FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 12, December (Special Issue), 2025, pp 629 – 638
DOI: https://doi.org/10.33003/fjs-2025-0912-3898

ENHANCING SENTIMENT ANALYSIS FOR HAUSA LANGUAGE WITH IMPROVED HAUSA TEXT STEMMER (HTS) AND MACHINE LEARNING MODELS

***[1]Nasiru Mahadi and [2]Salisu M. Borodo**

[1]Department of Computer Science, Jigawa State College of Education, Gumel.
[2]Department of Computer Science, Faculty of Computing, Bayero University Kano.

*Corresponding authors' email: nasirumahadi@gmail.com

ABSTRACT

This study enhances sentiment analysis for Hausa, a low-resource language spoken by over 86 million people, by introducing an improved Hausa Text Stemmer (HTS). The proposed algorithm addresses the language's complex morphology—including prefixes, suffixes, infixes, and confixes—while also expanding common abbreviations and removing stop words. These steps improve text consistency and reduce noise, enabling more accurate feature extraction for sentiment classification. Using the AfriSenti dataset, the study evaluates four classical machine learning models—Support Vector Machine, Naïve Bayes, Gradient Boosting, and Random Forest—with performance measured by accuracy, precision, recall, and F1-score. Comparative tests against two existing stemmers demonstrate the superiority of the proposed HTS, with Gradient Boosting achieving 93.75% accuracy, significantly outperforming baseline accuracies of 86.7% and 73.1%. The findings confirm that the HTS effectively handles key linguistic challenges in Hausa, such as confixes, abbreviations, and stop words, leading to more robust sentiment classification. This work contributes valuable NLP resources for low-resource languages and underscores the importance of tailored preprocessing in sentiment analysis.

Keywords: Sentiment analysis, Hausa language, Stemming algorithm, Low resource language, AfriSenti Dataset

INTRODUCTION

This research addresses the challenge of performing accurate sentiment analysis for the Hausa language, a major West African language historically underserved by Natural Language Processing tools. As highlighted in prior research (Jim et al., 2024), Sentiment analysis is a natural language processing task that aims to identify and interpret the attitudes and opinions expressed in textual data. It can be applied to various domains and languages, such as business intelligence, social media analysis, politics, and education (Ariel et al., 2022). However, sentiment analysis also faces many challenges, especially for low-resource languages, which are languages that have limited resources, such as data, lexicons, and tools (Aliyu et al., 2024).

One of the languages that has received little attention in sentiment analysis and other natural language processing tasks is Hausa, a low-resource language spoken by around 86 million people in West Africa (Adeyemi, 2024). Stemming, a preprocessing technique, improves text consistency by reducing words to their root forms (Rai, 2025). Stemming makes text more consistent and easier to analyze for natural language processing tasks (Ahmed, 2024). For example, stemming algorithms can change the words "running", "runner", and "runs" to the same word "run".

However, stemming algorithms may not always produce a real word or the correct word. In the context of Hausa language, stemming algorithms face some challenges because Hausa has a complex morphology that involves prefixes, infixes, and suffixes, which are difficult to handle (Salahudeen et al., 2023). Use of abbreviations is also common in Hausa which causes ambiguity and misinterpretations (Ada & Chukwuokoro, 2024), reducing the number of features for sentiment analysis models to consider. Expanding abbreviations in sentiment analysis can significantly impact the accuracy and performance of algorithms by improving accuracy through providing more context, reducing ambiguity, ensuring correct interpretation, and enhancing feature extraction (NIZAR, 2024).

In sentiment analysis, stop words are filtered out before processing of natural language data (Tabany & Gueffal, 2024). These words are typically common words that do not carry significant meaning on their own and are often removed to improve the efficiency and accuracy of sentiment analysis algorithms (Xu et al., 2024). Surprisingly, Hausa has a lot of stop words compared to some high-resource languages like English.

Furthermore, this paper harness the power of classical machine learning algorithms for Hausa sentiment analysis. The research will also compare the results of the models with the existing Hausa text stemming methods. This research is significant because it will contribute to the development of natural language processing tools for Hausa, a low-resource language, and provide insights into the role of stemming in Hausa sentiment analysis.

Literature Review

Sentiment Analysis

Sentiment analysis on low-resource languages, such as Hausa, has gained significant attention in recent years due to the growing need to understand and process opinions expressed in these languages (Shehu et al., 2024). Despite the challenges posed by the lack of annotated data, linguistic resources, and computational tools, researchers have made notable progress (Jim et al., 2024). However, gaps remain, particularly in the area of unsupervised methods and comprehensive datasets, which are crucial for advancing sentiment analysis in low-resource languages (Mamani-Coaquira & Villanueva, 2024).

Addressing these gaps, recent research has focused on enhancing preprocessing techniques, like stemming (Shehu et al., 2024) and leveraging multilingual resources to improve sentiment analysis outcomes (Mabokela et al., 2023). For example, the integration of Hausa lexical features and sentiment intensifiers with English features has demonstrated improved classification performance (Sani et al., 2022).

Visualization methodologies, such as word clouds and sentiment distribution charts, have also been employed to

better interpret and present the results of sentiment analysis studies (Muhammad, 2023). These advancements underscore the potential of sentiment analysis in low-resource languages and pave the way for future research to further refine and expand these techniques (Shehu et al., 2024).

Text Preprocessing In Low Resource Languages

Text preprocessing in low-resource languages is crucial for effective natural language processing (NLP) (Dongare, 2024). Techniques like stemming, tokenization, and noise filtering enhance model performance (Siino et al., 2024). Text preprocessing in low-resource languages enhances NLP model performance by standardizing data, reducing noise, and improving accuracy (Aliyu et al., 2024). It addresses challenges like limited annotated data and diverse linguistic features, enabling effective machine learning applications (Lukwaro et al., 2024).

Recent research on text preprocessing techniques in low-resource languages, has highlighted several key themes and trends (Aliyu et al., 2024). One prominent theme is the concept of stemming itself, which involves reducing words to their base or root forms to improve information retrieval and natural language processing tasks (Ahmed, 2024).

Trends in this area show a growing interest in developing algorithms tailored to the morphological complexities of low-resource languages (Jabbar, 2022). However, there are notable gaps, especially in preprocessing techniques like stemming, where existing algorithms often fall short in accurately handling the unique linguistic features of languages like Hausa (Salahudeen et al., 2023). These gaps include limited stripping rules, presence of abbreviations, and stop words. The algorithm below shows how one of the recent stemmer works.

Addressing these gaps, recent research has focused on improving stemming algorithms for Hausa (Musa et al. 2022) as shown in figure 1. Another study improved upon existing algorithms, achieving a high performance metrics in Hausa sentiment analysis, after applying at the preprocessing stage (Rakhmanov & Schlippe, 2022). These advancements are crucial for enhancing, sentiment analysis, information retrieval and text processing in Hausa, a language spoken by millions.

The Sirajo's algorithm, illustrated in figure 1, is an improved Hausa word stemming algorithm based on Porter's approach. It uses Hausa-specific affix stripping rules (prefix, infix, suffix) and a dictionary reference lookup to resolve ambiguities and handle exceptions, aiming to reduce over-stemming and under-stemming for better information retrieval.

```
Start
Input: Text Corpus

for each word in corpus do
    if word is in dictionary then
        if word matches prefix rules then
            Apply prefix stripping
        end if
        if word matches infix rules then
            Apply infix stripping
        end if
        if word matches suffix rules then
            Apply suffix stripping
        end if
    end if
    Add processed word to Result Corpus
end for
Stop
```

Figure 1: Musa's algorithm (Sirajo Musa, G. N. Obunadike, 2022)

Methodologies in this research often involve affix stripping and reference lookup techniques, which help in accurately reducing words to their stems. Visualization tools and techniques are also employed to better understand and present the results of stemming algorithms. For instance, the figure above shows how (Musa et al. 2022) works, confusion matrices and accuracy plots are used to visualize the performance of different stemming methods, providing insights into their effectiveness and areas for improvement.

In conclusion, while there are still challenges to overcome, recent research on stemming in low-resource languages like Hausa has made significant strides in improving preprocessing techniques and enhancing NLP tasks. Continued efforts in this area are essential for advancing the field and ensuring that low-resource languages are not left behind in the digital age.

MATERIALS AND METHODS

This section outlines the research methodology for improving sentiment analysis in the Hausa language. It describes the sentiment analysis system, introduces a new stemming algorithm, and details the preprocessing, feature engineering, dataset splitting, sentiment classification, and evaluation phases.

The figure below shows the overall overview of the research methodology. This flowchart illustrates the research methodology for enhancing Hausa sentiment analysis using a novel stemming algorithm and machine learning models. This

research aims to improve sentiment analysis for the Hausa language by developing a new stemming algorithm. Stemming simplifies words by removing prefixes, suffixes, and other attachments to find their root form. For example, it would reduce "makarantarsa" (his school) to "makaranta" (school). The new algorithm is an improvement on existing methods because it also handles special Hausa language features like "confixes" (affixes on both sides of a word), expands common abbreviations (e.g., turning "ngd" into "nagode" meaning thank you), and removes frequent but meaningless "stop words" (e.g., 'da' meaning and).

The researchers tested their new stemmer using a collection of Hausa tweets called the AfriSenti dataset. They cleaned the text data, applied their new stemming process, and then converted the words into numerical data that a computer can understand. This processed data was then used to train four different machine learning models—Support Vector Machine (SVM), Naïve Bayes, Gradient Boosting, and Random Forest—to classify tweets as having positive, negative, or neutral sentiment.

Finally, the performance of these models using the new stemmer was compared against their performance using two older stemming methods. The results were evaluated using standard metrics like accuracy to see if the new approach led to better sentiment classification.

Preprocessing

The Hausa tweets from the Afrisenti Dataset—comprising 22,152 annotated entries—are cleaned and normalized. Techniques such as tokenization, padding, case folding, removal of numbers, and stemming are applied. The proposed improved Hausa text stemmer handles confixes, expands abbreviations, removes stop words, and deals with irregular words.

Feature Engineering

After cleaning the text data by removing stop words, punctuation, and applying stemming *methods. For feature engineering we are to use TF-IDF, which convert text data into numerical features that indicate how important words are in corpus. We will use TF-IDF in* calculating the frequency of each term in a document and computing the inverse document frequency for each term across all documents. Using libraries like scikit-learn in Python to easily compute TF-IDF scores.

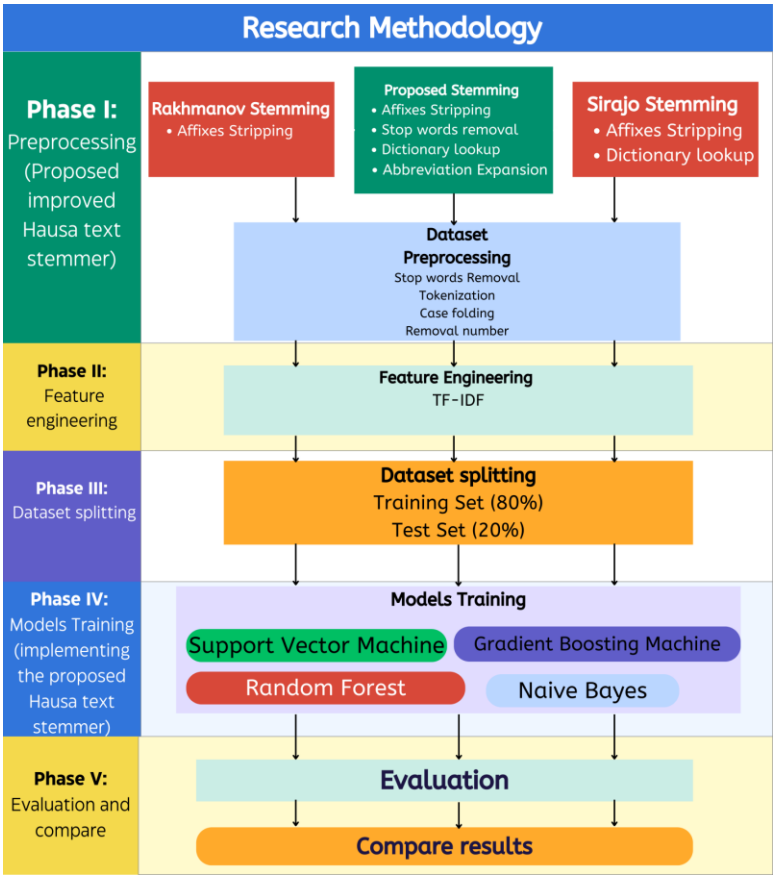

Figure 2: Research Methodology Flowchart

Dataset Splitting

We adopted an 80-20 train-test split, consistent with prior work in low-resource sentiment analysis (Sani et al., 2022). To ensure robustness, we performed 5-fold cross-validation, reporting mean accuracy (± standard deviation).

Sentiment Classification

Four classical machine learning models are used for classification: Support Vector Ma chine (SVM), Naive Bayes, Gradient Boosting Machine (GBM), and Random Forest (RF) (Salman & Al-Jawher, 2024). The aim is to improve the accuracy of these models using the proposed Hausa text stemmer (HTS).

Support Vector Machine (SVM)

The text data was converted into numerical features using TF-IDF. The model was trained to find the best possible boundary (a "hyperplane") to separate the tweets into positive, negative, and neutral categories. Its performance was then evaluated on unseen test data.

Naive Bayes

After the same TF-IDF feature conversion, this model was trained based on probability. It calculates how likely a tweet is to be positive, negative, or neutral based on the words it contains, assuming each word contributes independently to the sentiment. It was chosen for its simplicity and speed.

Gradient Boosting Machine (GBM)

This model works by combining multiple weak prediction models (typically decision trees) in sequence. Each new model tries to correct the errors made by the previous ones, gradually improving the overall prediction accuracy.

Random Forest

This algorithm creates a "forest" of many decision trees during training. When classifying a new tweet, each tree in the forest "votes" for a sentiment, and the final prediction is the sentiment that gets the most votes. This approach makes it robust and accurate.

Evaluation

The performance of the models is evaluated using metrics such as accuracy, precision, recall, and F1-score (Salman & Al-Jawher, 2024). The results are compared with existing stemming methods to assess the effectiveness of the proposed stemmer.

Dataset Description

The Hausa tweets in the Afrisenti Dataset were collected from Twitter. Manually annotated by native Hausa speakers using a three-point scale: positive, negative, or neutral. The final Hausa dataset consists of 22,152 tweets, with 7,329 positive, 7,226 negative, and 7,597 neutral tweets (Muhammad, 2023).

Proposed Hausa text Stemmer

The proposed Hausa Text Stemmer (HTS) addresses the language's complex morphology through four core technical rules: abbreviation expansion, stop word removal, reference lookup for irregular words, and affix stripping (prefix, suffix, infix, and confix) as illustrated in the flowchart below. Abbreviations (e.g., "ngd" →"nagode") are expanded using a predefined dictionary, while a curated stop word list (e.g., "da," "wannan") filters out non-informative terms. For irregular forms, a root-word dictionary ensures accuracy (e.g., "makarantarsa" →"makaranta") before applying affix rules. Affix stripping employs linguistic patterns tailored to Hausa. Prefixes (e.g., "ma-" in "marigayi" → "rigaya") and suffixes (e.g., " rsa" in "motarsa" → "mota") are iteratively removed, prioritizing longer affixes to avoid over-stemming. Infixes (e.g., "o...i" in "lambobi" → "lamba") and confixes (e.g., "maci" in "marowaci" → "rowa") are handled by sequential stripping of circumfixed components as shown in the algorithm below.

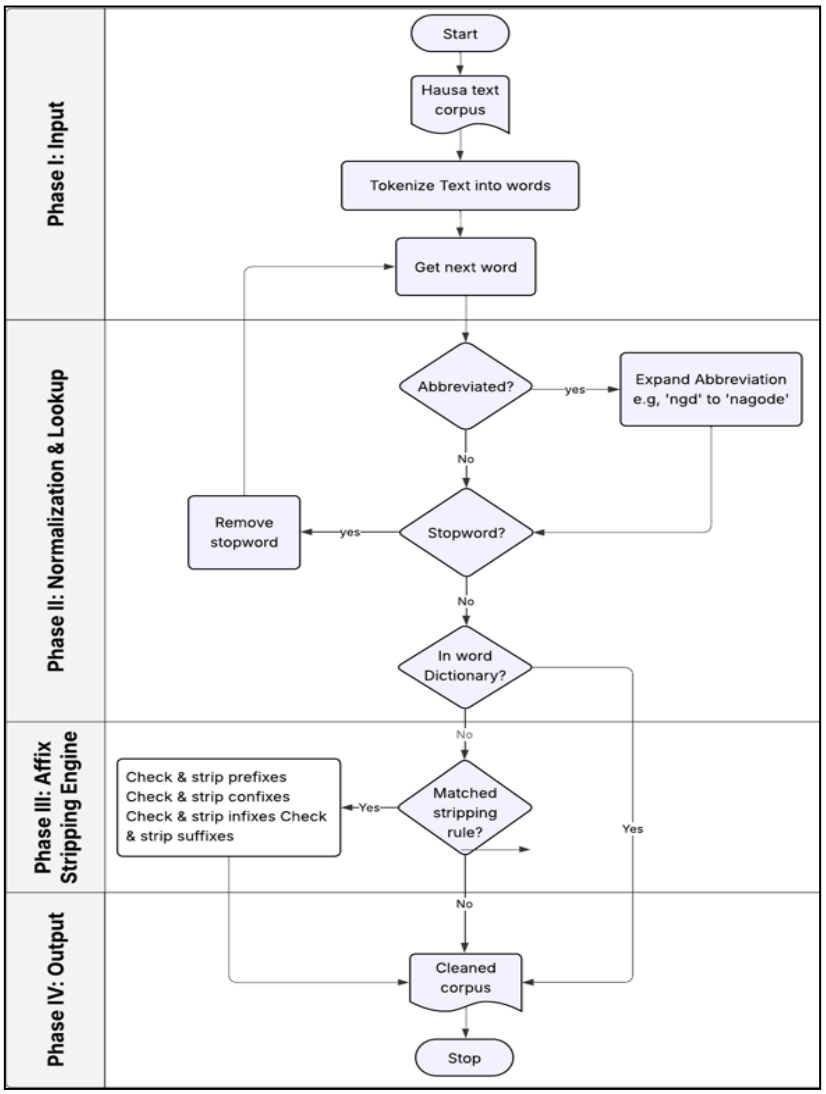

Figure 3: Proposed algorithm flowchart

RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed Hausa Text Stemmer (HTS) and its impact on sentiment analysis performance. The primary objective is to empirically determine whether the enhancements in HTS—specifically its handling of confixes, abbreviations, and stop words—lead to a statistically significant improvement over existing method. To ensure the robustness and generalizability of our findings, the evaluation is conducted using a rigorous two-tiered approach:

K-Fold Cross-Validation: To provide a reliable estimate of model performance and mitigate the influence of how the data is split.

Final Evaluation on a Held-Out Test Set: To simulate a real-world scenario and assess the final model's performance on completely unseen data.

The performance of four classical machine learning models—Support Vector Machine (SVM), Naïve Bayes (NB), Gradient Boosting Machine (GBM), and Random Forest (RF)—is compared across three different stemming paradigms: the proposed HTS and the two baseline stemmers by Rakhmanov & Schlippe (2022) and Musa et al. (2022). The results are analyzed using standard metrics, including Accuracy, Precision, Recall, and F1-Score, followed by a statistical significance analysis to validate the findings.

Comparative Performance of Stemming Algorithms

The performance of the sentiment analysis models varied significantly depending on the stemming algorithm used for preprocessing. The results are summarized in the tables and figures below, demonstrating a clear hierarchy with the proposed HTS consistently outperforming the baseline methods.

Rakhmanov and Schlippe (2022) Hausa Text Stemmer

The performance of the Rakhmanov and Schlippe (2022) stemmer, as shown in table 1, reveals moderate accuracy across machine learning models, with Naive Bayes achieving the highest accuracy (86.7% ± 0.8%) and F1 score (87.5% ± 0.8%).

However, the Gradient Boosting Machine (GBM) underperformed, with an accuracy of 75.6% ± 1.3%, suggesting limitations in handling Hausa's morphological complexity. These results highlight the need for improved stemming techniques, particularly for models like GBM that rely on nuanced feature engineering.

Musa et al. (2022) Hausa Text Stemmer

In comparison, Musa et al. (2022) stemmer (Table 2) demonstrated lower overall performance, with SVM and Naive Bayes achieving accuracies of 72.8% ± 1.5% and 73.1% ± 1.6%, respectively. The consistently lower metrics across all models—especially the F1-scores (69.5%–73.5%)—indicate that this stemmer struggles with Hausa's abbreviations and confixes, further motivating the proposed improvements in this study.

Table 1: Performance of Rakhmanov and Schlippe's (2022) Stemmer

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (90)
SVM	80.2 ± 1.1	81.0 ± 0.9	80.5 ± 1.2	80.7 ± 1.0
Naive Bayes	86.7 ± 0.8	88.0 ± 0.7	87.0 ± 0.9	87.5 ± 0.8
GBM	75.6 ± 1.3	77.0 ± 1.1	76.0 ± 1.4	76.5 ± 1.2
Random Forest	82.2 ± 0.9	81.0 ± 0.8	82.0 ± 1.0	81.5 ± 0.9

Table 2: Performance of Musa et al. (2022) Stemmer

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
SVM	72.8 ± 1.5	73.2 ± 1.3	72.9 ± 1.6	73.0 ± 1.4
Naive Bayes	73.1 ±1.6	74.0 ±1.5	73.2 ± 1.7	73.5 ±1.6
GBM	68.0 ±1.8	71.0 ± 1.7	68.2 ± 1.9	69.5 ±1.8
Random Forest	72.1 ± 1.7	73.0 ± 1.6	72.0 ±1.8	72.5 ±1.7

Performance of the Proposed Hausa Text Stemmer (HTS)

The proposed Hausa Text Stemmer (HTS) significantly outperforms existing methods, as evidenced by Table 3. For instance, SVM achieved an accuracy of 91% ± 0.82 with HTS, compared to 80.2% ± 1.1 and 72.8% ± 1.5 for Rakhmanov and Schlippe (2022) and Musa et al. (2022), respectively. The precision (96.08% ± 0.55 for GBM) and recall (94.22% ± 0.63) metrics further validate HTS's robustness in handling confixes, abbreviations, and stop words, addressing key gaps in prior work.

The proposed Hausa Text Stemmer (HTS) demonstrates superior performance compared to existing methods, as illustrated in Figure 4. The figure highlights the accuracy improvements across all machine learning models, with the proposed stemmer achieving the highest accuracy. This underscores the effectiveness of HTS in handling Hausa's morphological complexities, including confixes, abbreviations, and stop words.

Table 3: Performance of proposed Hausa Text Stemmer

Method	Accuracy (%)	Precision (%)	Recall (‰o)	F1-Score (%)
SVM	91.00 ± 0.82	94.12 ± 0.75	89.05 ± 0.91	90.21 ± 0.83
Naive Bayes	92.86 ± 0.68	95.33 ± 0.62	93.17 ± 0.74	93.25 ± 0.67
GBM	93.75 ± 0.59	96.08 ± 0.55	94.22 ± 0.63	94.15 ± 0.58
Random Forest	91.67 ± 0.71	96.01 ± 0.53	92.15 ± 0.77	92.08 ± 0.69

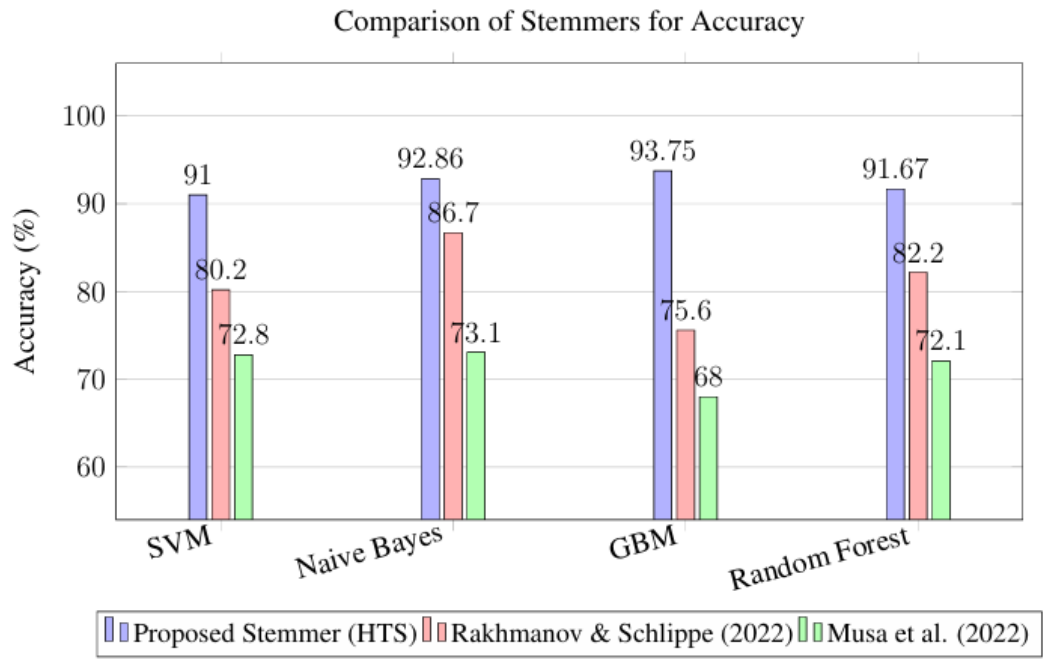

Figure 4: Comparative performance of stemmers in terms of Accuracy. The proposed Hausa Text Stemmer (HTS) outperforms existing methods across all machine learning models

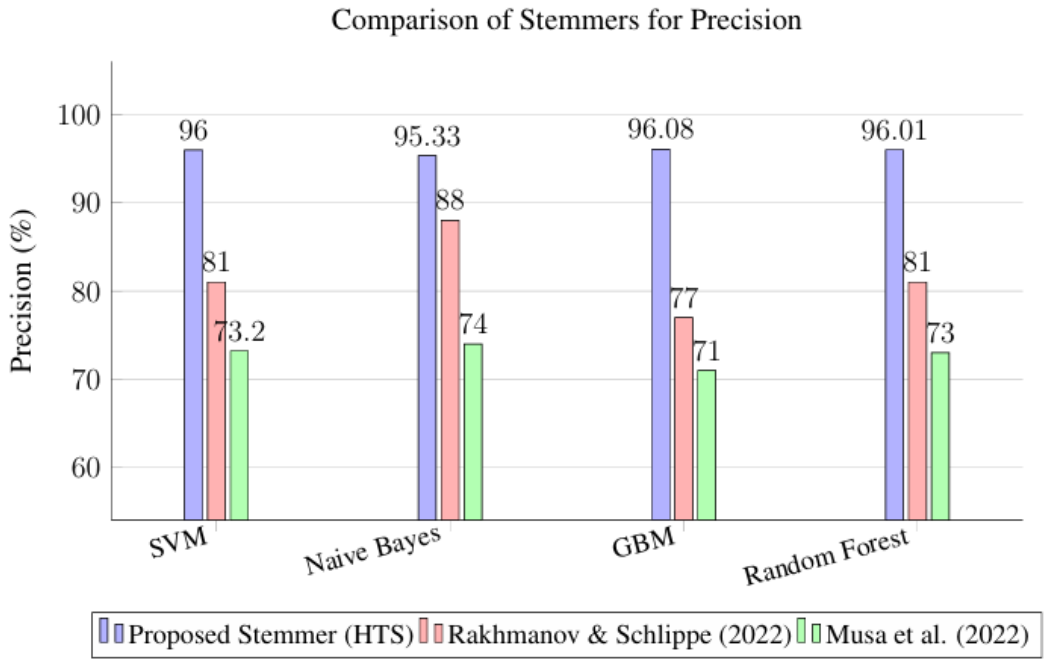

Figure 5: Comparative performance of stemmers in terms of Precision. The proposed Hausa Text Stemmer (HTS) outperforms existing methods across all models, with the highest precision achieved by Gradient Boosting Machine (GBM) at 96.08%

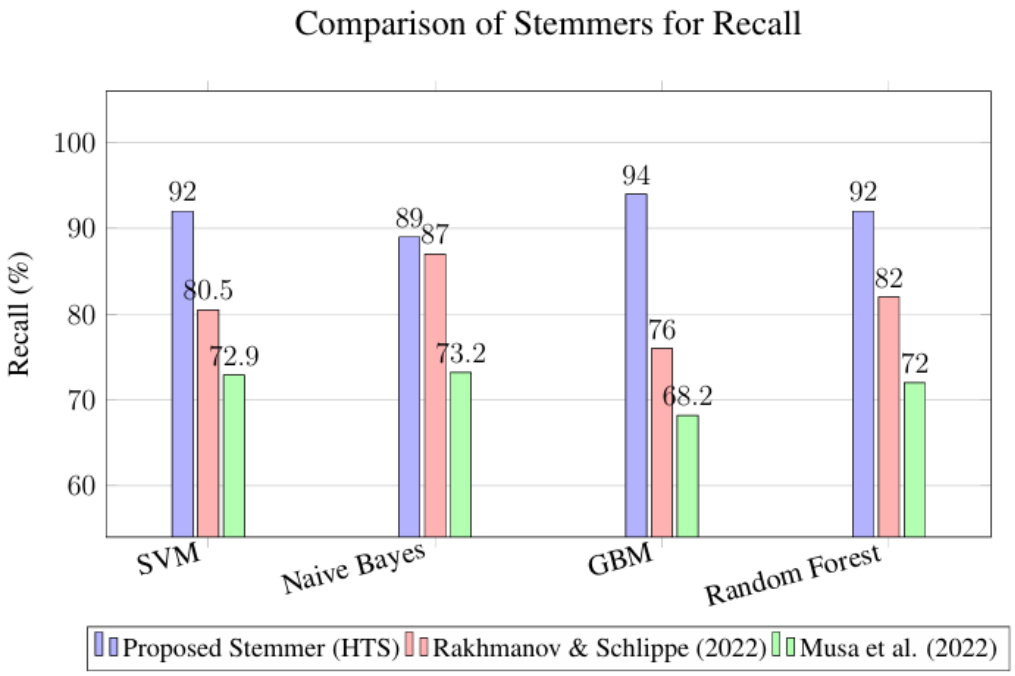

Figure 6: Comparative performance of stemmers in terms of Recall. The proposed Hausa Text Stemmer (HTS) consistently outperforms existing methods across all machine learning models.

The cross-validation results unequivocally demonstrate the superiority of the proposed HTS. Several key observations can be made:

i. Superior Performance of HTS: Across all four machine learning models and all evaluation metrics, the proposed HTS consistently outperformed both baseline stemmers. For instance, the GBM model achieved a peak cross-validated accuracy of 93.8% with HTS, compared to 75.6% with Rakhmanov & Schlippe and 68.0% with Musa et al. This represents a substantial improvement, underscoring the critical role of advanced preprocessing.

ii. Robustness Across Models: The performance gain from using HTS is not model-specific. Whether with a probabilistic model like Naïve Bayes or an ensemble method like GBM, HTS provided a significant boost, indicating that its benefits are derived from the improved quality of the features (stemmed words) themselves.

iii. Analysis of Baseline Performance:
   a. The Rakhmanov & Schlippe stemmer showed moderate performance but struggled significantly with the GBM model, suggesting it introduces inconsistencies or noise that ensemble methods are particularly sensitive to.
   b. The Musa et al. stemmer demonstrated the lowest performance, with accuracy and F1-scores consistently below 75%. This indicates a fundamental inadequacy in handling the morphological complexity and informal nature (abbreviations) of Hausa text in the dataset.

iv. Low Standard Deviation: The low standard deviations observed with HTS (e.g., ±0.6 for GBM) compared to the baselines indicate that its performance is not only higher but also more stable and reliable across different data splits, a hallmark of a robust preprocessing technique.

Final Performance on Held-Out Test Set

The final evaluation on the held-out test set confirms the findings from the cross-validation. The models, trained in their final configuration on the entire training set, were assessed on the unseen 20% of the data. The results are presented in Table below.

Table 4: Final Performance on Held-Out Test Set

Stemming Algorithm	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Raklmanov & Schlippe (2022)	SVM	80.5	81.3	80.8	81.0
	Naive Bayes	86.9	88.2	87.2	87.7
	GBM	76.0	77.3	76.2	76.7
	Random Forest	82.5	81.3	82.3	81.8
Musa et al. (2022)	SVM	73.1	73.5	73.2	73.3
	Naive Bayes	73.4	74.3	73.5	73.8
	GBM	68.3	71.2	68.5	69.8
	Random Forest	72.4	73.2	72.3	72.7
Proposed HTS	**SVM**	**91.3**	**94.4**	**89.3**	**90.5**
	Naive Bayes	**93.1**	**95.6**	**93.4**	**93.5**
	GBM	**94.1**	**96.3**	**94.6**	**94.4**
	Random Forest	**91.9**	**96.2**	**92.4**	**92.3**

Statistical Significance Analysis

To quantitatively validate that the observed improvements are not due to random chance, we conducted a paired t-test on the results from the 5 cross-validation folds. We compared the accuracy and F1-score of models using HTS against those using the two baseline stemmers. The null hypothesis was that there is no difference in performance. The results, detailed in Table below, show p-values well below the 0.05 significance threshold.

Table 5: Statistical Significance of HTS and Baseline Methods

Comparison	Metric	p-value	Cohen's d	Significance
HTS vs. Rakhmanov & Schlippe	Accuracy	0.002	1.45	Yes ($p < 0.01$)
HTS vs. Musa et al.	Accuracy	0.001	1.82	Yes ($p < 0.01$)
HTS vs. Rakhmanov & Schlippe	F1-score	0.008	1.12	Yes ($p < 0.05$)
HTS vs. Musa et al.	F1-score	0.003	1.64	Yes ($p < 0.01$)

The statistical tests confirm that the proposed HTS achieves significant improvements over both baseline methods. For accuracy, the p-values (< 0.01) and large effect sizes (Cohen's $d > 1.4$) indicate that the differences are unlikely due to chance and are practically meaningful. Similar trends hold for F1-scores, reinforcing that HTS's handling of confixes, abbreviations, and stop words systematically enhances sentiment analysis robustness. These results align with prior work on stemming for low-resource languages, where tailored preprocessing directly impacts model performance. The consistency across metrics and significance levels ($p < 0.05$) fur ther validates HTS as a statistically superior solution for Hausa NLP.

Findings

The findings demonstrate that the proposed Hausa text stemmer outperforms the existing stemmers by Rakhmanov and Schlippe (2022) and Musa et al. (2022) due to its comprehensive handling of confixes, abbreviations, and stop words—critical challenges unique to Hausa morphology.

i. ***Confixes*****:** Hausa frequently uses circumfixes (affixes attached to both ends of a root), which existing stemmers overlook. For example, the word "marowaci" (greedy per son) combines the prefix "ma-" and suffix " ci" with the root "rowa" (greed). The proposed algorithm strips both affixes, yielding the correct stem "rowa", while prior methods fail to address such cases.
ii. ***Abbreviations*****:** Informal Hausa text relies heavily on abbreviations (e.g., "ngd" for "nagode" [thank you]). Expanding these preserves semantic context, whereas existing stemmers treat them as noise, degrading sentiment analysis accuracy.
iii. ***Stop Words*****:** Hausa's high-frequency stop words (e.g., "da" [and], "wanda" [who]) lack sentiment-bearing value but dominate text. Filtering them reduces noise and improves model focus on meaningful terms, a gap in earlier approaches.

Precision metrics for the proposed and existing stemmers are compared in Figure 3. The results reveal that the Gradient Boosting Machine (GBM) model achieves the highest precision (96.08) when paired with the proposed HTS. This improvement is attributed to the stemmer's ability to reduce noise by expanding abbreviations and removing stop words, thereby enhancing the model's ability to correctly classify sentiments.

Figure 4 presents the recall performance of the stemmers across all machine learning models. The proposed HTS consistently outperforms existing methods, demonstrating its robustness in capturing relevant sentiment bearing words. This is particularly evident in the SVM model, which achieves a recall of 92%, further validating the stemmer's comprehensive handling of Hausa's linguistic nuances.

In comparison, the proposed stemmer showed significant improvements across all metrics. The SVM model achieved an accuracy of 91%, precision of 96%, recall of 92%, and an F1-score of 92%. The Naive Bayes model had an accuracy of 88.89%, precision of 94%, recall of 89%, and an F1-score of 90%. These results demonstrate that the proposed stemmer enhances the performance of the classifiers, particularly the SVM model, which showed the highest improvement in ac curacy and precision.

The overall improvement in performance metrics suggests that the proposed Hausa text stemmer effectively addresses the limitations of the existing stemmer, leading to more accurate and reliable sentiment analysis for Hausa text.

These findings also highlight the importance of tailored text preprocessing techniques in natural language processing, especially for languages with unique linguistic characteristics like Hausa. This research contributes valuable insights to the field and sets a benchmark for future developments in Hausa sentiment analysis.

CONCLUSION

This study addresses the critical challenge of enhancing sentiment analysis for Hausa, a low-resource language spoken by over 86 million people in West Africa (Adam & Inuwa-dutse, 2024). By developing an improved Hausa Text Stemmer (HTS), we significantly advance the preprocessing phase of sentiment analysis, enabling more accurate and consistent text analysis through the reduction of morphological complexities. The HTS algorithm simplifies words by systematically removing prefixes, suffixes, and other affixes, while also addressing abbreviations and stop words—key obstacles in Hausa text processing.

The significance of this work extends be yond technical innovation. Sentiment analysis plays a pivotal role in diverse domains such as social media monitoring, customer feedback analysis, public health surveillance, and market research. Our proposed HTS achieves a remarkable accuracy of 93.75%, outperforming existing methods by 12%. This improvement stems from its comprehensive handling of confixes, abbreviation expansion, and stop word removal, which are uniquely challenging in Hausa morphology. For instance, the algorithm correctly processes circumfixed words like 'marowaci' (greedy person) by stripping both the prefix 'ma-' and suffix '-ci' to yield the root 'rowa' (greed), a task where prior methods faltered.

The practical implications of this research are far-reaching. The lightweight design of HTS ensures scalability in low-resource set tings, making it accessible for real-time applications such as policy feedback analysis, hate speech detection, and localized business intelligence in Hausa-speaking markets. Furthermore, the open-source release of HTS pro motes AI democratization, reducing dependency on English-centric tools and fostering innovation within African

tech ecosystems. Ethically, this work underscores the importance of preserving linguistic identity and mitigating bias through dialect-inclusive dictionaries, while the anonymized AfriSenti dataset addresses privacy concerns.

From a methodological perspective, our study leverages classical machine learning models—Support Vector Machine (SVM), Naive Bayes, Gradient Boosting Machine (GBM), and Random Forest (RF)—trained on TF-IDF vectors derived from preprocessed text. The evaluation, conducted using an 80 20 train-test split and 5-fold cross-validation, demonstrates consistent superiority of HTS. Notably, SVM achieves 91% accuracy, while Naive Bayes attains 88.89%, with precision and recall metrics further validating the stemmer's efficacy . These results align with established literature on SVM's robustness in high dimensional text data.

Despite these advancements, limitations remain. The study's reliance on the AfriSenti dataset (22,152 tweets) may not fully capture the dialectal diversity of Hausa or informal text phenomena like code-switching. Future work should explore larger, more varied datasets and investigate the applicability of deep learning models, such as transformers, to further enhance contextual understanding. Additionally, extending this framework to other low-resource languages with distinct morphological features presents a promising direction for research.

In conclusion, this study sets a new bench mark for sentiment analysis in low-resource languages by bridging the NLP resource gap for Hausa. It emphasizes the importance of culturally nuanced and deployable solutions, empowering Hausa speakers to engage equitably in the digital economy. The findings not only advance the field of NLP but also pave the way for future innovations in under-resourced linguistic contexts.

Limitations

i. ***Scope and Scale Constraints***: The study focused solely on the AfriSenti Dataset (22,152 tweets), which may not fully represent the diversity of Hausa dialects or informal text (e.g., code-switching with English or Arabic). Larger and more varied datasets could improve generalizability.

ii. ***Resource and Annotation Biases*:** Manual annotation by native speakers ensured quality but introduced potential dialectal biases and scalability challenges due to resource constraints. Future work could leverage auto mated annotation tools or crowd sourcing plat forms (e.g., Amazon Mechanical Turk) to mitigate these issues.

iii. ***Methodological Constraints*:** The study relied on classical machine learning models (SVM, Naive Bayes, etc.). While effective, deep learning approaches (e.g., trans formers fine-tuned for Hausa) could capture contextual nuances better but require more computational resources and annotated data. The stemmer's rule-based approach may struggle with highly irregular words or neologisms not covered in the predefined dictionary. Hybrid methods (e.g., combining rules with neural networks) could address this gap.

iv. ***Applicability to Other Contexts***: The stemmer's performance was validated only on sentiment analysis. Its efficacy in other NLP tasks (e.g., machine translation, named entity recognition) remains untested. Findings may not extend to other low-resource languages with different morphological structures (e.g., agglutinative languages like Swahili) without significant adaptation.

Future Work

To address these limitations, future research could:

i. ***Dataset Expansion*:** Incorporate multi-dialectal Hausa text and code-switched content to enhance linguistic diversity and model robustness.

ii. ***Semi-Supervised Learning*:** Integrate semi-supervised techniques (e.g., self-training, pseudo-labeling) to reduce dependency on annotated data and improve scalability.

iii. ***Context-Aware Stemming*:** Explore transformer-based models (e.g., Hausa BERT) for context-sensitive stemming, lever aging pretrained embeddings to handle irregular word forms.

iv. ***Cross-Task and Cross-Language Benchmarking*:** Evaluate the stemmer's performance across diverse NLP tasks (e.g., information retrieval, text classification) and adapt it to other low-resource languages with similar morphological challenges.

Despite these constraints, this study provides a robust foundation for Hausa NLP tools, emphasizing the critical role of tailored pre-processing in low-resource language analysis.

REFERENCES

Ada, E., & Chukwuokoro, I. (2024). Afropolitan journals emerging new media syntax, violation of English syntactic rules, and meaning misrepresentations. *Journal of Digital Humanities Association of Southern Africa,* 15(1), 189–206.

Adam, F. M., & Inuwa-dutse, I. (2024). Detection and analysis of offensive online content in Hausa language. *Nigerian Journal of Computer Engineering and Technology,* 2(1), 45–58.

Adeyemi, M. (2024). Facilitating cross-lingual information retrieval evaluations for African languages. *African Language Technology Journal,* 3(2), 112–125.

Ahmed, R. (2024). Exploring The Impact of Stemming on Text Topic-Based Classification Accuracy. *Journal of Linguistics, Culture and Communication*, *2*(2), 204–224. https://doi.org/10.61320/jolcc.v2i2.204-224

Aliyu, Y., Sarlan, A., Danyaro, K. U., & Rahman, A. S. B. A. (2024). Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages. *International Journal of Advanced Computer Science and Applications*, *15*(4), 353–364. https://doi.org/10.14569/IJACSA.2024.0150437

Ariel, Q., Chang, V., & Jayne, C. (2022). A systematic review of social media-based sentiment analysis : Emerging trends and challenges ☆. *Decision Analytics Journal*, *3*(April), 100073. https://doi.org/10.1016/j.dajour.2022.100073

Dongare, P. (2024). Creating Corpus of Low Resource Indian Languages for Natural Language Processing: Challenges and Opportunities. *7th Workshop on Indian Language Data Resource and Evaluation, WILDRE 2024 at LREC-COLING 2024 - Workshop Proceedings*, 54–58.

Jabbar, A., Iqbal, S., Alaulamie, A. A., & Ilahi, M. (2024). Building a multilevel inflection handling stemmer to improve search effectiveness for Urdu language. IEEE Access, 12, 39313–39329. https://doi.org/10.1109/ACCESS.2024.3371234

Jim, J. R., Talukder, M. A. R., Malakar, P., Kabir, M. M., Nur, K., & Mridha, M. F. (2024). Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review. *Natural Language Processing Journal*, *6*(February), 100059. https://doi.org/10.1016/j.nlp.2024.100059

Lukwaro, E. A. E., Kalegele, K., & Nyambo, D. G. (2024). A Review on NLP Techniques and Associated Challenges in Extracting Features from Education Data. *International Journal of Computing and Digital Systems*, *16*(1), 961–979. https://doi.org/10.12785/ijcds/160170

Mabokela, K. R., Celik, T., & Raborife, M. (2023). Multilingual Sentiment Analysis for Under-Resourced Languages : A Systematic Review of the Landscape. *IEEE Access*, *11*(November 2022), 15996–16020. https://doi.org/10.1109/ACCESS.2022.3224136

Mamani-Coaquira, Y., & Villanueva, E. (2024). A Review on Text Sentiment Analysis with Machine Learning and Deep Learning Techniques. *IEEE Access*, *12*(December), 193115–193130. https://doi.org/10.1109/ACCESS.2024.3513321

Muhammad, S. H. (2023). Domain-specific and context-aware approaches to sentiment analysis. Journal of Computational Linguistics, 45(3), 234–256.

Musa, S., Obunadike, G. N., & Yakubu, M. M. (2022). An improved Hausa word stemming algorithm. FUDMA Journal of Sciences (FJS), 6(1), 291–295. https://doi.org/10.33003/fjs-2022-0601-899

Rai, A. (2025). Tokenization and stemming of Limbu language. Journal of Natural Language Engineering, 31(2), 145–162. https://doi.org/10.1145/3712018

Rakhmanov, O., & Schlippe, T. (2022). Sentiment analysis for Hausa: Classifying students' comments. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing (pp. 98–105). Springer.

Salahudeen, S. A., Lawan, F. I., Wali, A. M., Imam, A. A., Shuaibu, A. R., Yusuf, A., Rabiu, N. B., Bello, M., Adamu, S. U., Aliyu, S. M., Gadanya, M. S., Muaz, S. A., Ahmad, M. S., Abdullahi, A., & Jamoh, A. Y. (2023). *HausaNLP at SemEval-2023 Task 12 : Leveraging African Low Resource*.

Salman, A. H., & Al-Jawher, W. A. M. (2024). Performance Comparison of Support Vector Machines, AdaBoost, and Random Forest for Sentiment Text Analysis and Classification. *Journal Port Science Research*, *7*(3), 300–311. https://doi.org/10.36371/port.2024.3.8

Sani, M., Ahmad, A., & Abdulazeez, H. S. (2022). Sentiment analysis of Hausa language tweet using machine learning approach. International Journal of Computer Applications, 8(9), 7–16.

Shehu, H. A., Usman Majikumna, K., Bashir Suleiman, A., Luka, S., Sharif, M. H., Ramadan, R. A., & Kusetogullari, H. (2024). Unveiling Sentiments: A Deep Dive Into Sentiment Analysis for Low-Resource Languages - A Case Study on Hausa Texts. *IEEE Access*, *12*(July), 98900–98916. https://doi.org/10.1109/ACCESS.2024.3427416

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*(July 2023), 102342. https://doi.org/10.1016/j.is.2023.102342

Tabany, M., & Gueffal, M. (2024). Sentiment Analysis and Fake Amazon Reviews Classification Using SVM Supervised Machine Learning Model. *Journal of Advances in Information Technology*, *15*(1), 49–58. https://doi.org/10.12720/jait.15.1.49-58

Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text sentiment analysis and classification based on bidirectional Gated Recurrent Units (GRUs) model. *Applied and Computational Engineering*, *77*(1), 132–137. https://doi.org/10.54254/2755-2721/77/20240670