Science Discovery Artificial Intelligence
2026, Vol. 1, No. 1, pp. 14–26
https://doi.org/10.11648/j.sdai.20260101.13

Research Article

# Performance Evaluation of Hybrid Bert Model on Code-mixed for Hausa-English Using Adapted Pre-trained Data

**Ali Garba Jakwa[1, *], Faseki Ngozi Franscisca[1], Abubakar Yunusa Ahmad[1], Musa Ibrahim[2]**

[1]Department of Computer Science, Nigerian Army University Biu, Biu, Nigeria

[2]Department of Cyber Security, Nigerian Army University Biu, Biu, Nigeria

## Abstract

This research evaluates the potentials of using BERT (Bidirectional Encoder Representations from Transformers) language model on code-mixed for English-Hausa Language code-mixed using adapted pre-trained dataset. The main aim of this research was to unveil the potential benefits of using pre-trained models for handling code-mixed data to improved language understanding and context sensitivity in relation to Hausa-English-Language, the objective of this research was achieved by developing a BERT model that is capable of handling Hausa-English code-mixed dataset exploring different machine learning language models by training the chosen model with the adapted English-Hausa Language code-mixed. What necessitates this research was due to low data corpus on the language domain of Hausa-English code-mixed while other languages were explored like English-Hindu Code-Mixed. The model was developed using python transformer library. The adapted pre-trained dataset was first pre-processed, tokenized and fine-tuned in order to fit into the BERT model, the dataset was normalized in the context of code-mixed conversation based on annotate language labels to distinguish between English and Hausa Language segments in the code-mixed text, appropriate parameter for training were set with different optimization strategies for fine-tuning, adjusted learning rate, batch sizes and training epochs for performance optimization. The model was evaluated based on accuracy, F1-score, precision and recall for Code-Mixed tasks, the results of HauBERT our proposed model showed more than 90% accuracy, the result was compared with state-of-the-art BERT language models, and the study recommended that this adapted pre-trained model should be applied in large language model for language understanding and context sensitivity.

## Keywords

Multilingual, Machine Learning, Conversational Code-mixed, Local Large Language Model

*Correspondence: Ali Garba Jakwa (alijakwa@gmail.com)

**Received:** 8 October 2025; **Accepted:** 27 January 2026; **Published:** 21 February 2026

# 1. Introduction

Multilingualism has become increasingly important in globalized world today according to different sources, including the United Nations. With the ease of internet and travel, people from different country and cultures are interacting with one another more frequently, and the ability to communicate effectively without any language barriers is very important [1]. In the world increased connectivity, where multilingual conversation are becoming the order of the day, the need for efficient and effective language processing models is more important than ever. According to [2], Code-mixing is a traditional linguistic phenomenon in multilingual communities, where speakers interchange between two or languages within single conversation or between two or more sentences. Natural Language Processing (NLP) development techniques for code-mixed data is a challenging task due to the nature of complexity and changing nature of code-mixed language use [3].

BERT (Bidirectional Encoder Representation from Transformer), have displayed outstanding performance in different areas, including text classification, machine translation and sentiment analysis [4, 28]. According to study by [5] which presented L3CubeHingCorpus, and was observed as the first large-scale real Hindi-English code-mixed data In a Roman script, the model consisting of 52.93 million sentences and 1.04 billion tokens obtained from Twitter. BERT models was introduced in the study for pre-training on code-mixed Hing corpus with the objective of using masked language modelling, in which the study showed effectiveness in downstream tasks like code-mixed sentiment analysis, Named Entity Recognition (NER), POS tagging, and LID from the GLUECoS benchmark.

The use of multiple language in the same text in a given writing is termed as code mixed. It is clear that there are a lot of adaptions of exploring code-mixed data, especially combination of regional language and English language during conversation on social media platforms [6].

Code-mixed question answering dataset has been created and discussed by different researchers [7], which include 1,694 Hinglish, 2,848 Tamlish, and 1,391 Tenglish factoid questions together with their answers, these datasets were used for code-mixed question answering challenges, the data was used by seven teams that took the data from the database [4]. However, it was observed BERT models was not mentioned or research for Hausa-English code-mixed data in this area. in essence, this study, proposed to unveil the potential of BERT model for Hausa-English code-mixed conversational data as an ongoing area of research and also to improve the inclusion of Hausa as a low resource language in the domain of QA dataset [8].

The unique challenges posed to the Hausa-English code-mixed conversational data is low-resource languages, due to the minimum or less training data and the nature of the difficulty of language mixing, in order to take care of these challenges, researchers have extracted different methods which includes multilingual contrastive training and learning transfer to improve the evaluation and performance of question answering systems in the area of low-resource language settings [9].

The development of KenSwQuAD for Swahili, question answering datasets which was purposely designed for low-resource languages, has greatly contributed to the advancement of machine understanding in these languages [10]. Valuable resources were provided by these datasets for training and evaluating question answering models, this gesture helps researchers to deal with new approaches and methods for handling the difficulties of code-mixed conversational data in low-resource languages.

To create a reliable Question-answering system, this research will train and optimize these models using sizable corpus of code-mixed conversational data. Language specialist, academics, and developers working in the field of machine learning will eventually find the Insights from this study to be a useful resource as they help create more precise and effective language processing models for multilingual conversations.

The aim of this paper is to unveil the potentials of BERT model on English-Hausa Code-mixed using adapted pre-trained data.

The major contributions of this work are to;

1) Develop a dataset for Hausa-English language that will be used to test the potential of BERT models and evaluate their capabilities in Hausa language understanding.
2) Design a BERT model that can interpret Hausa-English language code-mixed dataset by Injecting text identification tag in the traditional BERT model.
3) Evaluate the model and compare it with the state-of-the-art BERT models used for code-mixed conversational question answering, however, the study presents the results in the context of English-Hausa code-mixing and the idea hopes to be extended generically to any other code-mixed language.

*Feature of the code-mixed dataset*

*Original text:* what is abin dake gefen hagu of the window.

*Preprocessed Text:*

*Interleaved Word Language*

*Table 1. Language identification tag.*

| what is | abin dake gefen hagu | of the window |
|---|---|---|
| ENGLISH | HAUSA | ENGLISH |

*Approach:* what is EN abin dake gefen hagu HA of the window EN

| | | |
|---|---|---|
| 16 | what is opposite to the talabijin | sofa |
| 17 | what is abin da ke bayan hagu of the sofa | game_table |
| 18 | what is abin da ke gefen hagu | bed |
| 19 | what is that da ke bayan television | remote_control |
| 0 | what is that kayan ado close to the wall | wall_decoration |
| 1 | what is abin da ke gefen hagun books din | television |
| 2 | which object ne babba | sofa |
| 3 | what is that tsakanin kujeru and sofa | table |
| 4 | what is the kalar sofa | brown |
| 5 | menene aka gudanar a cikin coffee machine | coffee_pot |
| 6 | what is sama da fan | window |

*Figure 1. CSV file of the code-mixed conversational text for Hausa-English.*

## 2. Paper Organization

This paper is organized based on the following sections, section one presented the introduction and general background of the paper including the concepts of natural language processing, section two presented organization of the paper, section three presented the literature review of the research work, section four presented the methodology on how the research was carried out with the experiment and design, section five presented the experimental results and their discussions and section six presented the recommendations and conclusion.

## 3. Literature Review

This section presents a literature review and related work from different authors on natural language processing and code-mixed with their accompanying experiment and discussions.

Conversational Machine Comprehension (CMC) literature aimed at highlighting the importance of multi-turn CQA, which has achieved advantage due to the advances in natural language understanding and the large-scale availability of conversational datasets [11]. Proposed and reviewed on a comprehensive overview of CQA, based on consolidated scattered knowledge in the AI domain, their review focused on common trends over recently published models. Furthermore, the review emphasized on a generic platform for CQA models while paying more emphasis on the need for streamlining future area of research.

In a proposed framework and published research by [12] on the techniques and challenges for creating an effective dialogue system for AI, as in the case of Chabot voice assistants, or conversational agents, in which they emphasized that these systems generally depend on user inputs, the ability to produce understandable and relevant responses, depends on the ability to maintain dialogue history and state. And they observed that developing a dialogue system for AI, requires some number of methods, which include the use of dialogue blueprint or scripts to achieve user needs and develop understanding [13].

According to published work by [14] Argued that a lot of work on conversational AI and machine understanding underscore the higher level of interest and potential field. multi-turn CQA are more focused on, there are a lot of challenges and methods for designing effective dialogue systems, this area and domain are labelled with high consolidation of knowledge that are indicative of having important strides in being developed in advancing conversational AI and machine understanding.

Code-mixing is the technique of switching between two or more languages which has privilege in bi and multi-lingual communities. This technique is usually available in natural language processing (NLP) applications, they are generally designed with thinking of single interaction language and are being struggled with code-mixed pronunciations, and the above discussion is being hindered with a lot of difficulties which account into the development of research and emergence of code-mixed question answering (CMQA). [15] observed that a lot of challenges in the area of code-mixed question answering and it requires a significant effort to address the complexity posed to question answering therefore they proposed a research on linguistically motivated techniques for code-mixed question generation (CMQG) and neural network-based architectures to handle code-mixed question answering (CMQA), their model have been evaluated on benchmark datasets,

like SQuAD and MMQA, and model demonstrated effectively with a promising results, but model was developed to handle only question answering conversations.

The mechanism designed for handling conversation in the system is the dialog manger (DM) and is the most important component of a dialog system, it has the capability for handling the process and movement of the conversation. Human communication are collected and are generally converted to some kind of speech or text acts, the dialogue agents are used to decide the next action and predict the state of the conversation history. In other aspect of language pre-training [16] Observed that the essential component in Chatbot is dialogue management that is used to conduct contextual communications, and it helps in comprehension aspects that makes the agent design in the dialogue management to be right.

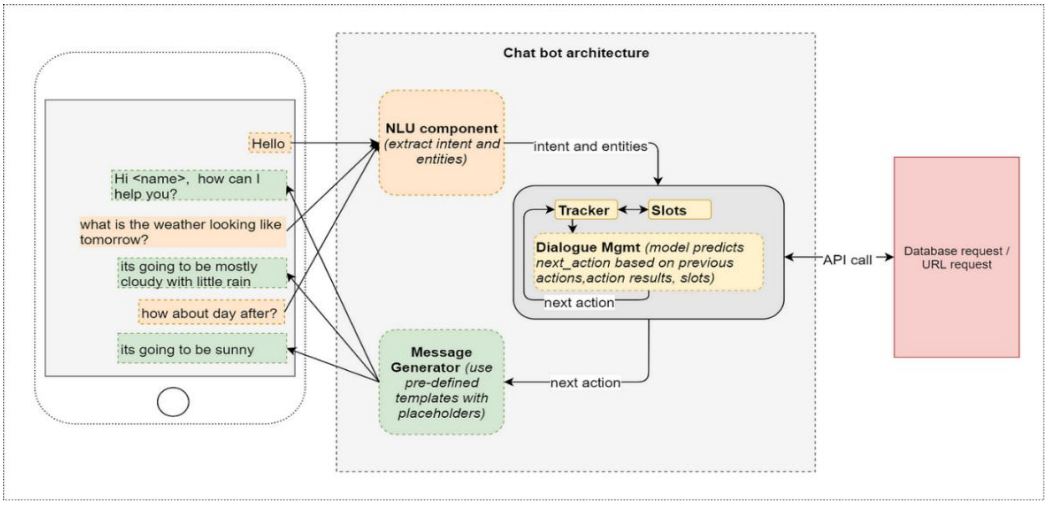


***Figure 2.*** *Chat bot CQA Architecture. Source: [29].*

According to [16] so many design approaches for dialogue structure were envisaged by exploring different approaches and frameworks like VoiceXML, AIML, Facade, ChatScript, CDM, and TuTalk, and used to Improve speech dialogs and chat-bots, control devices and other tutorial dialogs etc, they observed that these dialog structures are generally described as state chart diagrams categorized by formal languages like SCXML.

In a research by [16] they proposed an improved dialog management system by enhancing the scalability in dealing with scarcity of data and improved efficiency in training, in their model, they adopted deep learning methods in which Alexa conversation was used as a dialog manager to accept inputs, these techniques allows developers to have experience on overall design focus [17].

Consequently, [18] proposed a multilingual conversational question answering model by leveraging the performance of the pre-trained language model (PLM) for question answering purpose, their model performed well on large language model thereby improved the performance of natural language processing task on low-resource languages.

Another effort by [19] on adapting multilingual PLM and CQA, the study proposed a multilingual adaptive fine-tuning (MAFT) which brought about changing the architecture of PLM by using pre-raining objective their approach have shown great effectiveness in adapting across new languages and promote transfer learning across languages [20]. Furthermore, it has been observed multilingual frameworks for CQA are based on pre-trained multilingual transformers like mBERT, or mT5 [21].

In a study by [22] argued that the important of requirement for evaluating conversational model for Hausa-English code-mixed conversational data is availability of appropriate and suitable dataset in which they proposed in their research, they observed that use of other datasets for evaluating the model as provided by CC-Net repository can only evaluate mono and multilingual models, therefore they proposed additional development on the code-nixed conversational corpus which will help in evaluating QA models in code-mixed languages.

Based on cited work by [22], it can be observed that evaluating pre-trained language framework involves evaluating the performance of the model on some number of tasks which includes QA and sentiment analysis in code switched language pairs. Empirical techniques have been envisaged to access the switching accuracy in code-switched sentences, with a clear indication of better performance and improved natural language understanding tasks.

In a research by [6] proposed a model by examining the performance of BERT based models on low-resource code-mixed Hindi-English Datasets through experimenting language augmentation approaches, their approach demonstrated and assessed the performance of code-mixed on five different code-mixed Hindi-English classification datasets based on

hate speech detection, sentiment analysis, and emotion detection. Their model was evaluated using metrics such as accuracy, precision, recall, and F1 score. Their proposed findings on the language augmentation approach presented better results on the chosen models but research was restricted on one language code-mixed data without exploiting other datasets in their experiment.

*Summary of the Literature Review*

In this review, the study has investigated the effectiveness of using adapted pre-trained models for code-mixed conversational data, and different frameworks that could be tailored between Hausa and English. The literature review analyzed various existing Conversational models and assessed their performance in handling code-mixed data. The subsequent section presents the methods used in carrying out the research.

# 4. Methodology

This research proposed a HauBERT model that can handle Hausa-English code-mixed effectively, as it can be observed from the literature which shows there were no suitable dataset for low-resource languages like Hausa-English code-mix, therefore, this research developed dataset for Hausa-English code-mixed, and subsequently developed a BERT model that can handle Hausa-English language code-mixed by injecting language identification tag for Hausa language in the traditional BERT architecture.

## 4.1. Problem Formulation

The purpose of this research was to design an efficient BERT model that can handle Hausa-English code-mixed data effectively, in this system, a language identification tag was proposed and injected into the traditional BERT model architecture in order to keep track of the Hausa language character which has different layouts, so as task specification will be improved.

### 4.1.1. Existing System

*Traditional BERT architecture*

BERT model is a two-way encoder and decoder model in which text or dataset are processed in both directions in order to analyze language inputs through the aid of transformer capability.

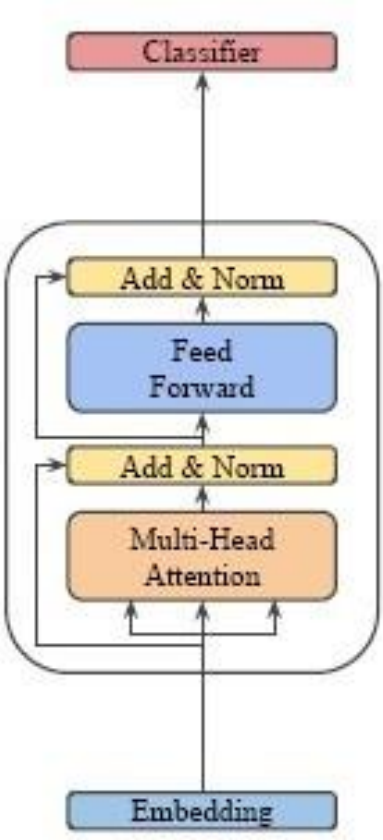


***Figure 3.*** *BERT architecture [23].*

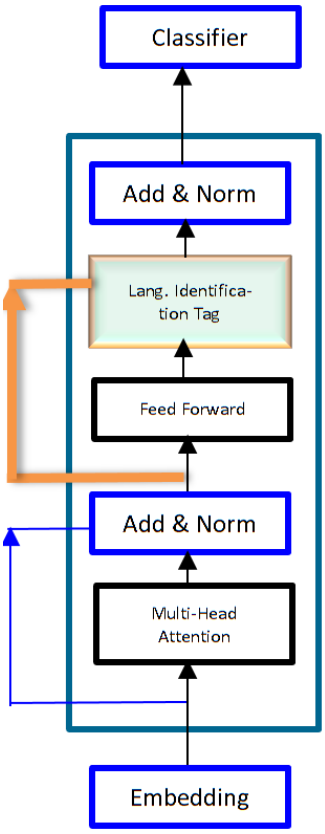


***Figure 4.*** *HauBERT architecture.*

Figure 3. above shows the architecture of BERT model used for natural language processing, the traditional BERT works by accepting code-mixed text and process based on the language argumentation strategy, the model accepts embedded text with different annotations and normalized based on the system requirement, further more feed forwarded for classification with the aid of classifier [23].

### 4.1.2. Proposed System

*Modified BERT architecture*

The proposed system modified BERT is a BERT model proposed by the modification of the traditional BERT model by this in mind, model was modified at the encoding stage in which coded text or dataset are being identified by the aid of language identification tag after following the formal normalization process, after successful tagging the text will be normalized for further classification as shown in Figure 4.

## 4.2. System Implementation

The model was designed and implemented using Python transformer library; this library is an open-source large language model for simulating natural language processing in machine learning environment.

## 4.3. Sequence Diagram

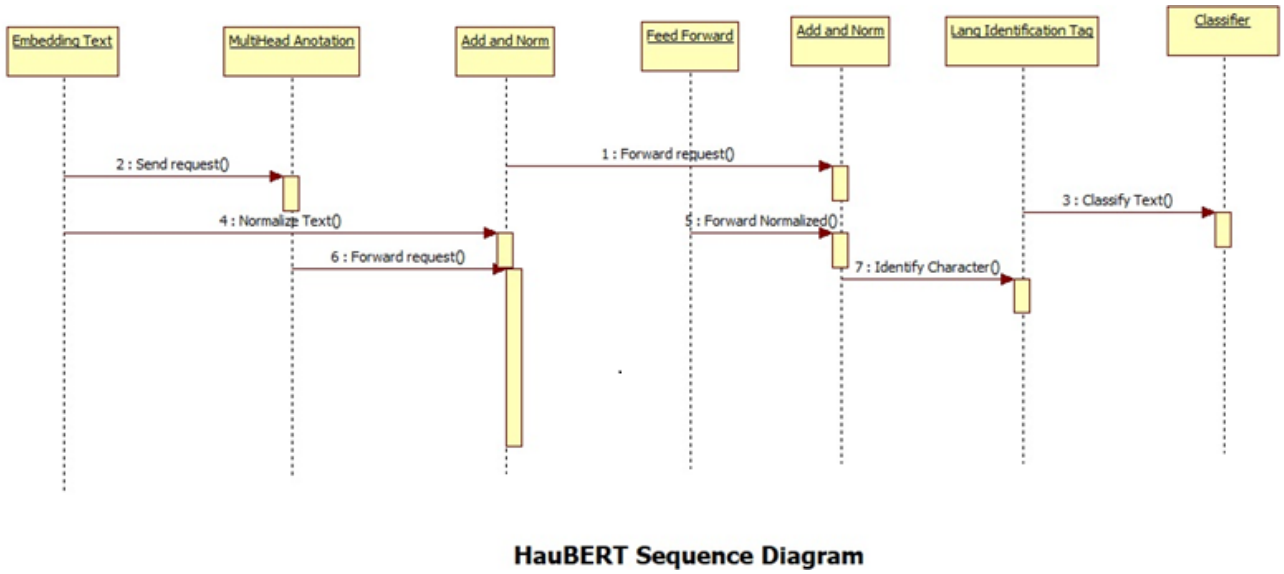


***Figure 5.*** *Sequence Diagram.*

The Figure 5 above depicts the work flow of the proposed BERT model which shows the following objects as the main parameters of the model; parameters include text embedding, Multi-Head Attention, text normalization, feed forward processing, Language Identification and classifier as specified by the sequence diagram above Figure 5.

This study employs language identification tags to unveil the potential of BERT model on English-Hausa language code-mixed by following these laid down approaches.

## 4.4. Data Preprocessing

Text Normalization and Tokenization: The code-mixed text Standardized and tokenized to ensure consistency in representation, accounting for language-specific linguistic variations for better suit in to the model.

Language Identification: language identification algorithms Implemented to identify and label the language components within the code-mixed sentences in an appropriate language model for processing.

### 4.4.1. Model Selection and Preprocessing

Pre-trained Language Models: Choose appropriate pre-trained language models such as BERT, mBERT, or XLM-R, capable of handling multilingual and code-mixed data. The data was normalized by converting it to lowercase text in which Python's regex and emoji libraries were used for text preprocessing. Model loading and Adaption; the dataset was loaded in the pre-trained language model and adapted to the Hausa-English code-mixed context using fine-tuning techniques.

### 4.4.2. Fine-tuning and Training

Fine-tuning Setup: Split the annotated dataset into training, validation, and test sets. Utilize the training set to fine-tune the pre-trained models for code-mixed question answering. Training Parameters; Define appropriate hyper-parameters and optimization strategies for fine-tuning, adjusting learning rates, batch sizes, and training epochs to optimize performance.

## 4.5. Model Evaluation

Performance Metrics: the model performance was evaluated based on the following evaluation metrics: accuracy, F1 score, precision, and recall for Conversational tasks. Comparative Analysis; the performance of the HauBERT model was compared with existing results of other Code-Mixed BERT models on standard evaluation datasets to assess the performance and identify areas for enhancement and observing the potentials of HauBERT model in handling Code-Mixed for English-Hausa Language. The comparison was performed on three dataset types, which includes original dataset developed during this research, online hate speech dataset, and Sentiment dataset. The results for the training, testing and validation are presented in the next sections.

## 4.6. Datasets

***Table 2.*** *Dataset train-test-validation split.*

| S/NO | Dataset name | Total size | Train size | Test size | Eval. Size | Labels |
|---|---|---|---|---|---|---|
| 1 | Developed dataset | 1960 | 1372 | 294 | 294 | 2 |
| 2 | Hate speech dataset | 2250 | 1575 | 338 | 338 | 2 |
| 3 | Sentiment dataset | 9974 | 6982 | 1496 | 1496 | 6 |


***Figure 6.*** *Dataset train-test-validation split.*

Table 2 shows the size of the available corpus, which are

divided into testing, training, and validation splits, and accompanying labels associated to each dataset. The model performance was evaluated on the adapted datasets was trained and evaluated based on three parameters; developed dataset, hate speech and sentiment analysis.

1) Sentiment dataset

This dataset focused on emotions that talks about happiness or sadness in the text adopted, whether expressing sadness, surprise, anger, disgust, fear and happiness the dataset was divided into three which includes; testing, training and validation with ratio of 15:70:15, respectively.

2) Hate speech dataset

This dataset focused on emotions that talks about abuse in the text adopted, whether expressing bad comment or bullying, the dataset was divided into three which includes; testing, training and validation with ratio of 15:70:15, respectively.

3) The developed dataset

This dataset focused on traditional code-mixed of basic question answering scheme in Hausa with corresponding English language counter-part as depicted in Figure 1, CSV file, the dataset was divided into three which includes; testing, training and validation with ratio of 15:70:15, respectively as shown in Figure 6.

### 4.7. Models Used

The following BERT models were used in the research as a benchmark in carrying out this research. [6].

1) HingBERT is a BERT model proposed in 2022 which is publicly available and accessible which was trained on benchmark dataset provided by L3Cube, this model was powerful and robust on evaluating datasets.
2) HingRoBERTa: this model was created in 2022 it was based on RoBERTa model which was fine-tuned on combination of Hindu and English codes in Roman text [28].
3) HingRoBERTa-Mixed: this model was proposed in 2022 by Nayak and Joshi, It Is a BERT model that combines English and Hindi codes together In Roman scripts and it was fine-tuned In Devanagari code code-mixed text [26].
4) Multilingual BERT: mBERT model was proposed in 2018 by Devlin et al, it was also a version of BERT which was trained on a large multilingual dataset.
5) AlBERT: A Lite BERT (ALBERT) model was proposed in 2019 by Ian et al, is model used for self-supervised learning in language representation, this model exhibits same architectural feature as BERT, but main purpose was to improve performance [31].
6) BERT: was proposed by [30], this model is a self-supervised transformer and was formerly pre-trained on large corpus of multilingual data.
7) RoBERTa: was proposed in 2019 which was also known as Robustly Optimized BERT Pre-training Approach, this model was based on Google's BERT model, it differs from normal BERT in the sense that it removes next sentence steps of pre-training [25].
8) HauBERT Is the proposed Large Language model designed for the purpose of this research, what differentiate this model with the traditional BERT model is the injection of language identification tag for Hausa language characters in order to Identifier Hausa-English code-mixed effectively. The model was designed and implemented using Python transformer library; this library Is an open source large language model for simulating natural language processing in machine learning environment.

## 5. Results

This section presented the simulated results of the models used; these tables displayed the raw results of the training, testing and validation followed by discussion of the results accordingly.

***Table 3.*** *Results displaying F1 score.*

| Models | Hate Speech | Emotions | Developed Dataset |
|---|---|---|---|
| HingBERT | 0.96854 | 0.94586 | 0.84657 |
| HingRoBERTa | 0.84557 | 0.95564 | 0.75487 |
| HingRoBERTa-Mixed | 0.95645 | 0.96584 | 0.87654 |
| Multilingual BERT | 0.98548 | 0.99584 | 0.87546 |
| AlBERT | 0.98567 | 0.96584 | 0.76548 |
| BERT | 0.89754 | 0.86594 | 0.76546 |
| RoBERTa | 0.98654 | 0.99546 | 0.87657 |
| HauBERT | 0.98765 | 0.98675 | 0.98764 |

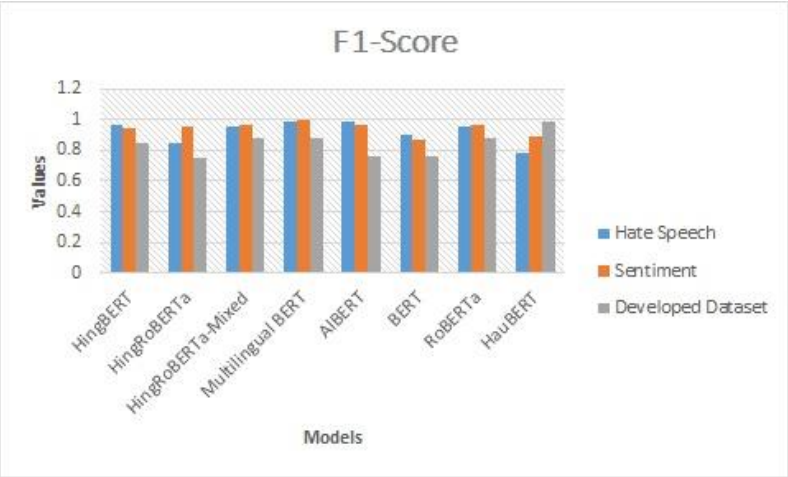


***Figure 7.*** *F1-score.*

Table 3 and figure above presented the F1 score of the three categories of datasets used, from the table it can be observed the models were trained, tested and validated based on the percentage of parameter presented in the previous section, the result displayed for F1 score when utilizing HauBERT with better results with regards to the developed datasets followed by Multilingual BERT (mBERT) with regards to Hate speech and Sentiment dataset of which HingRoBERTa model also showed a promising result in terms of sentiment with slight difference than other models, this is evident probably due to the size of the dataset used, in contrast with RoBERTa and HingBERT models displayed a better result with slight differences when compared with other models as presented in Table 3 and Figure 7 respectively.


***Figure 8.*** *Developed dataset results.*

***Table 4.*** *Developed dataset results.*

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| HingBERT | 0.76858 | 0.76757 | 0.85437 |
| HingRoBERTa | 0.82548 | 0.83425 | 0.88775 |
| HingRoBERTa-Mixed | 0.82657 | 0.84453 | 0.76386 |
| Multilingual BERT | 0.85731 | 0.84523 | 0.88765 |
| AlBERT | 0.75735 | 0.94523 | 0.71035 |

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| BERT | 0.81738 | 0.74423 | 0.87651 |
| RoBERTa | 0.71734 | 0.84525 | 0.87652 |
| HauBERT | 0.98765 | 0.90765 | 0.96764 |

The above Table 4 and figure above showed the performance evaluation of the HauBERT dataset, after exclusive simulation experiment the result of the experiment HauBERT model indicated greater than 90% outcome with regards to the developed dataset as compared to other models in which better accuracy and recall as shown in Table 4 and Figure 8 respectively.

***Table 5.*** *Hate speech dataset results.*

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| HingBERT | 0.88547 | 0.64423 | 0.85437 |
| HingRoBERTa | 0.92548 | 0.83425 | 0.95637 |
| HingRoBERTa-Mixed | 0.92657 | 0.84453 | 0.94537 |
| Multilingual BERT | 0.85731 | 0.94523 | 0.80037 |
| AlBERT | 0.85735 | 0.94523 | 0.91035 |
| BERT | 0.81738 | 0.84423 | 0.85025 |
| RoBERTa | 0.91734 | 0.94525 | 0.93542 |
| HauBERT | 0.98765 | 0.98765 | 0.94764 |

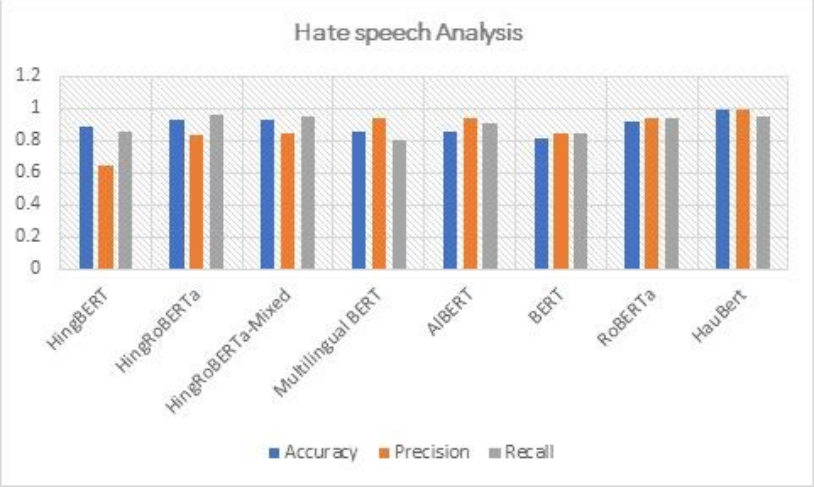


***Figure 9.*** *Hate Speech Analysis.*

Table 5 above showed the accuracy, precision and recall for the dataset used for hate speech analysis code-mixed, from the table it can be observed the models were trained, tested and validated based on the percentage of parameter presented in the previous section, the result displayed for when utilizing HauBERT, Multilingual BERT (mBERT) and HingRoBERT-Mixed models showed a promising result with slight difference than other models with regards to accuracy, precision and recall, this is evident due to the size of the dataset used, in contrast with the Hate speech dataset, RoBERTa and HingBERT models displayed a better recall result with slight difference when compared with other models as presented in Table 5 and Figure

9 respectively in which similar result was observed by [27].

*Table 6. Sentiment Analysis results.*

| Models | Accuracy | Precision | Recall |
|---|---|---|---|
| HingBERT | 0.88547 | 0.64444 | 0.85422 |
| HingRoBERTa | 0.92548 | 0.83445 | 0.95641 |
| HingRoBERTa-Mixed | 0.92657 | 0.84456 | 0.94544 |
| Multilingual BERT | 0.85731 | 0.94234 | 0.80022 |
| AlBERT | 0.85735 | 0.94565 | 0.98031 |
| BERT | 0.81738 | 0.84412 | 0.95234 |
| RoBERTa | 0.91734 | 0.94564 | 0.98542 |
| HauBERT | 0.98765 | 0.98763 | 0.96761 |

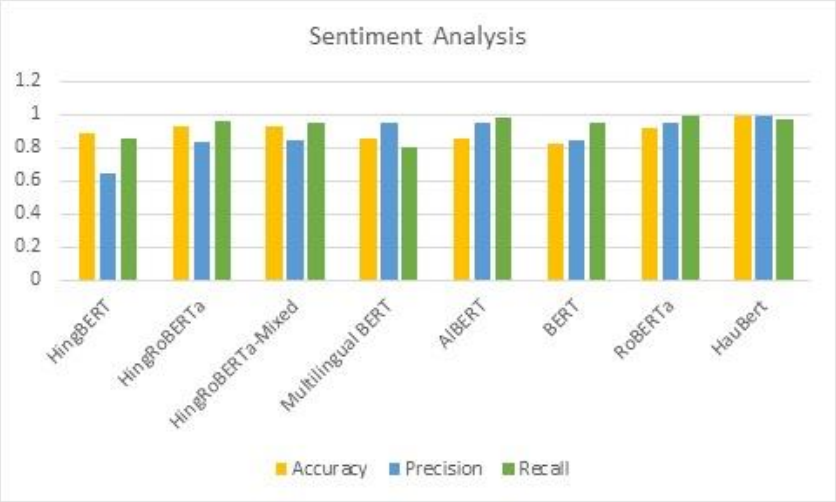


*Figure 10. Sentiment Analysis.*

Table 6 above showed the accuracy, precision and recall for the dataset used for sentiment analysis code-mixed, from the table it can be observed better accuracy and precision was obtained when the models were trained, tested and validated based on the percentage of parameter presented in the previous section, the result displayed accuracy and precision when utilizing HauBERT model with relative closer results as in Multilingual BERT (mBERT) with better precision and HingRoBERT-Mixed model with good recall, which showed a promising result with slight difference than other models, in terms of accuracy, precision and recall, this is evident due to the size of the dataset used, in contrast with the RoBERTa and AlBERT models displayed a better precision result with slight difference when compared with other models as presented in Table 6 and Figure 10 respectively, this result also Indicated a closer relationship with the work proposed by [24, 29].

## 6. Conclusion

This study evaluated the performance of HauBERT model on conversational question answering for English-Hausa language Code-Mixed with the aim to improve language understanding and context sensitivity among languages, the study introduced language identification in the regular BERT flow to improve the BERT model capability for handling Hausa language character set. HauBERT model was proposed and implemented on python transformer library, the model was put to test differently and Its behaviors was observed by subjecting it to the adapted pre-trained dataset, after several training, testing and validation, the results revealed that the HauBERT model presented a better result which is greater than 90% with regards to F1 score, accuracy, precision and recall, our proposed model was compared with other benchmark BERT

models like multilingual BERT (mBERT) and HingRoBERT-Mixed models by using sentiment analysis dataset, Hate speech dataset and the developed dataset, while In contrast to using Hate speech, developed dataset a better F1 score was achieved with HauBERT with slight differences to RoBERTa and HingRoBERT models this is evident due to the size of the dataset used in carrying out the experiment, with this outcome, the results indicated the potential of using BERT model in conversational question answering code-mixed with respect to English-Hausa language, hence it can be concluded that HauBERT model can be applied in large language model for language understanding and context sensitivity, hence the objectives of this research has been mate, and the study recommended that further research will conducted on decoding end of the BERT model since this research modified the traditional BERT model at the encoding stage to ascertain the performance of the BERT model.

## Abbreviations

| | |
|---|---|
| BERT | Bidirectional Encoder Representations from Transformers |
| mBERT | Multilingual Bidirectional Encoder Representations from Transformers |
| HauBERT | Hausa Bidirectional Encoder Representations from Transformers |
| HingRoBERTa | Hindi and English Robustly Optimized Bidirectional Encoder Representations from Transformers Pre-training Approach |
| HingBERT | Hindi and English Bidirectional Encoder Representations from Transformers |
| AlBERT | A Lite Bidirectional Encoder Representations from Transformers |
| RoBERTa | Robustly Optimized Bidirectional Encoder Representations from Transformers Pretraining Approach |

## Conflicts of Interest

The authors declare no conflicts of interest.

## References


[1] Stein-Smith, K. (2022). Multilingualism And Its Purposes–Interdisciplinary Applications In Language Education And Advocacy. *Journal of Languages for Spec (JLSP)*, 67.

[2] Shashirekha, H. L., Balouchzahi, F., Anusha, M. D., & Sidorov, G. (2022). CoLI-machine learning approaches for code-mixed language identification at the word level in Kannada-English texts. *arXiv preprint arXiv: 2211.09847*.

[3] Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T., & Barzilay, R. (2024). Conformal Language Modeling. 12th International Conference on Learning Representations, ICLR 2024, 1–28.

[4] Chandu, K., Loginova, E., Gupta, V., Genabith, J. V., Neumann, G., Chinnakotla, M.,... & Black, A. W. (2019). Code-mixed question answering challenge: Crowd-sourcing data and techniques. In *Third Workshop on Computational Approaches to Linguistic Code-Switching* (pp. 29-38). Association for Computational Linguistics (ACL).

[5] Nayak, R., & Joshi, R. (2022). L3Cube-HingCorpus and HingBERT: A Code-Mixed Hindi-English Dataset and BERT Language Models. *arXiv preprint arXiv: 2204.08398*.

[6] Takawane, G., Phaltankar, A., Patwardhan, V., Patil, A., Joshi, R., & Takalikar, M. S. (2023). Language augmentation approach for code-mixed text classification. Natural Language Processing Journal, 5(October), 100042. https://doi.org/10.1016/j.nlp.2023.100042

[7] Zhang, Y., Shen, B., & Cao, X. (2022). Learn a prior question-aware feature for machine reading comprehension. *Frontiers in Physics*, *10*, 1311.

[8] Zhang, W., Majumdar, A., & Yadav, A. (2024). Code-mixed LLM: Improve Large Language Models' Capability to Handle Code-Mixing through Reinforcement Learning from AI Feedback. http://arxiv.org/abs/2411.09073

[9] Kumar, Gokul Karthik, Abhishek Singh Gehlot, Sahal Shaji Mullappilly, and Karthik Nandakumar. "Mucot: Multilingual contrastive training for question-answering in low-resource languages." *arXiv preprint arXiv: 2204.05814* (2022).

[10] Wanjawa, B. W., Wanzare, L. D., Indede, F., McOnyango, O., Muchemi, L., & Ombui, E. (2023). KenSwQuAD—A Question Answering Dataset for Swahili Low-resource Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), 1-20.

[11] Gupta, S., Rawat, B. P. S., & Yu, H. (2020). Conversational machine comprehension: a literature review. *arXiv preprint arXiv: 2006.00671*.

[12] Balouchzahi, F., Butt, S., Hedge, A., Ashraf, N., Shashirekha, H. L., Sidorov, G., & Gelibukh, A. (2022). Overview of CoLI-Kanglish: Word Level Language Identification in Code-mixed Kannada-English Texts at ICON 2022. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts* (pp. 38-45).

[13] McTear, M. (2021). Introducing Dialogue Systems. In *Conversational AI: dialogue systems, conversational agents, and chatbots* (pp. 11-42). Cham: Springer International Publishing.

[14] Baradaran, R., Ghiasi, R., & Amirkhani, H. (2022). A survey on machine reading comprehension systems. *Natural Language Engineering*, *28*(6), 683-732.

[15] Gupta, D., Lenka, P., Ekbal, A., & Bhattacharyya, P. (2018, October). Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 119-130).

[16] Dai, Y., Yu, H., Jiang, Y., Tang, C., Li, Y., & Sun, J. (2020). A survey on dialog management: Recent advances and challenges. *arXiv preprint arXiv: 2005.02233*.

[17] Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q. & Pettigrue, A. (2018). Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv: 1801.03604*.

[18] Gupta, D. (2022). Learning to Answer Multilingual and Code-Mixed Questions. *arXiv preprint arXiv: 2211.07522*.

[19] Nowakowski, K., Ptaszynski, M., Murasaki, K., & Nieuważny, J. (2023). Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, *60*(2), 103148.

[20] Alabi, B. (2022). Basic intuition of Conference Question Answering Systems (CQA), AR/VR Jouney: Augmented & Virtual Reality Magazine.

[21] Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems*, *64*(12), 3151-3195.

[22] Das, R., Ranjan, S., Pathak, S., & Jyothi, P. (2023, July). Improving Pretraining Techniques for Code-Switched NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1176-1191).

[23] Patil, A., Patwardhan, V., Phaltankar, A., Takawane, G., & Joshi, R. (2023). Comparative Study of Pre-Trained BERT Models for Code-Mixed Hindi-English Data.

[24] Deriu, J., Rodrigo, A., Otegi, A., Echegoyen, G., Rosset, S., Agirre, E., & Cieliebak, M. (2021). Survey on evaluation methods for dialogue systems. *Artificial Intelligence Review*, *54*, 755-810.

[25] Gessler, L. (2023). *Low-Resource Monolingual Transformer Language Models* (Doctoral dissertation, Georgetown University).

[26] Janarthanam, S. (2017). *Hands-on chatbots and conversational UI development: build chatbots and voice user interfaces with Chatfuel, Dialogflow, Microsoft Bot Framework, Twilio, and Alexa Skills*. Packt Publishing Ltd.

[27] Joshi M, Choi E, Weld DS, Zettlemoyer L (2017) TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Minneapolis, Minnesota, pp 1601–1611. https://doi.org/10.18653/v1/N19-1028

[28] Phade, A., & Haribhakta, Y. (2021, November). Question Answering System for low resource language using Transfer Learning. In *2021 International Conference on Computational Intelligence and Computing Applications (ICCICA)* (pp. 1-6). IEEE.

[29] Raychawdhary, N., Das, A., Dozier, G., & Seals, C. D. (2023, July). Seals_Lab at SemEval-2023 Task 12: Sentiment Analysis for Low-resource African Languages, Hausa and Igbo. In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)* (pp. 1508-1517).

[30] Thara, S., Sampath, E., & Reddy, P. (2020, June). Code mixed question answering challenge using deep learning methods. In *2020 5th international conference on communication and electronics systems (ICCES)* (pp. 1331-1337). IEEE.

[31] Yusuf, A., Sarlan, A., Danyaro, K. U., & Rahman, A. S. B. (2023, August). Fine-tuning Multilingual Transformers for Hausa-English Sentiment Analysis. In *2023 13th International Conference on Information Technology in Asia (CITA)* (pp. 13-18). IEEE.