IJAAR PUBLISHING®

Global Journal of Artificial Intelligence and Technology Development | ISSN: 3027-0502
Vol. 3, Issue 1 (2025) | www.ijaar.org

Enhancing Sentiment Analysis for Hausa Language Using Dataset Augmentation and Machine Learning Models

Nasiru Mahadi[1*], Ahmad Usman[1], Yakubu Magaji[1], and Haisal Dauda Abubakar[1]
[1]Department of Computer Science, Jigawa State College of Education, Gumel

Abstract: *This study focuses on enhancing sentiment analysis for the Hausa language, a low-resource language, through dataset augmentation and machine learning models. Hausa faces significant challenges, including limited annotated datasets, linguistic complexity, and cultural nuances, which hinder the development of robust sentiment analysis systems. To address these issues, this research employs dataset augmentation techniques- synonym replacement and back-translation to create a larger, more diverse dataset. The augmented dataset is then used to train and evaluate various machine learning models, including traditional methods-Support Vector Machines (SVM) and Naive Bayes, as well as advanced deep learning models like BERT and LSTM. The results demonstrate that dataset augmentation significantly improves model performance, with BERT achieving the highest accuracy (90.22%) and F1-score (92.46%). Deep learning models consistently outperformed traditional methods, highlighting the importance of leveraging contextual embeddings and sequential learning for low-resource languages. The study also identifies challenges in handling sarcasm, irony, and regional dialectal variations, underscoring the complexity of sentiment analysis in Hausa. This research contributes to the advancement of NLP for Hausa by creating an augmented dataset, comparing machine learning models, and emphasizing the effectiveness of dataset augmentation. The findings provide valuable insights for future work in Hausa NLP and offer strategies adaptable to other low-resource languages, promoting inclusive language technologies and bridging the digital divide.*

Keywords: Sentiment Analysis, Hausa Language, Dataset Augmentation, Machine Learning Models, Low-Resource Languages.

Mahadi, N., Usman, A, Magaji, Y., & Abubakar, H. D. (2025). Enhancing Sentiment Analysis for Hausa Language Using Dataset Augmentation and Machine Learning Models. *Global Journal of Artificial Intelligence and Technology Development. 3*(1), 81-95. DOI: https://doi.org/10.5281/zenodo.15744573

1 Introduction

Sentiment analysis, a subfield of natural language processing (NLP), plays a critical role in understanding and interpreting human emotions, opinions, and attitudes expressed in text [1][2]. With the exponential growth of user-generated content on social media, review platforms, and other digital channels, sentiment analysis has become an indispensable tool for extracting actionable insights from unstructured text data [3]. By automatically classifying text as positive, negative, or neutral, sentiment analysis enables organizations to gauge public opinion, monitor brand reputation, and make data-driven decisions [2][4][5]. For instance, businesses use sentiment analysis to analyze customer feedback, improve products, and enhance customer satisfaction, while governments and NGOs leverage it to understand public sentiment on policies or social issues [6].

In addition to its practical applications, sentiment analysis contributes significantly to advancing NLP research [7]. It serves as a benchmark task for evaluating the effectiveness of machine learning and deep learning models, particularly in understanding context, sarcasm, and nuanced language [8][9]. Recent advancements in transformer-based models like BERT and GPT have demonstrated remarkable performance in sentiment analysis tasks, showcasing their ability to capture complex linguistic patterns [10][11][12]. These models have also been adapted for low-resource languages, highlighting the importance of sentiment analysis in bridging the gap between high-resource and underrepresented languages [13][14].

Moreover, sentiment analysis has become a cornerstone for interdisciplinary research, combining insights from linguistics, psychology, and computer science [15][16]. It helps researchers study emotional expression, cultural differences in communication, and the impact of language on decision-making. For example, sentiment analysis has been used to analyze mental health trends on social media, providing valuable insights into public well-being during crises like the COVID-19 pandemic [17].

Sentiment analysis for low-resource languages like Hausa faces significant challenges due to the lack of linguistic resources and research attention [14][18]. Unlike high-resource languages such as English, Hausa suffers from a scarcity of annotated datasets, which are essential for training accurate machine learning models [19]. This scarcity stems from limited digital content in Hausa and the absence of large-scale text corpora tailored for sentiment analysis. Without sufficient labeled data, developing robust sentiment analysis systems becomes inherently difficult [20].

Another challenge is the linguistic complexity of Hausa, which includes unique morphological and syntactic structures. For instance, Hausa employs agglutination and tonal variations, which are not present in many high-resource languages [21]. These features complicate text preprocessing and feature extraction, as standard NLP tools designed for English often fail to handle Hausa's linguistic nuances effectively [19][22]. Additionally, the lack of pre-trained language models for Hausa further exacerbates the problem, as researchers cannot leverage transfer learning techniques that have proven successful for other languages [23].

Cultural and contextual differences also pose challenges. Sentiment expressions in Hausa are often context-dependent and culturally specific, making it difficult to generalize sentiment labels [20]. For example, idiomatic expressions and proverbs, which are common in Hausa, may not translate well to sentiment categories without deep cultural understanding. This requires careful annotation and domain expertise, which are often lacking.

Unlike high-resource languages such as English, Hausa has limited publicly available corpora for tasks like sentiment analysis, machine translation, and text classification [19]. This scarcity stems from insufficient investment in computational linguistics for African languages, as well as the complexity of annotating data for low-resource languages [14]. The absence of standardized datasets hinders the development and evaluation of machine learning models, making it difficult to achieve state-of-the-art performance in Hausa NLP tasks [18][19] [22].

Recent studies highlight the critical need for annotated datasets to advance Hausa language processing. For instance, research by [22] and [24] emphasize the challenges of creating sentiment analysis datasets for Hausa due to the language's rich morphology and dialectal variations. Additionally, the lack of pre-trained language models tailored for Hausa further exacerbates the problem, as most existing models are optimized for high-resource languages [14]. Efforts to address these gaps, such as crowd-sourced annotations and dataset augmentation, are still in their infancy. Without substantial investment in linguistic resources, the development of robust NLP tools for Hausa will remain constrained, limiting its digital inclusion and representation in global AI systems [25].

Existing research on Hausa sentiment analysis is limited, often relying on small, non-diverse datasets that hinder model generalization [19]. Additionally, the absence of standardized tools and preprocessing techniques for Hausa exacerbates the challenge [18][26]. Dataset augmentation techniques, such as back-translation and synthetic data generation, have shown promise in low-resource languages but remain underutilized for Hausa [27]. Advanced machine learning models, including transformer-based architectures, have not been extensively applied to Hausa sentiment analysis, leaving a significant gap in achieving state-of-the-art performance [14][20].

This research aims to enhance sentiment analysis for Hausa by leveraging dataset augmentation to create a larger, more diverse dataset. By employing synonym replacement and back-translation, we address the scarcity of annotated data. Furthermore, we evaluate the performance of various traditional machine learning models, on the augmented dataset. The primary contributions of this work include the creation of an augmented Hausa sentiment analysis dataset and a comprehensive comparative analysis of machine learning models, providing insights into the most effective approaches for Hausa sentiment analysis. These contributions aim to bridge the resource gap and advance NLP research for the Hausa language.

2. Literature Review

2.1 Sentiment Analysis

Sentiment analysis, a subfield of natural language processing (NLP), focuses on identifying and extracting subjective information from text, such as opinions, emotions, and attitudes [2]. It is widely used in applications like customer feedback analysis, social media monitoring, and market research [28]. The process typically involves text preprocessing, feature extraction, and classification using machine learning or deep learning models [29][30]. Recent advancements in sentiment analysis have leveraged contextual embeddings and transformer-based models like BERT, which capture nuanced linguistic patterns [31]. Despite its success in high-resource languages, sentiment analysis for low-resource languages remains underexplored, posing challenges due to limited datasets and linguistic resources [32].

2.2 Sentiment Analysis for Low-Resource Languages

Sentiment analysis for low-resource languages, such as African languages, faces significant challenges due to the scarcity of annotated datasets and linguistic tools [14][33]. Recent studies have explored techniques like transfer learning and multilingual models to address these limitations [34]. For instance, research on Amharic and Swahili sentiment analysis has demonstrated the effectiveness of leveraging pre-trained multilingual models like mBERT and AfriBERTa [35][36]. However, Hausa, a major African language, has received limited attention in sentiment analysis research. Efforts to build annotated datasets and develop language-specific models are crucial for advancing sentiment analysis in these languages and bridging the digital divide [19][22].

2.3 Dataset Augmentation Techniques

Dataset augmentation is a critical strategy for improving model performance in low-resource settings [37][38]. Techniques like back-translation, where text is translated to another language and back, help generate diverse training samples. Synonym replacement and word embeddings-based substitution are also effective for creating variations in the dataset [39][40]. Recent research has explored synthetic data generation using generative models like GPT-3 to augment small datasets. These methods have shown promise in enhancing model generalization and robustness, particularly for languages with limited annotated data [41]. However, careful validation is required to ensure the quality and relevance of augmented data [42].

2.4 Machine Learning Models for Sentiment Analysis

Traditional machine learning models like Support Vector Machines (SVM), Naive Bayes, Gradient boosting machine and Random forest have been widely used for sentiment analysis due to their simplicity and efficiency [43][44]. These models rely on handcrafted features such as n-grams and TF-IDF [45]. In contrast, deep learning models like Long Short-Term Memory (LSTM) networks and transformer-based architectures (e.g., BERT) have revolutionized sentiment analysis by capturing contextual information and long-range dependencies [8][31]. Pretrained models like BERT and its variants (e.g., RoBERTa, DistilBERT) have achieved state-of-the-art performance across multiple languages, making them suitable for low-resource settings through fine-tuning [14][46][47].

2.5 Hausa Language NLP

Research on Hausa language processing has gained momentum in recent years, focusing on tasks like machine translation, text classification, and sentiment analysis [48]. Studies have explored the use of rule-based and statistical methods for Hausa text processing, but deep learning approaches remain underexplored [49]. The lack of large-scale annotated datasets is a major bottleneck [19][50]. Recent efforts have focused on creating Hausa corpora and leveraging multilingual models for tasks like sentiment analysis. For example, researchers have fine-tuned mBERT for Hausa sentiment classification, achieving promising results [14][51]. However, more work is needed to develop robust, language-specific models and datasets.

3 Methodology

This part details the sentiment analysis system, introduces the technique for augmenting the dataset, and elaborates on the preprocessing, feature engineering, dataset splitting, sentiment classification, and evaluation stages. The accompanying figure provides a complete overview of the research methodology.

3.1 Dataset Collection

For this research, a dataset of 2,500 Hausa tweets will be augmented. The original tweets were collected from Twitter using keywords related to politics, sports, entertainment, religion, and culture. The dataset was filtered to eliminate duplicates, retweets, and non-Hausa content. Native Hausa speakers manually annotated the remaining tweets, categorizing them on a three-point scale: positive, negative, or neutral. This annotated dataset will serve as the foundation for further analysis and augmentation in the study.

Sentiment	Number of Tweets	Percentage
Neutral	858	34.3%
Negative	800	32.6%
Positive	842	33.1%

Table 1: Distribution of Sentiment in Tweets

3.2 Data Preprocessing

Preprocessing involved cleaning, and normalization of the Hausa text. Cleaning re moved irrelevant characters, such as emojis, URLs, and punctuation, while preserving meaningful text. Tokenization split sentences into words. Normalization included converting text to lowercase, removing diacritics, and standardizing spelling variations to ensure consistency. The preprocessed text was then converted into a format suitable for feature extraction, ensuring that the input data was clean and standardized for machine learning models

3.3 Dataset Augmentation

Dataset augmentation was achieved through synonym replacement and back-translation of the original dataset. Back-translation involved translating Hausa text to English and back to Hausa using tools like Google Translate, generating paraphrased versions. Synonym replacement utilized Hausa word embeddings to swap words with their synonyms, enhancing lexical diversity. These methods aimed to expand the dataset while maintaining its linguistic integrity.

Algorithm 1 Synonym Replacement for Dataset Augmentation

Require: Original dataset D with Hausa text, Hausa word embeddings E

Ensure: Augmented dataset D' with synonym-replaced text

Initialize $D' \leftarrow \emptyset$
for each text $t \in D$ **do**
  Tokenize t into words $w_1, w_2, \ldots, w_n$
  for each word w_i in t **do**
    if w_i has synonyms in E **then**
      Replace w_i with a synonym s_i from E
    end if
  end for
  Add the modified text t' to D'
end for
return D'

Algorithm 2 Back-Translation for Dataset Augmentation

Require: Original dataset D with Hausa text, Translation tool T (Google Translate)

Ensure: Augmented dataset D' with back-translated text

Initialize $D' \leftarrow \emptyset$
for each text $t \in D$ **do**
  Translate t from Hausa to English using T to get t_{en}
  Translate t_{en} back to Hausa using T to get t'_{ha}
  Add the back-translated text t'_{ha} to D'
end for
return D'

The augmented dataset underwent rigorous validation. Native speakers manually reviewed samples to ensure quality and sentiment consistency. Automated checks measured text similarity using cosine similarity and verified label preservation. This two-step validation process ensured the augmented data remained accurate and representative of the original dataset, supporting robust model training.

Algorithm 3 Validation of the Original Dataset

Require: Original dataset D, Native speakers S, Similarity tool A

Ensure: Validated dataset D_v

$D_v \leftarrow \emptyset$
for each $t \in D$ **do**
  Validate t with S and check similarity using A
  if t passes validation **then**
    Add t to D_v
  end if
end for
return D_v

3.4 Feature Extraction

After cleaning the text data by removing stop words, punctuation, and applying stemming, we use TF-IDF (Term Frequency-Inverse Document Frequency) for feature engineering. TF-IDF converts text into numerical features, highlighting the importance of words within a corpus. It calculates the frequency of each term in a document and its inverse frequency across all documents, ensuring more relevant terms are weighted higher.

To implement TF-IDF, we utilize Python's scikit-learn library, which simplifies the computation of TF-IDF scores. This approach transforms raw text into structured numerical data, enabling effective analysis and modeling for machine learning tasks.

3.5 Machine Learning Models

In this phase, classical machine learning and deep learning models classified the original and augmented datasets into positive, negative, or neutral sentiments based on their features. Two classical models—Support Vector Machine (SVM) and Naive Bayes—along with deep learning models, BERT and LSTM, were implemented for sentiment analysis. The goal was to evaluate their performance in accurately categorizing sentiments.

The researchers aimed to enhance the accuracy of these models for Hausa sentiment analysis by leveraging the augmented dataset. By integrating the augmented data vectors into the algorithms, the study sought to improve the models' ability to capture sentiment nuances, ultimately refining their predictive capabilities.

3.6 Evaluation Metrics

This phase assesses the performance of four models using accuracy, and F1-score evaluation metrics. It examines the impact of augmentation techniques on the original dataset by employing machine learning algorithms for sentiment analysis. The results are compared with

those from the original dataset to determine the effectiveness of the augmentation methods in improving model performance.

The study aims to identify the most effective model and evaluate the significance of data augmentation in sentiment analysis tasks. By analyzing the performance metrics, it provides insights into how augmentation techniques influence the accuracy and reliability of sentiment analysis models, offering valuable guidance for future improvements in the field.

3.7 Experimental Setup

The dataset was divided into training (80%) and test (20%) sets for fair evaluation. Hyperparameters, such as learning rate, batch size, and layer count, were optimized using grid and random search. Experiments were conducted on a GPU-equipped high-performance computing cluster to meet deep learning demands, ensuring reproducibility and fair model comparison.

4. Results and Discussion

4.1 Performance of Augmented Dataset

The performance of machine learning models significantly improved when trained on the augmented dataset compared to the original dataset. For instance, the accuracy of the LSTM model increased from 77.54% to 88.56%, while the F1-score of the BERT model improved from 78.33% to 92%. Similarly, the SVM and Naive Bayes models also demonstrated notable improvements in accuracy and F1-score. The table below summarizes the performance metrics of the models on the augmented dataset. These consistent improvements across all models high light the effectiveness of dataset augmentation in addressing the scarcity of labeled Hausa sentiment data. The augmented dataset provided a more diverse and representative sample of Hausa language usage, enabling models to generalize better to unseen data.

The improvement was particularly notable in handling rare or complex sentiment expressions, which were underrepresented in the original dataset. Augmentation techniques such as back-translation and synthetic data generation introduced variations in sentence structure and vocabulary, reducing overfitting and enhancing model robustness. These results underscore the critical role of dataset augmentation for low-resource languages like Hausa, where labeled data is limited.

Methods	Accuracy (%)	F1-Score (%)
SVM	71.00	68.00
$SVM_{augmented}$	86.00	80.66
Naive Bayes	76.00	74.00
$Naive\ Bayes_{augmented}$	86.67	87.00
LSTM	77.54	73.68
$LSTM_{augmented}$	88.56	90.00
BERT	75.00	78.33
$BERT_{augmented}$	90.22	92.46

Table 2: Performance comparison of different methods and their augmented versions.

4.2 Model Comparison

The evaluation of machine learning models revealed that deep learning models, particularly BERT and LSTM, outperformed traditional models like SVM and Naive Bayes. BERT achieved the highest accuracy (90.22%) and F1-score (92.46%), followed by LSTM with an accuracy of 88.56% and an F1-score of 90%. SVM and Naive Bayes achieved accuracies of 71% and 76%, respectively, which improved to 86% and 86.67% after augmentation. The superior performance of BERT can be attributed to its ability to capture contextual information and semantic relationships in text, which are critical for sentiment analysis.

LSTM also performed well, leveraging its sequential learning capability to model dependencies in text. However, its performance was slightly lower than BERT due to its inability to capture long-range dependencies as effectively. Traditional models, while simpler and faster to train, struggled with the nuances of Hausa sentiment expressions, highlighting the need for more advanced architectures in low-resource language settings.

4.3 Error Analysis

Error analysis revealed that models often failed in cases involving sarcasm, irony, or mixed sentiments. For example, a sentence like "Wannan abu ya yi kyau sosai, amma ban yarda ba" (This thing is very good, but I don’t agree) was frequently misclassified due to its contradictory nature. Additionally, models struggled with rare or dialectal expressions that were not well-represented in the training data. For instance, regional variations in Hausa vocabulary led to misclassifications in some cases.

Another common source of error was the misinterpretation of context. For example, the word "dadi" (sweet) could express positive sentiment in some contexts but neutral or negative sentiment in others, depending on the surrounding words. These errors highlight the need for more sophisticated context-aware models and further refinement of the dataset to include diverse linguistic variations.

4.4 Impact of Dataset Augmentation

Dataset augmentation had a profound impact on model performance and generalization. By increasing the size and diversity of the training data, augmentation reduced overfitting and improved the models’ ability to handle unseen data. For example, the F1-score for the BERT model improved from 78.33% to 92.46% on the augmented dataset, demonstrating the effectiveness of techniques like back-translation and synthetic data generation. Augmentation also addressed the issue of class imbalance, which was prevalent in the original dataset. By generating additional samples for underrepresented sentiment classes, augmentation ensured a more balanced distribution, leading to fairer and more accurate predictions. These results highlight the critical role of dataset augmentation in enhancing sentiment analysis for low-resource languages like Hausa.

4.5 Limitations

This study highlights several limitations despite its promising results. Firstly, the augmented dataset, though larger, may still lack sufficient representation of regional dialects and colloquial expressions, limiting the models’ ability to generalize across all Hausa speaking communities. Secondly, the quality of synthetic data generated through back-translation and synonym replacement could introduce noise or bias, as these methods depend on the accuracy of translation tools and the availability of high quality synonym resources. Thirdly, the high computational cost of training deep learning models like BERT and LSTM may restrict accessibility for researchers with limited re sources

Additionally, the reliance on manual annotation for dataset validation introduces subjectivity, and inter-annotator disagreements may affect label consistency. Future work should explore more robust validation techniques, such as automated consistency checks and the use of multiple annotators with diverse linguistic backgrounds. Incorporating additional linguistic resources, like sentiment lexicons and dialect-specific corpora, could further enhance dataset quality and representativeness. Overall, the study underscores the importance of dataset augmentation and advanced machine learning models in improving sentiment analysis for low-resource languages like Hausa. While significant progress has been made, continued research and collaboration are essential to address these limitations and develop more robust, inclusive language technologies. The findings provide valuable insights for future advancements in sentiment analysis for Hausa and other underrepresented languages

5. Conclusion

This study demonstrated the effectiveness of dataset augmentation and advanced machine learning models in enhancing sentiment analysis for the Hausa language. The augmented dataset, created using techniques like back-translation and synonym replacement, significantly improved model performance, with BERT achieving the highest accuracy (90.22%) and F1-score (92.46%). Deep learning models, particularly BERT and LSTM, consistently outperformed traditional methods such as SVM and Naive Bayes, highlighting the importance of leveraging contextual embeddings and sequential learning for low-resource languages. The results also revealed challenges in handling sarcasm, irony, and regional dialectal variations, underscoring the complexity of sentiment analysis in Hausa. The findings emphasize the critical role of dataset quality and diversity in training robust sentiment analysis models. By addressing the scarcity of labeled data and improving model generalization, this research provides a foundation for future work in Hausa NLP.

The success of augmentation techniques and deep learning models offers valuable insights for similar efforts in other low-resource languages. Additionally, the study highlights the need for more sophisticated context-aware models and further refinement of datasets to include diverse linguistic variations, ensuring more accurate and inclusive sentiment analysis systems for Hausa and other underrepresented languages.

5.1 Contributions

This research makes several key contributions to the field of sentiment analysis for the Hausa language. First, it introduces a comprehensive augmented dataset that addresses the scarcity of labeled Hausa sentiment data, enabling more accurate and reliable model training. Second, it provides a comparative analysis of traditional and deep learning models, demonstrating the superiority of contextual embeddings and sequential learning architectures like BERT and LSTM. Third, the study highlights the importance of dataset augmentation techniques, such as back-translation and synthetic data generation, in improving model performance and generalization.

Additionally, this work contributes to the broader NLP community by showcasing effective strategies for sentiment analysis in low-resource language settings. The findings and methodologies can be adapted to other African languages, fostering advancements in multilingual NLP research. By addressing the unique challenges of Hausa sentiment analysis, this research paves the way for more inclusive and equitable language technologies.

5.2 Future Work

Future research should explore more advanced augmentation techniques, such as adversarial training and domain-specific data generation, to further enhance dataset quality and diversity. Leveraging multilingual models like XLM-R and AfriBERTa could also improve performance by incorporating cross-lingual transfer learning. Additionally, efforts should focus on creating more comprehensive linguistic resources for Hausa, including sentiment lexicons and annotated corpora, to support model training and evaluation.

Another promising direction is the development of context-aware models that can better handle sarcasm, irony, and mixed sentiments. Incorporating multimodal data, such as text and emojis from social media, could also improve sentiment analysis accuracy. Finally, expanding the scope to include regional dialects and colloquial expressions will ensure that models are more representative of the diverse Hausa-speaking population. These advancements will contribute to more robust and inclusive sentiment analysis systems for Hausa and other low-resource languages.

6. References

[1] A. Berrajaa, "Natural Language Processing for the Analysis Sentiment using a LSTM Model," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 5, pp. 777–785, 2022, doi: 10.14569/IJACSA.2022.0130589.

[2] M. Wankhade, A. C. S. Rao, and C. Kulkarni, *A survey on sentiment analysis methods, applications, and challenges*, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.

[3] A. S. George and T. Baskar, "Partners Universal International Innovation Journal (PUIIJ) Leveraging Big Data and Sentiment Analysis for Actionable Insights : A Review of Data Mining Approaches for Social Media Partners Universal International Innovation Journal (PUIIJ)," no. August, pp. 39–59, 2024, doi: 10.5281/zenodo.13623777.

[4] "Marcus Vinicius Estrela Borges Data-Driven Marketing : A sentiment analysis study in Nonprofit Marketing," no. October, 2022.

[5] T. Anderson, S. Sarkar, and R. Kelley, "Analyzing public sentiment on sustainability: A comprehensive review and application of sentiment analysis techniques," *Nat. Lang. Process. J.*, vol. 8, no. August, p. 100097, 2024, doi: 10.1016/j.nlp.2024.100097.

[6] M. G. Moldovan, D. C. Dabija, L. Stanca, and C. B. Pocol, "A Qualitative Study on the Consumer Behaviour Related to Food Waste: Romanian Perspectives through Word Cloud and Sentiment Analysis," *Sustain.* , vol. 16, no. 10, 2024, doi: 10.3390/su16104193.

[7] H. Huang, A. A. Zavareh, and M. B. Mustafa, "Sentiment Analysis in E-Commerce Platforms: A Review of Current Techniques and Future Directions," *IEEE Access*, vol. 11, no. July, pp. 90367–90382, 2023, doi: 10.1109/ACCESS.2023.3307308.

[8] N. A. Sharma, A. B. M. S. Ali, and M. A. Kabir, "A review of sentiment analysis: tasks, applications, and deep learning techniques," *Int. J. Data Sci. Anal.*, 2024, doi: 10.1007/s41060-024-00594-x.

[9] A. D. Yacoub, S. O. Slim, and A. E. Aboutabl, "A Survey of Sentiment Analysis and Sarcasm Detection: Challenges, Techniques, and Trends," *Int. J. Electr. Comput. Eng. Syst.*, vol. 15, no. 1, pp. 69–78, 2024, doi: 10.32985/ijeces.15.1.7.

[10] K. Kheiri and H. Karimi, "SentimentGPT: Exploiting GPT for Advanced Sentiment Analysis and its Departure from Current Machine Learning," 2023, [Online]. Available: http://arxiv.org/abs/2307.10234

[11] G. Yenduri *et al.*, "GPT (Generative Pre-Trained Transformer) - A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions," *IEEE Access*, vol. 12, no. March, pp. 54608–54649, 2024, doi: 10.1109/ACCESS.2024.3389497.

[12] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-of-the-art review," *Nat. Lang. Process. J.*, vol. 6, no. February, p. 100059, 2024, doi: 10.1016/j.nlp.2024.100059.

[13] T. Zhong *et al.*, "Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research," pp. 1–31, 2024, [Online]. Available: http://arxiv.org/abs/2412.04497

[14] Y. Aliyu, A. Sarlan, K. Usman Danyaro, A. S. B. A. Rahman, and M. Abdullahi, "Sentiment Analysis in Low-Resource Settings: A Comprehensive Review of Approaches, Languages, and Data Sources," *IEEE Access*, vol. 12, no. April, pp. 66883–66909, 2024, doi: 10.1109/ACCESS.2024.3398635.

[15] R. Kataishi and D. Ph, "The Technological Trajectory of Semantic Analysis : A Historical-Methodological Review of NLP in Social Sciences".

[16] M. Kumar, L. Khan, and H. T. Chang, "Evolving techniques in sentiment analysis: a comprehensive review," *PeerJ Comput. Sci.*, vol. 11, pp. 1–61, 2025, doi: 10.7717/PEERJ-CS.2592.

[17] D. Valdez, M. ten Thij, K. Bathina, L. A. Rutter, and J. Bollen, "Social media insights into US mental health during the COVID-19 pandemic: Longitudinal analysis of twitter data," *J. Med. Internet Res.*, vol. 22, no. 12, 2020, doi: 10.2196/21418.

[18] I. Adebara, "Towards Afrocentric natural language processing," no. February, 2024, [Online]. Available: https://open.library.ubc.ca/collections/24/items/1.0440415

[19] I. Inuwa-Dutse, "NaijaNLP: A Survey of Nigerian Low-Resource Languages," vol. 1, no. 1, 2025, [Online]. Available: http://arxiv.org/abs/2502.19784

[20] T. A. Jimoh, T. De Wille, and N. S. Nikolov, "Bridging Gaps in Natural Language Processing for Yor\`ub\'a: A Systematic Review of a Decade of Progress and Prospects," pp. 1–45, 2025, [Online]. Available: http://arxiv.org/abs/2502.17364

[21] I. Adebara, A. R. Elmadany, M. Abdul-Mageed, and A. Alcoba, "SERENGETI: Massively Multilingual Language Models for Africa," *Proc. Annu. Meet. Assoc. Comput. Linguist.*, pp. 1498–1537, 2023, doi: 10.18653/v1/2023.findings-acl.97.

[22] H. A. Shehu *et al.*, "Unveiling Sentiments: A Deep Dive into Sentiment Analysis for Low-Resource Languages – A Case Study on Hausa Texts," *IEEE Access*, vol. PP, pp.

1–1, 2024, doi: 10.1109/access.2024.3427416.

[23] ANANDA MUHAMAD TRI UTAMA, “No 主観的健康感を中心とした在宅高齢者における 健康関連指標に関する共分散構造分析Title,” vol. 9, pp. 356–363, 2022.

[24] R. Y. Zakari, “A Systematic Literature Review of Hausa Natural Language Processing,” no. September, pp. 4–11, 2021, doi: 10.24203/ijcit.v10i4.86.

[25] D. L. Ofori-addo, “INFORMATION AND COMMUNICATIONS TECHNOLOGY IN SUB-SAHARAN PR,” 2024.

[26] H. A. Shehu *et al.*, “Unveiling Sentiments: A Deep Dive Into Sentiment Analysis for Low-Resource Languages - A Case Study on Hausa Texts,” *IEEE Access*, vol. 12, no. July, pp. 98900–98916, 2024, doi: 10.1109/ACCESS.2024.3427416.

[27] R. L. Myoya, A. Public, and T. User, “Analysing Public Transport User Sentiment,” 2024.

[28] M. Salampasis, G. Paltoglou, and A. Giachanou, “Using social media for continuous monitoring and mining of consumer behaviour,” *Int. J. Electron. Bus.*, vol. 11, no. 1, pp. 85–96, 2014, doi: 10.1504/IJEB.2014.057905.

[29] V. Maslej-Krešňáková, M. Sarnovský, P. Butka, and K. Machová, “Comparison of deep learning models and various text pre-processing techniques for the toxic comments classification,” *Appl. Sci.*, vol. 10, no. 23, pp. 1–26, 2020, doi: 10.3390/app10238631.

[30] Q. Li *et al.*, “A Survey on Text Classification: From Traditional to Deep Learning,” *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, 2022, doi: 10.1145/3495162.

[31] “A COHERENT KNOWLEDGE-DRIVEN DEEP LEARNING MODEL FOR IDIOMATIC- AWARE SENTIMENT ANALYSIS OF UNSTRUCTURED TEXT USING BERT TRANSFORMER BASHAR M A TAHAYNA (COMPUTER SCIENCE) FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY UNIVERSITI TUNKU ABDUL RAHMAN,” no. March, 2023.

[32] N. Pahari, “Sentiment Analysis on Code Switched and Low Resource Settings”.

[33] Y. Aliyu, A. Sarlan, K. U. Danyaro, and A. S. B. A. Rahman, “Comparative Analysis of Transformer Models for Sentiment Analysis in Low-Resource Languages,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 4, pp. 353–364, 2024, doi: 10.14569/IJACSA.2024.0150437.

[34] A. R. Jafari, B. Heidary, R. Farahbakhsh, M. Salehi, and M. Jalili, “Transfer Learning for Multi-lingual Tasks -- a Survey,” 2021, [Online]. Available: http://arxiv.org/abs/2110.02052

[35] M. Gokani, “Exploring Sentiment Analysis in Low-resource Languages,” 2023.

[36] S. H. Muhammad *et al.*, “SemEval-2023 Task 12: Sentiment Analysis for African Languages (AfriSenti-SemEval),” *17th Int. Work. Semant. Eval. SemEval 2023 - Proc. Work.*, pp. 2319–2337, 2023, doi: 10.18653/v1/2023.semeval-1.315.

[37] M. Xia, X. Kong, A. Anastasopoulos, and G. Neubig, "Generalized data augmentation for low-resource translation," *ACL 2019 - 57th Annu. Meet. Assoc. Comput. Linguist. Proc. Conf.*, pp. 5786–5796, 2020, doi: 10.18653/v1/p19-1579.

[38] M. Seo, J. Baek, J. Thorne, and S. J. Hwang, "Retrieval-Augmented Data Augmentation for Low-Resource Domain Tasks," 2024, [Online]. Available: http://arxiv.org/abs/2402.13482

[39] A. A. ElSabagh, S. S. Azab, and H. A. Hefny, *A comprehensive survey on Arabic text augmentation: approaches, challenges, and applications*, vol. 0123456789. Springer London, 2025. doi: 10.1007/s00521-025-11020-z.

[40] J. Kapusta, D. Držik, K. Šteflovič, and K. S. Nagy, "Text Data Augmentation Techniques for Word Embeddings in Fake News Classification," *IEEE Access*, vol. 12, no. March, pp. 31538–31550, 2024, doi: 10.1109/ACCESS.2024.3369918.

[41] I. N. Gomez, "Using Generative AI in Research," *Philipp. J. Allied Heal. Sci.*, vol. 7, no. 2, 2024, doi: 10.36413/pjahs.0702.001.

[42] S. Gholami and M. Omar, "Does Synthetic Data Make Large Language Models More Efficient?," 2023, [Online]. Available: http://arxiv.org/abs/2310.07830

[43] Q. Siddique, "Comparative Analysis of Sentiment Classification Techniques on Flipkart Product Reviews : A Study Using Logistic Regression , SVC , Random Forest , and Gradient Boosting," *J. Digit. Mark. Digit. Curr.*, vol. 1, no. 1, pp. 21–42, 2024, doi: 10.47738/jdmdc.v1i1.4.

[44] M. Ahmad, S. Aftab, S. S. Muhammad, and S. Ahmad, "Machine Learning Techniques for Sentiment Analysis: A Review," *Int. J. Multidiscip. Sci. Eng.*, vol. 8, no. 3, pp. 27–32, 2017, [Online]. Available: http://www.ijmse.org/Volume8/Issue3/paper5.pdf

[45] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, "Text Classification: How Machine Learning Is Revolutionizing Text Categorization," *Inf.*, vol. 16, no. 2, 2025, doi: 10.3390/info16020130.

[46] V. K. Agbesi *et al.*, "Pre-Trained Transformer-Based Models for Text Classification Using Low-Resourced Ewe Language," *Systems*, vol. 12, no. 1, pp. 1–22, 2024, doi: 10.3390/systems12010001.

[47] A. S. Alammary, "BERT Models for Arabic Text Classification: A Systematic Review," *Appl. Sci.*, vol. 12, no. 11, 2022, doi: 10.3390/app12115720.

[48] K. R. Mabokela, T. Celik, and M. Raborife, "Multilingual Sentiment Analysis for Under-Resourced Languages : A Systematic Review of the Landscape," *IEEE Access*, vol. 11, no. November 2022, pp. 15996–16020, 2023, doi: 10.1109/ACCESS.2022.3224136.

[49] G. Bhatia *et al.*, "UBC-DLNLP at SemEval-2023 Task 12: Impact of Transfer Learning on African Sentiment Analysis," 2023.

[50] D. I. Adelani *et al.*, "A Few Thousand Translations Go A Long Way! Leveraging Pre-trained Models for African News Translation," *NAACL 2022 - 2022 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Conf.*, no. 1, pp. 3053–

3070, 2022, doi: 10.18653/v1/2022.naacl-main.223.

[51] S. A. Sani, S. H. Muhammad, and D. Jarvis, “Investigating the Impact of Language-Adaptive Fine-Tuning on Sentiment Analysis in Hausa Language Using AfriBERTa,” 2025, [Online]. Available: http://arxiv.org/abs/2501.11023