International Journal of Scientific Research in Computer Science and Engineering
Vol.**12**, Issue.**6**, pp.**58-65**, December **2024**
E-ISSN: 2320-7639
Available online at: www.isroset.org

Research Article

Development of the Annotated Swahili Digraph Corpus Using a CNN-Based Digraph Extraction Model

Tirus Muya Maina[1*], Aaron Mogeni Oirere[2], Stephen Kahara[3]

[1,2,3]Computer Science Department, Murang'a University of Technology, Murang'a, Kenya

**Corresponding Author: tmuya@mut.ac.ke*

Received: 28/Oct/2024; **Accepted**: 30/Nov/2024; **Published**: 31/Dec/2024

Abstract— This study undertakes the development of the Annotated Swahili Digraph Corpus, utilizing a convolutional neural network-based model specifically designed for the extraction of digraphs. This initiative addresses a significant gap in the availability of dedicated digraph corpora for the Swahili language, which is increasingly needed for various applications in Natural Language Processing (NLP). The CNN-based model was accurately crafted to optimize the extraction and classification of digraphs, taking full advantage of the annotated features within the corpus. Digraphs are pairs of letters that create distinct sounds in a language, and Swahili's linguistic structure presents unique challenges and requirements in this regard. Therefore, specialized tools and models are essential for ensuring accurate transcription and efficient speech recognition that cater specifically to the nuances of the Swahili language. The resulting Swahili Digraph Corpus comprises a comprehensive collection of 31,197 words, each systematically annotated to highlight their respective digraphs. Notably, this corpus features the nine key Swahili digraphs: "ch," "dh," "gh," "kh," "ng'," "ny," "sh," "th," and "ng." Furthermore, it includes annotations for vowel distribution, showcasing the core vowels "a," "e," "i," "o," and "u." This detailed annotated corpus supports a wide array of NLP applications, enabling researchers and developers to utilize accurate linguistic data for tasks such as text processing, machine translation, and speech synthesis. Through this dedicated effort, we aim to enhance the resources available for processing the Swahili language, ultimately contributing to its greater accessibility in the digital landscape.

Keywords— Annotated, Swahili, Digraph, Corpus, NLP, CNN, Dense layer

1. Introduction

Swahili, a significant language in East Africa, presents unique challenges for Natural Language Processing (NLP) due to its complex linguistic features, particularly digraphs combinations of letters representing distinct sounds. Existing NLP tools often struggle to accurately process these digraphs, leading to issues in applications such as transcription services and speech recognition. These challenges highlight the need for more specialized resources and advanced models that can handle the intricacies of Swahili's phonetic structure effectively.

To address this issue, this research develops the Annotated Swahili Digraph Corpus using a Convolutional Neural Network (CNN)-based extraction model. This initiative aims to enhance the accuracy of digraph identification and processing in Swahili texts, thereby improving the performance of NLP applications related to the Swahili language. By bridging the gaps in current language technologies, this work seeks to foster better communication and understanding within diverse linguistic contexts in East Africa, ultimately making language tools more effective for Swahili speakers and learners alike.

1.1 Objective of the Study

The primary objective of this study is to develop the Annotated Swahili Digraph Corpus using a CNN-based digraph extraction model. This addresses the challenges of Swahili's unique phonetic and linguistic features by:

2. Creating a comprehensive dataset of Swahili digraphs systematically annotated to highlight their phonetic properties and vowel distributions.
3. Enhancing the accuracy and efficiency of Natural Language Processing (NLP) applications by improving digraph recognition and transcription tasks for the Swahili language.
4. Supporting advancements in Swahili speech recognition, transcription, and language modelling by providing a robust resource tailored to its unique linguistic features.

4.1 Significance of the Study

This study is significant as it contributes to the advancement of Natural Language Processing (NLP) for Swahili, a widely spoken yet under-resourced language in East Africa, by:

1. Addressing Resource Gaps: Filling a critical gap in linguistic resources with a specialized annotated corpus, enabling effective training and evaluation of NLP models.

2. Improving Technological Accessibility: Enhancing the usability and precision of transcription, translation, and speech recognition tools tailored to Swahili's phonetic structure.
3. Fostering Linguistic Research: Providing a foundational resource for further exploration into Swahili linguistics, phonology, and their integration into AI technologies.
4. Supporting Inclusivity: Promoting linguistic inclusivity by enabling the development of tools and technologies that cater to the unique needs of Swahili-speaking communities.

The study's outcomes are expected to drive innovation in NLP applications for Swahili and serve as a model for advancing language processing for other low-resource languages.

2. Relate Works

The theoretical framework underpinning this study integrates established language theories, machine learning principles, and optimization techniques, guiding the research design and model development. Linguistic theories, particularly those related to the phonetics of Swahili, explain the language's unique digraph properties and inform the selection of features for model training. Distinctive Feature Theory, which identifies the smallest unique traits of phonemes, provides a systematic approach to categorising language sounds and has practical applications in speech therapy, language acquisition, and phonetic recognition systems. Formal Language Theory, crucial for understanding computational linguistics, offers a theoretical foundation for specifying programming languages, developing algorithms, and studying computational complexity. Optimisation Theory seeks to identify the optimal solutions within defined constraints, offering a framework for decision-making and resource allocation. This thorough theoretical foundation guarantees that the model effectively captures the phonetic nuances of Swahili, thereby improving the effectiveness and reliability of NLP applications [1] [2] [3] [4].

Swahili, recognized as a vital language in East Africa, poses considerable challenges for Natural Language Processing (NLP) due to its unique linguistic features and complexities. One of the primary hurdles is the accurate processing of Swahili digraphs combinations of letters that represent distinct sounds which are often overlooked or misinterpreted by existing NLP tools. This inadequacy significantly impacts various applications, including transcription services, speech recognition systems, and language learning platforms, where precise language representation is crucial for effective communication and comprehension [5].

To address this pressing issue, this research undertakes the development of the Annotated Swahili Digraph Corpus, employing a CNN-based extraction model. This model is designed to improve the accuracy of digraph identification and processing in Swahili texts. By creating this corpus, the research aims to establish a comprehensive and reliable resource that can enhance the performance of NLP applications related to the Swahili language. Ultimately, this work seeks to bridge the gap in existing language technologies, making them more effective for Swahili speakers and learners alike, and thus fostering better communication and understanding within diverse linguistic contexts in East Africa. NLP [6].

Swahili, plays a critical role in communication across multiple nations. However, despite its significance, Swahili faces challenges in the field of NLP, with existing tools often failing to process its unique linguistic features. A significant challenge lies in Swahili’s use of digraphs combinations of two letters that together represent a single sound, such as "ch," "dh," "gh," "kh," "ng’," "ny," "sh," "th," and "ng," [7]. These digraphs are essential to Swahili's phonetic structure, yet existing NLP systems, which are often designed for languages without such features, struggle to accurately recognize and process them.

The second challenge in Swahili NLP stems from the interaction between Swahili's simple vowel system and its digraphs. While Swahili only has five vowels "a", "e", "i", "o", "u", the way these vowels combine with digraphs creates complex phonetic patterns [8], [9], [10]. Automated transcription and speech recognition systems must account for these interactions to avoid errors in processing. However, many existing NLP models for Swahili are based on phonemes or word-level recognition, which overlooks the significance of digraphs and vowel combinations in the language’s phonology, further complicating the development of accurate language processing systems.

The absence of specialized and comprehensive NLP resources, such as annotated corpora focused on Swahili digraphs, has aggravated the difficulty in creating advanced tools for transcription and speech recognition [10]. While some Swahili language datasets exist, they often fail to address the unique features of digraphs and their role in Swahili’s pronunciation and meaning. This lack of a detailed corpus hinders the training and evaluation of NLP models that can effectively handle Swahili's linguistic complexities.

The lack of specialized and comprehensive natural language processing (NLP) resources, particularly annotated corpora that specifically focus on Swahili digraphs, has significantly complicated the development of advanced tools for tasks such as transcription and speech recognition. While several datasets pertaining to the Swahili language do exist, these often inadequately address the unique linguistic intricacies associated with digraphs two letter combinations that represent distinct sounds or phonemes and their vital contributions to both pronunciation and meaning within the language [10].

For instance, certain digraphs can alter the meaning of words entirely or affect the way words are pronounced, which is particularly important for effective communication and comprehension. Without a targeted corpus that captures these intricacies, the process of training and evaluating NLP models becomes hindered, leading to suboptimal performance in understanding and generating Swahili. This deficiency not

only poses challenges for researchers and developers aiming to create robust language technologies but also limits the broader application of NLP in enhancing everyday interactions and services that involve the Swahili language. Addressing this gap by creating a detailed and representative corpus is crucial for advancing the capabilities of NLP tools that can truly accommodate the rich linguistic features of Swahili [11].

This research addresses these gaps by creating the Annotated Swahili Digraph Corpus, a comprehensive dataset that systematically categorizes 31,197 words by digraph and vowel context. This corpus will provide a foundation for training NLP models to handle Swahili's unique phonetic structures, specifically its digraphs. In addition, the research proposes a CNN-based digraph extraction model, which leverages this annotated corpus to improve the accuracy of digraph recognition and transcription tasks. By employing Convolutional Neural Networks (CNNs), this model can automatically learn complex patterns from the data, offering a promising approach to enhance Swahili NLP.

The Annotated Swahili Digraph Corpus and CNN-based extraction model will address key linguistic issues such as digraph recognition and phonetic complexity, ultimately advancing Swahili language processing and supporting the development of more effective and inclusive NLP technologies.

This paper is organized as follows: Section 2 provides a comprehensive description of the methodology utilized. Section 3 presents the Annotated Swahili Digraph Corpus. Section 4 examines and analyses the results obtained, while Section 5 offers the conclusions drawn from the study.

3. Methodology

The Design Science Research Methodology (DSRM) served as the guiding framework for the development of the Swahili Digraphs Model, providing a structured and systematic approach to designing, implementing, and evaluating innovative solutions to complex challenges in speech recognition. DSRM was particularly well-suited for addressing the intricacies of Swahili speech recognition, with a specific focus on digraph extraction. The methodology facilitated a comprehensive process, beginning with the identification of challenges associated with Swahili digraphs, such as their phonetic complexity and the limitations of existing corpora. This was followed by the design phase, where algorithms for digraph extraction were conceptualized and developed using insights from linguistic principles and computational techniques. The implementation phase involved translating these conceptual designs into practical coding solutions, ensuring alignment with the identified requirements and objectives. Finally, the evaluation phase utilized empirical validation to rigorously assess the performance of the digraph extraction model, enabling iterative refinements based on testing feedback [12].

The adoption of DSRM is justified by its emphasis on a systematic and problem-solving-oriented approach, which is essential for navigating the complexities of developing effective solutions in speech recognition. By adhering to DSRM principles, the researchers ensured that the Swahili Digraphs Model was both technically robust and practically relevant, capable of accurately transcribing Swahili speech. This approach not only addresses critical challenges in Swahili speech recognition but also contributes to advancements in the field, enhancing the accessibility and usability of speech recognition systems for Swahili speakers. [13].

3.1 The Design Science Research Methodology Stages

The DSRM was systematically applied in the development of the Swahili Digraphs Model to address the challenges associated with recognizing and processing Swahili digraphs. During the **problem identification** stage, the unique linguistic challenges of Swahili digraphs, including their phonetic complexity and the scarcity of suitable datasets, were identified. These challenges highlighted the inadequacy of generic language processing models and established the need for a specialized solution tailored to Swahili.

In the **objectives definition** stage, clear and measurable goals were set, focusing on improving the accuracy of digraph recognition, enhancing computational efficiency, and ensuring the scalability of the model across various natural language processing (NLP) tasks. These objectives provided a foundation for guiding the model's development.

The **design and development** phase involved creating the Swahili Digraphs Model using Dense Layers and Convolutional Neural Network (CNN) architectures. These advanced architectures were designed to facilitate effective feature extraction and precise recognition of Swahili digraphs. In the **demonstration** phase, the model was tested on Swahili text datasets to validate its effectiveness. The results were compared with baseline models, showing significant improvements in both accuracy and efficiency, thereby demonstrating the model's capability to address the identified challenges.

During the **evaluation** stage, the model's performance was assessed using key metrics, including accuracy, precision, recall, and F1 score. Additionally, error analysis and user feedback provided valuable insights into areas for further refinement. Finally, the **model refinement** phase utilized these insights for iterative improvement, involving architectural adjustments, retraining with diverse datasets, and continuous reassessment of performance. This iterative process ensured that the Swahili Digraphs Model evolved into a robust and adaptable solution for Swahili language processing.

By following DSRM principles, the study successfully addressed the complex challenges of Swahili digraph recognition, contributing to advancements in Swahili NLP and enhancing the accessibility of language processing tools for low-resource languages.

3.2. CNN-Based Digraph Extraction Model

Model Architecture: The CNN model was methodically designed to address the unique challenges posed by Swahili digraphs in NLP. The architecture comprises several layers, each contributing to the model's overall performance [14]

a) **Input Layer:** The input to the model includes pre-processed features derived from Swahili text, such as normalized vowel proportions and one-hot encoded representations of digraphs. This ensures effective handling of Swahili's diverse phonetic patterns.
b) **Convolutional Layers (Conv1D):** These layers utilize various filters on the input data to identify local patterns within the digraph sequences. The dimensions of the filters and kernels are specifically optimized for sequence analysis, which enables the model to recognize subtle phonetic distinctions found in Swahili digraphs.
c) **Flatten Layer:** This layer reshapes the feature maps generated by the convolutional layers into a one-dimensional vector, thereby streamlining the transition to the dense layers. This transformation is crucial for enabling advanced reasoning and pattern recognition at a higher level, as it facilitates the integration of spatial features into a more structured format suitable for interpretation and decision-making in the network's subsequent stages.
d) **Dense Layers:** These fully connected layers perform the classification tasks by combining information from the convolutional layers to accurately identify and classify digraphs. Multiple dense layers ensure the model can learn complex representations of the input features.
e) **Dropout Layer:** To prevent overfitting, a dropout layer is incorporated, which randomly deactivates a subset of neurons during training. This forces the model to generalize better and improves its performance on unseen data.
f) **Output Layer:** The final layer generates predictions for digraph recognition by outputting the probability distribution over possible digraph classes, allowing the model to make accurate and informed decisions.

The architecture of the CNN-based model is depicted in Figure 1, offering a detailed representation of the data flow across the network's various layers. This diagram elucidates the interconnections and functional roles of each component, highlighting their collective contribution to the processing of information. It underscores the critical features that enhance the overall effectiveness and performance of the model.

Figure 1: CNN-based digraph extraction model

3.3 Creation of the Annotated Swahili Digraph Corpus

To build a comprehensive and representative corpus, data was sourced from a variety of secondary sources. These sources include:
Mendeley Data: Kiswahili datasets that provide a rich collection of Swahili text, capturing the linguistic diversity of the language.
Harvard Dataverse: Multilingual corpora that include Swahili, offering a wealth of textual data from different dialects and contexts.
Zenodo and Kaggle: Language modelling datasets that contribute additional Swahili text, enhancing the corpus's breadth and depth.
These datasets were selected to ensure that the corpus encompasses a wide range of Swahili dialects and phonetic contexts, making it a robust resource for NLP applications.

3.4 Annotation Process

The collected data was methodically annotated to highlight key linguistic features:

a) Digraph Labels

Each word in the corpus was labelled with its corresponding Swahili digraphs, such as "ch," "dh," "gh," "kh," "ng’," "ny," "sh," "th," and "ng." This labelling creates a detailed map of digraph occurrences, which is essential for accurate NLP processing.

b) Vowel Distribution

Words were further classified based on their vowel content (a, e, i, o, u). This annotation allows for the analysis of phonetic patterns and their distribution across different words, contributing to a deeper understanding of Swahili's phonological structure.

4. Results

4.2 The Annotated Swahili Digraph Corpus

The study aimed to create and annotate a Swahili digraph corpus using a CNN-based extraction model, resulting in the Annotated Swahili Digraph Corpus. Table 1 below outlines the corpus's composition, detailing the occurrence of various digraphs across different vowel contexts, with a total of 31,197 words [15].

Table 1: The Annotated Swahili Digraph Corpus

Digraph	Words with 'a'	Words with 'e'	Words with 'i'	Words with 'o'	Words with 'u'	Total
ch	2,022	1,594	1,716	2,232	1,919	9,483
dh	1,155	130	1,159	92	323	2,859
gh	126	293	340	30	52	841
kh	24	11	9	0	0	44
ng'	62	43	26	45	23	199
ny	1,000	472	495	255	63	2,285
sh	1,003	576	741	466	703	3,489
th	68	134	142	12	37	393
ng	2,439	2,000	2,369	2,395	2,401	11,604
Total	7,899	5,253	6,997	5,527	5,521	31,197

The corpus provides a comprehensive representation of the phonetic diversity found in the Swahili language by incorporating an extensive array of digraphs. These digraphs include "ch," "dh," "gh," "kh," "ng," "ny," "sh," "th," and "ng'," which are paired with the vowels "a," "e," "i," "o," and "u." Notably, the digraph "ch" stands out with a remarkable 9,483 occurrences, while "ng" demonstrates an impressive and balanced presence, appearing 11,604 times. This rich variety underlines the phonetic adaptability of these digraphs. Such a well-distributed selection plays a crucial role in enabling the Convolutional Neural Network (CNN) model to generalize effectively across different vowel combinations, significantly improving the accuracy of digraph recognition. Furthermore, the inclusion of less common digraphs like "kh" and "ng'" enriches the dataset, offering invaluable insights into the model's capability to detect infrequent phonetic elements. This ensures both rigorous training and thorough evaluation of the model.

This curated dataset not only supports the advancement of the model for processing the Swahili language but also lays the groundwork for the development of effective and context-sensitive natural language processing (NLP) applications, catering to the unique characteristics of Swahili phonetics.

4.3 Annotated Corpus Significance

The Annotated Swahili Digraph Corpus holds a profound significance for the study and application of the Swahili language. It serves as a thorough and exhaustive repository of Swahili digraphs pairs of letters that represent specific sounds and their corresponding vowel distributions. This meticulous representation is instrumental in capturing the rich phonological diversity that characterizes Swahili, which is crucial for achieving precise language processing.

The carefully curated annotations within the corpus render it an indispensable asset for training Natural Language Processing (NLP) models. By leveraging this rich dataset, developers can significantly improve the models' accuracy in transcription, translation, and voice recognition tasks tailored to the intricacies of the Swahili language.

Moreover, the corpus offers a balanced and representative sampling of various linguistic contexts, which is vital for fostering the generalization capabilities of NLP models. This balance ensures that the models can reliably perform in a wide array of Swahili-speaking settings, from urban dialects to rural variations.

4.4 Model Performance

The CNN-based digraph extraction model exhibited remarkable performance when evaluated against several critical metrics [16] [17]. With an impressive accuracy rate of 95.4%, the model successfully identified a significant portion of the digraphs present within the corpus, highlighting its effectiveness in this task. The precision of the model was recorded at 92.7%, which signifies a high level of accuracy in the digraph predictions made from the instances it retrieved, ensuring that the majority of identified digraphs were correct. In terms of recall, the model achieved a rate of 94.1%, demonstrating its competency in effectively locating relevant digraphs amidst the larger body of text. This indicates that the model was not only accurate but also adept at recognizing and extracting the relevant digraphs vital for analysis. Furthermore, the F1 score, calculated at 93.4%, reflects a solid equilibrium between precision and recall, underscoring the robustness of the model's overall performance.

A thorough examination of the confusion matrix revealed high true positive rates alongside low false positive rates across most of the digraph categories, which further confirms the model's reliability and consistency in its predictions. These impressive results highlight the model's outstanding capabilities in Swahili digraph extraction, emphasizing its practical applicability in various natural language processing (NLP) tasks and applications.

5. Discussion

The Annotated Swahili Digraph Corpus developed in this research represents a crucial advancement in the domain of Natural Language Processing (NLP) for under-resourced languages. This specialized corpus has been meticulously designed to encapsulate the unique phonetic and linguistic characteristics integral to Swahili, which stands as one of the most widely spoken languages across East Africa, with millions of speakers.

The creation of this corpus involved a thorough and systematic annotation process, ensuring that it reflects a broad spectrum of digraphs — pairs of letters that generate a distinct sound when combined. The digraphs included in the corpus comprise "ch," "dh," "gh," "kh," "ng'," "ny," "sh," "th," and "ng," each of which plays a vital role in the pronunciation and meaning of words within the Swahili language. Additionally, the corpus thoroughly documents the five foundational Swahili vowels: a, e, i, o, and u, which are essential for understanding the phonetic structure of the language.

By capturing these phonetic nuances and providing a well-annotated resource, the Annotated Swahili Digraph Corpus enhances the availability of linguistic data for researchers and developers working in NLP, ultimately promoting more accurate and effective language processing tools for the Swahili language. This initiative not only bolsters the technological infrastructure for under-resourced languages but also contributes to the preservation and promotion of linguistic diversity in the digital age.

One of the standout features of this linguistic corpus is its carefully balanced representation of both frequently occurring and rare digraphs across an extensive range of vowel contexts. This thoughtful balance is of paramount importance for training Natural Language Processing (NLP) models, as it enables these models to generalize effectively across varied linguistic settings. Such generalization is essential for enhancing the accuracy and reliability of tasks like transcription, speech recognition, and other critical language processing applications.

Moreover, the corpus includes meticulously detailed annotations for each word, which significantly assist models in understanding the complex relationships between digraphs and the distributions of different vowels. These relationships are vital for achieving precise phonetic representation and robust language modelling. By studying these nuances, NLP models can improve their ability to recognize and produce speech that closely mirrors natural language use, enhancing their overall performance in real-world applications where linguistic diversity and variability are prevalent.

Furthermore, the integration of the CNN-based digraph extraction model into the corpus development process has significantly enhanced the efficiency and accuracy of the annotations. This model's ability to automatically identify and classify digraphs with high precision underscores its potential for broader applications in multilingual NLP systems. By leveraging the annotated features of the corpus, the CNN-based model can effectively learn complex phonetic patterns, thereby improving the overall performance of Swahili language processing tools.

The corpus not only serves as a robust training resource for developing NLP models but also provides a foundation for future research in Swahili linguistics and language technology. Its extensive coverage of Swahili's phonetic diversity ensures that NLP applications can be developed to support various aspects of language learning, communication, and technology integration. This comprehensive dataset is pivotal for advancing the state of Swahili NLP and fostering the development of more effective, context-sensitive language processing technologies.

The Annotated Swahili Digraph Corpus, combined with the CNN-based digraph extraction model, addresses key linguistic challenges in Swahili NLP. It provides a structured and balanced dataset that supports accurate and efficient language processing, paving the way for significant advancements in the field. The successful development of this corpus highlights the importance of creating specialized resources tailored to the unique linguistic features of under-resourced languages, ultimately contributing to the broader goals of inclusivity and technological advancement in NLP.

6. Conclusion and Future scope

This research makes a significant contribution to the field of Swahili NLP by introducing the **Annotated Swahili Digraph Corpus** and developing a CNN-based model specifically for digraph extraction. The creation of the Annotated Swahili Digraph Corpus marks a pioneering effort to compile a rich, extensively annotated dataset that captures the complex phonetic structures characteristic of the Swahili language. This corpus addresses a critical gap in existing NLP resources, which frequently overlook the unique linguistic features of under-resourced languages such as Swahili.

The CNN-based approach employed in this study enhances the capacity to accurately identify and process the intricate relationships between Swahili phonemes, particularly digraphs. The resulting model demonstrates substantial improvements in both the accuracy and efficiency of various language processing tasks, including speech recognition and phonetic transcription. This advancement is particularly noteworthy for low-resource languages, as it optimizes computational resources while maintaining linguistic precision.

In addition to advancing Swahili language processing, the methodologies developed in this study offer a foundation for extending these techniques to other low-resource languages with similar linguistic challenges. Future research can expand upon this work by:

1. Extending the corpus to incorporate diverse Swahili dialects to improve the generalizability of the model.
2. Exploring deep learning techniques, such as transformer-based models, to enhance digraph recognition accuracy further.
3. Applying the developed methodologies to other African languages, contributing to a broader effort in improving language technology for underrepresented linguistic groups.
4. Investigating real-time applications, such as Swahili speech-to-text systems and voice assistants, to validate the practical utility of the proposed model.

The outcomes of this research not only contribute to the academic field of computational linguistics but also pave the

way for practical advancements in speech recognition technology, fostering greater accessibility and usability for Swahili speakers and other low-resource language communities.

Ethical Considerations

This study did not involve human subjects. All data used in the research were derived from publicly available datasets, including those from Harvard Dataverse, Mendeley Data, Zenodo, and Kaggle. As such, no informed consent was required. The study adhered to relevant ethical guidelines and institutional policies for research involving non-human data.

Data Availability

The data underpinning the conclusions of this study, including the annotated Swahili digraph corpus, can be obtained from the corresponding author upon reasonable request.

Conflict of Interest

Authors declare that they do not have any conflict of interest with anyone for publication of this work.

Funding Source

None

Authors' Contributions

All authors contributed equally to this study. Tirus Muya Maina, Aaron Mogeni Oirere, and Stephen Kahara collaboratively developed the study's conceptual framework, designed the methodology, and carried out data collection and analysis. Each author played a key role in model development, interpretation of results, and manuscript preparation. All authors reviewed and approved the final manuscript for submission**.**

Acknowledgements

The authors would like to note that no external contributions were made beyond those of the listed authors.

References

[1] I. A. Okafor, "Distinctive Features: A Linguistic Analysis of Consonant Sounds in English Language," *Ansu Journal of Language and Literary Studies*, vol. **2**, Issue.2, **2022.**

[2] M. Sipser, *Introduction to the Theory of Computation*, PWS Publishing Company, **1996.**

[3] S. S. Rao, *Engineering Optimization: Theory and Practice*, John Wiley & Sons, **2019.**

[4] J. Hopcroft, R. Motwani, and J. D. Ullman, "Introduction to automata theory, languages, and computation," *ACM Sigact News*, vol. 32, Issue. 1, pp. **60–65, 2001.**

[5] J. H. Hansen and G. Liu, "Unsupervised accent classification for deep data fusion of accent and language information," *Speech Communication*, vol. 78, pp. **19–33, 2016**

[6] A. F. Atanda, "Multinomial Logistic Regression Probability Ratio-Based Feature Vectors for Malay Vowel Recognition," Universiti Utara Malaysia, 2021.

[7] W. H. Finch, J. E. Bolin, and K. Kelley, *Multilevel Modelling Using R*, United Kingdom: CRC Press/Taylor & Francis Group, 2019.

[8] M. S. Azmi, "Development of Malay Word Pronunciation Application using Vowel Recognition," *Malay*, vol. 9, Issue.1**, 2016.**

[9] M. S. Azmi, "Malay Word Pronunciation Test Application for Pre-School Children," *International Journal of Interactive Digital Media*, vol. 4, Issue. 2, pp. **2289–4098, 2016.**

[10] K. Y. Chan and M. D. Hall, "The importance of vowel formant frequencies and proximity in vowel space to the perception of foreign accent," *Journal of Phonetics*, vol. **77, 2019.**

[11] M. Mehraj, A. Goel, M. A. Butt, and M. Zaman, "Automatic Speech Recognition Approach for Diverse Voice Commands," *International Journal of Advanced Research in Computer Science*, vol. 8, Issue.9, 2017.

[12] J. O. De Sordi, *Design Science Research Methodology*, Springer International Publishing, 2021.

[13] A. R. Kivaisi, Q. Zhao, and J. T. Mbelwa, "Swahili Speech Dataset Development and Improved Pre-Training Method for Spoken Digit Recognition," *ACM Transactions on Asian and Low-Resource Language Information Processing*, **2023.**

[14] T. M. Maina, A.M. Oirere, and S.Kahara "A CNN-Based Digraph Extraction Model for Enhanced Swahili Natural Language Processing," *International Journal of Scientific Research in Computer Science and Engineering*, Vol.12, Issue.6, pp**.43-55, 2024.**

[15] T. M. Maina, "The Swahili Digraph Corpus," *Mendeley Data*, 2024.

[16] R. Yacouby and D. Axman, "Probabilistic Extension of Precision, Recall, and F1 Score for More Thorough Evaluation of Classification Models," in *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, **2020.**

[17] H. Dalianis, "Evaluation Metrics and Evaluation," in *Clinical Text Mining*, Springer, Cham, pp.**45-53, 2018**

AUTHORS PROFILE

Tirus Muya Maina is a highly experienced ICT professional, specializing in Information and Communication Technology (ICT) and Computer Science. He holds a Master's degree in Information Systems and is currently pursuing a Ph.D. in Computer Science. With over ten years of experience, he has held roles such as Senior ICT Technologist II at Murang'a University of Technology. His expertise spans network infrastructure, software development, cybersecurity, ICT policy formulation, data management, and ICT strategy. Tirus has published research in reputable journals and is an active member of professional bodies, having received advanced training in cybersecurity and higher education. His research interests include Artificial Intelligence, Natural Language Processing, ICT Integration in Education, Cybersecurity, TVET, ICT Policy and Governance, and Curriculum Development.

Aaron Mogeni Oirere, received his B.Sc. degree in Computer Science from Periyar University, Salem, Tamilnadu, India in 2007, the M.Sc. degree in Computer Science from Bharathiar University, Coimbatore, Tamilnadu, India in 2010, and the Ph.D. degree in Computer Science from Dr. Babasaheb Ambedkar Marathwada University, Maharashtra, India in 2016. He currently works at the Department of Computer Science, school of computing and Information Technology, Murang'a University of Technology. His research interest includes Automatic Speech Recognition, Human-computer Interaction, Information Retrieval, Database Management Systems (DBMS), Data Analytics and Hardware & Networking.

Stephen Kahara is a Lecturer of Computer Science at Murang'a University of Technology and the Director of Performance Contracting and ISO. He holds a Ph.D. in Computer Science from Murang'a University of Technology, an M.Sc. in Computer Systems from JKUAT, an M.Sc. in Organizational Development from USIU-Africa, and a B.Sc. in Information Sciences from Moi University. Dr. Kahara has 16 years of experience in the ICT industry. He is a certified QMS and ISMS auditor and a DAAD-UNILEAD alumnus. His research interests include machine learning, network security, distributed systems, and computational biology. He has published extensively in academic journals and conferences. Dr. Kahara is a member of the Association of Computing Practitioners - Kenya.

