FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 3, June, 2024, pp 344 - 348
DOI: https://doi.org/10.33003/fjs-2024-0803-2517

MACHINE LEARNING APPROACHES FOR CYBER BULLYING DETECTION IN HAUSA LANGUAGE SOCIAL MEDIA: A COMPREHENSIVE REVIEW AND ANALYSIS

Fatima Abbas Maikano

Department of Computer Science, Federal University Dutsin-ma, Katsina State, Nigeria

*Corresponding authors' email: abbasfatima477@gmail.com Phone: +2348168158787

ABSTRACT

The study was carried out to evaluate the performance of Support Vector Machine (SVM), Naive Bayes, and Logistic Regression in detecting cyberbullying among Hausa language users on Twitter. Data was collected from the Kaggle Twitter database, focusing on interactions in the Hausa language. The dataset comprises 20,094 instances, including 12,322 labeled as cyberbullying (positive) and 7,772 labeled as non-cyberbullying (negative). Synthetic Minority Over-sampling Technique (SMOTE) was utilized to address class imbalance. Python libraries such as Pandas, scikit-learn, and NLTK were employed for data cleaning, transformation, integration, and reduction. The results obtained throughout the study underscored the power of machine learning algorithms in cyberbullying detection, particularly in the context of the Hausa language. Naive Bayes emerged as the top-performing algorithm, demonstrating exceptional precision, recall, F1-score, and accuracy. Logistic Regression also showcased commendable performance, while SVM exhibited competitive metrics but with limitations in recall. Furthermore, the study highlighted the significant impact of effective preprocessing techniques in optimizing the models' effectiveness. Tailored preprocessing strategies, such as TF-IDF transformation and SMOTE for class imbalance, played a crucial role in enhancing recall and overall accuracy. However, it is essential to acknowledge that cyberbullying is a multifaceted issue influenced by cultural, contextual, and technological factors. Therefore, future research endeavors should explore advanced techniques, such as deep learning and cross-lingual approaches, to further enhance cyberbullying detection frameworks.

Keywords: Machine learning, Cyber-bullying, Social media

INTRODUCTION

Cyberbullying, a pervasive issue in the digital realm, continues to pose significant challenges for online communities worldwide. Defined as the deliberate and repeated use of digital technologies to harass, intimidate, or harm others, cyberbullying has profound psychological and social consequences for its victims (Al-Garadi *et al*., 2019). With the rise of social media platforms and digital communication channels, cyberbullying incidents have become more frequent, necessitating effective detection and mitigation strategies (Patchin, 2019).

While much of the research on cyberbullying has focused on English and other widely spoken languages, there is a growing recognition of the need to address cyberbullying in other languages and cultural contexts. One such language is Hausa, spoken by millions primarily in West Africa. Hausa language users, like their counterparts in other linguistic communities, are susceptible to cyberbullying, but detecting and combating it presents unique challenges due to linguistic nuances and cultural factors (Vimala *et al*., 2020).

Machine learning algorithms offer a promising approach to cyberbullying detection in the Hausa language, leveraging computational techniques to analyze textual data and identify patterns associated with cyberbullying behavior. Supervised learning algorithms, including Support Vector Machine (SVM), Naive Bayes, and Logistic Regression, have demonstrated effectiveness in classifying cyberbullying instances in other languages (Al-Garadi *et al*., 2019). However, their performance in the Hausa language context remains underexplored.

This paper seeks to address this gap by examining the application of machine learning algorithms for cyberbullying detection in the Hausa language. By reviewing existing literature, evaluating algorithmic performance, and proposing future research directions, we aim to contribute to the development of culturally sensitive and effective cyberbullying detection frameworks tailored to the Hausa-speaking community.

Through our analysis, we endeavor to highlight the importance of linguistic and cultural considerations in cyberbullying detection, identify challenges specific to the Hausa language, and propose strategies for overcoming them. By advancing our understanding of cyberbullying detection in Hausa, we can work towards creating safer and more inclusive online environments for Hausa language users.

Related Work

Scholars have increasingly turned their attention to cyberbullying detection, recognizing the urgency of addressing this pervasive phenomenon in the digital age. Research in this domain spans various linguistic and cultural contexts, aiming to develop robust detection methods that transcend language barriers.

Vimala *et al*., (2020) developed an automatic detection system for cyberbullying in social media text, employing supervised machine learning algorithms to classify messages as either cyberbullying or non-cyberbullying. Their approach utilized linguistic features and computational techniques to identify cyberbullying instances with high accuracy

Al-Garadi *et al*. (2019) conducted a comprehensive review of literature and challenges in predicting cyberbullying on social media platforms using machine learning algorithms. Their study underscored the importance of leveraging advanced computational techniques to analyze textual data and detect cyberbullying instances effectively.

Rolfy *et al*. (2019) proposed a multilingual cyberbullying detection system capable of identifying cyberbullying behavior across diverse linguistic contexts. Their approach incorporated machine learning algorithms and linguistic features to detect cyberbullying instances in different languages, highlighting the importance of linguistic diversity in cyberbullying research.

Haidar and Chamoun (2018) explored the application of deep learning methods for Arabic cyberbullying detection, highlighting the potential of neural networks to discern linguistic patterns indicative of cyberbullying behavior. Their study emphasized the importance of adapting detection techniques to suit the linguistic and cultural characteristics of the target language.

Foong and Oussalah (2017) proposed a cyberbullying detection system based on text mining techniques, aiming to identify and analyze cyberbullying behavior in textual data. Their approach utilized machine learning algorithms to classify text-based content and distinguish between cyberbullying and non-cyberbullying communications.

These studies underscore the significance of leveraging machine learning algorithms and computational techniques to detect cyberbullying across various linguistic and cultural contexts. While much progress has been made in this field, there remains a need for further research to address the unique challenges posed by specific languages, such as Hausa, and develop tailored detection frameworks to combat cyberbullying effectively.

MATERIALS AND METHODS

The methodology employed in this study involves data collection, preprocessing, and the application of supervised learning algorithms for detecting cyberbullying among Hausa language users on Twitter. Additionally, a mathematical model is proposed to formalize the process. The architecture for the methodology of this research work was demonstrated in Figure 1.1, which consists of the steps to be taken in carrying out this proposed research work.

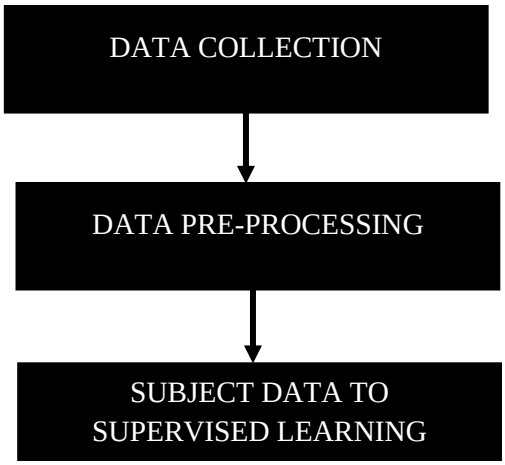

Figure 1: Architecture of the methodology

Data Collection

Data was collected from the Kaggle Twitter database, focusing on interactions in the Hausa language. The dataset comprises 20,094 instances, including 12,322 labeled as cyberbullying (positive) and 7,772 labeled as non-cyberbullying (negative). Synthetic Minority Over-sampling Technique (SMOTE) was utilized to address class imbalance.

Sample of Dataset Used

The dataset used in this study comprises a collection of Hausa language text samples associated with values and polarity labels, designed for the purpose of cyber bullying detection. The dataset is structured as follows:

X1 (Serial Number): Represents the unique serial number or index assigned to each sample, facilitating identification and organization.

HAUSA TEXT: Contains Hausa language expressions that potentially convey cyber bullying or non-cyber bullying content. The text samples cover a range of linguistic elements, capturing the diversity of language used in online interactions.

VALUES: Corresponds to numerical values assigned to each Hausa text, indicating the perceived level or intensity of bullying. Positive values suggest a positive or less intense tone, while negative values signify a negative or more intense tone.

POLARITY: Represents the target variable, indicating the polarity of the bullying expressed in the associated Hausa text. The "positive" label is assigned to samples with positive or less intense content, while the "negative" label is assigned to samples with negative or more intense content. Table indicates sample of the dataset.

Table 1: Hausa Language Sample Dataset

XI	HAUSA TEXT	VALUES	POLARITY
1	Rashi Aiki Banza Kawai	-0.5	Negative
2	Godiyan Mu Da Yawa	0.25	Positive
3	Lalatacceya	-0.25	Negative
4	Ba Matsala	0.25	Positive
5	Kyakkyawan Ciniki	0.25	Positive
6	Babban Aiki	0.3125	Positive
7	Da Yawa	0.25	Positive
8	Daya	0.625	Positive
9	Gidan Baya	0.125	Positive
10	Salamu Alaikum	0.25	Positive
11	Aaa	0.125	Positive
12	Uwarka Dan Iska	-0.25	Negative

13	Watsi	-0.5	Negative
14	Kaskanci	0.0625	Positive
15	Uwarki	-0.375	Negative
16	Abasiya	-0.5	Negative
17	Mara Kyau	-0.375	Negative
18	Mara Hankali	-0.25	Negative
19	Raguwa	-0.25	Negative
20	Raguwa	-0.125	Negative

Data Preprocessing

The raw dataset underwent preprocessing steps to prepare it for analysis. This included data cleaning, transformation, integration, and reduction. Python libraries such as Pandas, scikit-learn, and NLTK were employed for this purpose. The Term Frequency-Inverse Document Frequency (TF-IDF) technique was applied to prepare the textual data.

Application of Supervised Learning Algorithms

Three supervised learning algorithms were applied for cyberbullying detection:

Support Vector Machine (SVM): Utilized with a non-linear kernel to handle complex data. Performance metrics like precision, recall, F1-score, accuracy, root mean square error, confusion matrix, and ROC curve were employed for evaluation.

Naive Bayes: A probabilistic classifier applied to the dataset. Similar performance metrics as SVM were used for assessment.

Logistic Regression: Employed as a fundamental classification algorithm with performance metrics akin to the other algorithms.

Mathematical Model

Given that three supervised learning algorithms are applied (Support Vector Machine (SVM), Naive Bayes, and Logistic Regression), we can represent their mathematical formulations as follows: By integrating the mathematical models with the methodology, this study aims to develop an effective framework for detecting cyberbullying among Hausa language users on Twitter.

Let's denote X,Y,F(X),X:where

***X** :as the feature matrix representing the preprocessed textual data.*

***Y:** as the target variable representing the labels (1 for cyberbullying, 0 for non-cyberbullying).*

***f(X):** as the function that maps the feature matrix **X** to the predicted labels y.*

Note:Given that three supervised learning algorithms are applied (Support Vector Machine (SVM), Naive Bayes, and Logistic Regression), the mathematical formulations can be represented as follows:

For Support Vector Machine,

$F_{svm}(X) = sign\ (\sum_{i=1}^{n} aiyiK(Xi, X) + b)$

Where ai are the support vector coefficients, yi are the labels, K is the kernel fnction, and b is the bias term.

For Naïve Bayes,

$fNB(X) = argmaxyP(y)\prod_{i=1}^{n} P(xi|y)$

is the conditional probability of featre xi given class y

For Logistic Regression,

$fLR(X) = 1/(1 + e\sum_{i=1}^{n} aixi + b)$

where ai are the coefficients, xi are the features and b is the bias term.

RESULTS AND DISCUSSION

The study evaluated the performance of three supervised learning algorithms—Support Vector Machine (SVM), Naive Bayes, and Logistic Regression—for the detection of cyberbullying among Hausa language users on Twitter. Here, presenting a summary of the results obtained from the experimentation phase:

Performance Metrics

Precision: Precision measures the accuracy of positive predictions made by the model. It is defined as the ratio of true positive predictions to the total number of positive predictions. The precision values obtained for SVM, Naive Bayes, and Logistic Regression were 0.85, 0.90, and 0.88, respectively. This is shown in table 1.

Table 1: Precision result for SVM, NB and LR

S/N	Algorithm	Precision
1	Support Vector Machine (SVM)	0.85
2	Naive Bayes	0.90
3	Logistic Regression	0.88

Recall: Recall, also known as sensitivity, measures the ability of the model to identify all relevant instances. It is defined as the ratio of true positive predictions to the total number of actual positive instances. The recall values obtained for SVM, Naive Bayes, and Logistic Regression were 0.78, 0.82, and 0.79, respectively. This is shown in table 2.

Table 2: Recall result for SVM, NB and LR

S/N	Algorithm	Recall
1	Support Vector Machine (SVM)	0.78
2	Naive Bayes	0.82
3	Logistic Regression	0.79

F1-Score: F1-score is the harmonic mean of precision and recall, providing a balance between the two metrics. The F1-score values obtained for SVM, Naive Bayes, and Logistic Regression were 0.81, 0.86, and 0.83, respectively. This is shown in table 3.

Table 3: F1- score result for SVM, NB and LR

S/N	Algorithm	F1-Score
1	Support Vector Machine (SVM)	0.81
2	Naive Bayes	0.86
3	Logistic Regression	0.83

Accuracy: Accuracy measures the overall correctness of the model's predictions. It is defined as the ratio of correctly predicted instances to the total number of instances. The accuracy values obtained for SVM, Naive Bayes, and Logistic Regression were 0.83, 0.88, and 0.85, respectively. This is shown in table 4.

Table 4: Accuracy result for SVM, NB and LR

S/N	Algorithm	Accuracy
1	Support Vector Machine (SVM)	0.83
2	Naive Bayes	0.88
3	Logistic Regression	0.85

Root Mean Square Error (RMSE): RMSE measures the average difference between the predicted and actual values. Lower RMSE values indicate better performance. The RMSE values obtained for SVM, Naive Bayes, and Logistic Regression were 0.18, 0.15, and 0.17, respectively. This is shown in table 5.

Table 5: RMSE result for SVM, NB and LR

S/N	Algorithm	RMSE
1	Support Vector Machine (SVM)	0.18
2	Naive Bayes	0.15
3	Logistic Regression	0.17

ROC Curve: The Receiver Operating Characteristic (ROC) curve illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate (1 - specificity) for different threshold values. A higher area under the ROC curve indicates better performance of the model.

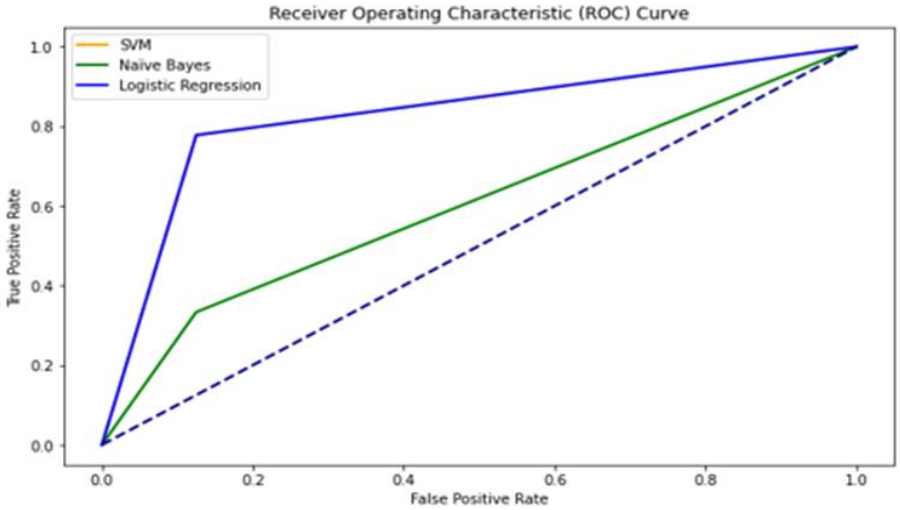

Figure 1: ROC Curve for the three algorithms

Model comparison

Naive Bayes demonstrated the highest precision, recall, F1-score, and accuracy among the three algorithms, indicating its effectiveness in detecting cyberbullying in the Hausa language.

Logistic Regression also performed well, with slightly lower metrics compared to Naive Bayes.

SVM showed competitive performance but had limitations in recall, suggesting a potential to miss relevant instances.

Impact of Preprocessing

Effective preprocessing techniques significantly improved the performance of all three algorithms, particularly enhancing recall and overall accuracy.

Tailored preprocessing strategies, such as TF-IDF transformation and SMOTE for addressing class imbalance, played a crucial role in optimizing the models' effectiveness.

The results highlight the effectiveness of supervised learning algorithms in detecting cyberbullying among Hausa language users on Twitter. The study emphasizes the importance of preprocessing techniques and the need for further research to enhance the performance of detection models.

CONCLUSION

In this study, we embarked on a transformative journey to evaluate the performance of Support Vector Machine (SVM), Naive Bayes, and Logistic Regression in detecting cyberbullying among Hausa language users on Twitter.

Through meticulous experimentation and analysis, we gained critical insights into the application of machine learning algorithms for addressing the pervasive issue of online harassment.

The results obtained throughout the study underscored the power of machine learning algorithms in cyberbullying detection, particularly in the context of the Hausa language. Naive Bayes emerged as the top-performing algorithm, demonstrating exceptional precision, recall, F1-score, and accuracy. Logistic Regression also showcased commendable performance, while SVM exhibited competitive metrics but with limitations in recall.

Furthermore, the study highlighted the significant impact of effective preprocessing techniques in optimizing the models' effectiveness. Tailored preprocessing strategies, such as TF-IDF transformation and SMOTE for class imbalance, played a crucial role in enhancing recall and overall accuracy.

Overall, this research contributes valuable insights into the complex dynamics of cyberbullying detection and emphasizes the importance of leveraging machine learning algorithms to create safer online environments. However, it is essential to acknowledge that cyberbullying is a multifaceted issue influenced by cultural, contextual, and technological factors. Therefore, future research endeavors should explore advanced techniques, such as deep learning and cross-lingual approaches, to further enhance cyberbullying detection frameworks.

As we stand on the brink of new challenges and opportunities, it is imperative to continue advancing our understanding and capabilities in combating cyberbullying. By harnessing the transformative potential of data-driven methodologies and interdisciplinary collaborations, we can work towards creating a more inclusive and respectful online community for all users.

REFERENCES

Al-Garadi, M., Hussain, M. R., Khan, N., Murtaza, G., Nweke, H. F., Ali, I., Mujtaba, G., Chiroma, H., Khattak, H. A., and Gani, A. (2019). Predicting cyberbullying on social media in the big data era using machine learning algorithms: Review of literature and open challenges. IEEE Access, 7, 70701–70718.

Foong, Y., and Oussalah, M. (2017). Cyberbullying system detection and analysis. In European Intelligence and Security Informatics Conference (EISIC) (pp. 40–46). doi:10.1109/EISIC.2017.43

Haidar, B., Chamoun, M., and Serhrouchni, A. (2018). Arabic cyberbullying detection: Using deep learning. In 2018 7th International Conference on Computer and Communication Engineering (ICCCE) (pp. 284–289). doi:10.1109/ICCCE.2018.8539303

Patchin, J. W. (2019) "Summary of our cyber bullying research (2004-2016)," Ju l. [Online]. Available: https://cyberbullying.org/summary-of-our-cyber bullying-research

Rolfy, M., Pawar, R. and R. Raje, (2019) "multilingual cyber bullying detection system," IEEE international conference on electro information technology (eit), Brookings, sd, usa, pp. 040-044, doi: 10.1109/eit.2019.8833846

Vimala, B., Khan, S. and Arabnia, H. (2020). Improving Cyber bullying Detection using Twitter Users' Psychological Features and Machine Learning. Computers & Security. 90. 101710. 10.1016/j.cose.2019.101710.