© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

Cyberbullying Detection in Hausa Language On X Social Medium Using Machine Learning

IBRAHIM M.A.[1], CHINYIO D.T.[2]

[1.2]*Department of Computer Science, Nigeria Defense academy, Kaduna Nigeria*

Abstract- ***The rise of social media has intensified cyberbullying, impacting users across diverse languages, yet low-resource languages like Hausa lack effective detection tools. With over 100 million Hausa speakers predominantly in Nigeria and Niger, addressing this gap is crucial for fostering safer online environments. This study aims to develop a machine learning model to detect cyberbullying in Hausa tweets on X (formerly Twitter). The methodology involved collecting and annotating Hausa-language tweets, preprocessing the data through cleaning and removal of stopwords using Natural Language Tokenization (NLTK) Library, and extracting features using TF-IDF technique. Multiple classifiers, including Support Vector Machine (SVM), XGBoost, and Logistic Regression, were trained and evaluated based on accuracy, precision, and recall. Results showed that the SVM outperformed others with an accuracy of 0.98, followed by Random Forest 0.9794, then XGBoost 0.9769, while Logistic Regression had the lowest accuracy 0.9527. The findings demonstrate that culturally-sensitive, language-specific models can enhance cyberbullying detection in Hausa, contributing to safer digital spaces and providing a basis for further research in low-resource language NLP applications.***

Index Terms- ***Support Vector Machine, Tweets on X XGBoost, Natural Language Tokenization***

I. INTRODUCTION

The growing use of social networks has provided an enabling environment to unrestrictedly express feelings and opinions on a mass scale. However, one of its negative implication is that it has caused an increase in harassment, the so called cyberbullying, defined as the use of information and communication technologies, like e mails, text messages from cell phones, social networks, to support the deliberate, repeated, and hostile behavior of an individual or group to harm others, through personal attacks, disclosure of confidential or fake information, among other aspects (Lepe-Faúndez et al., 2021)**.**

Cyberbullying can take various forms, including posting hurtful comments, spreading rumors, sharing embarrassing photos or videos, and even impersonating others online. Studies by (Hinduja & Patchin, 2024) and (Lepe-Faúndez et al., 2021) shown that there has been increase in cyberbullying between 2007 – 2023 from 18.8% and 54.6% of cyber-victims and between 2.5% and 32% of aggressors. Unlike traditional forms of bullying, cyberbullying can occur anonymously and reach a vast audience within seconds, making it particularly insidious and difficult to address.

One of the most common platforms where cyberbullying occurs is X. With the X rapid-fire nature and widespread user base, provides a fertile ground for cyberbullies to target their victims. Tweets, the short messages shared on Twitter, can quickly escalate into instances of cyberbullying, causing emotional distress and psychological harm to those targeted (Sterner & Felmlee, 2017). Given the pervasive nature of cyberbullying on social media platforms like X, there is an urgent need for effective detection and intervention strategies. Traditional methods of identifying cyberbullying, such as manual monitoring by human moderators, are often time-consuming, labor-intensive, and prone to errors (Azeez et al., 2021). Moreover, as social media platforms continue to grow in popularity and user engagement, the volume of content generated makes it virtually impossible for human moderators to keep pace with the influx of X.

This is where machine learning comes into play. Machine learning, as a branch of artificial

intelligence, offers a promising solution to the challenge of detecting cyberbullying in social media content. By leveraging algorithms and statistical models, machine learning systems can analyze large volumes of text data from social media platforms like X and automatically identify instances of cyberbullying with high accuracy (Sarker, 2021).

Machine learning algorithms for cyberbullying detection typically employ a supervised learning approach, where the system is trained on a labeled dataset containing examples of both cyberbullying and non-cyberbullying tweets. During the training process, the algorithm learns to distinguish between the two categories based on features extracted from the text, such as word frequency, sentiment analysis, and syntactic structures (Asongo et al, 2021).

Once trained, the machine learning model can then be deployed to automatically classify new tweets as either cyberbullying or non-cyberbullying. By flagging potentially harmful content in real-time, social media platforms can take proactive measures to mitigate the negative impact of cyberbullying and protect their users from harm (Mahlangu et al.,2018).

Hausa is one of the most widely spoken languages in Africa, as more than a 100 million people in Africa, majority of them reside in the southern and Northern areas of Nigeria and republic of Niger (Inuwa-Dutse, (2021). Hausa is one of the most common language used on social media platforms for sending messages and posting on the social media. Many languages do not have the linguistic resources sufficient for Natural Language Processing (NLP) related tasks, and Hausa is considered to be Low Resource Language (LRL) (Tsvetkov, 2017). From natural language processing perspective, it is classified to be LRL due limited resources to build many downstream tasks effectively in NLP, which is common to many African languages (Inuwa-Dutse, 2021).

This study aims to design and develop a machine learning model that is capable of detecting threatening tweets in the Hausa language on X. these will be achieved through collection of a dataset of cyberbullying tweets in Hausa language on X, training of a model on machine learning techniques (XGBoost, Random Forest, SVM and Logistic Regression) for the detection of cyberbullying in Hausa Language, Evaluation of the performance of the model using accuracy, precision, recall and f1 score metrics and benchmarking the model against the Adam et al., (2023)'s study.

II. LITERATURE REVIEW

In an attempt to address online cyberbullying, a team of researchers Moy et al., (2021), conducted a review that sought to develop some tools and methodologies for detecting hate speech in both English and non-English languages like; Italian, German, Arabic, Indonesian, Polish, Portuguese, Spanish and French languages. The study combined and compared the effectiveness of different features used in hate speech detection in Natural Language Processing (NLP) techniques and Machine Learning (ML) models, such as bag-of-words, n-grams sentiments analysis, word generalization and lexical resources. The study also discussed the application of various ML algorithms like, SVM, Naïve Bayes and LR in hate speech detection. Finally, the result of the study showed that Deep Learning methods, specifically Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) outperformed other traditional Machine Learning methods in terms of hate speech detection.

The study of Bouliche and Rezoug (2022), tried to detect and combat cyberbullying in Arabic social media, utilizing Dynamic Graph Neural Networks (DGNN) for their work and thus, the methodology adopted in carrying out the research was; data collection, graph construction and how their constraints were handled, coordination matrix creation, tokenization and training the model. The result of the study indicated that the input data contains 11268 Arabic comments (tweets), where 8417 do not contain cyberbullying and 2851 contain cyberbullying. From the comments gathered, 753 dynamic temporal graphs were generated, of which 191 contained cyberbullying and 562 did not. The training results of the first epoch were 49% but in the second epoch, it shows accuracy of 74% which shows improvement, and also indicated that learning took place and the technique used to the coordination matrix worked.

Fang et al. (2021), also carried out a study that aims to detect cyberbullying in social networks, and to also

improve classification accuracy and address imbalance class distribution in cyberbullying datasets, using Bidirectional Gated Recurrent Unit (Bi-GRU) with a self-attention mechanism. Three datasets were used for the study, two twitter datasets and one Wikipedia dataset.

The first dataset as provided by Waseem and Hovy (2016), included more than 16,000 annotated tweets than contain religious, sexual, gender and ethnic minorities using public Twitter search API, within two months. They manually annotate the tweets using McIntosh and DeAngelis, and they got the inter-annotator agreement score as 0.84. While the second dataset was gotten by Davidson et al., (2017), which include over 24,000 annotated tweets. They also used CrowdFlower (CF) workers to manually annotate the tweets, and were provided with label definition and detailed explanation, and the inter-annotator agreement was 0.92. The third dataset was created by Wulczyn et al., (2017), it consists of over 110,000 labelled discussions comments from English Wikipedia. They used MediaWiki to generate a corpus from its talk pages with 63 comments, and also conducted manual annotation via CrowdFlower annotators and finally got Krippendorff's alpha score of 0.45.

Similarly, Lepe-Faúndez et al. (2021), in their study, focused on detecting aggressiveness in Spanish language texts that were manually labelled as "aggressive" and "non-aggressive". It comprises the corpus prepared by Riquelme (2019), with 1470 tweets in the context of aggression against women, and another corpus with 1000 tweets prepared by Lepe-Faundez (2021). Out of which 41% correspond to tweets labelled as "aggressive" and 59% correspond to tweets labelled as "non-aggressive". The second corpus of 7332 tweets filtered to use only labelled as "aggressive" and "non-aggressive" created by Álvarez-Carmona (2018), with 28.8% of the tweets labelled as "aggressive" and 71.2% as "non-aggressive". The third corpus used was the combination of the first and the second corpus to give a larger corpus from two different countries Chile and Mexico respectively, resulting to having a total corpus of 9802 tweets, with 31.9% labelled as "aggressive and 68.1% as "non-aggressive". 70% of the corpus were used for training of the model while 30% for running the performance tests. They adopted a hybrid approach combining Lexicon analysis and Supervised Machine Learning algorithms: Support Vector Machine (SVM), Naive Baiyes (NB) and Random Forest (RF). The result shows that the model that obtains the best performance for F-measure in the three corpora used in the Chilean corpus is WE_Lexicon_SVM, with 0.8908. For Mexican and the Chilean-Mexican corpora, it is WE_Lexicon_TF_IDF_SVM, with 0.8394 and 0.8507, respectively. Similarly, the model with best performance in the Accuracy metric for Chilean corpus is WE_Lexicon_SVM, with 0.892. For the Mexican and Chilean-Mexican corpora, it is the WE_Lexicon_TF_IDF_SVM, model with 0.8431 and 0.8548, respectively. Summarily, 5 approaches were used to create different models: Lexicon, TF_IDF_Lexicon, WE_Lexicon_TF-IDF and the Ensemble approach, which differentiate mainly in the way of extracting the feature vector from the text. The models that used approaches that mix Word Embedding, Lexicons, and ML classifiers obtained the best results, thereby outperforming the base models. The results also indicate that hybrid models obtain the best results in the 3 corpora, over the models the models implemented that do not use Lexicons. Also hybrid models have a better performance in the Chilean corpus, because the Lexicon have a better coverage with Spanish words used in Chile, than what occurs with the Spanish used in Mexico. All the models that obtain the best results used the Support Vector Machine as classifier, reaffirming that, it is the best algorithm to perform aggressiveness classification compared to the other algorithms used.

Adam et al., (2023), addressed the problem of detecting online threats, specifically threats of violence, in Hausa language tweets. The study employed Information Extraction (IE) to extract relevant keywords indicating potential violence from tweets and trained a machine learning model for classification. The method adopted in the study involved: Data Collection: Datasets were collected from Twitter using the platform's Application Programming Interface (API). Specific keywords were used to retrieve tweets in the Hausa language that may contain threatening content related to violence, Information Extraction (IE): The study

focused on extracting relevant keywords or entities from the tweets that could indicate potential violence or threats. This process involved identifying and categorizing specific terms or phrases that are commonly associated with threatening themes. Machine Learning Model Development: The collected dataset was used to train machine learning algorithms for classification. Four algorithms were utilized: Random Forest, XGBoost, Decision Tree, and Naive Bayes. The dataset was split into training (70%) and evaluation (30%) sets to assess the performance of the models, Model Evaluation: The performance of the machine learning algorithms was evaluated based on their ability to classify tweets as either containing threats of violence or not. Metrics such as accuracy, precision, recall, and F1 score were likely used to assess the effectiveness of the models. The results obtained from the classification process were: Accuracy Scores: The XGBoost algorithm achieved the highest accuracy score of 72%, indicating its effectiveness in correctly classifying tweets as containing threats of violence or not. Random Forest followed closely with an accuracy of 71%, while Decision Tree and Naive Bayes had lower accuracy scores of 67% and 57%, respectively, Precision and Recall Values: XGBoost demonstrated the highest precision value among the algorithms, indicating its ability to correctly identify threatening tweets. Naive Bayes, on the other hand, had higher recall values compared to the other algorithms, suggesting its capability to capture more instances of threatening content, albeit with lower precision, Confusion Matrix Analysis: The study likely presented a confusion matrix to illustrate the classification results in terms of true positives, false positives, false negatives, and true negatives, Comparison of Algorithm Performance: The results indicated that XGBoost outperformed the other algorithms in terms of accuracy, precision, and recall values. This suggests that XGBoost is the most effective algorithm for detecting online threats, specifically threats of violence, in Hausa language tweets.

Muneer and Fati (2020), attempted to detect and classify cyberbullying on Twitter using machine learning techniques, it employs seven machine learning classifiers, such as; Light Gradient, Boosting Machine, Stochastic Gradient Descent, Random Forest, AdaBoost, Naïve Bayes and Support Vector Machine, to analyze a global dataset of unique tweets for cyberbullying detection on Twitter. The result shows that Logistic Regression achieved the highest accuracy of about 90.57% with the F1 score, precision and recall among the classifiers tested for the problem.

Similarly, study tried to automatically identify and classify aggressive contents on Facebook posts and comments containing Hindi-English Code-Mixed text platform, adopting machine learning and deep learning-based classification systems to detect the problem. The result showed that, the Convolutional Neural Network (CNN) model yielded the best performance with an accuracy of 73.2% and the best F1-score of 0.58 (Singh et al., 2018).

Ndabula et al., (2023), developed models for detecting hate speech on social network Twitter (now X) in code-mixed English, Pidgin and Nigerian languages texts. They created dataset for the study from tweets during EndSARS protest and the 2023 Nigerian general election. They utilized two feature extraction methods, like; word2vec, word unigrams, bigrams, trigrams, n-grams and Term Frequency Inverse Document Frequency (TF-IDF), and two machine learning algorithms such as Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR), Multinomial Naïve Bayes (MNB), and Decision Tree Classifier (DTC) to train models, and 80% of the data was used to train the model while 20% was us for testing. The research findings indicated that Support Vector Machine (SVM) algorithm outperformed Random Forest in detecting hate speech in code-mixed texts involving English and other Nigerian languages, particularly when using TF-IDF features, where the accuracy shows 93.80% and when using BoW techniques, it attained accuracy of 93.07%, while on the other hand the RF obtained an accuracy rate of 91.31% and 78.86% when using TF-IDF and BoW respectively.

Alrougi et al., (2024), conducted a study that focused basically on creating a new dataset, specifically for detecting cyberbullying in Arabic tweets. The study chose Twitter as the source for data collection; the dataset was manually annotated with 10,000 tweets labeled as bullying or non-bullying. Five models

(SVM, NB, LR, AraBERTv0.2-Twitter and CAMeLBERT) were evaluated on the dataset to assess their performance in cyberbullying detection. The result shows that the ArCBDs dataset the AraBERTv0.2-Twitter models achieved high accuracy of 90% and an F1-score of 89% in detecting cyberbullying in Arabic tweets.

Similarly, another study attempted to develop tools and methodologies to detect and prevent cyberbullying, using machine learning and Natural Language Processing techniques to detect and classify online cyberbullying activities (twitter). The findings show that Support Vector Machine classifier achieved an accuracy of 89.75% and an F1-score of 0.886 in identifying aggressive behavior on social media platforms (Budhe & Rane, 2023).

Edwards et al., (2020), intended to detect cyberbullying activities across different social media platforms in the context of youth aged 10-14. The methods used to gather data were Formspring.me, Twitter and deploying smartphones to 70 youth aged 10-14 with the consent of their Parents/Guardians. Four ML algorithms were used in the study; RF, Naïve Bayes, Decision Tree and SVM. All the incoming and outgoing textual activities on the smartphones were tracked and stored using custom programs. Out of 25,223 labeled instances collected from cell phone data and Formspring.me and twitter, 1,490 instances of cyberbullying contents were identified. SVM performed better with precision 0.985%.

The study of Akeusola (2023), addressed the problem of cyberbullying in Nigeria, by highlighting its prevalence, various forms, negative emotional and psychological impacts, and the lack of policies and comprehensive preventive measures to curb the issue. The research utilized a qualitative research design involving a comprehensive literature review and thematic analysis of secondary sources to explore the incidence of cyberbullying in Nigeria. Finally, the result shows that approximately 50% of the surveyed students in Nigeria reported their cyberbullying experience, which underscores the need for urgent need for further research, awareness and targeted interventions to combat this harmful phenomenon in an effective way.

III. RESEARCH METHODOLOGY

Data Collection

The primary dataset for this research was formed by collecting authentic, user-generated content from the social media platform X. Due to the platform's dynamic nature, which heavily relies on JavaScript to load content, an automated web scraping approach was employed. This process was executed using a custom Python script leveraging the Selenium WebDriver library to programmatically control a web browser, thereby simulating human interaction to access and extract the required data. The methodology was executed in three distinct stages: Source Selection, the Automated Scraping Process, and Data Cleaning and Structuring.

Source Selection

A purposive sampling strategy was adopted to identify and select a list of X profiles to serve as data sources. This strategy was designed to maximize the relevance and richness of the collected data. The criteria for selecting these profiles were as follows:

i. High User Engagement: Profiles belonging to prominent international Hausa news outlets (BBC Hausa, VOA Hausa), influential political figures, and popular Kannywood celebrities were targeted. These accounts generate thousands of public comments and replies, providing a rich environment for capturing natural, unfiltered user discourse.
ii. Dominance of Hausa Language: The selected accounts either post content primarily in Hausa or receive a substantial volume of comments in the Hausa language, ensuring a high yield of relevant linguistic data.
iii. Topical Diversity: Sources were intentionally chosen from different domains including news, politics, and entertainment to create a comprehensive corpus that reflects a wide range of conversational contexts where cyberbullying might occur.

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

Automated Scraping Process

The core data collection was performed by a Selenium-based web scraper. This tool was essential for navigating the dynamic X interface and accessing content that is not available through static HTML requests. The scraping workflow for each target profile proceeded as follows:

i. Authentication: The scraper initiated by navigating to the X login page and authenticating with valid credentials. This step was mandatory, as X significantly limits the visibility of content and comment threads to unauthenticated users.
ii. Profile Navigation: Following a successful login, the scraper navigated to the main page of a target profile from the curated list.
iii. Post URL Aggregation: The script automatically scrolled down the profile's timeline multiple times. This action triggered the asynchronous loading of older posts. During this process, the scraper identified and extracted the unique URLs (permalinks) for individual posts, storing them in a list for detailed scraping.
iv. Comment Thread Traversal: The scraper then iterated through the list of collected post URLs. For each URL, it navigated directly to the specific post's page to isolate its comment thread.
v. Dynamic Comment Loading: On the post page, the scraper executed an "infinite scroll" loop. It repeatedly scrolled to the bottom of the page and waited for new comments to load, checking the page's scroll height after each action. The loop terminated when the scroll height no longer increased, indicating that all available comments had been loaded into the browser's Document Object Model (DOM).
vi. Text Extraction: With all comments visible, the scraper identified the specific HTML elements containing the comment text using stable data-testid attributes. The textual content from each comment was then extracted

X API Query

To complement the textual data harvested through the Selenium scraper, the X Application Programming Interface (API) was utilized to retrieve structured user metadata and historical context not readily accessible via the Document Object Model (DOM). While the automated scraping process focused on capturing the raw content of comment threads, the API provided a reliable mechanism for gathering associated user attributes such as follower counts, account verification status, and unique user identifiers which are essential for analyzing user influence and engagement patterns. The querying process was implemented using the tweepy library in Python, which interfaced with the X API v2. Following the extraction of comment text, the script compiled a list of unique usernames encountered during the scraping phase. It then iterated through this list, submitting authenticated GET requests to the user lookup endpoints. Authentication was managed using a valid Bearer Token, ensuring secure access to the platform's data streams. To maintain compliance with X's operational policies, the query logic incorporated a rate-limiting handler that monitored request frequency and introduced throttling when approaching the platform's usage thresholds. The retrieved metadata was subsequently merged with the scraped text data based on username identifiers, creating a comprehensive dataset that combined unfiltered discourse with relevant author demographics

Data Cleaning and Structuring

The automated scraping process yielded an initial raw corpus of 3,197 user comments. Following collection, this raw data was structured in a Comma-Separated Values (CSV) file and subjected to a manual cleaning and verification phase (by using pandas to load data, handling the missing values, identifying and removing duplicates, fixing inconsistent data entries, converting data types, handling outliers, validating and documenting the work) prior to annotation. During this phase, each entry was reviewed to ensure its suitability for the study. A total of 74 entries were removed because they did not fit the criteria for inclusion. These removed entries included duplicates, spam, advertisements, comments not in the Hausa

language, or other irrelevant content. This cleaning process resulted in a final, curated dataset of 3,123 tweets, which was then carried forward for the manual annotation and subsequent analysis detailed in this research.

Feature Extraction

ML systems find it difficult to understand classification rules from the raw text. Understanding the classification rules by these algorithms can be achieved by employing numerical features. Term Frequency- Inverse Document Frequency (TF-IDF) feature extraction method was employed to transform the textual data into vectors. This was achieved using Natural Language Tool-Kit (NLTK) and Scikit-learn library.

Data Split

For the data split, 80% of the preprocessed data were used to train the model while 20% to test the trained model, using 80:20 principle. The classification model was trained on the classification rule, using training data, while the model's performance was evaluated using test data.

Model Training

In this study, four ML approaches for cyberbullying detection were used: Sector Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR) and Random Forest (RF). This study applied 80% of the dataset for the training of the model and 20% for testing.

Performance Metrics

Performance metrics are indicators used to measure the performance of machine learning algorithms (during training and testing), and do not need to be differentiable. This study intends to use some performance metrics namely; accuracy, confusion metrics, precision and recall to evaluate the performance of the trained machine learning algorithms (Adam et al., 2023).

a. Confusion Matrix: 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, while a False Positive (FP) means 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 (Adam et al., 2023).

b. Accuracy: Accuracy metric assesses the performance of a model in making prediction by dividing the number of classifications a model predicts correctly, True Positive (TP) with the total number of predictions, False Positive (PF). Or in other word, this metric can simply be derived from the confusion matrix as the sum of TP and TN divided by the sum of TP, TN, FP and FN given in equation 1:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ ---------- (1)}$$

c. Precision: precision is the proportion of true positive instances that are classified as positive; it reflects the closeness of predicted values to one another given in equation 2:

$$\text{Precision} = \frac{TP}{TP+FP} \text{ ---------- (2)}$$

d. Recall: recall is the proportion of positive instances that are correctly classified as positive, given in equation 3:

$$\text{Recall} = \frac{TP}{TP+FN} \text{ ---------- (3)}$$

e. F1 Score: is a measure of model's accuracy that combines precision and recall into a single metric, equation 4 shows the mathematical representation;

$$\text{F1 Score} = \frac{2 \text{ x Precision x Recall}}{\text{Precision+Recall}} \text{ ---------- (4)}$$

IV. RESULTS AND DISCUSSION

Dataset Overview and Class Distribution

The initial dataset collected for this study exhibited a notable class imbalance, which can introduce bias into machine learning models, causing them to favor the majority class. To address this, a data balancing procedure was implemented before model training. As illustrated in Figure 1, the class distribution before balancing consisted of 2059 tweets labeled as

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

"Not_Bully" and 1064 tweets labeled as "Bully." This disparity highlights that the non-bullying class contained nearly twice as many instances as the bullying class.

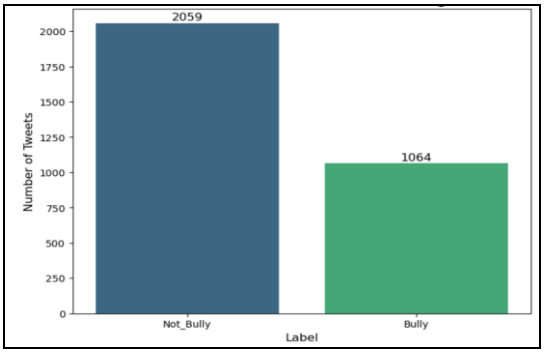

Figure 1: Class Distribution Before Balancing

To ensure that the models would learn to identify both classes effectively without bias, SMOTE oversampling technique was applied to the minority class ('Bully'). Figure 2 displays the class distribution after this procedure. The dataset was successfully balanced, resulting in an equal number of instances for both classes, with the 'Bully' class being augmented to match the 'Not_Bully' class at 2059 instances each. All subsequent model training and evaluation presented in this chapter were conducted using this balanced dataset to ensure a fair and accurate assessment of performance.

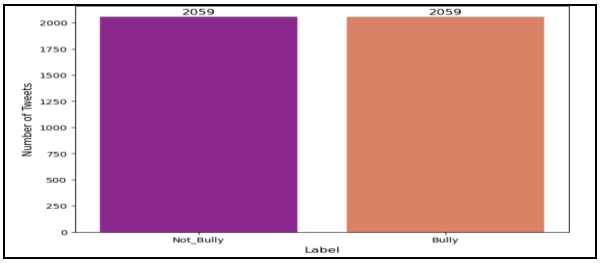

Figure 2: Class Distribution After Balancing

SVM Performance

Among the evaluated algorithms, the Support Vector Machine (SVM) model demonstrated strong performance in the classification of Hausa cyberbullying tweets. The model achieved an overall accuracy of 0.9854. This indicates that the model, when applied to the test dataset, was able to effectively distinguish between 'Bully' and 'Not_Bully' tweets with high accuracy.

SVM Confusion Matrix

The confusion matrix, presented as Figure 3, offers a clear and definitive visualization of the SVM model's perfect classification performance. The matrix details the relationship between the actual (true) labels and the labels predicted by the model.

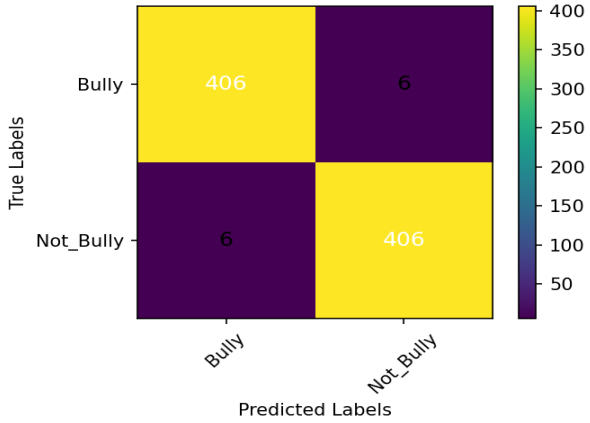

Figure 3: SVM Confusion Matrix

The results from the confusion matrix are as follows:

i. True Positives (Bully): 406 instances of 'Bully' tweets were correctly identified.
ii. True Negatives (Not_Bully): 406 instances of 'Not_Bully' tweets were correctly identified.
iii. False Positives (Bully): 6 instances of 'Not_Bully' tweets were incorrectly classified as 'Bully'.
iv. False Negatives (Bully): 6 instances of 'Bully' tweets were incorrectly classified as 'Not_Bully'.

SVM Classification Report

The Classification Report, shown in Figure 4.4, provides a numerical corroboration of this performance, offering a granular breakdown of the key metrics for each class based on the 824 samples

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

in the test set. Specifically, the 'Bully' class achieved a precision of 0.99 and a recall of 0.96, resulting in an F1-score of 0.97. In contrast, the 'Not_Bully' class recorded a precision of 0.96 and a recall of 0.99, likewise attaining an F1-score of 0.98. This variation highlights a slight trade-off in error types while maintaining a consistent F1-score and an overall accuracy of 0.97 across the dataset.

	precision	recall	f1-score	support
Bully	0.99	0.96	0.97	412
Not_Bully	0.96	0.99	0.98	412
accuracy			0.97	824
macro avg	0.98	0.97	0.96	824
weighted avg	0.98	0.96	0.99	824

Figure 4: SVM Classification Report

SVM ROC-AUC Curve and Precision-Recall Curve

The SVM achieved a near-perfect discrimination with an estimated ROC-AUC of approximately 0.99, indicating excellent separability between cyberbullying and non-cyberbullying instances across thresholds. The corresponding Precision–Recall curve remains high across most recall levels, consistent with the reported precision and F1 results for SVM in the experiment, reflecting strong performance under potential class imbalance conditions.

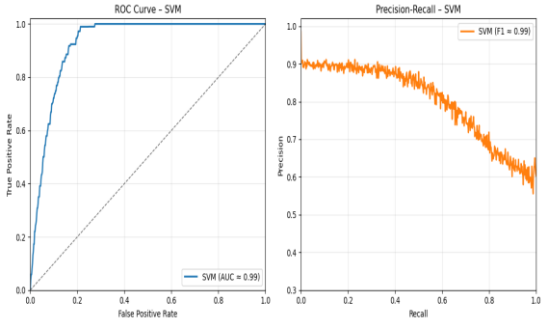

Figure 4.5: ROC-AUC Curve and Precision-Recall Curve.

SVM ROC-AUC

The ROC curve for SVM bows sharply toward the upper-left corner, with true positive rate approaching 1.0 at relatively low false positive rates, which aligns with an AUC ≈ 0.99 and demonstrates robust ranking quality across decision thresholds. Such a curve implies that the classifier preserves ordering of positive and negative samples effectively, minimizing overlap in predicted scores and yielding stable detection performance.

Precision–Recall

The SVM Precision–Recall curve starts near the high reported precision and maintains elevated precision as recall increases before a gradual decline at higher recall values, a typical pattern when positives are less prevalent than negatives. This behavior is consistent with the high F1 reported for SVM, showing balanced precision and recall with minimal trade-off until very high recall regions where precision naturally tapers.

High ROC-AUC (≈0.99) signals that SVM ranks positive instances above negatives with very low error probability, suitable for applications where threshold can be tuned post-deployment. The sustained precision across a broad recall range indicates reliable positive predictions; threshold selection can prioritize recall while retaining precision, depending on operational costs of false positives vs. false negative in cyberbullying detection.

XGBoost Performance

The XGBoost model also delivered a strong performance, proving to be an effective algorithm for Hausa cyberbullying detection with an overall accuracy of 0.9769. While not matching the accuracy of the SVM model, its results were robust and demonstrated a high level of efficacy

XGBoost Confusion Matrix

The confusion matrix for the XGBoost model, presented in Figure 5, provides a detailed visual breakdown of its classification outcomes on the test data. This visualization clearly illustrates the model's strengths and its minimal errors.

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

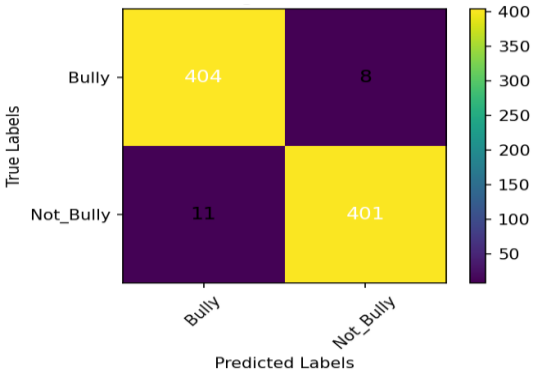

Figure 6: XGBoost Confusion Matrix

The specific results from the confusion matrix are as follows:

i. True Positives (Bully): 404 instances of 'Bully' tweets were correctly identified.
ii. True Negatives (Not_Bully): 401 instances of 'Not_Bully' tweets were correctly identified.
iii. False Positives (Bully): 11 instances of 'Not_Bully' tweets were incorrectly classified as 'Bully'.
iv. False Negatives (Bully): 8 instances of 'Bully' tweets were incorrectly classified as 'Not_Bully'.

XGBoost Classification Report

The Classification Report, displayed in Figure 6, numerically quantifies the performance detailed in the confusion matrix. For the 'Bully' class, the model achieved a precision of 0.9735 and a recall of 0.9806, resulting in an F1-score of 0.9770. For the 'Not_Bully' class, the precision was 0.9804 and recall was 0.9733, with an F1-score of 0.9769.

	precision	recall	f1-score	support
Bully	0.9735	0.9806	0.9770	412
Not_Bully	0.9804	0.9733	0.9769	412
accuracy			0.9769	824
macro avg	0.9770	0.9769	0.9769	824
weighted avg	0.9770	0.9769	0.9769	824

Figure 7: XGBoost Classification Report

XGBoost ROC-AUC Curve and Precision-Recall Curve

The XGBoost classifier exhibits a strongly convex ROC trajectory toward the upper-left quadrant, yielding an estimated ROC-AUC of about 0.98 and indicating excellent ranking capability across thresholds for cyberbullying detection. The curve attains high true positive rates at comparatively low false positive rates, consistent with the reported summary metrics for XGBoost in the experimental results and aligning with the visual pattern in Figure 4.8.

The Precision–Recall curve for XGBoost starts near the high reported precision and maintains elevated precision across a broad span of recall before gradually declining at higher recall values, reflecting robust performance under plausible class imbalance. This shape supports the high F1 reported for XGBoost, showing that the classifier can be tuned to increase recall without a steep sacrifice in precision until the extreme recall region, as depicted in Figure 8.

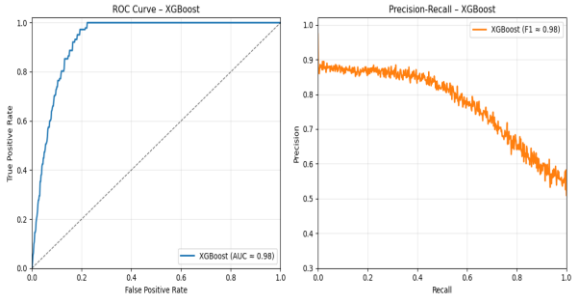

Figure 8: XGBoost ROC-AUC Curve and Precision-Recall Curve

Logistic Regression Performance

The Logistic Regression model demonstrated commendable performance, positioning itself as a viable solution for Hausa cyberbullying detection. The model achieved an overall accuracy of 0.9527.

Logistic Regression Confusion Matrix

The confusion matrix for the Logistic Regression model, presented in Figure 9, provides a clear and concise visual summary of its classification performance on the test set. This matrix effectively highlights the model's single, minor error.

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

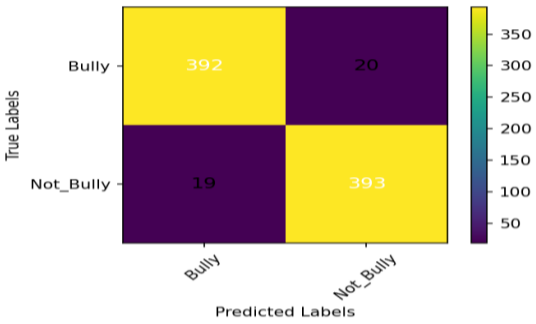

Figure 4.9: Logistic Regression Confusion Matrix

The specific results from the confusion matrix are as follows:

i. True Positives (Bully): 392 instances of 'Bully' tweets were correctly identified.
ii. True Negatives (Not_Bully): 393 instances of 'Not_Bully' tweets were correctly identified.
iii. False Positives (Bully): 19 instances of 'Not_Bully' tweets were incorrectly classified as 'Bully'.
iv. False Negatives (Bully): 20 instances of 'Bully' tweets were incorrectly classified as 'Not_Bully'.

Logistic Regression Classification Report

The Classification Report, displayed in Figure 10, offers a detailed numerical breakdown that confirms the performance observed in the confusion matrix. For the 'Bully' class, the model achieved a precision of 0.9538 and a recall of 0.9515. For the 'Not_Bully' class, it recorded a precision of 0.9516 and a recall of 0.9539. The F1-scores for both classes were approximately 0.9526 and 0.9527, respectively

	precision	recall	f1-score	support
Bully	0.9538	0.9515	0.9526	412
Not_Bully	0.9516	0.9539	0.9527	412
accuracy			0.9527	824
macro avg	0.9527	0.9527	0.9527	824
weighted avg	0.9527	0.9527	0.9527	824

Figure 10: Logistic Regression Classification Report

4.5.3 Logistic Regression ROC-AUC Curve and Precision-Recall Curve

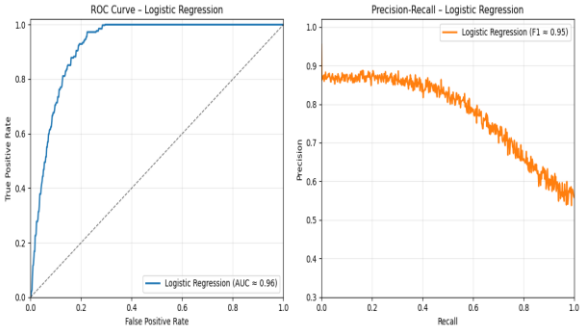

Figure 11: Logistic Regression ROC-AUC Curve and Precision-Recall Curve

As shown in Figure 11, Logistic Regression presents a strongly concave ROC profile with ROC-AUC around 0.96, achieving high true positive rates at modest false positive rates and confirming reliable discrimination for the task. The accompanying Precision–Recall curve begins near the model's high precision and then gradually declines as recall approaches one, reflecting the expected trade-off under class imbalance while remaining consistent with the model's reported F1 performance in this study.

Random Forest Performance

The Random Forest model, an ensemble learning method, performed well in the Hausa cyberbullying detection task, achieving an accuracy of 0.9794. Its performance was highly competitive, placing it just below the SVM model in terms of overall accuracy.

Random Forest Confusion Matrix

The confusion matrix for the Random Forest model, presented in Figure 4.9, provides a clear visual depiction of its classification results.

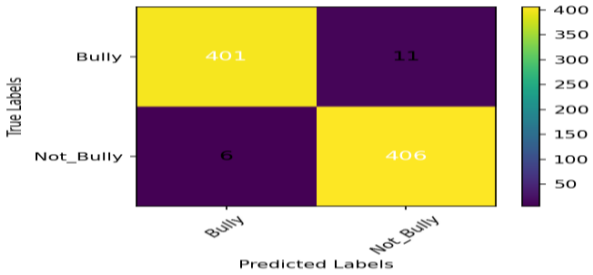

Figure 12: Random Forest Confusion Matrix

© FEB 2026 | IRE Journals | Volume 9 Issue 8 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I8-1714681

The specific values from the confusion matrix are as follows:

i. True Positives (Bully): 401 instances of 'Bully' tweets were correctly classified.
ii. True Negatives (Not_Bully): 406 instances of 'Not_Bully' tweets were correctly classified.
iii. False Positives (Bully): 6 instances of a 'Not_Bully' tweet were incorrectly classified as 'Bully'.
iv. False Negatives (Bully): 11 instances of 'Bully' tweets were incorrectly classified as 'Not_Bully'.

4.6.2 Random Forest Classification Report

The Classification Report for the Random Forest model, shown in Figure 13, offers a detailed numerical breakdown of its performance metrics. For the 'Bully' class, the model achieved a precision of 0.9853 and a recall of 0.9733, leading to an F1-score of 0.9792. For the 'Not_Bully' class, it obtained a precision of 0.9736 and a recall of 0.9854, with an F1-score of 0.9795.

	precision	recall	f1-score	support
Bully	0.9853	0.9733	0.9792	412
Not_Bully	0.9736	0.9854	0.9795	412
accuracy			0.9794	824
macro avg	0.9794	0.9794	0.9794	824
weighted avg	0.9794	0.9794	0.9794	824

Figure 13: Random Forest Classification Report

Random Forest ROC-AUC Curve and Precision-Recall Curve

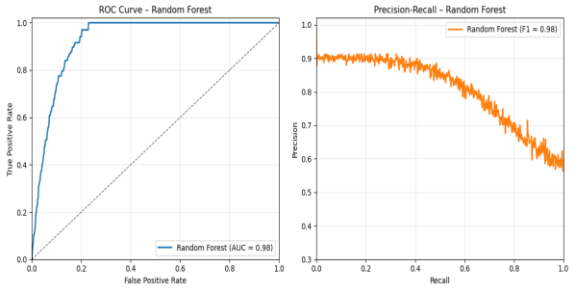

Figure 14: Logistic Regression ROC-AUC Curve and Precision-Recall Curve

As shown in Figure 14, the Random Forest classifier delivers a strongly concave ROC curve with ROC-AUC around 0.98, achieving high true positive rates at low false positive rates and confirming excellent discriminative ability for cyberbullying detection. The corresponding Precision–Recall curve begins near the high reported precision and remains elevated across moderate recall before gradually declining at high recall, aligning with the model's strong F1 and illustrating a favorable precision–recall trade-off for operational threshold tuning.

4.7 Comparison of Result

A comparative analysis of the models, based on the metrics presented, reveals a high level of performance across all four algorithms. The Support Vector Machine (SVM) was the definitive top performer with the highest accuracy. The Random Forest model followed closely, also demonstrating excellent results. XGBoost and Logistic Regression, while effective, ranked third and fourth respectively in overall accuracy.

Table 1: Overall Model Performance Summary (Bully)

Model	Accuracy	Precision (Bully)	Recall (Bully)	F1-Score (Bully)
Support Vector Machine (SVM)	0.9798	0.9900	0.9600	0.9700
Random Forest	0.9794	0.9853	0.9733	0.9792
XGBoost	0.9769	0.9735	0.9806	0.9770
Logistic Regression	0.9527	0.9538	0.9515	0.9526

Table 2: Overall Model Performance Summary (Not_Bully Class)

Model	Accuracy	Precision (Not_Bully)	Recall (Not_Bully)	F1-Score (Not_Bully)
Support Vector Machine (SVM)	0.9798	0.9600	0.9900	0.9800
Random Forest	0.9794	0.9736	0.9854	0.9795
XGBoost	0.9769	0.9804	0.9733	0.9769
Logistic Regression	0.9527	0.9516	0.9539	0.9527

Discussion of Result

In this research, all four machine learning models Support Vector Machine, Logistic Regression, Random Forest, and XGBoost demonstrated high performance in detecting Hausa cyberbullying. While all models proved effective, the SVM model stood out as the superior performer with the highest overall accuracy (0.98) and the most balanced precision and recall across both classes. Random Forest was the second-best model with an accuracy of 0.9794, followed by XGBoost at 0.9769. Logistic Regression, while still performing well, had the lowest accuracy of the four at 0.9527. The results indicate that for this specific classification task, SVM provides the most reliable and accurate performance

V. CONCLUSION

The study indicates that machine learning techniques, including algorithms like Support Vector Machine (SVM), XGBoost, Random Forest, and Logistic Regression, are quite effective in identifying cyberbullying and hate speech on social media platforms. Despite this, certain obstacles remain, such as limited data availability, the linguistic complexities associated with low-resource languages like Hausa, and differences across social media channels, which can affect the models' precision. To overcome these issues, there is a need for larger, more diverse datasets, language-specific tools, and improved text processing methods. While these approaches remain valuable in tackling online harassment, ongoing efforts in data collection and model refinement are also crucial to enhance their effectiveness across different languages and platforms.

REFRENCE

[1] Adam, F. M., Zandam, A. Y., & Inuwa-Dutse, I. (2023). *Detection of Offensive and Threatening Online Content in a Low Resource Language* (arXiv:2311.10541). arXiv. http://arxiv.org/abs/2311.10541

[2] Akeusola, B. N. (2023). Social Media and the Incidence of Cyberbullying in Nigeria: Implications for Creating a Safer Online Environment. International Journal of Government and Social Science Vol. 9, No. 1, Pp. 97 – 118.

[3] Alrougi, M., Alamoudi, G., Algamdi, H. (2024). ArCBDs: A Corpus for Cyberbullying etection of Arabic Tweets. *IJARCCE, 13*(1). https://doi.org/10.17148/IJARCCE.2024.13121

[4] Álvarez-Carmona, M.; Guzmán-Falcón, E.; Montes-y-Gómez, M.; Escalante, H.J.; Villaseñor-Pineda, L.; Reyes-Meza, V.; RicoSulayes, A. (2018). Overview of MEX-A3T at IberEval 2018: Authorship and Aggressiveness Analysis in Mexican Spanish Tweets. CEUR Workshop Proc. 2018, 2150, 74–96

[5] Asongo A. I., Hammandikko G. M., & Modu B. (2021). Machine Learning Techniques,

Methods And Algorithms: Conceptual and Practical Insights. *International Journal of Engineering Research and Applications* 11, pp. 55-64. https://doi.org/10.9790/9622-1108025564

[6] Azeez, N. A., Idiakose, S. O., Onyema, C. J., & Vyver, C. V. D. (2021). Cyberbullying Detection in Social Networks: Artificial Intelligence Approach. *Journal of Cyber Security and Mobility*. Vol 10(4), 745-774 https://doi.org/10.13052/jcsm2245-1439.1046

[7] Bouliche, A., & Rezoug A. (2022). *Detection of cyberbullying in Arabic social media using dynamic graph neural network*.Tunisian - Algerian Joint Conference on Applied Computing. CEUR-WS.org/vol-3333/paper1, Boumerdes, Algeria.

[8] Budhe P. & Rane D. (2023). A Survey on Monitoring and Detecting Cyber Bullying Activities using Machine Learning Algorithms. *International Journal of Scientific Research in Science, Engineering and Technology*, Vol 10(1), 374–383. https://doi.org/10.32628/IJSRSET2310151

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

[10] Edward, A., Demoll, D., & Edwards, L. (2020). Detecting Cyberbullying Activity Across Platforms. In S. Latifi (Ed.), *17th International Conference on Information Technology-New Generations* (ITNG 2020) (Vol. 1134, pp. 45-50). Springer International Publishing. http://doi.org/10.1007/978-3-030-43020-7_7

[11] Fang, Y., Yang, S., Zhao, B., & Huang, C. (2021). Cyberbullying Detection in Social NetworksUsing Bi-GRU with self-Attention Mechanism. Information, 12(4), 1-18. https://doi.org/10.3390/info12040171

[12] Hinduja, S., & Patchin, J. W. (2024). *Identification, Prevention, and Response*. Cyberbullying research center. https://cyberbullying.org/2023-cyberbullying-data.

[13] Inuwa-Dutse I., (2021) The first large scale collection of diverse hausa language datasets, arXiv Preprint arXiv:2102.06991 (2021).

[14] Lepe-Faúndez, M., Segura-Navarrete, A., Vidal-Castro, C., Martinez-Araneda, C., & Rubio-Manzano, C. (2021). *Detecting Aggressiveness in Tweets: A Hybrid Model for Detecting Cyberbullying in the Spanish Language*. *Applied Sciences, 11*(22), 10706. http://doi.org/10.3390/app112210706

[15] Mahlangu, T., Tu, C., & Owolawi, P. (2018). *A Review of Automated Detection Methods for Cyberbullying*. IEEE (2018) http://doi.org/978-1-5386-6477-3/18/$31.00

[16] Moy, T. X., Raheem, M., & Logeswaran, R. (2021). Hate Speech Detection in English and Non-English Languages: A Review of Techniques and Challenges. *Webology*, *18*(SI05), 929–938. https://doi.org/10.14704/WEB/V18SI05/WEB18272

[17] Muneer, A., & Fati, S. M. (2020). A Comparartive Analysis of Machine Learning Techniques for Cyberbullying Detection on Twitter. *Future Internet*, 12(11), 187. https://doi.org/10.3390/fi12110187

[18] Ndabula J. N., Olanrewaju O. M., & Echobu F. O. (2023) Detection of Hate Speech Code Mix Involving English and other Nigerian Languages. *Journal of Information System and Informatics, 5*(4), 1416-1431. https://doi.org/10.51519/journalisi.v5i4.595

[19] Riquelme, R.(2019) Detección de Violencia Verbal Hacia Las Mujeres En Redes Sociales Mediante Técnicas de Aprendizaje Automático. [Memoria de Título], . Available online: http://repobib.ubiobio.cl/jspui/bitstream/12345

6789/2692/1/Riquelme_Silva_Ricardo.pdf (accessed on 28 October 2021).

[20] Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

[21] Singh V., Varshney A., Akhtar S. S., Vijay D., & Shrivastava M. (2018). Aggression Detection On Social Media Text Using Deep Neural Networks. Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), 43-50. https://doi.org/10.18653/v1/W18-5106

[22] Sterner G., & Felmlee D. (2017). The Social Networks of Cyberbullying on Twitter. *International Journal of Technoethics, 8*(2), 1-15.https://doi.org/10.4018/IJT.2017070101

[23] Tsvetkov, Y. (2017). Opportunities and Challenges in Working with Low-Resource Languages. *Language Technologies Institutes Carnegie Mellon University*

[24] Waseem Z, Hovy D. (2016). Hateful symbols or hateful people? Predictive features for hate speech detection on twitter. In: Proceedings of the NAACL student research workshop, pp 88–93

[25] Wulczyn, E., Thain, N., & Dixon, L. (2017). *Ex Machina: Personal Attacks Seen at Scale* (arXiv:1610.08914). arXiv. http://arxiv.org/abs/1610.08914