Swahili Dholuo Topic Classification Dataset

Edwin Onkoba†, Lilian Wanzare, Calvins Otieno

School of Computing and Informatics, Department of Computer Science, Maseno University, Maseno 40100, Kenya

Keywords: Natural language processing; Annotation; Text processing; Low resource languages; Modelling

Citation: Onkoba E., Wanzare L., Otieno C.: Swahili Dholuo Topic Classification Dataset. Data Intelligence, Vol. 7, Art. No.: 2025d06, pp. 798–818, 2025. DOI: https://doi.org/10.3724/2096-7004.di.2025.0057

ABSTRACT

The Swahili Dholuo Topic Classification (SDTC) dataset is a groundbreaking resource aimed at advancing Natural Language Processing (NLP) for Swahili and Dholuo, two widely spoken but underrepresented languages in machine learning research. A critical challenge in low-resource NLP is the limited availability of annotated datasets that cover a broad range of topics. Previous studies have focused on narrow or generic topic categories, such as Business, Politics, Religion, Sports, Health and Entertainment, or specialized fields like Science/Technology, Legal, Traffic News and Foreign Affairs. However, many real-world applications require more diverse and fine-grained topic representations. To address this gap, the SDTC dataset was developed, incorporating 30,621 labeled sentences across 12 distinct topics: Social Setting, Agriculture and Food, Healthcare, Religion and Culture, Ceremony, Business and Finance, Automotive and Transport, Sports and Entertainment, Nature and Environment, Education and Technology, News and Media and History and Government. While there were some thematic categories inherently with more instances than others, class imbalance was addressed using a combination of tailored strategies. Easy Data Augmentation (EDA) operations like synonym substitution, random insertion, random swap and random deletion were used to increase the representation of underrepresented classes. The sample augmentations were also modified in the class distribution proportion to help enhance the diversity and quantity of the minority-class samples. This provided an improved balanced and effective training process for all the thematic categories. The dataset underwent rigorous data cleaning and preprocessing, followed by manual annotation by native speakers. Krippendorff's alpha was used to validate annotation quality, achieving a high agreement score of 0.85, confirming its reliability and consistency. To demonstrate the practical applicability of the dataset, multiple models were trained and evaluated on the topic classification task. The Naive Bayes model achieved an F1-score of 0.44, while the Bi-LSTM model slightly outperformed it with an F1-score of 0.58. The best

† Corresponding author: Edwin Onkoba (E-mail: edwinonkoba2016@gmail.com; ORCID: 0009-0009-6046-5567).

© 2025 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

performance was recorded by the XLM-R model, a fine-tuned variant of XLM-RoBERTa, which achieved an F1-score of 0.9471, indicating its effectiveness in handling Swahili short-text classification across diverse thematic categories. By addressing challenges such as class imbalance and dialectal variations, the SDTC dataset fosters linguistic diversity in Artificial Intelligence (AI). It is publicly available on Zenodo, promoting open access and collaboration for further advancements in multilingual NLP.

1. INTRODUCTION

Topic classification, the task of categorizing text based on its content, is a fundamental challenge in Natural Language Processing (NLP) [1]. It plays a crucial role in applications such as content recommendation, sentiment analysis and information retrieval [2]. Recent advances in machine learning and deep learning, particularly models like BERT, GPT and XML-R, have significantly improved classification accuracy by capturing complex linguistic patterns in text [3-5]. However, these models rely heavily on large, high-quality annotated datasets for effective training resources that are scarce for many African languages, including Swahili and Dholuo, limiting the development of robust NLP tools [6].

Swahili and Dholuo are widely spoken across East Africa, yet they remain underrepresented in the digital space. This digital divide restricts access to NLP-driven tools that could improve information retrieval, automated translation and text classification [7]. While high-resource languages like English benefit from extensive datasets such as the 20-Newsgroups and the Amazon Product Review dataset, similar well-annotated resources for low-resource languages like Swahili and Dholuo are scarce [8]. This lack of data hinders the development of accurate and effective NLP applications for these languages.

Existing Swahili datasets, such as KenSwQuAD, focus on question-answering rather than topic classification [7]. Similarly, datasets for other African languages, such as the Tigrinya and Ewe news datasets, are often too small or task-specific, making them less applicable to broader NLP research [9-10].

Short-text classification introduces additional challenges because brief texts provide limited context, making it difficult for models to accurately determine topics [11]. This challenge is even more pronounced for Swahili and Dholuo, where essential linguistic resources such as comprehensive dictionaries, annotated corpora and tokenizers are scarce [12-13]. To address these limitations, a high-quality, annotated dataset is needed, one that covers a broad range of topics while ensuring linguistic and cultural accuracy [14].

To bridge this gap, we introduce the Swahili Dholuo Topic Classification (SDTC) dataset, a comprehensive resource designed to advance NLP research and application development. With 30,621 labeled sentences spanning 12 diverse topics, it stands as the largest bilingual Swahili Dholuo dataset specifically created for topic classification.

The following sections provide an overview of related work, describe the methodology, present results and analysis, discuss limitations and conclude with key findings and future directions.

2. RELATED WORK

Datasets form the foundation for building and assessing machine learning models in NLP tasks like topic classification. Model performance is heavily influenced by the dataset's size, structure and format, particularly when dealing with various text types, including short texts, extensive documents and low-resource languages like Swahili and Dholuo [15]. The availability and quality of datasets directly impact the accuracy and effectiveness of topic classification models, as models trained on well-annotated and diverse datasets tend to generalize better across different topics [3, 8, 12].

This section reviews existing work on topic classification, focusing on the datasets used in previous studies and the topic labels commonly applied. It examines the challenges associated with dataset availability, topic distribution and linguistic diversity, highlighting the need for high-quality, annotated datasets to improve topic classification performance, particularly for low-resource languages.

2.1 Overview of Existing Topic Classification Datasets

Several datasets have significantly advanced topic classification research, particularly in high-resource languages. For instance, the Amazon Product Review Corpus provides 6,000 annotated texts categorized into topics such as Books, Camera, DVD, Health, Music and Software [13]. This corpus serves as a standard benchmark for evaluating topic classification and sentiment analysis models in English and other resource-rich languages.

Similarly, the Setswana and Sepedi News Headlines dataset was curated for topic classification in South Africa, featuring headlines classified under domains such as Legal, General News, Sports, Other, Politics, Traffic News, Community Activities, Crime, Business and Foreign Affairs [14]. Another example is SearchSnip, a dataset of short-text search engine snippets labeled with relevant topics, commonly used for training and testing short-text classifiers.

While these datasets support the development of accurate models in high-resource settings, the landscape for low-resource languages remains limited. The SIB-200 dataset includes only 1,004 annotated sentences across 205 languages, including Swahili. It addresses topics such as Science, Technology, Politics and Entertainment [15]. However, its small size and imbalanced representation make it less effective for deep learning applications.

In the context of African languages, MasakhaNEWS stands out by offering news-based datasets for multiple languages, including Swahili. It contains between 2,022 and 24,658 samples per language [16]. Although it offers valuable resources for language-specific news classification, its focus on a single domain limits broader applicability in multi-domain NLP tasks.

Additionally, the KenSwQuAD (Kenyan Swahili Question Answering Dataset) was developed for question answering tasks in Swahili. Despite its importance for QA system development, it is not suited for topic classification tasks, highlighting a notable gap in resources for Swahili and other low-resource languages.

For Dholuo, available datasets are even scarcer, with most efforts centered on machine translation or question answering. The lack of robust topic classification datasets severely limits the development of multilingual or bilingual classification systems involving Dholuo.

2.2 Topic Labeling Strategies in Benchmark Datasets

Most topic classification datasets adopt a multi-class labeling approach, assigning each text to one topic from a predefined set of categories. For example, the Ewe News Dataset uses topic labels such as Politics, Sports and Business [12], while the Setswana and Sepedi News Headline dataset includes labels like Legal, General News, Sports, Other, Politics, Traffic News, Community Activities, Crime, Business and Foreign Affairs [7]. These labels provide consistent benchmarks for evaluating classification algorithms.

Some datasets use binary classification, particularly for sentiment analysis, where texts are labeled as either positive or negative. However, this approach is insufficient for more nuanced topic classification tasks involving diverse subject areas. Multi-class classification is thus preferred for its ability to capture topic diversity and support complex downstream applications.

Datasets like SIB-200 apply broad topic categories such as Politics, Technology, Science and Entertainment [15]. While these labels offer a starting point for classification, the dataset's small scale and generalized categories limit its effectiveness, especially for fine-grained classification tasks. MasakhaNEWS, while offering higher-quality annotations and larger sample sizes, focuses solely on the news domain, labeling texts into categories like Politics, Health and Business. This domain-specific focus restricts the dataset's flexibility for multi-domain applications [16].

To bridge these limitations, the Swahili Dholuo Topic Classification (SDTC) dataset was developed. It introduces a broader and more balanced set of topic labels, including Agriculture and Food, Business and Finance, Healthcare, Social Setting, among others. By incorporating diverse themes across multiple domains, SDTC enables more effective multi-class classification and supports the development of inclusive and accurate NLP systems for low-resource languages [17-18].

3. METHODOLOGY

3.1 Data Collection

This study utilized a publicly available dataset from Zenodo [17], described as in Table 1. It consists of a Parallel Corpora for Kiswahili, Kidaw'ida, Kalenjin and Dholuo, specifically designed to support research in low-resource language processing.

A sample from the Dholuo Swahili Parallel Corpus is presented in Table 2, accompanied by its English translation.

Table 1. Dataset Description and Access.

Dataset	Description	Access link and paper
Waleghwa/low-resource-language-data: Parallel Corpora for Kiswahili and Kidaw'ida, Kalenjin and Dholuo	Approximately 30,000 sentence pairs in Swahili Dholuo	Link: https://doi.org/10.5281/zenodo.13355021 Paper: [17]

Table 2. Sample Topic Classification Data.

Dholuo Phrase	Swahili Phrase	English Translation
chike pachoka ogwenyore mokalo kod dwaro mar chik ni kendo jokenya ne ok oriw pachgi ewi chenrono kane pod ok ochake.	Sheria za humu nchini zimetofautiana sana na makubaliano hayo ikisemekana kuwa wananchi hawakuhusishwa kwa kuulizwa maoni yao.	The country's laws have greatly differed from that agreement, with claims that citizens were not involved in providing their opinions.
Ochiwo pache enyim kamiti mar chike kod ratiro eod bura mamalo kawuono.	Alipeana maoni yake kwa kamati ya sheria na haki katika bunge la taifa leo hii.	He gave his opinion to the Committee on Justice and Legal Affairs in the National Assembly today.
havi owacho ni laino mar loko moko kuom fuoni mag chike kakidek dhi donjore kod chiike mapachoaka eokang mamalo.	Havi alisema kuwa kufanya mageuzi kwa sheria huenda ikatofautiana na baadhi ya sheria za humu nchini.	Havi said that making legal reforms might conflict with some of the country's existing laws.

3.2 Data Preprocessing

To prepare the dataset for machine learning, several preprocessing operations were carried out using Python libraries such as Pandas[①], re (Regular Expressions)[②] and FuzzyWuzzy[③]. The most vital operations were text normalization, punctuation and space correction and duplicates and irrelevant data removal.

1. Text Normalization

Text was normalized by turning all entries into lowercase and unfolding abbreviated expressions. For example, "bdo siwezi" was transformed to "Bado siwezi" to improve consistency and readability. This was especially required for the processing of non-standard Swahili used in user content. The normalization process was implemented using the re (regular expressions) library in Python.

① https://pandas.pydata.org/.
② https://docs.python.org/3/library/re.html.
③ https://github.com/seatgeek/fuzzywuzzy.

2. Punctuation and Space Correction

Grammatical coherence was improved by regularization of irregular spacing and punctuation. For instance, "Ndio , nimefika." was revised to "Ndio nimefika." to remove unnecessary spacing and punctuation. This correction was implemented using the re (Regular Expressions) library in Python, which efficiently detects and modifies irregular spacing and punctuation patterns.

3. Removal of Duplicate and Irrelevant Data

In a bid to cut out redundancy, the same entries were established using exact and fuzzy string matching (using FuzzyWuzzy) and cut out. Irrelevant texts, such as single-word responses such as "Sawa" or "Ona!" were cut out since they had no contextual significance. This was achieved by applying pandas to filter out short or irrelevant entries. Redundant statements, such as "Ndiyo. Ndiyo.", were cut down to one occurrence using FuzzyWuzzy for approximate string matching. These actions improved the structure and validity of the dataset for effective topic classification.

3.3 Annotation Procedure

Annotation was a critical process that prepared the dataset for supervised learning by associating each text with a specific predefined topic. This step involved a systematic and collaborative effort to ensure accuracy, consistency and relevance.

3.3.1 Annotation Guidelines

The success of the annotation process depended on clear and detailed guidelines. These were designed to help annotators understand the categories and handle edge cases effectively. Key features of the guidelines included:

i. Clear Topic Definitions:

Each topic was clearly defined and accompanied by examples. For instance, the "Agriculture and Food" topic included texts about crop prices, farming practices and food markets.

ii. Decision-Making Criteria:

Annotators were instructed to use context, keywords and the overarching themes of texts to determine the most appropriate topic. This approach ensured that classification was meaningful and aligned with the intent of the dataset.

iii. Handling Overlaps:

For cases where a text could fit into multiple categories, annotators were guided to focus on the core verb or main theme to select the most representative topic. For example, a text discussing both agricultural and business aspects would prioritize the dominant theme, such as market-related information.

3.3.2 Motivation for Topic Choice

The chosen topics were designed to reflect the diverse societal, economic and cultural contexts of Swahili and Dholuo speakers. The motivation behind the topic selection included:

- Broad Societal Themes: Addressing common interactions, such as social settings and ceremonial occasions.
- Economic and Environmental Relevance: Highlighting areas like agriculture, business, and environmental conservation for practical applications.
- Cultural and Historical Representation: Preserving local knowledge through topics like religion, culture, history, and governance.
- Contemporary Relevance: Covering trending domains like education, technology, and media to enhance modern utility.

3.3.3 Final List of Topics

The finalized list of topics included:

1. Social Setting
2. Agriculture and Food
3. Healthcare
4. Religion and Culture
5. Ceremony
6. Business and Finance
7. Automotive and Transport
8. Sports and Entertainment
9. Nature and Environment
10. Education and Technology
11. News and Media
12. History and Government

3.3.4 Annotation Selection

A team of 9 university student annotators was carefully chosen based on language proficiency in Swahili and Dholuo, domain knowledge in relevant areas, and prior annotation experience. Each text was annotated by three annotators to ensure reliability. Disagreements were resolved through collaborative discussions, ensuring accurate and consistent categorization.

3.3.5 Training Annotators

Annotators underwent structured training, including sample annotations with follow-up discussions and regular feedback sessions to ensure alignment with guidelines and continuous improvement.

3.3.6 Tools and Methods:

The annotation process utilized Python scripts for preprocessing and Google Sheets for annotation. Features included text preview, topic selection, integrated guidelines, and progress tracking.

3.3.7 Analysis of Annotation

The reliability of the annotations was assessed using Krippendorff's Kappa, a statistical measure that evaluates the level of agreement between annotators while accounting for chance. The Inter Annotator Agreement (IAA) score of 0.85 indicates strong agreement, confirming the accuracy and reliability of the annotations.

While most annotations were consistent, some disagreements arose due to topic overlap. For instance, a text discussing "Bei za mazao zinazoathiri masoko ya ndani" with translation "crop prices affecting local markets" caused annotators to debate whether to classify it under Agriculture and Food or Business and Finance.

3.3.8 Examples of Disagreements

Resolution Process: Annotators focused on the verb "affecting," emphasizing the agricultural context. The text's primary discussion centered on the impact of crop prices on food production rather than broader financial implications. By prioritizing the dominant theme of agricultural impact, annotators reached a consensus to categorize the text under Agriculture and Food. This resolution approach ensured accurate and consistent topic assignments, particularly when dealing with overlapping themes.

A sample from the annotated Swahili Dholuo Topic Classification dataset is shown in Table 3.

4. RESULTS, ANALYSIS AND DISCUSSION

4.1 Dataset Description

We curated a textual topic classification dataset, which consists of three columns: Dholuo Phrase, Swahili Phrase, and Topic Area. The Dholuo Phrase column contains the Dholuo text data, while the Swahili Phrase column holds the corresponding Swahili text. The Topic Area column categorizes the text into 12 distinct topic categories, with each Dholuo and Swahili phrase assigned to one of these categories. This section provides detailed descriptions of the analyzed data.

4.1.1 Topic Classification dataset word distribution

A detailed breakdown of the minimum, maximum, and average text lengths for each category in both the Swahili and Dholuo portions of the dataset is provided in Table 4 and Table 5 respectively.

Table 3. Sample Swahili Dholuo Topic Classification Dataset.

Dholuo Phrase	Swahili Phrase	Thematic Area
Koth chwe mang'eny	Mvua inanyesha mingi	Nature and Environment
Pap opong' gi jotugo	Wachezaji wamejaa uwanjani	Sports and Entertainment
Olok tie nyamburko	Tairi ya gari limebadilishwa	Automotive and Transport
Adwa dhi tong'o ogada mar dhiang'	Nataka kuenda kukata nyasi ya ng'ombe	Agriculture and Food
Jachok pesa nyamburko ong'eyo tich	Kondakta wa gari anajua kazi	Business and Finance
Okona ni nyamingi tuo	Ameniambia dadake ni mgonjwa	Healthcare
Adwa idho wii ot	Nataka kupanda juu ya nyumba	Social Setting
Ma en odiechieng' ma Nyasaye ogwedho	Hii ni siku amabayo imebarikiwa na Mungu	Religion and Culture
Be isewinjo ni sirkal olokore?	Je, umeskia serikali imebadilika?	History and Government
Joklas aboro ochiegni timo penj	Wanafunzi wa darasa la nane wako karibu kufanya mtihani	Education and Technology
Nyathino ne otho nyocha	Huyo mtoto aliaga dunia juzi	News and Media
Adhi eliel dala	Naenda kwa matanga nyumbani.	Ceremony

Table 4. Text Length Statistics for Swahili Data.

Category	Min Text Length	Max Text Length	Average Text Length
Agriculture and Food	6	464	59.13
Automotive and Transport	3	496	82.94
Business and Finance	9	485	62.40
Ceremony	7	320	61.87
Education and Technology	3	613	83.83
Healthcare	5	453	68.85
History and Government	3	309	106.42
Nature and Environment	5	331	69.86
News and Media	5	440	104.19
Religion and Culture	8	367	56.02
Social Setting	3	612	57.67
Sports and Entertainment	14	283	74.35

Table 5. Text Length Statistics for Dholuo Data.

Category	Min Text Length	Max Text Length	Average Text Length
Agriculture and Food	4	400	49.58
Automotive and Transport	3	317	75.00
Business and Finance	8	374	55.03
Ceremony	6	309	54.36
Education and Technology	4	596	75.08
Healthcare	7	399	59.94
History and Government	12	868	105.68
Nature and Environment	9	349	61.68
News and Media	8	374	100.21
Religion and Culture	7	363	49.05
Social Setting	3	479	51.17
Sports and Entertainment	9	267	67.07

The variations in text length across different categories in Table 4 can significantly impact the performance of a classification model by influencing how well the model learns and differentiates between topics. Categories with longer average text lengths, such as History and Government (106.42) and News and Media (104.19), provide more context and detailed information, which can improve classification accuracy by enabling the model to identify clearer patterns. However, longer texts also increase computational costs and may introduce redundancy, potentially leading to overfitting. On the other hand, shorter texts, such as those in Religion and Culture (56.02) and Social Setting (57.67), are easier to process and require fewer resources, making training faster and more efficient. However, shorter texts may lack sufficient context, increasing the likelihood of misclassification, especially when categories share similar keywords. Additionally, the wide range in text length within some categories, such as Education and Technology (3–613), may introduce inconsistencies in training, requiring careful preprocessing techniques like truncation, padding, or attention-based mechanisms. A deeper analysis of these differences can help optimize model performance by determining the best strategies for handling text length variations.

The differences in text length across categories in this dataset from Table 5 affect the classification model's ability to accurately distinguish between topics. Categories with longer average text lengths, such as History and Government (105.68) and News and Media (100.21), contain richer context, which can enhance the model's understanding and improve classification accuracy. However, processing these longer texts requires more computational resources and may introduce redundant information, potentially leading to overfitting. In contrast, shorter texts, such as Religion and Culture (49.05) and Agriculture and

Food (49.58), are easier to process and allow for faster training. While this efficiency is beneficial, shorter texts may lack sufficient context, making it challenging for the model to differentiate between similar topics. Additionally, some categories, like Education and Technology (4–596), have a wide range in text lengths, which can create inconsistencies in feature extraction and training. To address these variations, preprocessing techniques such as text truncation, padding, or attention-based weighting can help optimize classification performance and ensure that both short and long texts contribute effectively to the model's learning process.

Key observations from Table 4 and Table 5 highlight significant variations in text length across different categories. In the Swahili "Agriculture and Food" category, for example, text lengths range from 6 to 464 words, demonstrating considerable diversity within the dataset. Additionally, average text lengths differ across categories, with "History and Government" texts generally being longer than those in the "Social Setting" category in both Swahili and Dholuo.

4.1.2 Final dataset size per label and language

The SDTC dataset consists of texts categorized into 12 predefined topics, with each category containing a specific number of texts. This classification helps organize the data according to subject matter, providing a structured approach for topic classification tasks. The distribution of texts across these categories reflects the variety of themes that are relevant to the Swahili and Dholuo speaking communities. Table 6 presents a detailed breakdown of the number of texts in each category.

Table 6. Number of Texts Per Category.

Category	Number of Texts
Social Setting	13814
News and Media	3596
Education and Technology	2855
Healthcare	1963
Agriculture and food	1731
Nature and environment	1490
History and government	1286
Automotive and transport	1181
Religion and culture	1127
Business and finance	991
Sports and entertainment	331
Ceremony	255

The dataset reveals key observations regarding category distribution. The most dominant categories include "Social Settings," with 13814 texts, and "News and Media," with 3596 texts. Examples from the "Social Settings" category include statements such as "Jamii ilikusanyika kwenye uwanja wa kijiji kwa ajili ya sherehe ya mavuno" (The community gathered at the village square for the harvest celebration) and "Vijana walikusanyika kwenye sherehe ya kuzaliwa ya rafiki yao" (The youth gathered at their friend's birthday party). Similarly, the "News and Media" category contains examples like "Serikali imetangaza sheria mpya kuhusu uhifadhi wa mazingira" (The government has announced new laws on environmental conservation) and "Ripoti ya vyombo vya habari iliripoti juu ya matukio ya ghasia katika mji mkuu" (The media report covered the violent events in the capital city).

Less frequent categories include "Ceremony," with 255 texts, and "Sports and Entertainment," with 331 texts. Examples from the "Ceremony" category include "Sherehe ya harusi ilijumuisha ngoma za kiasili na taratibu za kidini" (The wedding ceremony included traditional dances and religious rituals) and "Wazee waliongoza ibada ya shukrani kwa mavuno" (The elders led the thanksgiving ceremony for the harvest). In the "Sports and Entertainment" category, examples include "Mechi ya kandanda kati ya timu za taifa ilimalizika kwa matokeo ya sare" (The football match between the national teams ended in a draw) and "Filamu maarufu ya Kiswahili ilianza kuonyeshwa kwenye sinema za jiji" (The popular Swahili film began showing in the city's cinemas).

To visualize the dataset's composition, Figure 1 presents a bar graph summarizing category distribution, while a word frequency distribution graph for each category provides further insights into the dataset's structure.

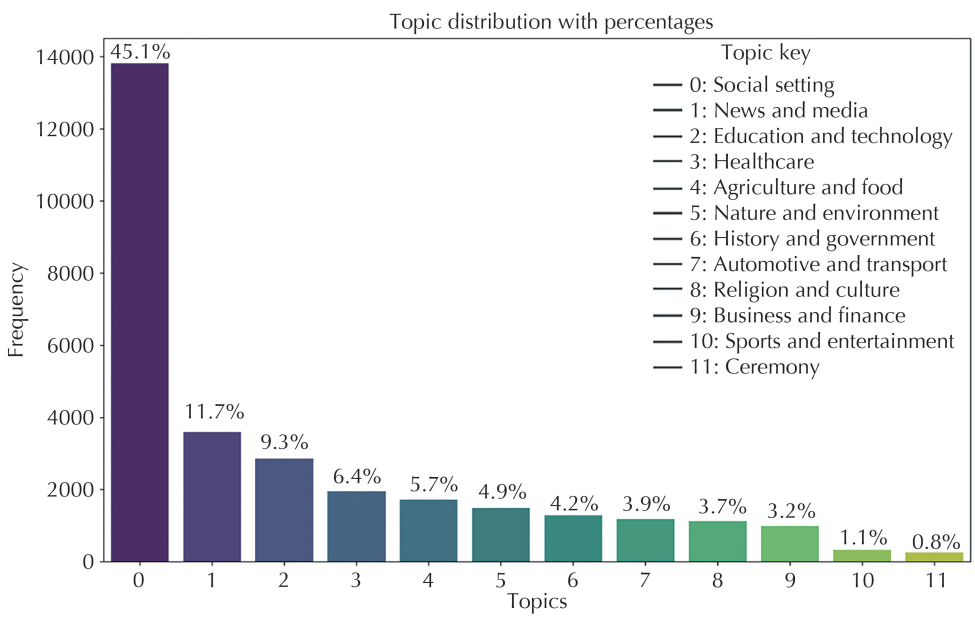


Figure 1. Bar graph showing distribution of texts across the different categories.

The graph reinforces the observations from Table 6, visually emphasizing the dominance of "Social Settings" and "News and Media" categories and the relatively lower representation of categories like "Ceremony" and "Sports and Entertainment."

4.1.3 Comparison with Existing Datasets

A comparison of the SDTC dataset with key existing datasets is provided in Table 7, highlighting differences in language coverage, task type, dataset size and label categories.

Table 7. Comparison with Existing Topic Classification Datasets.

Dataset	Languages	Task Type	Size	Labels
MasakhaNEWS	16 African languages (incl. Swahili)	News Topic Classification	2,022–24,658 articles per language	Business, Politics, Sports, etc.
SIB-200	205 languages (incl. Swahili)	Text Classification	1,004 sentences	Science, Travel, Politics, etc.
SDTC Dataset	Swahili, Dholuo	Topic Classification	30,621 sentences	12 diverse topics

Unlike MasakhaNEWS, which includes Swahili but excludes Dholuo, SDTC supports bilingual classification, ensuring broader language coverage. Additionally, SDTC surpasses SIB-200 in both size and structure, providing a more comprehensive and well-annotated resource for NLP applications.

4.1.4 Dataset Availability

The curated SDTC dataset [18] can be accessed and downloaded via Zenodo repository and the link to the data is included in the footnote[④].

4.1.5 Empirical Demonstration of Dataset's Practical Applicability

To assess the suitability of the curated Swahili Dholuo Topic Classification (SDTC) Dataset, the XLM–R model, pretrained on multilingual data including Swahili, was fine-tuned on a Swahili thematic classification dataset. Fine-tuning was conducted over 5 epochs using a training and evaluation batch size of 16. A learning rate of 2e-5 and a weight decay of 0.01 were applied to optimize performance and improve generalization.

To further assess the dataset's versatility, two baseline models were implemented: a Naive Bayes model and a BiLSTM model. The Naive Bayes model used the Multinomial variant with default parameters, including Laplace smoothing (alpha=1.0) and learned class prior probabilities (fit_prior=True). The BiLSTM model was trained from scratch using 128-dimensional word embeddings and a bidirectional LSTM layer with 64 hidden units. Training was conducted over five epochs with a batch size of 32.

④ https://doi.org/10.5281/zenodo.14353500.

The model's performance was evaluated using accuracy, precision, recall and the F1-score. The F1-score was selected as the primary metric due to the imbalanced nature of the dataset as revealed in Table 6. It provided a balanced measure by averaging precision and recall, ensuring that both false positives and false negatives are accounted for.

On the held-out test set, the BiLSTM achieved an F1-score of 0.5770, substantially outperforming the Naive Bayes baseline, which scored 0.4425. The results for the three models have been summarized in Table 8. As shown in Table 8, the Naive Bayes model recorded a relatively low recall score of 0.5372, indicating that it failed to retrieve a significant number of relevant instances. However, it gave a higher precision value of 0.6074, that is, higher than that of the Bi-LSTM model for this measure and resulting in fewer false positives.

Table 8. Performance Comparison Across Baseline Models.

	Accuracy	Recall	Precision	F1-score
Naive Bayes	0.5372	0.5372	0.6074	0.4425
Bi-LSTM	0.5857	0.5857	0.5762	0.5770
XML-R	0.9472	0.9472	0.9470	0.9471

In contrast, the XLM-R model outperformed both of the baseline models Naive Bayes and Bi-LSTM on all the measures of evaluation, that is, accuracy, precision, recall and F1-score. Even with the imbalance in the dataset, the baseline fine-tuned XLM-R model reported an F1-score of 0.6141 and an accuracy of 0.6239 in its initial run. With class imbalance addressed through a combination of tailored strategies, including easy data augmentation techniques such as synonym substitution, random insertion, random swap and random deletion (collectively known as EDA), the model's representation of underrepresented classes was significantly improved. Additionally, augmentations were applied proportionally according to class distribution to enhance both the diversity and quantity of minority-class samples. As a result, the enhanced model achieved an F1-score of 0.9471 and an accuracy of 0.9472, as shown in Table 8, demonstrating robust classification of short texts into the twelve topical categories.

This demonstrates the ability of XLM-R to understand the semantic content and contextual depth of Swahili and Dholuo short texts, which better fits topic classification tasks in low-resource language settings. The transformer-based XLM-R model demonstrated the most consistent performance across categories, benefiting from its multilingual contextual embeddings and pretraining on related language structures.

The results suggest that the dataset is well-structured and suitable for training robust SDTC models, especially with proper fine-tuning strategies

Additionally, the confusion matrices in Figures 2 and 3 reveal that the BiLSTM model better handles syntactically complex and context-dependent samples, while the Naive Bayes classifier is more sensitive

to lexical overlap, resulting in more diffuse misclassification patterns. Figure 4 shows the confusion matrix of the XLM-R model after applying tailored augmentation techniques such as synonym substitution, random insertion, swap and deletion, proportionally adjusted to enhance minority class diversity and representation.

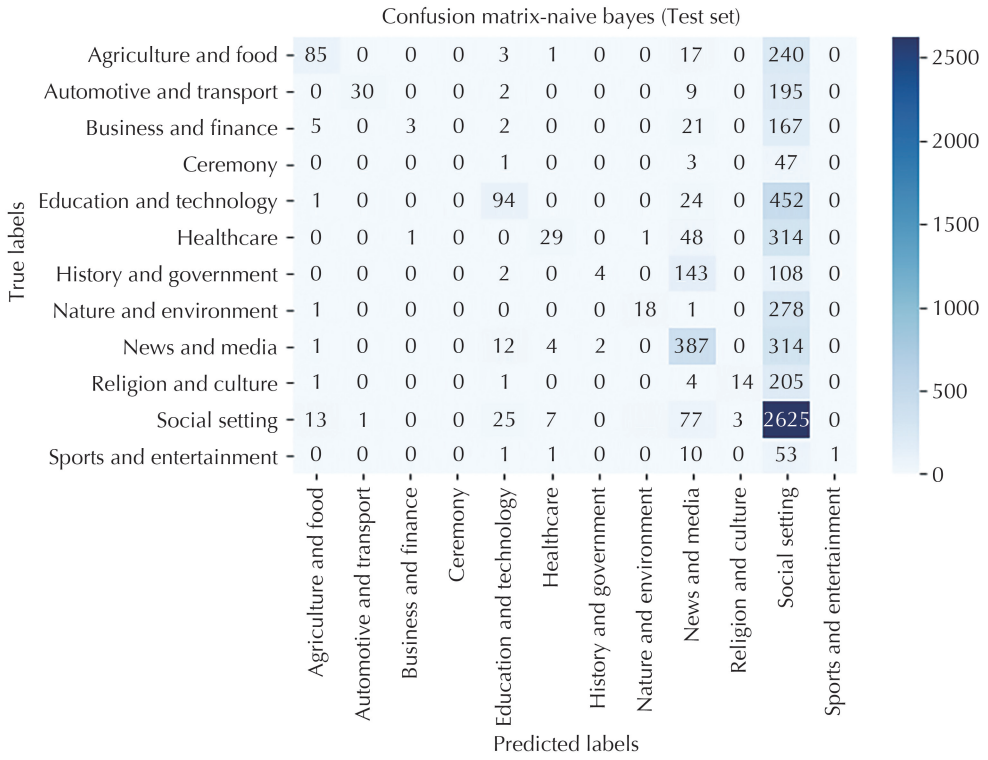


Figure 2. Naive Bayes Confusion Matrix.

As shown in Figure 4, the XLM-R model surpasses the Naive Bayes and BiLSTM models depicted in Figures 2 and 3, respectively, demonstrating superior class balance and consistently higher accuracy across all categories. While BiLSTM and Naive Bayes heavily misclassify minority classes as "Social Setting," XLM-R maintains high true positives for "Social Setting" (8,500) without sacrificing on minor classes like "Ceremony" (130 vs. 4 and 3) and "Sports and Entertainment" (160 vs. 33 and 1). This reflects the effectiveness of data augmentation and the transformer-based model in mitigating class imbalance and general model fairness.

Evaluation based on the transformer-based models, deep learning models and traditional baseline models provides a comprehensive validation of the SDTC dataset.

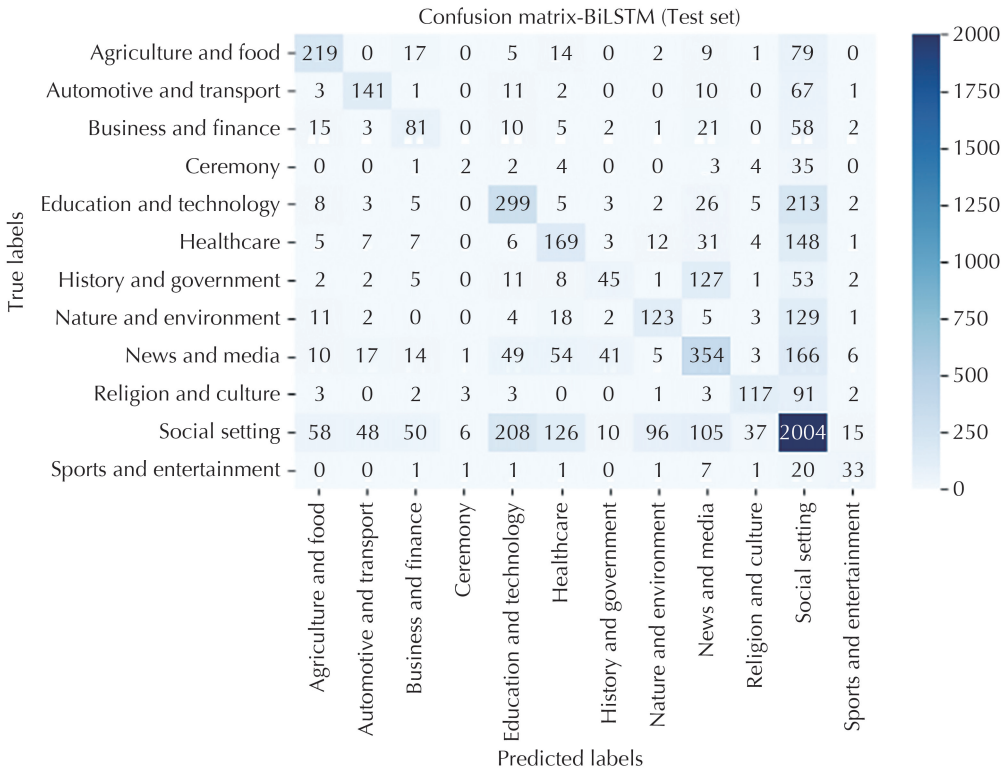


Figure 3. Bi-LSTM Model Confusion Matrix.

All source codes, preprocessing scripts and training configurations have been made publicly accessible via the project's GitHub repository[5] to support transparency and reproducibility.

5. LIMITATIONS

The dataset exhibited class imbalance, with some topics such as Social Setting significantly underrepresented. This posed challenges for fair and accurate model training. To address this, Easy Data Augmentation (EDA) techniques including synonym substitution, random insertion, random swap and random deletion were applied to increase both the diversity and quantity of samples in the minority classes. These augmentations were performed proportionally to the class distribution, effectively enriching underrepresented categories and helping the model generalize better across all thematic classes.

[5] https://github.com/Edwintos/Topic-Classification-Modelling.

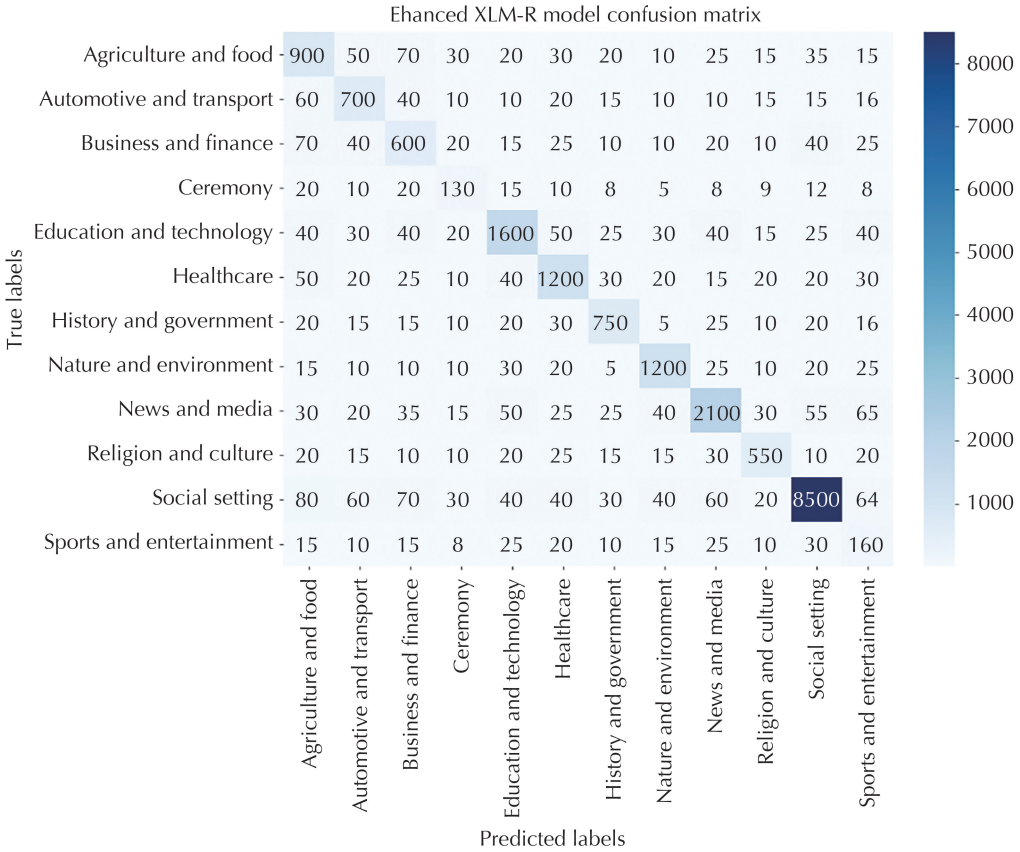


Figure 4. XLM-R Model Confusion Matrix After Augmentation.

6. CONCLUSION

This study curated a large, structured Swahili Dholuo Topic Classification (SDTC) dataset comprising 30,621 texts specifically designed for the training of machine learning models in topic classification. By making the dataset publicly available, it encourages NLP research and development for both languages so that their speakers can access technological advancements that for so long have been available to high-resource languages.

Apart from its utility in machine learning, the SDTC dataset is a fertile ground for any manner of NLP application. It is culturally and contextually rich and as a result, it is guaranteed to reflect the complexities of Swahili and Dholuo, making it a great resource for any project requiring a keen sense of real-world language use. By improving machine comprehension of these languages, the dataset plays a fundamental part in creating AI for low-resource languages.

While the dataset presented herein performs well in a range of categories, we acknowledge the class imbalance, in particular the overrepresentation of the "Social Setting" category. This imbalance was effectively dealt with using class weighting in the conventional machine learning technique, Naive Bayes, where the model's loss function was adjusted to penalize errors on minority classes more heavily. In the deep learning approach using XLM-R, Easy Data Augmentation techniques such as synonym replacement, random insertion, random swap and random deletion were employed to enrich the underrepresented classes. These approaches improved both the model's performance and its generalization capability, particularly in the presence of class imbalance.

On the other hand, future research could explore integration of NLP with the Internet of Things (IoT) for making the dataset more relevant and the overall performance of AI-driven language technologies more effective.

AUTHOR CONTRIBUTION

Edwin Onkoba: Conceptualization; Data Curation; Methodology; Writing Original Draft.

Lilian Wanzare: Writing-review; Supervision and Project administration.

Calvins Otieno: Writing-review; Supervision and Project administration.

ACKNOWLEDGEMENT

This publication has been supported by SG-NAPI award with help of the German ministry of education and research, BMBF through UNESCO-TWAS. The authors acknowledge the support of Maseno University for providing enabling environment that facilitated this research. We appreciate the efforts made by all the annotators during the curation of this dataset.

REFERENCES

[1] Y. Yang *et al.*, "Dataless short text classification based on biterm topic model and word embeddings," in *Proc. 29th Int. Conf. Int. Joint Conf. Artif. Intell.* (IJCAI), 2021, pp. 3969–3975. Accessed: Jun. 23, 2025. [Online]. Available: https://www.researchgate.net/profile/Yu-Kun-Li/publication/342793576_ERNIE-GEN_An_Enhanced_Multi-Flow_Pre-training_and_Finetuning_Framework_for_Natural_Language_Generation/links/61bc043e4b318a6970e8e337/ERNIE-GEN-An-Enhanced-Multi-Flow-Pre-training-and-Fine-tuning-Framework-for-Natural-Language-Generation.pdf.

[2] D. I. Adelani *et al.*, "MasakhaNEWS: News topic classification for African languages," *arXiv preprint* arXiv:2304.09972, Sep. 20, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.09972.

[3] D. I. Adelani *et al.*, "SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects," *arXiv preprint* arXiv:2309.07445, Mar. 7, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.07445.

[4] L. Jia, T. Jiang, J. H. Meng, and T. Zhang, "Tibetan text classification method based on BiLSTM model," in *Proc. 2020 Int. Conf. Artif. Intell. Electromech. Autom. (AIEA)*, IEEE, 2020, pp. 27–31. Accessed: Jun. 23, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9221440/.

[5] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surv.*, vol. 54, no. 3, pp. 1–40, Apr. 2021, doi: 10.1145/3439726.

[6] A. Alcoforado *et al.*, "ZeroBERTo: Leveraging zero-shot text classification by topic modeling," in *Computational Processing of the Portuguese Language*, V. Pinheiro, P. Gamallo, R. Amaro, C. Scarton, F. Batista, D. Silva, C. Magro, and H. Pinto, Eds., *Lecture Notes in Computer Science*, vol. 13208, Cham: Springer, 2022, pp. 125–136. doi: 10.1007/978-3-030-98305–5_12.

[7] B. W. Wanjawa, L. D. A. Wanzare, F. Indede, O. Mconyango, L. Muchemi, and E. Ombui, "KenSwQuAD—A question answering dataset for Swahili low-resource language," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, vol. 22, no. 4, pp. 1–20, Apr. 2023, doi: 10.1145/3578553.

[8] C. Babirye *et al.*, "Building text and speech datasets for low-resourced languages: A case of languages in East Africa," in *Proc. 3rd Workshop on African Natural Language Processing*, 2022. Accessed: Jun. 23, 2025. [Online]. Available: https://openreview.net/forum?id=SO-U99z4U-q.

[9] A. Fesseha, S. Xiong, E. D. Emiru, M. Diallo, and A. Dahou, "Text classification based on convolutional neural networks and word embedding for low-resource languages: Tigrinya," *Information*, vol. 12, no. 2, p. 52, 2021.

[10] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data Cogn. Comput.*, vol. 4, no. 1, p. 1, 2020.

[11] D. Voskergian, B. Bakir-Gungor, and M. Yousef, "TextNetTopics Pro, a topic model-based text classification for short text by integration of semantic and document-topic distribution information," *Front. Genet.*, vol. 14, p. 1243874, 2023.

[12] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[13] M. Reid, J. Hu, G. Neubig, and Y. Matsuo, "AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages," *arXiv preprint* arXiv:2109.04715, 2021. Accessed: Jun. 23, 2025. [Online]. Available: https://arxiv.org/abs/2109.04715.

[14] V. Marivate et al., "Low resource language dataset creation, curation and classification: Setswana and Sepedi—Extended abstract," arXiv preprint arXiv:2004.13842, Mar. 30, 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2004.13842.

[15] M. A. Rahman and Y. A. Akter, "Topic classification from text using decision tree, K-NN and multinomial naïve Bayes," in *Proc. 2019 1st Int. Conf. Adv. Sci., Eng. Robot. Technol. (ICASERT)*, IEEE, 2019, pp. 1–4.

Accessed: Jun. 23, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8934502/.

[16] V. K. Agbesi, C. Wenyu, N. A. Kuadey, and G. T. Maale, "Multi-topic categorization in a low-resource Ewe language: A modern transformer approach," in *Proc. 2022 7th Int. Conf. Comput. Commun. Syst. (ICCCS)*, IEEE, 2022, pp. 42–45. Accessed: Jun. 23, 2025. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9846372/.

[17] A. Mbogho, Q. Awuor, A. Kipkebut, L. Wanzare, and V. Oloo, "Building low-resource African language corpora: A case study of Kidawida, Kalenjin and Dholuo," *arXiv preprint* arXiv:2501.11003, Jan. 19, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2501.11003.

[18] E. O. Kinanga, L. Wanzare, and C. Otieno, "SWAHILI DHOLUO TOPIC CLASSIFICATION DATASET [DS/OL]," *Zenodo*, vol. 1, Feb. 11, 2025. [Online]. Available: https://doi.org/10.5281/zenodo.14353500.

AUTHOR BIOGRAPHY



Edwin Onkoba is a final year Masters in Computer Science student at Maseno University, Kenya and holds a Bachelor's degree in Information technology from Kibabii University. He is a Natural Language Processing (NLP) and data science researcher working on low resource languages and has contributed to dataset curation for African languages such as Swahili. Apart from being a researcher, he is also a network administrator assistant at Maseno Uviversity.



Dr. **Lilian Wanzare** is the co-founder of KenCorpus and the Chairperson of the Department of Computer Science at Maseno University. She holds a Bachelor of Science in Computer Science (University of Nairobi), Master of Science in Computer Science (Free University of Bozen-Bolzano), Master of Science in Language Technology (Universität des Saarlandes) and Doctor of Philosophy in Computational Linguistics (Universität des Saarlandes). Dr. Lilian is one of the supervisors for Edwin Onkoba Kinanga and is skilled in Natural Language Processing, Text mining, Knowledge Extraction, Knowledge acquisition, Programming (Python, Java, R) and Machine Learning. Her interests and passions lie in Natural Language Processing (NLP), Natural Language Understanding (NLU), Data Science, Machine Learning, building chatbots and personal assistants.



Dr. **Calvins Otieno** is the Dean School of Computing at Maseno University. He holds a Bachelor of Science in Computer Technology (Jomo Kenyatta University of Agriculture and Technology), Master of Science inSoftware engineering (Jomo Kenyatta University of Agriculture and Technology) and Doctor of Philosophy in Information Technology (Jomo Kenyatta University of Agriculture and Technology). Dr. Calvins Otieno is one of the supervisors for Edwin Onkoba Kinanga and is skilled in Natural Language Processing and software engineering.