in the Hausa language. To maximise the chances of developing an effective detection system, we explore how existing translation engines and multilingual models perform in processing Hausa terms. Table 3 presents the most frequent terms found in the study data. Some of these terms are quite offensive, but difficult to translate with a powerful translation engine. To determine how effective Google's translation engine performs in translating offensive Hausa words into English, we took the following top samples for comparison.

Translation of Offensive Terms Some examples of incorrectly translated offensive terms using Google's translation engine ('Hausa term' $\longrightarrow$ *meaning*$\longrightarrow$ *Google translation*):

- ('dan iska' $\longrightarrow$ rascal $\longrightarrow$ *and iska*)
- ('yan kutumar uba' $\longrightarrow$ d***head $\longrightarrow$ *father's cousins*)
- ('wawa jaki' $\longrightarrow$ idiot, donkey $\longrightarrow$ *wow what*)
- ('dan jaka' $\longrightarrow$ jennet's offspring $\longrightarrow$ *and jackets*)
- ('sakarai' $\longrightarrow$ dull (male) $\longrightarrow$ *boy*
- ('sakara' $\longrightarrow$ dull (female) $\longrightarrow$ *connection*

Despite the common use of the terms mentioned (as seen in Table 3), translation engines, like Google's, which have access to abundant and varied data, do not perform well. In the examples given, the best results are limited to the literal meaning, and do not take into account the subtleties of the language to detect offensive contexts. This highlights the importance of enriching low-resource languages, especially in the current era of large language models (LLMs).

Term	Category	Proportion
dan iska	abusive	27.7%
kutumar uba	abusive	10.4%
shege/shegiya	abusive	9.0%
jaki	offensive	8.7%
uwarka	abusive	8.0%
dan kutumar uba	abusive	6.9%
wawa	offensive	6.9%
jahili	offensive	6.90%
ubanka	abusive	6.6%

Table 3: Some examples of the most frequent terms found in the data.

4.2 Detection System

The problem is modelled as a classification task consisting of two classes for both the detection tasks - offensive/non-offensive involving the HOC data. This section describes the development and evaluation of applicable models.

Model Building

To determine the best detection model, we explore the following groups of machine learning models:

- **Linear and kernel-based models:** This group consists of logistic regression, a collection of regression algorithms that convey the relationship between variables (dependent and independent) and support vector machines (SVM) that categorise information separately using hyperplane to maximise the margin between them.
- **Naive Bayes** Naive Bayes is one of the popular models for classification tasks, especially in NLP. It is based on Bayes' theorem and assumes that features are conditionally independent given the class label.
- **Ensemble models:** This group comprises random forest and XGBoost models. The random forest classifier employs a set of decision trees. Closely related to the random forest is the extreme gradient boosting algorithm (XGBoost), which is based on the gradient tree boosting

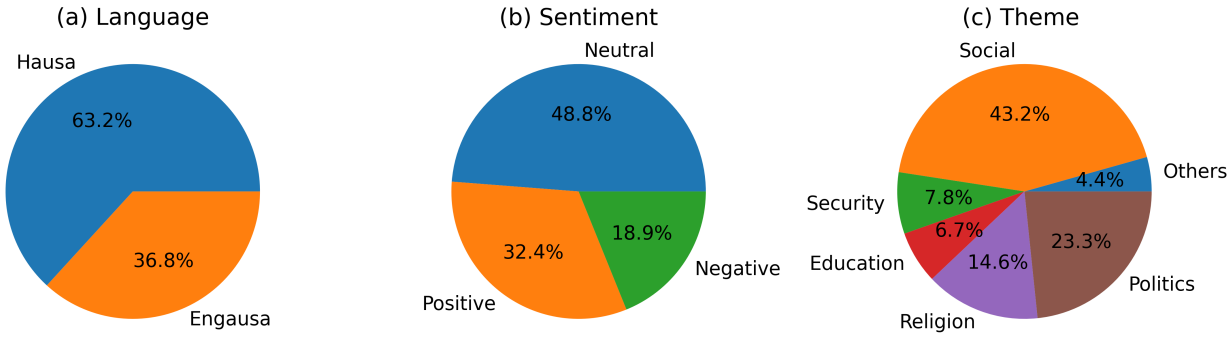

Figure 2: Statistics about language, sentiment and the major themes associated with the HOC dataset.

technique. These algorithms have been used for their fast learning and performance scalability.

- **Neural networks:** This group consists of multilayer perceptron (MLP) and convolutional neural network (CNN) models. The MLP is made up of an input layer, at least one hidden layer of computational neurones, and an output layer. A CNN is a deep neural network design consisting of convolutional and pooling or sub-sampling layers that feed input to a fully connected network.

Pretrained Multilingual Models In addition to the above standard ML models discussed earlier, we leverage state-of-the-art multilingual pretrained models that are widely used in NLP downstream tasks and align well with our research objectives. The models utilised in the study include:

- **XLM-RoBERTa Model:** we employ the Cross-lingual Language Model - Robustly Optimised BERT Pretraining Approach (XLM-RoBERTa) [Conneau *et al.*, 2019], implemented via Hugging Face[8]. Trained on a large-scale multilingual dataset comprising 100 languages sourced from CommonCrawl[9], this model benefits from extensive pretraining on 2.5TB of filtered data. This enables XLM-RoBERTa to capture rich cross-lingual representations, making it highly suitable for our application.
- **BERT multilingual base model (cased):** we also utilise the multilingual Bidirectional Encoder Representations from Transformers (mBERT) model [Devlin *et al.*, 2019], a transformer-based architecture designed to pretrain deep bidirectional contextual representations from large-scale unlabelled text corpora. Similar to XLM-RBERTa, we use Hugging Face[10] implementation, which has been pretrained on the top 104 languages with the largest Wikipedia presence using a masked language modelling (MLM) objective.

[8]https://huggingface.co/FacebookAI/xlm-roberta-base
[9]https://commoncrawl.org/
[10]https://huggingface.co/google-bert/bert-base-multilingual-cased

Feature Extraction and Training

The process of extracting relevant training features involves converting the raw data (textual) into numerical vectors that can be fed into the learning models for prediction. We apply both word embedding and the Term Frequency-Inverse Document Frequency (TF-IDF) techniques in transforming the cleaned version of the data utilised in training the aforementioned classic models. Word embedding maps words to a vector space for representation. This technique enables the models to process the data and capture the semantic relationships between words [Mikolov *et al.*, 2013]. Using the TFIDF technique, we tried unigrams, bigrams, and trigrams to maximise the models' prediction power.

Performance Metrics

Performance metrics are crucial to evaluating the quality and effectiveness of machine learning models. We chose the following quantifiable metrics to assess the efficacy of the models in building the detection system:

- *confusion matrix:* is a composite metric that gives the prediction output and quantification of how perplexed the model is. A true positive (TP) value signifies that the positive value is correctly predicted, a false positive (FP) means a positive value is falsely classified, a false negative (FN) means a negative value is incorrectly predicted, and a true negative (TN) means the negative value is correctly classified.
- *accuracy:* is the number of successfully categorised instances divided by the total number of instances. This metric can be derived from the confusion matrix as the sum of TP and TN divided by the sum of TP, TN, FP and FN.
- *precision and recall:* these are important metrics for binary classification problems.
  - Precision is the proportion of true positive instances that are classified as positive; it reflects the closeness of predicted values to one another given by: $precision = \frac{TP}{TP+FP}$.

  - On the other hand, recall is the proportion of positive instances that are correctly classified as positive, given by: $recall = \frac{TP}{TP+FN}$.

- *F1-score:* This metric combines both precision and recall into a single metric, giving equal weight to both.

In addition to the above objective metrics, we use the Google translation tool service (see Section 4.1) to evaluate its efficacy in translating some offensive content into the Hausa language.

5 Results and Discussion

This study explored the issues of offensive content detection in the Hausa language through user studies and predictive analysis. In this section, we present and discuss our main findings from user studies and the detection task.

5.1 Public Engagement

In this section, we present and discuss the main takeaway from the set of user studies. In total, we received responses from 180 participants who participated in the surveys. We asked the participants to rate the degree of offensiveness (Table 4) on some selected posts.

Participants' Take on Offensive Content

According to the demographic data (Table 1), a typical participant was a male between 25 and 35 years old who underwent a post-graduate study. In addition to the student category, the primary work place is the public sector followed by the self-employed. A large proportion (49%) of the respondents lamented the proliferation of offensive content, especially of abusive nature; 33% reported encountering offensive content quite frequently. About 83% of the respondents believe that young people tend to use abusive terms the most, and 21% reported the likelihood of commenting on abusive online posts. We attribute this to the use of abusive terms in both jovial and offensive contexts. Furthermore, about 41% of the respondents believe that abusive, hateful, or offensive content hurts online engagements. A substantial proportion (86%) of the respondents believe that political discourse is responsible for many abusive and hateful online content. In Table 4, about 75% rated example 1 as very offensive. There are mixed ratings for example 2; about 38% rated the post as not offensive despite the use of a term (*dan iska*) that is often considered offensive. Examples 3 and 4 also received mixed ratings with 28% and 31% of the participants labelling them offensive. Further detail is provided in the *Remarks* section of Table 4.

5.2 Detection Task

In this section, we present and discuss the efficacy of the trained models in the detection of offensive Hausa content.

Offensive Content Detection

Table 5 and Figure 3 show the performance of trained models in the detection of online offensive content in Hausa. XGBoost and CNN achieved the best result (with f-score values of 0.86% and 0.84%, respectively). This is followed by the SVM, MLP, and Logistic Regression. The use of trigrams greatly improves performance in all models. These results, especially those with higher recall and precision (Tables 5) show a promising start in the task of detecting offensive and abusive online content in the Hausa language. Overall, the pretrained models outperform all the models.

N-grams are contiguous sequences of n tokens in a given dataset, usually text or speech. To build effective detection systems, we tried various n-grams across the chosen models. We use bigrams and trigrams because of the prevalence of two- and three-adjacent words in both offensive and abusive terms. For example, *dan iska* is an abusive word that is used both offensively and jovially. The word is made up of two independent words, *dan* (son of or affiliated with) and *iska* (air). The use of bigrams or trigrams will maximise correct identification and prediction of the next word in a sentence. As demonstrated in Tables 5, and Figures 3 and the strategy of using ngrams, especially trigrams, culminated in a significant improvement in the prediction task. This has the effect of identifying common phrases and expressions in data collection.

Error Analysis To further validate our findings, we conducted a detailed error analysis focused on identifying and examining the mistakes made by our models, particularly those with lower performance (see Table 5). This analysis enables us to pinpoint specific areas of difficulty and understand the underlaying reasons for suboptimal performance. Table 6 presents examples of misclassified instances along with explanation for improving future work.

For the error analysis in Table 6, we selected the best-performing model, *bert-base-multilingual-cased*, as shown in Table 5. We observed that many misclassifications stemmed from the model's inability to interpret contextual meaning or its misunderstanding of certain terms. For instance, in some cases, abusive term was used within a post merely to report or describe inappropriate behaviour, yet the model labelled these posts as abusive (see **T1** in Table 6). This indicates that the model fails to distinguish between quoting or referencing abusive language versus actively using it. Such limitations highlight the need for more diverse and representative training data to improve contextual understanding. Moreover, the model struggles to differentiate between posts that describe abusive behaviour and those that actually exhibit it, suggesting an over-reliance on literal word presence. Another observed issue is the tendency to classify posts as abusive when they contain rare or unfamiliar words. For example, the post in **T2** uses a Hausa term *kanwa*[11], which was incorrectly flagged as abusive. Conversely, the model also failed to detect certain strongly abusive terms, underscoring its limited capacity to recognise the full range of offensive language (**T3**). Overall, these observations emphasise the model's difficulty in effectively identifying abusive content in a low-resource language such as Hausa. They also underscore the importance of curating larger, higher-quality, and more representative datasets to better capture nuanced language usage. Such improvements are crucial for ensuring safer online interactions and mitigating various forms of cyberbullying.

[11] *potash* in English

Type:	**On the use of offensive abusive terms**
Example 1:	*Dan gutsun uwarku watan maulid yaxo kuna bakin ciki don muna maulid amma duk tsinannan da yakarai mana happy new year sai munci uwarsa*
Ratings:	**very abusive (74.5%);** abusive (11.2%); somewhat abusive (7.1%); not abusive (2%)
Type:	**On the use of subtle abusive terms**
Example 2:	*wallahi abokina basan dan iska bane sai da naga yayi 5mins yana dariya*
Ratings:	very abusive (5.1%); abusive (4.1%); somewhat abusive (17.3%); might be abusive (15.3%); not abusive (20.4%); **not abusive at all (37.8%)**
Type:	**On disparaging remark**
Example 3:	*@user1 Shegiya me gudun dangi naga dae kema yar talakawa ce ko dan yanxu kin daena sae da awara da yaji iye*
Ratings	**very abusive (27.8%); abusive (24.7%)**; somewhat abusive (27.8%); might be abusive (10.3%); not abusive (5.2%); not abusive at all (4.1%)
Type:	**On the use of offensive and idiomatic expression**
Example 4:	*Bari ba shegiya bace ai, Azzamula kwai,Hakkinsa kuma yana nan Wallahi ranar da DSS basu da amfani bare wannan tsinannan Mulkin naku. @user2i gara dai kasa fir'auniyar matarka ta saki dan mutane dan Wallahi yanzu muka fara bayyana Zalincin da kukama Talakawa.*
Ratings:	**very abusive (31.3%); abusive (24%)**; somewhat abusive (26%); might be abusive (9.4%); not abusive (5.2%); not abusive at all (4.2%)
Remarks:	Example 1 is offensive and involves the use of strong abusive terms. Example 2 utilises abusive term but in a playful manner. Example 3 is derogatory and Example 4 combines subtle abusive terms and idiomatic expressions to lament on the state of leadership. Most of the participants' ratings agree with these observations.

Table 4: Summary of the results from the second part of the HOC user study ($n = 101$) about the public perceptions on the issue of offensive content in Hausa.

	N-grams	**Unigrams**				**Bigrams**				**Trigrams**			
	Metric	Acc.	Recall	Prec.	F-score	Acc.	Recall	Prec.	F-score	Acc.	Recall	Prec.	F-score
Models	Random Forest	0.75	0.75	0.75	0.75	0.77	0.77	0.77	0.77	0.80	0.80	0.80	0.80↑
	SVM	0.72	0.72	0.72	0.72	0.79	0.79	0.79	0.79	0.82	0.82	0.82	0.82↑
	Logistic Reg.	0.72	0.72	0.72	0.72	0.77	0.77	0.77	0.77	0.80	0.80	0.80	0.80↑
	XGBoost	0.75	0.75	0.75	0.75	0.81	0.81	0.81	0.81	**0.86**	**0.86**	**0.86**	**0.86**↑
	MLP	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.73	0.80	0.80	0.79	0.80↑
	CNN	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	xlm-roberta-base	—	—	—	—	—	—	—	—	0.84	0.84	0.87	0.84
	bert-base-multilingual-cased	—	—	—	—	—	—	—	—	0.86	0.86	0.89	0.86

Table 5: Performance of the models trained on the HOC datasets for the detection of offensive content. With the exception of the CNN, the trigrams strategy results in improved performance across all the models.

5.3 Discussion

There are many similarities between the use of both offensive and abusive terms; it is not always a straightforward approach to distinguish offensive terms in the context being used. For example, in Tables 3 and 7, there are some overlaps in the meanings associated with the terms. The distinguishing factor is the context and sentiment or polarity of the text. A high proportion of abusive terms and negative sentiment tend to point offensive content. To establish an effective distinction, manual moderation and a large collection of annotated datasets are required to discern offensive content tone.

Distribution of offensive and abusive content To support our observation about the frequency or dominance of offensive content in political and religious discourse, we identify the proportion of each theme in the whole data (see Figure 2(c)). This prevalence can be due to the following: political and religious discourse tends to be tense and often results in physical violence or offences. Table 3 shows the most frequently associated terms with offensive and abusive content in the distribution.

Topical issues vs offensive and abusive terms Building on the insight from the two user studies, our thematic analysis results show that political discourse, especially among young people, is at the forefront of attracting a high proportion of abusive terms (see Figure 2 and Table 7). For offensive terms, discourse on religion, ethnicity, and political issues tops the category.

Distinguishing jovial and offensive abusive terms Ambiguity represents a semantic condition in which words or phrases can seemingly carry multiple semantic interpretations or meanings. Discerning context is crucial since the meaning derived from a given text requires many factors to be considered. The problem of ambiguity involving the Hausa language has been analysed in [Ishaku *et al.*,]. The context in which an ambiguous statement is presented is crucial in deciphering its intended meaning. This is required since some

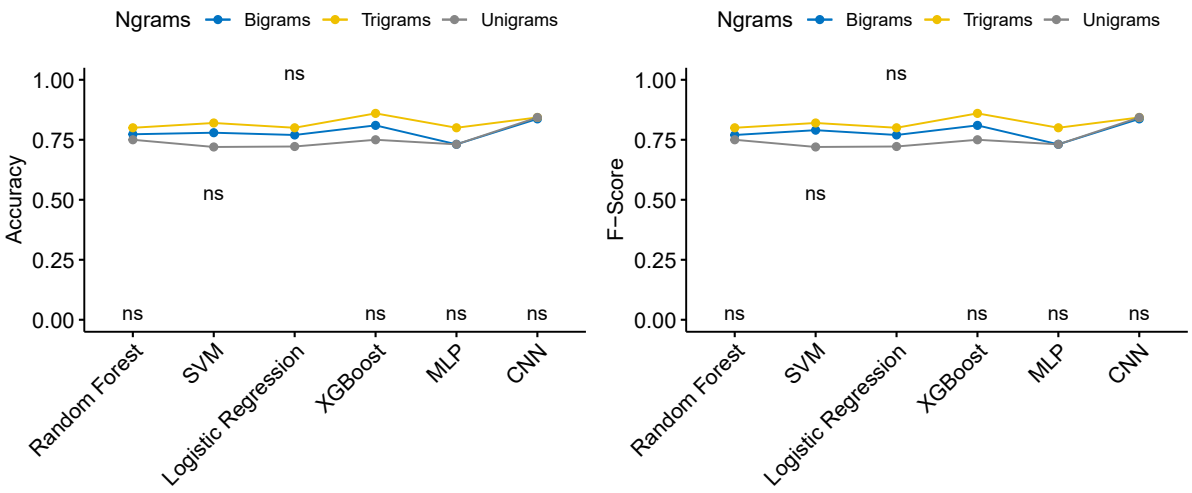

Figure 3: Accuracy and F-score performance of the trained models on the task of predicting offensive content (using the HOC dataset).

Models	Text	True Class	Predicted Class	Remark
bert-base-multilingual-cased	**T1**	Not Offensive	Offensive	focusing on single term
	T2	Not Offensive	Offensive	using unknown or rare term
	T3	Offensive	Not Offensive	using strong abusive terms

Examples:
T1: *Ohhh ni yar gidan mamana wallahi ina mugunjin haushin kalmar nan da wasu uwargidan ke fada wai AN AURAR MUSU MIJI Ohhh sannu me miji tinda mu zaman iskanci mukaxo ba sadaki yabayar ya auro mu ba, kofa a bude take gamasu xagi. Admin approved pls*
T2: *Ina yan boko yaude karya ta kare muku Kuxo kufdamin sunan kanwa da turanci*
T3: *D***N U**KA TAKA LAPIA*
T3: *??**shege dan iska** kawai abulfasadi Allah ya tsine mai albarka ??????????????punch lines paaa nie ????*

Table 6: Error analysis for offensive content detection models

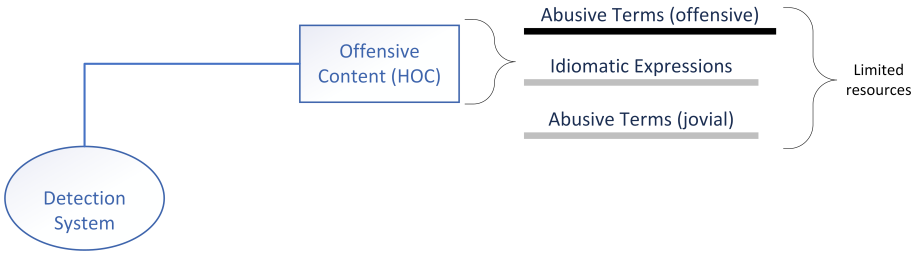

Figure 4: A summary of the areas to focus towards improving the detection system. Due to limited resources, the areas highlighted need further studies to build more effective and comprehensive detection systems for offensive content in the Hausa language.

of the abusive terms are often used in informal and jovial conversations among friends. To make the context clear and explain the distinction, we compute the sentiment associated with each post. By extracting the abusive terms and corresponding proportions in negative and positive (and neutral) cases, the distinction is made clear. Abusive terms associated with negative sentiment often point to offensive content. Although offensive in tone, abusive terms are commonly used in banter and playful engagement (see Tables 4, and 7 for examples).

Idiomatic expressions vs offensive content Idiomatic expressions are phrases that have a figurative, non-literal meaning. They often convey a specific idea, emotion, or concept that may not be immediately apparent to someone unfamiliar

Type	Sample Post	Category
	HOC Collection (Facebook and Twitter)	
abuse	dan shegiya	offensive
abuse	*Dan jaka shege dakiki*	offensive
abuse	dan iska	banter
abuse	Dan iska da fuska kamar ichen kabari taya bazatai block naka bah, malam you need madarar tamowa fah ...	social
abuse	@user1 @user2 @user3 Nace uwarka zai gyara anan, idan baka iya enigilishi ba kaje a fassara ma alaji.	social
abuse	@user1 @user2 Na rantse da Allah ni nasan su si kuwa wlhi sae dae su Mutu yan kutumar uba Munafukai	political
other	@user1 At all. Ai sai an dauki at least one and a half mins ana warm	other
abuse (allegation)	@user1 Thank you Dr. Jeffery. Atiku is corruption personified! You need to see how he made billions as chairman of National Council on Privatization during OBJ's tenure.Barawo BANSA Atiku.I am OBIDATTI.	political

Table 7: Sample comments from the participants.

with the language. Because they are often metaphorical and difficult to understand or translate for nonnative speakers, the set of words in idiomatic expressions has meanings that differ from the individual words. Some of these expressions can conceal offensive and abusive content in a more cryptic tone that will elude automated detection. To examine the interplay between idiomatic expressions conveying offensive or abusive tones, we focus on identifying the commonly used expressions and their meaning vis-a-vis offensive or abusive tones found in the data. Some examples of these expressions include: (1) *idan maye ya manta* (2) *shegiya bakar kaza*, (3) *bari ba shegiya bace* and (4) *abokin barawo barawo ne*. The above expressions can be used in different contexts and are subject to various interpretations. As shown earlier, Google's translation engine failed to correctly translate expressions (2). The bottom line is to collect a huge collection of rich data and explore various training strategies.

Impact of Offensive Terms
A tirade of online abuse and offense is likely to incite the public and precipitate physical violence. This is more pronounced in religious and political discourse. Unchecked cyberbullying could lead to continuous harassment both online and in physical space. When violence erupts, the use of abusive and offensive terms proliferate, thereby creating some sort of vicious cycle. At this juncture, it is pertinent to seek to understand the potential consequences of the forms of cyberbullying often encountered by online users.

6 Conclusion

Online social media platforms provide users with the opportunity to express their opinions, share messages, and socialise. However, offensive and abusive content often make the online space toxic and unwelcoming. While many detection strategies have been proposed, identifying cyberbullying-related content in low-resource languages is still a challenge. With focus on the most widely spoken Chadic languages, Hausa, this study contributed the following:

- The first annotated datasets consisting of offensive content to support downstream tasks in Hausa language.
- Detection systems to identify offensive and abusive posts in the Hausa language.
- Insights from two sets of user studies about the impact of offensive online content

We found that some offensive and abusive content have subtle relationships with idiomatic expressions that are expressed in the local tone, making them difficult to detect. This underscores the need to engage native speakers and resources to better understand local conventions and demographics in order to effectively combat issues that could amount to cyberbullying in the Hausa language.

Limitations and Future Work Limitations of the present work include (1) lack of larger annotated datasets (2) overlapping abusive terms with ambiguous meanings, and (3) the prevalence of idiomatic expressions with subtle offensive and abusive tones. We acknowledge that our current sample of the user study is biased towards male participants. In future work, we will recruit a more diverse group of annotators to mitigate biases and enhance the fairness and representativeness of the responses. Moreover, the data collection was conducted as part of a postgraduate research. The student received approval to proceed after their research proposal was reviewed and endorsed by a panel of academic staff. Should further ethical review be necessary, the project is referred to the institutional review board, which undertakes additional scrutiny and provides the requisite approval. The present ethical approval does not extend to engaging annotators with extensive offensive content without additional measures, such as mental health support and clear handling guidelines in place. Future work will focus on enriching the data to convey nuances and better contextualisation in the language, curating relevant idiomatic expressions, and using more powerful pre-trained language models (LLMs) for fine-tuning. Engagement with diverse stakeholders, such as the general public, experts (linguistics) and major news organisations, will be necessary to improve the training data. Figure 4 outlines the focus areas that need to be improved. We hope this will be a starting point to motivate further research into detecting various forms of cyberbullying in low-resource languages, particularly Hausa.

References

[Abubakar *et al.*, 2021] Amina Imam Abubakar, Abubakar Roko, Aminu Muhammad Bui, and Ibrahim Saidu. An enhanced feature acquisition for sentiment analysis of english and hausa tweets. *International Journal of Advanced Computer Science and Applications*, 12(9), 2021.

[Al-Garadi *et al.*, 2016] Mohammed Ali Al-Garadi, Kasturi Dewi Varathan, and Sri Devi Ravana. Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network. *Computers in Human Behavior*, 63:433–443, 2016.

[Aliyu *et al.*, 2023] Saminu Mohammad Aliyu, Idris Abdulmumin, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Saheed Abdullahi Salahudeen, Aliyu Yusuf, and Falalu Ibrahim Lawan. Hausanlp at semeval-2023 task 10: Transfer learning, synthetic data and side-information for multi-level sexism classification. *arXiv preprint arXiv:2305.00076*, 2023.

[Altay and Alatas, 2018] Elif Varol Altay and Bilal Alatas. Detection of cyberbullying in social networks using machine learning methods. In *2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*, pages 87–91. IEEE, 2018.

[Amin *et al.*, 2020] Mohammad Amin, Seyed Mahdi Nahavandi, and Mohammad Ghasemi. Various strategies for detecting online abusive content and combating cyberbullying. *IEEE Access*, 8:64595–64606, 2020.

[Antypas *et al.*, 2022] Dimosthenis Antypas, Asahi Ushio, Jose Camacho-Collados, Leonardo Neves, Vítor Silva, and Francesco Barbieri. Twitter topic classification. *arXiv preprint arXiv:2209.09824*, 2022.

[Balakrishnan *et al.*, 2023] Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N Govaichelvan. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–14, 2023.

[Basile *et al.*, 2019] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63, 2019.

[Bello *et al.*, 2023] Bello Shehu Bello, Muhammad Abubakar Alhassan, and Isa Inuwa-Dutse. # endsars protest: Discourse and mobilisation on twitter. *arXiv preprint arXiv:2301.06176*, 2023.

[Bully Facts, 2023] Bully Facts. Bullying statistics. http://www.bullyingstatistics.org/category/bully-facts, 2023. Accessed: 13-09-2023.

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

[Caselli *et al.*, 2020b] Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th language resources and evaluation conference*, pages 6193–6202, 2020.

[Chaure *et al.*, 2019] Ms Mrunali D Chaure, Jayant P Mehare, et al. Text classification and analysis with social media platform. *IJRAR-International Journal of Research and Analytical Reviews (IJRAR)*, 6(4):276–280, 2019.

[Conneau *et al.*, 2019] 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.

[Dambo *et al.*, 2022] Tamar Haruna Dambo, Metin Ersoy, Ahmad Muhammad Auwal, Victor Oluwafemi Olorunsola, and Mehmet Bahri Saydam. Office of the citizen: a qualitative analysis of twitter activity during the lekki shooting in nigeria's# endsars protests. *Information, Communication & Society*, 25(15):2246–2263, 2022.

[Davidson *et al.*, 2019] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[Dollarhide, 2021] M Dollarhide. Social media: Definition, effects, and list of top apps. *Retrieved from Investopedia: https://www. investopedia. com/terms/s/social-media. asp*, 2021.

[ECRI Policy, 2023] ECRI Policy. European commission against racism and intolerance. https://www.coe.int/en/web/european-commission-against-racism-and-intolerance/recommendation-no.15, 2023. Accessed: 10-09-2023.

[EU Code, 2023] EU Code. Code of conduct on countering illegal hate speech online. https://ec.europa.eu/commission/presscorner/detail/en/qanda_20_1135, 2023. Accessed: 15-09-2023.

[Founta *et al.*, 2018] Antigoni Maria 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. In *Twelfth International AAAI Conference on Web and Social Media*, 2018.

[Freelon *et al.*, 2016] Deen Freelon, Charlton D McIlwain, and Meredith Clark. Beyond the hashtags:# ferguson,#

blacklivesmatter, and the online struggle for offline justice. *Center for Media & Social Impact, American University, Forthcoming*, 2016.

[Haffner, 2018] Matthew Haffner. A spatial analysis of non-english twitter activity in houston, tx. *Transactions in GIS*, 22(4):913–929, 2018.

[Ibrahim *et al.*, 2022] Umar Adam Ibrahim, Moussa Mahamat Boukar, and Muhammed Aliyu Suleiman. Development of hausa dataset a baseline for speech recognition. *Data in Brief*, 40:107820, 2022.

[Inuwa-Dutse, 2021] Isa Inuwa-Dutse. The first large scale collection of diverse hausa language datasets. *arXiv preprint arXiv:2102.06991*, 2021.

[Ishaku *et al.*,] Joy Nanyisopwa Ishaku, Muhammad Mustapha, and Ubaidallah Muhammad Bello. Contrastive analysis of lexical and structural ambiguity between hausa and english languages.

[Kebriaei *et al.*, 2023] Emad Kebriaei, Ali Homayouni, Roghayeh Faraji, Armita Razavi, Azadeh Shakery, Heshaam Faili, and Yadollah Yaghoobzadeh. Persian offensive language detection. *Machine Learning*, pages 1–21, 2023.

[Kennedy, 2006] Helen Kennedy. Beyond anonymity, or future directions for internet identity research. *New media & society*, 8(6):859–876, 2006.

[Khan *et al.*, 2023] Anas Ali Khan, M Hammad Iqbal, Shibli Nisar, Awais Ahmad, and Waseem Iqbal. Offensive language detection for low resource language using deep sequence model. *IEEE Transactions on Computational Social Systems*, 2023.

[Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[Kovács *et al.*, 2022] György Kovács, Pedro Alonso, Rajkumar Saini, and Marcus Liwicki. Leveraging external resources for offensive content detection in social media. *AI Communications*, 35(2):87–109, 2022.

[Kumar *et al.*, 2014] Shamanth Kumar, Fred Morstatter, and Huan Liu. *Twitter data analytics*. Springer, 2014.

[Kusumawardani and Basri, 2017] RP Kusumawardani and MH Basri. Topic identification and categorization of public information in community-based social media. In *Journal of Physics: Conference Series*, volume 801, page 012075. IOP Publishing, 2017.

[Machuve *et al.*, 2022] Dina Machuve, Ciira Maina, and Wacuka Maina. Herdphobia: a dataset for hate speech against fulani in nigeria. 2022.

[Magee *et al.*, 2013] Rachel M Magee, Denise E Agosto, Andrea Forte, June Ahn, Michael Dickard, and Rebecca Reynolds. Teens and social media: Where are we now, where next? *Proceedings of the American Society for Information Science and Technology*, 50(1):1–4, 2013.

[Markov *et al.*, 2023] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15009–15018, 2023.

[Mathew *et al.*, 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, pages 14867–14875, 2021.

[Miao *et al.*, 2023] Zhenxiong Miao, Xingshu Chen, Haizhou Wang, Rui Tang, Zhou Yang, Tiemai Huang, and Wenyi Tang. Detecting offensive language based on graph attention networks and fusion features. *IEEE Transactions on Computational Social Systems*, 2023.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Miller, 2010] Harvey J Miller. The data avalanche is here. shouldn't we be digging? *Journal of Regional Science*, 50(1):181–201, 2010.

[Muhammad *et al.*, 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.

[Nemes and Kiss, 2021] László Nemes and Attila Kiss. Information extraction and named entity recognition supported social media sentiment analysis during the covid-19 pandemic. *Applied Sciences*, 11(22):11017, 2021.

[Nie *et al.*, 2020] Yuyang Nie, Yuanhe Tian, Xiang Wan, Yan Song, and Bo Dai. Named entity recognition for social media texts with semantic augmentation. *arXiv preprint arXiv:2010.15458*, 2020.

[Olteanu *et al.*, 2015] Alexandra Olteanu, Ingmar Weber, and Daniel Gatica-Perez. Characterizing the demographics behind the# blacklivesmatter movement. *arXiv preprint arXiv:1512.05671*, 2015.

[Oriola and Kotzé, 2020] Oluwafemi Oriola and Eduan Kotzé. Evaluating machine learning techniques for detecting offensive and hate speech in south african tweets. *IEEE Access*, 8:21496–21509, 2020.

[Oyewusi *et al.*, 2021] Wuraola Fisayo Oyewusi, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa. Naijaner: Comprehensive named entity recognition for 5 nigerian languages. *arXiv preprint arXiv:2105.00810*, 2021.

[Plaza-Del-Arco *et al.*, 2021] Flor Miriam Plaza-Del-Arco, M Dolores Molina-González, L Alfonso Ureña-López,

and María Teresa Martín-Valdivia. A multi-task learning approach to hate speech detection leveraging sentiment analysis. *IEEE Access*, 9:112478–112489, 2021.

[Puspitasari, 2017] Septy Puspitasari. Arab spring: A case study of egyptian revolution 2011. *Andalas Journal of International Studies (AJIS)*, 6(2):159–176, 2017.

[Rafiq *et al.*, 2018] Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, and Shivakant Mishra. Scalable and timely detection of cyberbullying in online social networks. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 1738–1747, 2018.

[Rajalakshmi *et al.*, 2023] Ratnavel Rajalakshmi, Srivarshan Selvaraj, Pavitra Vasudevan, et al. Hottest: Hate and offensive content identification in tamil using transformers and enhanced stemming. *Computer Speech & Language*, 78:101464, 2023.

[Rane and Salem, 2012] Halim Rane and Sumra Salem. Social media, social movements and the diffusion of ideas in the arab uprisings. *Journal of international communication*, 18(1):97–111, 2012.

[Saeed *et al.*, 2023] Ramsha Saeed, Hammad Afzal, Sadaf Abdul Rauf, and Naima Iltaf. Detection of offensive language and its severity for low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27, 2023.

[Sanguinetti *et al.*, 2018] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[Sani *et al.*,] Muhammad Sani, Abubakar Ahmad, and Hadiza S Abdulazeez. Sentiment analysis of hausa language tweet using machine learning approach.

[Schmidt and Wiegand, 2017] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[Simpson. and Weiner, 2019] John Simpson. and Edmund Weiner. *Oxford of Dictionary*. Oxford University Press, Oxford, UK, 2019.

[Sinyangwe *et al.*, 2023] Clement Sinyangwe, Douglas Kunda, and William Phiri Abwino. Detecting hate speech and offensive language using machine learning in published online content. *Zambia ICT Journal*, 7(1):79–84, 2023.

[Suleiman *et al.*, 2019] MA Suleiman, Muktar M Aliyu, and SI Zimit. Towards the development of hausa language corpus. *Int. J. Sci. Eng. Res*, 10:1598–1604, 2019.

[Tsvetkov, 2017] Yulia Tsvetkov. Opportunities and challenges in working with low-resource languages. *Slides Part-1*, 2017.

[Van Hee *et al.*, 2018] Cynthia Van Hee, Gilles Jacobs, Chris Emmery, Bart Desmet, Els Lefever, Ben Verhoeven, Guy De Pauw, Walter Daelemans, and Véronique Hoste. Automatic detection of cyberbullying in social media text. *PloS one*, 13(10):e0203794, 2018.

[Warf, 2018] Barney Warf. *The SAGE Encyclopedia of the Internet*. Sage, 2018.

[Yao *et al.*, 2019] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. Cyberbullying ends here: Towards robust detection of cyberbullying in social media. In *The World Wide Web Conference*, pages 3427–3433, 2019.

[Zandam *et al.*, 2023] Abubakar Yakubu Zandam, Fatima Adam Muhammad, and Isa Inuwa-Dutse. Online threats detection in hausa language. In *4th Workshop on African Natural Language Processing*, 2023.

[Zhong *et al.*, 2016] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. Content-driven detection of cyberbullying on the instagram social network. In *IJCAI*, volume 16, pages 3952–3958, 2016.

[Zishumba, 2019] Kudzai Zishumba. *Sentiment Analysis Based on Social Media Data*. PhD thesis, 2019.

[Zois *et al.*, 2018] Daphney-Stavroula Zois, Angeliki Kapodistria, Mengfan Yao, and Charalampos Chelmis. Optimal online cyberbullying detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2017–2021. IEEE, 2018.