Tanzania Journal of Science
Volume 51(3), 2025
journal homepage: https://commons.udsm.ac.tz/tjs/

A comparative study for classical machine learning models for swahili social media sentiment analysis

Mahadia Tunga[1*] and Davis David[2]

[1]*Department of Computer Science and Engineering, College of Information and Communication Technologies, University of Dar es Salaam, Dar es Salaam, Tanzania.* [2]*Tanzania Data Lab Organization (dLab), Dar es Salaam, Tanzania*

Keywords

Low-resource Language; Swahili Corpus; Natural Language Processing; Hyperparameter Optimization Technique

Abstract

Despite sentiment analysis being one of the most popular applications in Natural Language Processing (NLP), most studies are skewed towards languages with a rich corpus (language database). Less emphasis has been placed on low-resource languages like Swahili. Swahili is the official language of the African Union and of 4 countries in East Africa, and is spoken by many people on the African continent. This study performed sentiment analysis using 3,000 tweets hosted on the Zindi Africa platform. Data was processed using a term frequency-inverse document frequency vectorization method, and five classical machine learning algorithms (RandomForest, XgBoost, and CatBoost, HistogramGradientBoost, LightGradientBoos) were trained and evaluated using the collected tweets. We found that CatBoost produced the highest performance in general compared to other classical models, with 0.610 accuracy, 0.470 F1 score, 0.522 Precision and 0.462 Recall. The F1-score of 0.47 indicates modest performance and reflects the challenges posed by the small dataset and the complexity of Swahili sentiment analysis. This study offers a comprehensive overview of the relative performance of various classical machine learning models applied to Swahili social media sentiment data. These insights can help researchers make informed choices when selecting appropriate classical machine learning algorithms for sentiment analysis in a similar context.

Introduction

Sentiment analysis has gained much attention in recent years. Sentiment analysis is a subfield of Natural Language Processing (NLP) that deals with determining the emotional tone behind a piece of text, whether it is positive, negative, or neutral (Zhang et al. 2023). Sentiment analysis can be used in a variety of applications, including social media. Social media sentiment analysis can help businesses and organizations understand how Swahili speakers feel about their products, services, or brands and, consequently, make informed decisions.

Despite sentiment analysis being one of the most popular applications in NLP, most studies are skewed towards languages with rich language databases (Muhammad et al. 2022). Less emphasis has been placed on low-resource languages which have limited

* Corresponding editor: *mahadiatunga@udsm.ac.tz
Received 28 May 2025; Revised 17 Sep 2025; Accepted 9 October 2025; Published 30 October 2025
https://doi.org/10.65085/2507-7961.1051

ISSN 0856-1761, e-ISSN 2507-7961

data availability recorded in the language databases, like Swahili, an official language of the African Union and is spoken by 4 countries in East Africa including DRC, Kenya, Tanzania and Uganda (Chonka et al. 2023).

There are various approaches to sentiment analysis, including rule-based and machine-learning methods. Rule-based methods rely on a predefined set of rules to identify sentiment (Berka 2020). Thus, rule-based methods may be prone to overfitting, particularly if the dataset used to create the rules is not diverse or representative of the population (Kamal 2013). Machine Learning (ML) based methods use statistical algorithms to learn patterns in text data and make predictions about the sentiment expressed. Thus, can generalize well from the training data and can handle unseen data, making them well-suited for sentiment analysis tasks (Kursa and Rudnicki 2010, Ibrahim et al. 2020, Tanha et al. 2020). Also, Machine learning models can be trained on different datasets, enabling them to adapt to different writing styles, languages, and domains (Nguyen et al. 2023).

Standard sentiment analysis subtasks, such as polarity classification (positive, negative, neutral), are widely considered saturated and solved, with an accuracy of over 90% in certain languages and the focus has been on the state-of-the-art Deep learning techniques such as transformers and multilingual models (Muhammad et al. 2022). But so many challenges exist to date in low-resource African languages such as Swahili. Additionally, African languages present other difficulties for sentiment analysis, such as handling tone, code-switching, and digraphia (Adebara and Abdul-Mageed 2022). Existing work in sentiment analysis for African languages has therefore mainly focused on polarity classification to improve performance (Muhammad et al. 2022, Adebara and Abdul-Mageed 2022).

This paper presents a comparison of classical ML algorithms for Swahili sentiment analysis. Classical ML algorithms are a set of established and traditional techniques that were developed before the emergence of deep learning algorithms. Classical ML algorithms are used for various tasks such as classification, regression and clustering. Classical ML algorithms provide more interpretable results and are efficient with limited data compared to deep learning models (Gao and Guan 2023). The algorithm was selected after performing a series of ML experiments on the five classical ML algorithms: RandomForest, XgBoost, and CatBoost, HistogramGradientBoost, LightGradientBoost, using a dataset consisting of Swahili tweets. These five classical ML algorithms were selected based on their performance when trained on limited data (Zhang et al. 2023). Preprocessing techniques such as the TF-IDF (Term Frequency-Inverse Document Frequency) vectorization method and over-sampling method were applied to improve the quality of the dataset.

This paper aims to identify the most suitable classical ML model for social media sentiment analysis that will be capable of performing analysis under low resource constraints. The goal is to benchmark classical ML methods on Swahili data and demonstrate both challenges and the suitability of classical ML models for small datasets. The significance of this study lies in the potential to use classical ML algorithms to automatically comprehend Swahili public opinion on a variety of topics and make well-informed decisions based on social media sentiment. This study can be especially useful for organizations seeking an approach to assess consumer perception as well as for governments to monitor public opinion on particular policies. The study also provides future directions and recommendations.

Literature Review

Text Vectorization Methods

Text vectorization is a crucial step in the ML process when working with text data. It involves converting text data into numerical representations that can be used as input for ML models (Shahmirzadi et al. 2019). Count vectorization and TF-IDF (Term Frequency-Inverse Document Frequency) vectorisation are the two methods commonly used to

convert text data into numerical representations.

Count vectorization, also known as a bag of words, is a method of converting text data into numerical representations by counting the occurrences of each word in a document (Zhou 2021). This method creates a sparse matrix where each column represents a word, and each row represents a document. This method has been widely used in natural language processing tasks such as text classification and sentiment analysis (Das and Chakraborty 2018). Hence, this method can be applied to a large corpus of text data, making it highly scalable. However, Count vectorization only considers the frequency of words and does not capture semantic relationships between words, leading to a loss of context (Zhou 2021).

TF-IDF vectorization, short for the term frequency-inverse document frequency, is a method of converting text data into numerical representations by weighting the occurrences of each word in a document (Padurariu and Breaban 2019). This method also creates a sparse matrix, where the values in each cell represent the TF-IDF score of that word in the corresponding document. The TF-IDF score is calculated by multiplying the term frequency (***tf***) of a word in a document by the inverse document frequency (***idf***) of that word in the entire corpus, which, as a result, improves the performance of ML models in text classification tasks (Das and Chakraborty 2018).

Methods for Handling the Data Imbalance

An imbalance of data can affect the performance of an ML model (Kim and Kim 2018). For instance, when an ML model is trained on data with a target that has an unequal distribution of classes (imbalance), it can perform well on the majority classes but poorly on the minority classes presented in the same data (Padurariu and Breaban 2019). This could lead to the model's poor performance when deployed in real-world settings. Oversampling and Undersampling are common methods used to address the problem of dataset imbalance.

Oversampling methods tend to generate more data on the minority class to increase its representation in the dataset, as presented in Figure 1. This creates a balanced dataset, which contains an equal distribution of classes. A study by (Le 2022) found that the equal distribution of classes in Oversampling methods improves learning about the sample distribution in an efficient manner. Common oversampling method are random oversampling, SMOTE (Synthetic Minority Over-sampling Technique), and ADASYN (Adaptive Synthetic Sampling) (Haluska et al. 2022). SMOTE, which generates synthetic samples using interpolation between minority class instances, is widely used due to its simplicity and effectiveness. However, it has limitations, such as amplifying noise, generating unrealistic samples, and failing to capture the complex structures of minority classes. Despite many advanced SMOTE variants such as Borderline SMOTE, Safe-level SMOTE, ADASYN, SVM SMOTE, CDSMOTE, and Deep SMOTE, standard SMOTE remains popular for its competitive performance and efficiency (Rawat and Mishra 2024), which is why it was used in this study.

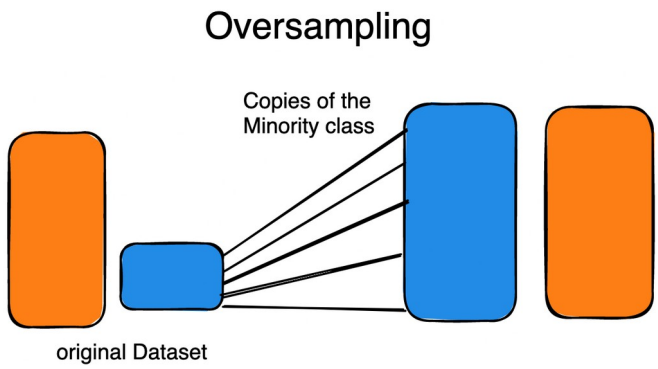

Figure 1: Oversampling Technique.

Undersampling methods tend to reduce the number of instances of the majority classes in a dataset, as presented in Figure 2. This can be done by randomly removing instances from the majority classes until the class distribution is balanced, or by using a specific algorithm to select a subset of instances to remove (Haluska et al. 2022). Thus, undersampling methods can lead to a loss of information (Rawat and Mishra 2024). In general, oversampling methods outperform undersampling because they balance the dataset without discarding valuable majority class data, resulting in improved model performance (Haluska et al. 2022).

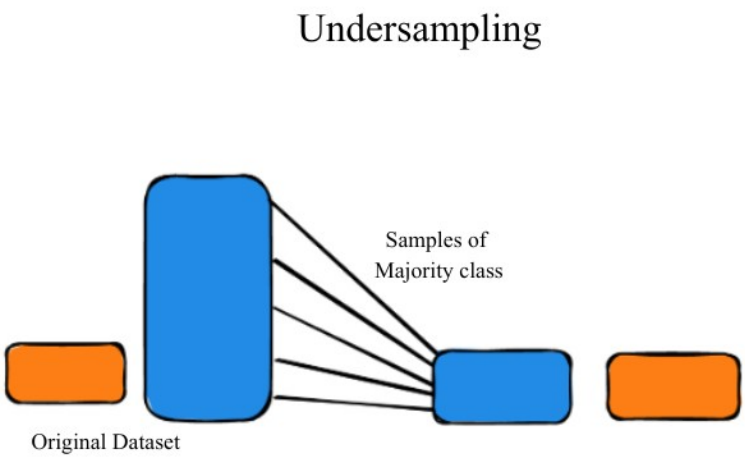

Figure 2:Undersampling Technique.

Classical Machine Learning Models

This section describes common classical ML algorithms: Random Forest, XGBoost, CatBoost, LightGradientBoost and HistogramGradientBoost for text classification and sentiment analysis.

Random Forest

Random Forest is an ensemble method that combines multiple decision trees to improve the accuracy and robustness of the predictions. Random Forest has the ability to utilise a combination of unigrams (single words), bigrams (pairs of words), and trigrams (triplets of words) as input features, and it can effectively handle high-dimensional data, making it particularly valuable in low-resource settings (Kursa and Rudnicki 2010; Khan et al. 2024b). However, Random Forest has limitations in handling data sparsity and informal usage, such as spelling variations, making it less effective for low-resource languages than context-aware models like BERT. But deep learning models like BERT need substantial labelled data for tasks like sentiment analysis or classification. In low-resource settings, annotated datasets are scarce or costly to produce, which limits BERT's performance and often leads to overfitting (Khan et al. 2024b).

Xgboost

The XGBoost algorithm, also known as Extreme Gradient Boosting, is an ML technique based on gradient boosting, which

combines several weak learners to form a strong ensemble model (Zhang et al. 2025). In low-resource languages like Tamil, Malay and Swahili, where large annotated datasets especially for tasks like speech or sentiment analysis are scarce, XGBoost handles limited data well, manages high-dimensional features, and captures complex relationships compared to training deep learning approaches. (Zhang et al. 2023; Zhang et al. 2025; Khan et al. 2024b). However, XGBoost is known to be prone to overfitting, especially when the data is noisy (Bentéjac et al. 2021). This can lead to poor generalization performance on unseen data.

CatBoost

CatBoost is a gradient boosting algorithm, specifically designed to handle categorical variables effectively. It performs well on small datasets and is particularly suitable for high-dimensional sparse data, such as text (Ibrahim et al. 2020). Its built-in support for handling class imbalances through the class_weights parameter often allows it to outperform many other machine learning algorithms in supervised learning tasks involving imbalanced data (Ibrahim et al. 2020). Additionally, CatBoost requires less manual feature engineering, which is especially valuable in low-resource settings. Its ordered boosting and Bayesian averaging further enhance performance when data is limited. However, optimizing model performance through parameter tuning can be challenging (Tanha et al. 2020). Parameter tuning is the process of selecting the best value for ML model's hyperparameters. This task is challenging due to complex hyperparameters ineteractions, large search spaces and high risk of overfitting (Tanha et al. 2020).

LightGradientBoost

LightGradientBoost is a ML algorithm that is based on the popular Gradient Boosting algorithm. Gradient boosting is effective for low-resource languages like Swahili because it can handle high-dimensional and imbalanced data without relying heavily on extensive NLP tools, which are often unavailable for low-resource languages (Khan et al. 2024a; Vitorino et al. 2023]. The LightGradientBoost algorithm works by building a series of decision trees, each of which is trained to correct the errors made by the previous tree. Hence, the LightGradientBoost can significantly outperform other machine-learning algorithms in terms of computational speed and memory consumption (Ke et al 2017). These benefits make it particularly appealing for low-resource language tasks, where limited computational resources and data availability pose significant challenges.

HistogramGradientBoost

HistogramGradientBoost is an algorithm that is used for gradient boosting in ML. It is a variation of the traditional gradient-boosting algorithm that uses histograms to represent the input data (Hang et al 2021). It divides the input data into small bins or intervals and creates a histogram to represent the data in each bin. Thus, the algorithm then uses these histograms to build a decision tree that can be used to make predictions. The algorithm can handle high-dimensional data, such as text data, which can be difficult to classify using traditional methods (Huyut et al. 2022). It also tends to be more robust to overfitting than other algorithms, as the histograms used as input are less prone to noise and outliers (Kim and Kim 2018). These features are particularly valuable in low-resource language settings, where data is often sparse and noisy. However, it can be computationally expensive, as the histograms need to be created and the gradient boosting classifier needs to be trained (Fan et al. 2024).

Methodology

Data Collection

This study used secondary data hosted on the Zindi Africa platform. The data was originally collected from Twitter by East Africa Zindi ambassadors. The dataset had a size of 3000 tweets before performing cleaning and preprocessing. The collected tweets were manually annotated using an overall polarity: positive (1), negative (-1) and neutral (0) as presented in Table 1.

Table 1: Distribution of Sentiment categories.

Category	Number of Tweets	Sample
Neutral (0)	1,765	telecom3 samahani wapendwa hiv inawezekanaje mtu kushwap laini mtu anaitumia
Positive (1)	913	telecom3tz hongereni mtandao kwanza kuongoza zoezi usajili alama vidole
Negative (-1)	322	mtandao telecom1 una tatizo simu siku tatu leo mambo hayasongi vizuri

Data Cleaning and Preprocessing

The Swahili tweets were cleaned by converting all characters to lowercase and removing punctuation, special characters, numbers, hashtags, emojis and URLs. This basic cleaning helped standardize the input and reduce irrelevant noise for the machine learning models. The TF-IDF text vectorization method was used to convert the cleaned Swahili tweets into a numerical representation by computing the Term Frequency (***TF***) and Inverse Document Frequency (***IDF***) of each token (word) in every tweet written in Swahili within the corpus. The method produced a transformed dataset that has 3,000 rows of data and 12,679 features in numerical representation.

Machine Learning Experiments

Multiple ML experiments were performed to find the ML algorithms for sentiment analysis that produce the best results.

All experiments were conducted on a Dell XPS laptop running Windows 10, equipped with an Intel Core i7 processor and 8 GB of RAM. In the first experiment, we use the default dataset with 3000 rows of sentiment data, which contains an unequal distribution of the target features (class imbalance). We trained five different ML algorithms, such as Random Forest, Xgboost, Catboost, HistogramGradientBoost and LightGradientBoost.

The ML training process was implemented using the cross-validation technique to avoid overfitting as presented in Figure 3. Cross-validation technique was used to evaluate the performance of the developed sentiment analysis model. In each iteration, involved dividing the dataset into five subsets, training the model on four subsets and testing it on one subset. The goal of cross-validation was to provide a more robust evaluation of the model's performance for Swahili sentiment analysis, as it accounts for the uncertainty caused by the specific data used for training and testing. It also helped to prevent overfitting by assessing the model's ability to generalize to new data.

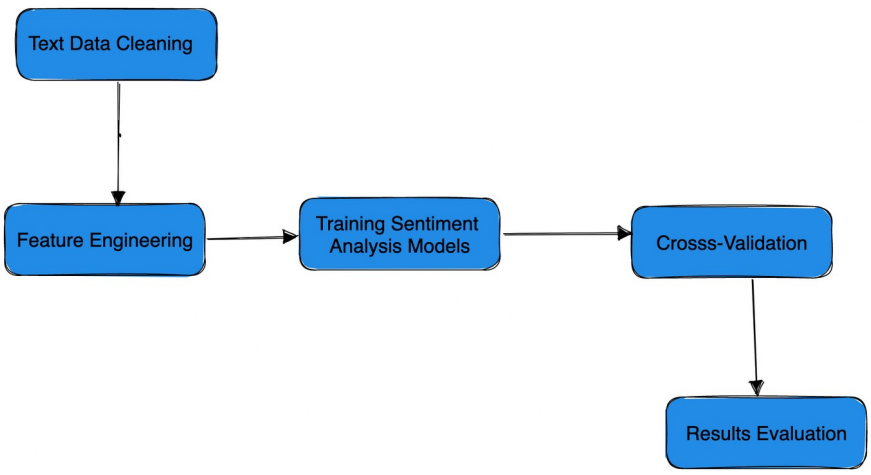

Figure 3: Development of the Sentiment analysis model and results.

In the second ML experiment, the first oversampling method, called random oversampling, was applied to generate more data on the preprocessed dataset in order to balance the number of classes on the target column from 3,000 samples to 4,236 samples. The random oversampling method generated new samples by randomly sampling with replacement from the current available samples. The final dataset was trained on the same ML algorithms for sentiment analysis using cross-validation technique with five folds.

In the third ML experiment, SMOTE (Synthetic Minority Oversampling Technique) was applied to generate more data on the preprocessed dataset, thereby balancing the number of classes in the target column. SMOTE balanced the class distribution by randomly replicating minority class examples from 3,000 samples to 4,236 samples, as presented in Table 2. The final dataset was trained on the same ML algorithms for sentiment analysis using cross-validation technique with five folds.

Table 2: Implementation of the SMOTE Method.

Target Classes	Original Distribution	SMOTE Results
Positive (1)	913	913
Negative (-1)	322	1,558
Neutral (0)	1,765	1,765
Total	**3,000**	**4,236**

In the last ML experiments, the third oversampling method, called ADASYN (Adaptive Synthetic Sampling Approach), was applied to generate more data on the preprocessed dataset. The ADAYSN generated different numbers of samples depending on an estimate of the local distribution of the class to be oversampled; the method produced a total of 4,207 samples. The final dataset was trained on the same ML algorithms for sentiment analysis using a cross-validation technique with five folds.

Model Evaluation Metrics

In this study, the accuracy and F1-score were used to evaluate the performance of the ML models for Swahili sentiment analysis. Accuracy represents the proportion of correct predictions made by the ML model. In this study, accuracy was calculated by dividing the number of correct predictions by the total number of predictions. In addition, the f1-score was used to evaluate the performance of the ML models for Swahili sentiment analysis because of the imbalanced nature of the dataset used in this study. F1-score is the harmonic mean of precision and recall as shown in Equation (1) (Vakili et al. 2020):

$$F1-score = 2 \star \frac{Precision \star Recall}{Precision + Recall} = \frac{2\,TP}{2\,TP + FP + FN} \quad (1)$$

Results

This section presents the results of the four ML experiments followed by the respective discussion of the results.

First Experiment

In the first ML experiment, 3,000 feature vectors were formed based on the 3,000 Swahili tweets. ML models were trained on the imbalance dataset using a cross-validation technique with five folds. The results across all five sentiment models had accuracy ranging from 0.563 to 0.596. Random Forest and Catboost have shown the highest score of 0.596 accuracy and Precision. The Histogramgradientboost performed the least with an accuracy score of 0.563 and Precision of 0.330. All five sentiment models had the same F1 score of 0.290 as presented in Table 3.

Table 3: Training models with an imbalanced dataset

Sentiment Analysis Model	Accuracy	F1 score	Precision	Recall
Random Forest	0.596	0.290	0.490	0.206
Xgboost	0.587	0.290	0.402	0.227
Catboost	0.596	0.290	0.490	0.206
Lightgradientboost	0.566	0.2901	0.471	0.209
Histogramgradientboost	0.563	0.290	0.330	0.259

As illustrated in Figure 4 below, the Random Forest model demonstrates strong predictive performance for class "0", correctly classifying 1,626 tweets. The model also achieves moderate accuracy for class "1", with 175 tweets accurately predicted. However, its performance significantly declines for class "-1", where only 2 tweets are correctly classified, indicating a limitation in predicting tweets for this class.

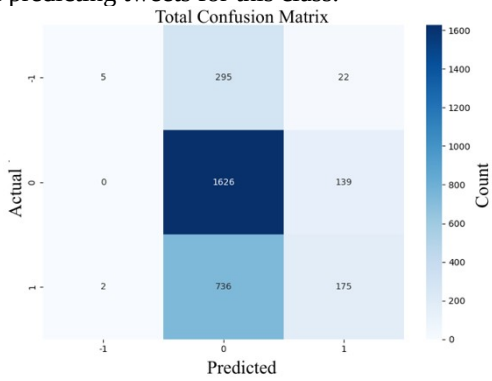

Figure 4: Confusion matrix table for Random Forest model.

Second Experiment

The second experiment used a balanced dataset using the RandomOversampling method. The resulting dataset had a total of 4,236 samples. Each class of the target column had an equal number of 1,412 samples; this is more data compared to the first experiment. The F1 score and accuracy improved with the increased size of the dataset. F1-score increased from 0.290 in the first experiment to 0.455 in the second experiment. Accuracy ranged from 0.548 to 0.613. The Xgboost outperformed other sentiment models with an F1-score of 0.455 while the Random Forest had the highest accuracy of 0.61 but the lowest F1-score of 0.400 as presented in Table 4.

Table 4: Training models with balanced dataset using RandomOversampling method.

Sentiment Analysis Model	Accuracy	F1 score	Precision	Recall
Random Forest	0.6133	0.400	0.290	0.336
Xgboost	0.560	0.455	0.290	0.559
Catboost	0.553	0.452	0.290	0.543
Lightgradientboost	0.548	0.445	0.290	0.507
Histogramgradientboost	0.566	0.447	0.290	0.514

Third Experiment

The SMOTE method was applied in the third experiment to increase the size of the minority classes in the dataset. The resulting dataset had a total of 4,236 samples. Each class had an equal number of 1,412 samples on the target column. The F1-score improved the performance of Catboost and

Lightgradientboost as compared to the previous experiments. Catboost had a higher F1-score of 0.47 and Lightgradientboost had an F1-score of 0.459 as presented in Table 5. On the other hand, Random Forest had a higher accuracy score of 0.62 but the lowest F1 score of 0.387.

Table 5: Training models with balanced dataset using SMOTE method.

Sentiment Analysis Model	Accuracy	F1 Score	Precision	Recall
Random Forest	0.620	0.387	0.601	0.379
Xgboost	0.598	0.4411	0.516	0.414
Catboost	0.5983	0.470	0.526	0.461
Lightgradientboost	0.588	0.459	0.496	0.434
Histogramgradientboost	0.553	0.440	0.473	0.428

Fourth Experiment

ADASYN method was applied in the fourth experiment to generate different numbers of samples depending on an estimate of the local distribution of the class. The resulting dataset had a total of 4,207 samples: 1,412 for the neutral class, 1,423 samples for the positive class and 1,372 samples for the negative class. The Catboost model performed better with an F1-score of 0.464 as presented in Table 6. Using accuracy as an evaluation metric, Random Forest, Xgboost and Catboost performed better with an accuracy score of 0.610.

Table 6: Training models with balanced dataset using ADASYN method.

Sentiment Analysis Model	Accuracy	F1 Score	Precision	Recall
Random Forest	0.610	0.407	0.641	0.403
Xgboost	0.610	0.445	0.577	0.439
Catboost	0.610	0.464	0.522	0.462
Lightgradientboost	0.570	0.430	0.492	0.426
Histogramgradientboost	0.563	0.439	0.501	0.430

Discussion

In this study, we compared the performance of five classical ML models to recommend a model that performs better for the Swahili sentiment analysis task as a low-resource language. The challenge we faced was the nature of the dataset which was small in size and imbalanced. The results indicated that CatBoost performed better than the other four models because of its features that handle text data well (Ibrahim et al. 2020) and also its ability to handle the imbalance of the target classes in the dataset using "class_weights" presented in the algorithm as compared to other Classical ML models (Ibrahim et al. 2020) .

Among the models tested, CatBoost consistently outperformed others across all four experiments, achieving F1-scores ranging from 0.452 to 0.470 and accuracy scores between 0.596 and 0.610. Its relative success can be attributed to its unique handling of categorical and textual data, built-in support for imbalanced classes through the "class_weights" parameter, and its robustness in working with small, high-dimensional sparse datasets. (Ibrahim et al. 2020; Siddiq et al. 2022).

Random Forest had better accuracy (0.610) in the second experiment because it is an ensemble of decision trees biased towards the majority class, thus exhibiting persistent bias favouring the majority class when trained on imbalanced datasets, contributing to better performance for that class (Wainberg et al. 2016). This accuracy score is misleading because the model performed well on the majority class (Neutral tweets) but poorly on the minority class (Negative tweets) with an average of F1-score 0.400.

The paper primarily relies on a relatively small dataset of 3,000 tweets, and this limitation raises concerns about the generalizability and robustness of the proposed models. No single Classical ML model in this paper has an F1 score of 0.5 and above. Classical ML models, especially in sentiment analysis, thrive on datasets of large sizes to capture the patterns and variations in human expression. On the other hand, the unavailability of large datasets limits the use of more robust approaches that could yield better results such as Deep learning models. If deep learning approaches were used, such as transformers and pre-trained models, they could lead to model overfitting. This study aimed to benchmark classical ML methods on Swahili data and demonstrate the current challenges, rather than claim to have achieved production-ready performance. It is crucial for future research to address this constraint by incorporating larger and more diverse datasets to enhance the reliability and effectiveness of sentiment analysis models in real-world scenarios.

Conclusion and Future Directions

This paper presented a comparison of the performance of various classical ML algorithms for Swahili sentiment analysis. We found that CatBoost performs better than the other four algorithms. The challenges associated with the study limited the size of the dataset used for training and the unequal distribution of classes in the target column as observed in the first experiment with the flat F1 score of 0.290 which was the lowest score among the four experiments performed. Implementation of Oversampling methods has a noticeable improvement in the performance of the ML models for Sentiment Analysis. Better values of FI-score and accuracy were recorded from the second experiments to the fourth experiments ranging from 0.387 to 0.470 and 0.553 to 0.620. Overall, the CatBoost model had the highest performance with an F1 score of 0.470 and an accuracy of 0.610. FI-score was used as a major determinant for the sentiment analysis model performance due to the imbalanced nature of the dataset used in this study. Accuracy may mislead because sentiment analysis model performance can be biased by majority classes. However, given that all models produced F1-scores below 0.5, their practical applicability remains limited. These low scores suggest the models would struggle to reliably detect minority sentiments in real-world scenarios, reinforcing the need for larger, more representative datasets and more robust modeling approaches.

The classical ML algorithms applied in this study offer significant advantages in terms of model interpretability and explainability, even though their performance may be low compared to state-of-the-art deep learning approaches. This transparency is particularly valuable for Swahili sentiment analysis applications where understanding the decision-making process is crucial, such as in social media monitoring, customer feedback analysis and content moderation. The interpretable nature of classical ML models allows practitioners to identify which features contribute most to sentiment predictions which enable better feature engineering and domain-specific insights that can inform future model improvements.

For the ML model for Swahili sentiment analysis to be transitioned from research settings to a real-world environment, investment in data collection is crucial. This study lays the groundwork for further research and underscores the need to enhance the Swahili corpus for improved Swahili sentiment analysis performance. Future studies could investigate multilingual transfer learning by incorporating labelled data from languages belonging to the same Greater Lake Bantu language sub-group, such as Luganda. Using pre-trained models fine-tuned on linguistically similar languages could help to address the issue of limited data size in low-resource settings like Swahili. This study serves as a steppingstone for advancing sentiment analysis in Swahili and other underrepresented African languages in NLP research.

Statements and Declarations

Funding Statement: No funding was received.

Competing Interests: None.

References

Adebara I and Abdul-Mageed M 2022 Towards Afrocentric NLP for African Languages: Where We Are and Where We Can Go. In: Muresan S, Nakov P and Villavicencio A (Eds*.) Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics*, 3814–3841.

Bentéjac C, Csörgő A and Martínez-Muñoz G 2021 A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* 54(3): 1937–1969

Berka P 2020 Sentiment analysis using rule-based and case-based reasoning. *J. Intell. Inf. Syst.* 55(3): 457–479.

Chonka P, Diepeveen S and Haile Y 2023 Algorithmic power and African indigenous languages: search engine autocomplete and the global multilingual Internet. *Media Cult. Soc*. 45(2): 246–265.

Das B and Chakraborty S 2018 An improved text sentiment classification model using TF-IDF and next word negation. arXiv:1806.06407.

Fan T, Chen W, Ma G, Kang Y, Fan L and Yang Q 2024 SecureBoost+: Large scale and high-performance vertical federated gradient boosting decision tree. *In: Proceedings of the 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2024)*, Springer, pp. 237–249.

Gao L and Guan L 2023 Interpretability of machine learning: Recent advances and future prospects. *IEEE MultiMedia* 30(4): 105–118. arXiv preprint arXiv:2305.00537.

Haluska R, Brabec J and Komárek T 2022 Benchmark of data preprocessing methods for imbalanced classification. In: *2022 IEEE International Conference on Big Data (Big Data), IEEE*, 2970–2979.

Hang H, Hung T, Cai Y and Lin Z 2021 Gradient boosted binary histogram ensemble for large-scale regression. DOI: 10.48550/arXiv.2106.01986.

Huyut MT, Velichko A and Belyaev M 2022 Detection of risk predictors of COVID-19 mortality with classifier machine learning models operated with routine laboratory biomarkers. *Appl. Sci.* 12(23): 12180.

Ibrahim AA, Ridwan RL, Muhammed MM, Abdulaziz RO and Saheed GA 2020 Comparison of the CatBoost classifier with other machine learning methods. *Int. J. Adv. Comput. Sci. Appl*. 11(11): 738–748.

Kamal A 2013 Subjectivity classification using machine learning techniques for mining feature-opinion pairs from web opinion sources. *Int. J. Comput. Sci. Issues* 10(5): 191–198. arXiv preprint arXiv:1312.6962**.**

Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu TY 2017 LightGBM: A highly efficient gradient boosting decision tree. In: Guyon I et al. (Eds.) *Advances in Neural Information Processing Systems* 30: 3146–3154. Available at: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf (Accessed: 18 September 2025).

Khan AA, Chaudhari O and Chandra R 2024a A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* 244: 122778.

Khan AA, Iqbal MH, Nisar S, Ahmad A and Iqbal W 2024b Offensive language detection for low resource language using deep sequence model. *IEEE Trans. Comput. Soc. Syst*. 11(4): 5210–5218.

Kim J and Kim J 2018 The impact of imbalanced training data on machine learning for author name disambiguation. *Scientometrics* 117(1): 511–526.

Kursa MB and Rudnicki WR 2010 Feature selection with the Boruta package. *J. Stat. Softw*. 36(11): 1–13.

Le T 2022 A comprehensive survey of imbalanced learning methods for bankruptcy prediction. *IET Commun*. 16(12): 1355–1378.

Muhammad SH, Adelani DI, Ruder S, Ahmad IS, Abdulmumin I, Bello BS, Choudhury M, Emezue CC, Abdullahi SS, Aremu A, Jorge A and Brazdil P 2022 NaijaSenti*: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis. In: Proceedings of the Thirteenth Language*

Resources and Evaluation Conference (LREC 2022), European Language Resources Association (ELRA), Marseille, France, 590–602. arXiv preprint arXiv: 2201.08277.

Nguyen T, Khadka R, Phan N, Yazidi A, Halvorsen P and Riegler MA 2023 Combining datasets to improve model fitting. In: *Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 1–9. arXiv preprint arXiv:2210.05165.

Padurariu C and Breaban ME 2019 Dealing with data imbalance in text classification. Procedia Comput. Sci. 159: 736–745.

Rawat SS and Mishra AK 2024 Review of methods for handling class imbalance in classification problems. In: *Proc. Int. Conf. Data Eng. Appl. (IDEA 2022)*, Lect. Notes Electr. Eng. 1146, Springer, Singapore.

Shahmirzadi O, Lugowski A and Younge K 2019 Text similarity in vector space models: A comparative study. In: *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, IEEE: 659–666.

Siddiq A, Shukla N and Pradhan B 2022 Spatio-temporal modelling of dengue fever cases in Saudi Arabia using socio-economic, climatic and environmental factors. Geocarto Int. 38(1): 2063080.

Tanha J, Abdi Y, Samadi N, Razzaghi N and Asadpour M 2020 Boosting methods for multi-class imbalanced data classification: an experimental review. Big Data 7(1): 47.

Vakili M, Ghamsari M and Rezaei M 2020 Performance analysis and comparison of machine and deep learning algorithms for IoT data classification. arXiv preprint arXiv:2001.09636.

Vitorino J, Praça I and Maia E 2023 Towards adversarial realism and robust learning for IoT intrusion detection and classification. arXiv preprint arXiv:2301.13122.

Wainberg M, Alipanahi B and Frey BJ 2016 Are Random Forests truly the best classifiers? J. Mach. Learn. Res. 17: 1–5. Available at: https://jmlr.org/papers/volume17/15-374/15-374.pdf (Accessed: 18th September 2025).

Zhang B, Abdul Latiff NA, Kan J, Tong R, Soh D, Miao X and McLoughlin I 2025 Automated evaluation of children's speech fluency for low-resource languages. arXiv preprint arXiv:2505.19671.

Zhang B, Abu Salem FK, Hayes MJ, Smith KH, Tadesse T and Wardlow BD 2023 Explainable machine learning for the prediction and assessment of complex drought impacts. Sci. Total Environ. 898: 165509.

Zhou X 2021 Sentiment Analysis of COVID-19 Tweets. Stanford University. Available at: http://cs230.stanford.edu/projects_spring_2021/reports/4.pdf (Accessed: 24 July 2025).