LLMs and African Language

1st Jason Wille
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
1352200@students.wits.ac.za

2nd Reece Lazarus
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
2345362@students.wits.ac.za

3rd Kaylyn Karuppen
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa
2465081@students.wits.ac.za

***Abstract*—This project explores the fine-tuning of XLM-RoBERTa [1], a multilingual transformer-based model, for two distinct natural language processing tasks in Swahili, an underrepresented African language. The first task involved masked language modeling (MLM) to enhance the model's understanding of Swahili by predicting masked tokens in a corpus. In the second task, the fine-tuned model was adapted for a classification task using a Swahili news dataset, aiming to categorize news articles into predefined classes. The performance of the fine-tuned model was compared against the base XLM-RoBERTa model to assess the impact of language-specific fine-tuning on classification accuracy. By focusing on Swahili, this work contributes to ongoing efforts in improving language models for African languages, which are often overlooked in NLP research. The results demonstrate the effectiveness of task-specific fine-tuning in improving model performance for low-resource languages, with implications for the broader use of multilingual models in African NLP applications.**

I. Introduction

Recent advancements in natural language processing (NLP) have led to the development of multilingual models capable of understanding and generating text in multiple languages. One such model, XLM-RoBERTa, extends the capabilities of transformer-based architectures to over 100 languages, making it a valuable tool for global NLP tasks. However, despite these advances, African languages, including Swahili, remain underrepresented in both research and practical applications. Swahili, spoken by millions of people across East Africa, presents a unique opportunity for further NLP development due to its widespread use and potential for various applications.

A key challenge in building effective models for low-resource languages like Swahili is the scarcity of large, high-quality datasets. While general-purpose multilingual models such as XLM-RoBERTa provide a strong foundation, fine-tuning these models on language-specific data can significantly improve performance. This project addresses this gap by fine-tuning XLM-RoBERTa on Swahili datasets, focusing on two tasks: masked language modeling (MLM) and text classification.

Masked language modeling is a common pre-training task in which certain tokens in a sentence are masked, and the model learns to predict these masked tokens. Fine-tuning XLM-RoBERTa using MLM on a Swahili corpus helps the model better capture the syntactic and semantic properties of the language. Following this, the model is further fine-tuned to perform classification on a Swahili news dataset, categorizing news articles into predefined categories.

This project contributes to the growing body of research on African languages in NLP by demonstrating the effectiveness of task-specific fine-tuning for Swahili. Through a comparison of the performance of the base and fine-tuned models, we aim to highlight the benefits of fine-tuning on low-resource languages.

II. Method

A. Dataset Description

Two datasets were utilized to achieve the final results of this project. The first dataset, a general unlabeled collection, was employed for the masked language modeling task. This dataset, sourced from Hugging Face's open datasets, is titled "uestc-swahili/swahili" [2]. The second dataset consists of labeled news articles that categorize content into various classifications, serving as the basis for the downstream classification task. This dataset, also obtained from Hugging Face's open datasets, is named "masakhane/masakhanews" [3].

1) uestc-swahili/swahili: This dataset comprises sentences gathered from a variety of Swahili online media platforms, covering a broad range of topics, including sports, general news, family, politics, and religion. The sentences have been divided into training, validation, and testing sets for language modeling tasks. The dataset contains 28k unique words. The training partition contains 6.84M words, validation contains 970k words and training contains 2M words. This roughly corresponds to a training, validation, test split ratio of 80:10:10. The entire dataset is lower-cased, devoid of punctuation marks, and includes start and end of sentence markers to facilitate easy tokenization during language modeling.

2) masakhane/masakhanews: This dataset is the largest available for news topic classification, encompassing 16 widely spoken African languages, including Swahili. The Swahili portion contains seven distinct topics, which serve as the classification labels. These topics include business, entertainment, health, politics, religion, sports, and technology. The dataset was divided into training, validation, and test sets with a split ratio of 70:10:20. This resulted in 1658 training articles, 237 validation articles, and 476 test articles. This differs from the stated design of doing binary classification, however, it still served the purpose of exploring the benefit of fine-tuning the base model on low-resource African language.

B. Model Architecture

In this project, we utilized XLM-RoBERTa, a variant of the RoBERTa architecture designed for cross-lingual tasks. XLM-RoBERTa builds on the success of RoBERTa by incorporating multilingual data, allowing it to handle over 100 languages, including Swahili. The architecture is based on the Transformer model, employing self-attention mechanisms and layer normalization to process sequential data efficiently. The model was sourced from Hugging Face and the model checkpoint name is *xlm-roberta-base*.

The base version of XLM-RoBERTa, which we fine-tuned, consists of 12 layers of transformers, each containing 768 hidden units and 12 attention heads. This architecture enables the model to capture complex syntactic and semantic relationships between words in different languages, making it particularly suited for tasks like masked language modeling (MLM) and text classification.

The model was first pretrained on a large, unlabeled Swahili dataset using the MLM objective, where the goal is to predict missing words in a sentence. Following this, the fine-tuned model was further adapted for a downstream task of news topic classification, leveraging the labeled Swahili news dataset. This two-step fine-tuning process allowed the model to learn general language features during pretraining, which were then refined for the specific classification task.

C. Training Setup

In this project, the fine-tuning of the XLM-RoBERTa model was conducted in two stages. The first stage involved masked language modeling (MLM) using an unlabeled Swahili dataset, while the second stage focused on a downstream task of news classification using a labeled Swahili news dataset.

1) Masked Language Modeling (MLM) Fine-tuning: The following parameters were used when fine-tuning XLM-RoBERTa for the MLM task:

- Training batch size: 8
- Evaluation batch size: 8
- Optimizer: AdamW
- Training duration: 5 epochs
- Evaluation metric(s): Loss and perplexity

2) News Topic Classification Fine-tuning: The following parameters were used when fine-tuning XLM-RoBERTa for the classification task:

- Training batch size: 8
- Evaluation batch size: 8
- Optimizer: AdamW
- Training duration: 5 epochs
- Evaluation metric(s): Accuracy, F1 score, precision, and recall.

The parameters used could have been fine-tuned further and were the default parameters provided by Hugging Face [4]. The small batch size used was originally due to memory limitation issues.

3) Systems Used for Training: The training and evaluation of the models were conducted on a high-performance computing cluster. Initially, there was a significant challenge in training the large transformer models with the compute resources available locally. The complexity of the models and the size of the datasets quickly exceeded the memory and processing capabilities of the local systems, resulting in extended training times and frequent memory allocation issues.

To address these limitations, a high-performance cluster was utilized. Specifically, a node known as bigbatch was employed for all training tasks. Each node is equipped with a single Intel Core i9-10940X CPU (14 cores), an NVIDIA RTX 3090 GPU with 24GB of memory, and 128GB of system RAM. This setup provided the necessary computational power to efficiently manage the memory and processing demands, enabling faster and more stable training.

III. Results

A. Masked Language Modeling Results

The results of the MLM fine-tuning task were evaluated based on the loss and perplexity. The training loss was measured every 10 steps while the evaluation loss was measured every 50 steps. The perplexity was measured before training (on the base model) and after training (on the fine-tuned model) using the test set. The following key observations were made:

1) Training and Evaluation Loss: Throughout the training process, both the training and evaluation loss were tracked to ensure that the model was performing as expected. Monitoring that the evaluation loss was decreasing along with the training loss ensured that the model was not overfitting. Originally the model was trained for 3 epochs and then later altered to train for 5 epochs. It can be seen that the loss still managed to decrease for both the training and evaluation sets from 3 to 5 epochs. Due to compute and time limitations further training was not possible but techniques such as early stopping could have been employed alongside a higher number of epochs to ensure that the optimal loss was achieved for the evaluation set. Please refer to figure number 1.

2) Perplexity: Perplexity, a measure of how well the language model predicts the masked tokens, was computed at the beginning and at the end of training on the test dataset. At the start of training, the model exhibited a higher perplexity, indicating that it struggled to predict masked tokens accurately in Swahili. However, by the end of the training process, the perplexity dropped significantly, reflecting the model's improved ability to predict masked tokens.

The initial perplexity was measured at 22.38 on the test set, which decreased to 4.04 by the end of the fine-tuning process (Table I). This reduction demonstrates a substantial

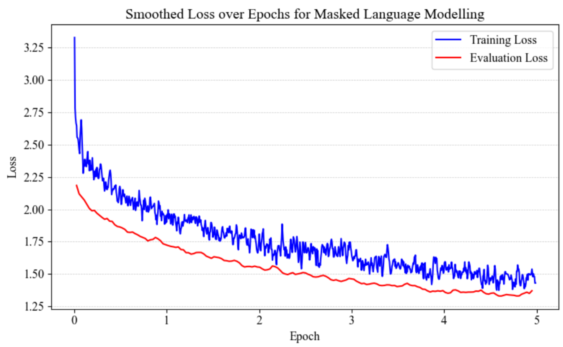

Fig. 1. Smoothed loss vs epochs for training and evaluation of the MLM task.

TABLE I
PERPLEXITY BEFORE AND AFTER MLM FINE-TUNING

	Perplexity
Base model (XLM-RoBERTa)	22.38
Fine-tuned model	4.04

improvement in the model's capacity to generate meaningful predictions for unseen text.

B. Classification Results

The classification tasks were evaluated on both the fine-tuned XLM-RoBERTa model and the base XLM-RoBERTa model. The performance of both models was assessed using four key metrics: accuracy, precision, recall, and F1 score. These metrics were tracked during evaluation steps throughout training and are visualized in the accompanying plots (Figure numbers 2, 3, 4, 5). Additionally, a final comparison of the models on the test set is provided in table II, showing the performance of each model before and after fine-tuning on the classification task.

1) Model Performance on the Test Set: The fine-tuned XLM-RoBERTa model outperformed the base model across all metrics. As shown in table II, the accuracy, precision, recall, and F1-score were consistently higher for the fine-tuned model once the model was trained for the downstream task. The starting metrics were fairly similar, but once training was complete the fine-tuned model outperformed the base model. This demonstrates the effectiveness of pretraining on a domain-specific dataset in improving downstream classification performance.

2) Evaluation During Training: During the evaluation steps of the training process, similar trends were observed. The fine-tuned model showed steady improvement across all metrics as training progressed, as illustrated in the plots (Figure numbers 2, 3, 4, 5). In contrast, the base model remained relatively stable with lower performance across the board.

Interestingly, it was noted that the accuracy and recall values were consistently identical during training and evaluation. This suggests that the model is classifying all positive instances correctly while maintaining strong overall performance, likely indicating a well-balanced dataset and effective model behavior in both identifying and distinguishing between positive and negative samples.

3) Attention Maps: In addition to evaluating the performance of the classification model using traditional metrics, attention maps were analyzed to better understand how the fine-tuned XLM-RoBERTa model processes Swahili text. Attention maps provide insight into which parts of a sentence the model focuses on when making predictions. By examining the attention heads across different layers of the model, several interesting patterns were observed.

It was difficult to determine interesting patterns in the attention heads and which heads picked up certain patterns. In order to do this a library called BertViz [5] was used. This library provided a view of the model which gave insight into all of the attention heads for a given input.

One interesting head that was located was head 1 in layer 12. This attention head seemed to pay a lot of attention to the words "corona" and "virus". The attention values for an extract from a sports article are shown in figure 6. Even though the model was paying attention to words that are usually associated with health related articles, it still manages to correctly classify this sentence as a sports article indicating that other attention heads are probably identifying other key patterns. This behavior of spotting complex patterns is to be expected from a higher layer attention head as it can pick up more advanced patterns. This suggests that the model has learned to emphasize key words that help to differentiate between different topics in the news dataset.

Another head that was probed was head 9 in layer 2. This head paid attention to the next word in the sentence. This is behavior that is to be expected from a lower layer attention head. The attention map that illustrates this is shown in figure number 7. Similarly, head 8 in layer 2 paid attention to the previous word.

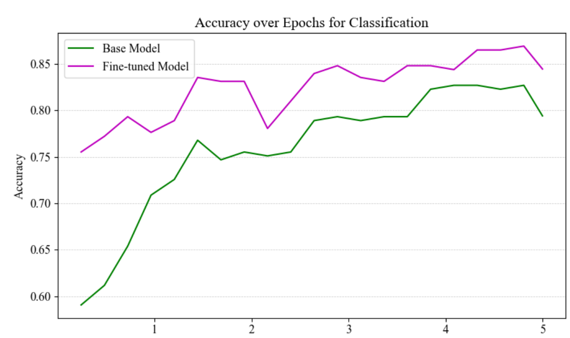

Fig. 2. Accuracy vs epochs for the classification task on the base and fine-tuned models.

IV. DISCUSSION

A. Impact of Fine-tuning

Fine-tuning played a critical role throughout this project. The initial fine-tuning of XLM-RoBERTa was performed on

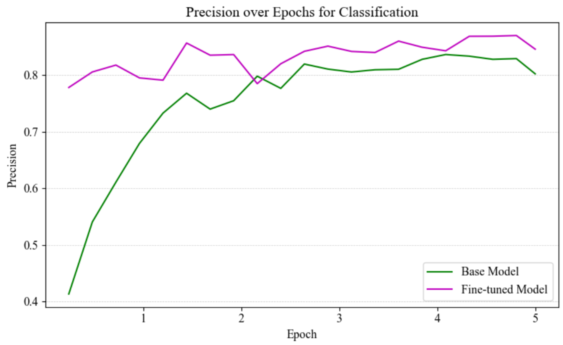

Fig. 3. Precision vs epochs for the classification task on the base and fine-tuned models.

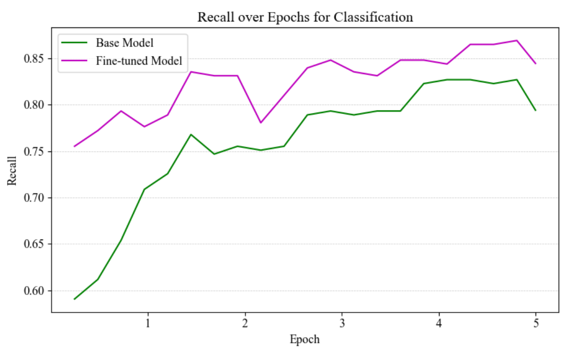

Fig. 4. Recall vs epochs for the classification task on the base and fine-tuned models.

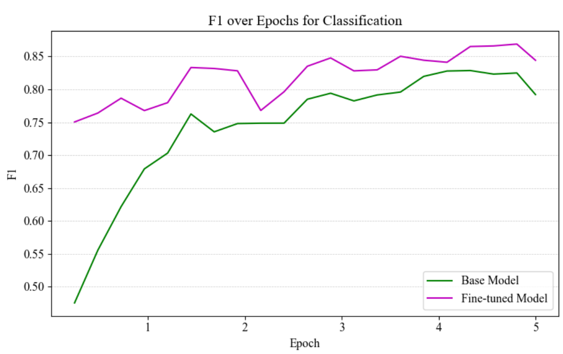

Fig. 5. F1 vs epochs for the classification task on the base and fine-tuned models.

TABLE II
CLASSIFICATION METRICS ON TEST SET BEFORE AND AFTER TRAINING

		Base Model	Fine-tuned Model
Accuracy	Before	0.21008	0.21428
	After	0.79411	0.84453
Precision	Before	0.044135	0.065372
	After	0.80215	0.84578
Recall	Before	0.21008	0.21428
	After	0.79411	0.84453
F1	Before	0.072945	0.082730
	After	0.79221	0.84428

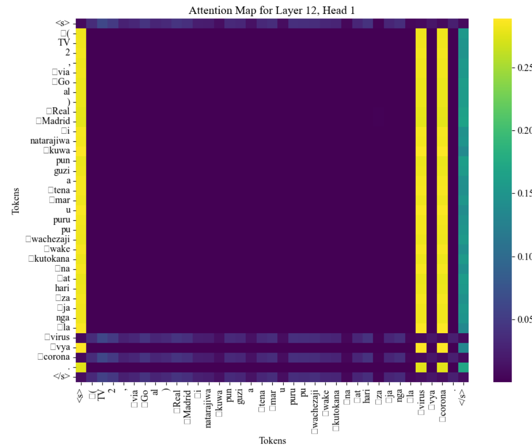

Fig. 6. Attention map highlighting the attention paid to the words "corona" and "virus".

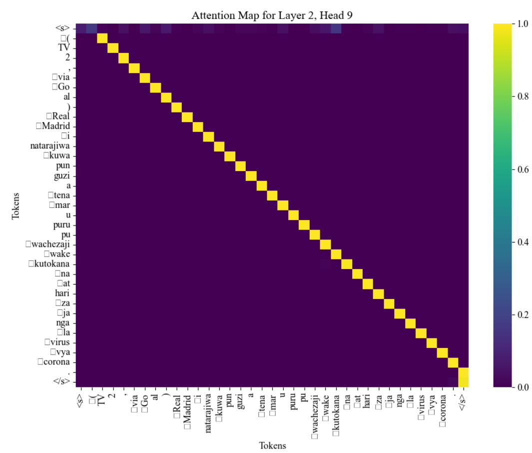

Fig. 7. Attention map highlighting the attention paid to the next word.

a general Swahili dataset for the task of masked language modeling. The goal of this step was to enhance the model's foundational understanding of the Swahili language. The primary metric used to evaluate the effectiveness of this process was perplexity. A significant reduction in perplexity was observed, indicating that the model had improved its ability to understand and generate Swahili text.

Subsequently, both the base XLM-RoBERTa model and the fine-tuned XLM-RoBERTa model were fine-tuned again, this time on a Swahili news classification dataset. The task involved predicting the correct label for news articles across seven possible categories. The hypothesis was that the previously fine-tuned model, with its improved language understanding (as indicated by the lower perplexity), would demonstrate superior classification performance compared to the base model. As shown in section III-B, this hypothesis was confirmed, with the fine-tuned model achieving higher accuracy, precision, recall and F1 score in classifying the Swahili news articles.

B. Challenges

This project faced several significant challenges, one is in relation to dataset quality and computational requirements. One of the primary obstacles was the difficulty in finding high-quality datasets for Swahili and other African languages. While some datasets are available, they are often limited in size or lack the richness required for large-scale pretraining and downstream tasks. This limitation constrained the potential performance gains from fine-tuning, as the model's ability to generalize and learn effectively is closely tied to the quality of the data it is trained on. Additionally, small datasets can contribute to issues such as overfitting, making it challenging to achieve robust and generalizable performance on the validation and test sets.

Another major challenge was the extensive computational resources required to train large models like XLM-RoBERTa. The masked language modeling task, in particular, demanded significant processing power (a GPU) and memory, which exceeded the capabilities of standard hardware. Access to a high-performance computing cluster was extremely helpful to meet these demands efficiently.

Furthermore, evaluating the performance of the model during the masked language modeling stage proved to be difficult. Unlike classification tasks, where metrics such as accuracy or F1 scores are used, masked language modeling relies on perplexity as a performance measure. Interpreting perplexity effectively is less straightforward, making it more challenging to assess the model's progress and success during training. This caused concern as the initial fine-tuning would be run and it would only be determined after the downstream task was completed if the initial fine-tuning was beneficial.

Lastly, preventing overfitting was a persistent concern, especially when fine-tuning large models on relatively small datasets. Overfitting can severely limit a model's ability to generalize to unseen data. This was checked by ensuring that the performance on the test set still portrayed the general trends that were observed during training on the training and validation sets.

References

[1] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2020. [Online]. Available: https://arxiv.org/abs/1911.02116

[2] C. S. Shikali and R. Mokhosi, "Enhancing african low-resource languages: Swahili data for language modelling," *Data in Brief*, vol. 31, p. 105951, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920308453

[3] D. I. Adelani, M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, O. Ogundepo, B. F. P. Dossou, A. Oladipo, D. Nixdorf, C. C. Emezue, S. S. al azzawi, B. K. Sibanda, D. David, L. Ndolela, J. Mukiibi, T. O. Ajayi, T. M. Ngoli, B. Odhiambo, A. T. Owodunni, N. C. Obiefuna, S. H. Muhammad, S. S. Abdullahi, M. G. Yigezu, T. Gwadabe, I. Abdulmumin, M. T. Bame, O. O. Awoyomi, I. Shode, T. A. Adelani, H. A. Kailani, A.-H. Omotayo, A. Adeeko, A. Abeeb, A. Aremu, O. Samuel, C. Siro, W. Kimotho, O. R. Ogbu, C. E. Mbonu, C. I. Chukwuneke, S. Fanijo, J. Ojo, O. F. Awosan, T. K. Guge, S. T. Sari, P. Nyatsine, F. Sidume, O. Yousuf, M. Oduwole, U. Kimanuka, K. P. Tshinu, T. Diko, S. Nxakama, A. T. Johar, S. Gebre, M. Mohamed, S. A. Mohamed, F. M. Hassan, M. A. Mehamed, E. Ngabire, , and P. Stenetorp, "Masakhanews: News topic classification for african languages," *ArXiv*, 2023.

[4] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2020. [Online]. Available: https://arxiv.org/abs/1910.03771

[5] J. Vig, "A multiscale visualization of attention in the transformer model," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 37–42. [Online]. Available: https://www.aclweb.org/anthology/P19-3007

APPENDIX

A. Contribution Statement

The contribution from the three members of the group was as follows:

- Jason Wille (1352200): 33%
- Kaylyn Karuppen (2465081): 33%
- Reece Lazarus (2345362): 34%

B. NeurIPS Paper Checklist

1) **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the contributions of the paper. The focus is on fine-tuning XLM-RoBERTa for masked language modeling and text classification in Swahili, and this is backed by experimental results throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2) **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in terms of dataset quality and computational requirements, which constrained performance gains. These limitations are explicitly mentioned in the section IV-B.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3) **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results, theorems, or proofs. It is focused on empirical analysis and fine-tuning of language models.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4) **Experimental Result Reproducibility**
Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
Answer: [Yes]
Justification: The paper provides details on the datasets used, model architecture, and training setup, allowing others to reproduce the main experimental results. Training batch sizes, optimizers, and epochs are clearly stated.
Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
    a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
    b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
    c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
    d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5) **Open access to data and code**
Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
Answer: [Yes]
Justification: The datasets used are sourced from open repositories on Hugging Face "uestc-swahili/swahili" [2], "masakhane/masakhanews" [3], and the models and training details are openly discussed. The code used to run the experiments is also available on GitHub[1].
Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6) **Experimental Setting/Details**

[1]GitHub Repository: LLMs and African Language

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
Answer: [Yes]
Justification: The paper specifies the training details, including hyperparameters, optimizer, batch sizes, and evaluation metrics. The system used for training (high-performance cluster) is also described.
Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7) **Experiment Statistical Significance**
Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
Answer: [No]
Justification: While the paper reports evaluation metrics (e.g., accuracy, F1-score), it does not report statistical significance or error bars. This could be improved in future work by adding statistical analysis.
Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8) **Experiments Compute Resources**
Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
Answer: [Yes]
Justification: The paper describes the high-performance computing cluster used, including the type of CPU, GPU, and memory available.
Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9) **Code Of Ethics**
Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?
Answer: [Yes]
Justification: The research conforms to the NeurIPS Code of Ethics. It discusses the implications of improving NLP for African languages and highlights the positive societal impact of expanding NLP tools for underrepresented languages.
Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10) **Broader Impacts**
Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
Answer: [Yes]
Justification: The paper discusses the positive societal impact of improving Swahili NLP and contributing to low-resource language research. However, it does not mention any potential negative impacts, which could be added for a more comprehensive discussion.
Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11) **Safeguards**
Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
Answer: [NA]
Justification: This paper does not describe safeguards as the models trained were not released.
Guidelines:
    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12) **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: The datasets used are sourced from Hugging Face and are properly credited and included in the references. The license and terms of use are respected. The original papers for Hugging Face and the datasets used are cited as well as any third party libraries used.
Guidelines:
    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13) **New Assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [NA]
Justification: The paper does not introduce new datasets or models, so this question is not applicable.
Guidelines:
    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14) **Crowdsourcing and Research with Human Subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [NA]
Justification: The paper does not involve crowdsourcing or research with human subjects.
Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.
15) **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**
Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [NA]
Justification: The paper does not involve human subjects, so IRB approval is not necessary.
Guidelines:
    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.