Extractive Swahili Question-Answering with

DistilBERT on KenSwQuAD ( Kencorpus Swahili Question Answering Dataset)

|  |  |
| --- | --- |
| Evelyn Wangui Wachira  C241-01-2305/2023  *Department of Computer Science*  *Dedan Kimathi University of Technology*  Nyeri, Kenya  evelyn.wachira23@students.dkut.ac.ke | Benson Kituku  *Department of Computer Science*  *Dedan Kimathi University of Technology*  Nyeri, Kenya benson.kituku@dkut.ac.ke |

***Abstract*—Although there are many non-English speaking persons in the globe, most of the question-answering (QA) systems available today are designed with English speakers in mind. This study explores the application of DistilBERT, a distilled version of the BERT model, for extractive question answering on Swahili, a low-resource language. The data was from Kencorpus Swahili Question Answering Dataset (KenSwQuAD), dataset is annotated from raw story texts of Swahili, a low-resource language that is predominantly spoken in eastern Africa. The data preparation process involved using the Haystack annotation tool to efficiently annotate and validate the Swahili question-answer pairs and adapting it into SQuAD format. DistilBERT, a smaller and faster variant of BERT, was chosen for its ability to achieve competitive performance with a confidence score of 0.77 against multilingual BERT which was 0.16 while tested using a Swahili example. The model was fine-tuned on customized KenSwQuAD dataset to learn patterns specific to the Swahili language. To evaluate the performance of the fine-tuned DistilBERT model, we employed two standard evaluation metrics: exact match and F1 score. With a performance of 6. 45 exact match and 35.4 F1 score. The low results were due to a small dataset. Our results demonstrate the effectiveness of using DistilBERT for extractive questions answering on Swahili, despite the limited availability of labeled data. .**

***Index Terms*—Extractive Question Answering, Low-Resource Languages, Swahili, DistilBERT, BERT-based Models**

# I. INTRODUCTION

The field of natural language processing (NLP) has witnessed significant advancements in recent years, particularly with the introduction of transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) (Kenton et al., 2019). Research on NLP models has focused on leveraging models like BERT, RoBERTa, ALBERT, and DistilBERT for various task-specific purposes, highlighting the versatility and adaptability of these models (Shidaganti, 2024). Comparative analyses of the models previously mentioned have provided valuable insights into their strengths and limitations, influencing their application in industry and research (ÖZKURT, 2024). Extractive question and answering¨ models, have become prominent for their accurate answer extraction capabilities (V, 2023). It has practical applications in various domains, including information retrieval, chatbots, and educational systems. DistilBERT(Sanh et al., 2019), a distilled version of BERT, has gained attention for its reduced size and faster inference speed while maintaining comparable performance to BERT. DistilBERT has been successfully applied to various NLP tasks, including question answering, and has shown potential for resource-constrained environments(Sanh et al., 2019). However, the majority of NLP research and resources have been focused on high-resource languages such as English, while low-resource languages have received comparatively less attention. While significant progress has been made in extractive question answering for high-resource languages, particularly with the use of BERT and its variants, research on this task for low-resource languages like Swahili is still limited.(Kenton et al., 2019).

Swahili, a Bantu language widely spoken in eastern and central Africa, plays a significant role in daily interactions among people of diverse ethnic backgrounds(Wanjawa et al., 2023). The historical background of Swahili has been influenced by various factors leading to diverging ways of speaking the language in different regions (Nassenstein & Dimmendaal, 2020). Despite its widespread use, Swahili has limited NLP resources and tools compared to high-resource languages such as English. In the realm of Natural Language Processing (NLP) question-answering tasks, the Swahili language presents distinctive challenges and opportunities. Swahili, characterized as a low-resource language, has attracted attention in NLP research due to the limited availability of data and resources for processing tasks (Masua & Masasi, 2020). The creation of language models tailored specifically for Swahili, SwahBert, has been developed to tackle the constraints faced by lowresource languages like Swahili in NLP applications (Martin et al., 2022). Furthermore, the existence of datasets like Kencorpus, encompassing Swahili, Dholuo, and Luhya languages, has facilitated the development of proof-of-concept systems for tasks like Swahili speech-to-text and machine learning based question-answering systems. Specialized datasets like KenSwQuAD for Swahili have been developed to meet the need for QA datasets in low-resource languages, enabling tailored solutions for information retrieval and question answering (Wanjawa et al., 2023). These efforts highlight the importance of tailored solutions and data resources in addressing the linguistic diversity and cultural nuances inherent in the Swahili language for precise processing in NLP question answering tasks.

Models like DistilBERT have shown promise in enhancing machine reading comprehension models and improving answer extraction accuracy(V, 2023). Cross-lingual studies have explored the transferability of monolingual representations in QA tasks, underscoring the importance of understanding linguistic nuances across different languages (Artetxe et al., 2020). This study draws inspiration from the development of chatbots for specific language speakers, such as the Central Kurdish (Sorani) dialect, emphasizing the importance of catering to diverse linguistic needs in NLP applications(Ahmed & Hussein, 2021). Given the limited research on extractive question answering for Swahili and the potential of transfer learning approaches, this study aims to investigate the application of DistilBERT for Swahili extractive question answering. By leveraging the knowledge learned from high-resource languages through the use of a multilingual DistilBERT model, this research explores the feasibility and effectiveness of transfer learning for improving question answering performance in a low-resource setting. The findings of this study contribute to the growing body of research on NLP for low-resource languages and provide insights into the challenges and opportunities of applying transfer learning techniques for Swahili extractive question answering. The outcomes of this research have the potential to advance the development of NLP applications for Swahili and other low-resource languages, enabling better access to information and knowledge for underserved communities. The remainder of this paper is subdivided as follows: Section 2 discusses some related work in question and answering . Section 3 describes the methodology used in data collection, preparation and model training for translation. Section 4 discusses the translation results of the model and evaluation of its performance. Finally, Section 5 concludes the work and provides an insight for future work.

# II. RELATED WORK

In recent years, the application of BERT-based models in extractive question answering tasks for low-resource languages has garnered significant attention in the Natural Language Processing (NLP) research domain. BERT, a groundbreaking language representation model introduced in 2018, pretrains deep bidirectional representations from unlabeled text by conditioning on both left and right context in all layers, revolutionizing the field of NLP and question answering (Kenton et al., 2019). The utilization of BERT pre-trained language model embeddings in question answering over knowledge bases has demonstrated the effectiveness of leveraging advanced language models for encoding question and answer contexts (Sai Sharath & Banafsheh, 2020). These models have shown promising results in enhancing language understanding and question-answering capabilities across diverse linguistic contexts. Multi-lingual BERT-based models, such as mBERT, have been utilized to transfer knowledge from high-resource languages to low-resource languages, showcasing the potential for cross-lingual applications in QA tasks (Kumar et al., 2022). Additionally, research has focused on utilizing advancements in NLP models, such as BERT, RoBERTa, ALBERT, and DistilBERT, for specific tasks, demonstrating the versatility and adaptability of these models in various NLP applications (Shidaganti, 2024). Challenges in domain adaptation related to tokenization and sub-word representations have been addressed to bolster the robustness of models like BERT in handling out-of-vocabulary words, thereby enhancing their performance across diverse linguistic contexts (Nayak et al., 2020). However, for languages like Thai, which face constraints in model choices due to limited datasets, strategies such as finetuning multi-lingual models like XLM-RoBERTa have been employed to address the scarcity of resources (Lowphansirikul et al., 2021). Moreover, recent advancements in transformer-based language models, such as BERT, have led to the development of specialized models like AraBERT for Arabic language understanding, showcasing the adaptability and efficiency of language-specific BERT-based models(Antoun et al., 2020).

Wanjawa et al. (2023) introduced the KenSwQuAD dataset, specifically designed for Swahili, which has played a crucial role in resourcing the Swahili language for QA tasks. This dataset has been instrumental in providing a foundation for research in Swahili QA systems. In their paper (Chaybouti, 2021) introduced EfficientQA, a RoBERTa-based question-answering system that utilizes BERT-based models to create dense representations of candidate answers based on SQuAD v1.1(Stanford Question Answering Dataset) and FQuAD(The French Question Answering Dataset) datasets. The study highlighted the model's efficiency in generating meaningful answers, leading to improved EM and F1 scores. (Cui et al., 2019) proposed a novel model called Dual BERT, leveraging large-scale training data from rich-resource languages to enhance reading comprehension performance in low-resource languages. The study focused on improving semantic relations between passages and questions in a bilingual context. (Debnath et al., 2021) explored the use of multilingual BERT for equitable question answering systems, outperforming existing approaches like DocumentQA. (Rajpurkar et al., 2018)introduced unanswerable questions for SQuAD, expanding the scope of question answering tasks. (Xu & Khanna, 2020) highlighted the importance of the single-span task formulation in extractive question answering, emphasizing the adaptability of models to address complex questions.

In the context of cross-lingual representation learning, models like XLM-Roberta have shown significant improvements in low-resource languages, enhancing accuracy and performance metrics for languages like Swahili and Urdu(Conneau et al., 2020). The development of deep learning models, such as DC-BERT, has addressed efficiency challenges by proposing a decoupled contextual encoding framework, incorporating online and offline BERT models for optimized contextual encoding in QA systems(Zhang et al., 2020). (Ojokoh et al., 2023) presented a graph model with an integrated pattern and query-based technique for extracting answers to questions in community question answering systems. This approach can offer valuable insights into answer extraction methodologies that may be applicable to Swahili Extractive Question Answering. (Farea, 2024) conducted an experimental design of Extractive Question-Answering Systems, focusing on the influence of error scores and answer length.

This section highlights the diverse strategies, models, and datasets employed in the development of extractive QA systems using BERT-based models for low-resource languages. These studies underscore the importance of leveraging pre-trained language models, cross-lingual representations, and domain-specific adaptations to enhance the effectiveness and applicability of QA systems in linguistic contexts with limited resources. This study provides insights into the evaluation of EQA systems using DistilBERT, where there are limited resources to train a custom model thus the use of a lightweight version of BERT and scarcity of question-and-answer dataset for a low resource languages like Swahili.

# III. METHODOLOGY

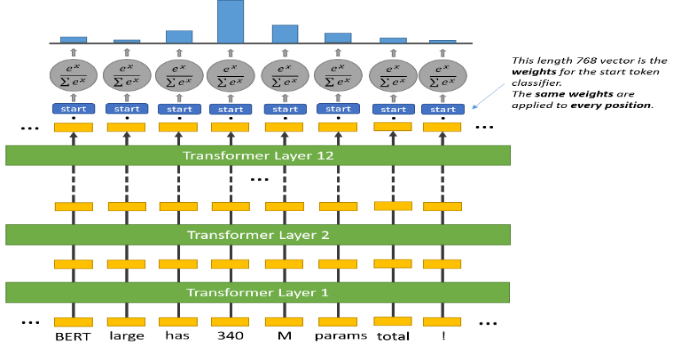
In this section, we outline the steps involved in developing a robust system that can accurately extract answers from Swahili text based on the KenSwQuAD dataset. This task requires leveraging the capabilities of DistilBERT, a distilled version of BERT, known for its efficiency in machine reading comprehension tasks. This includes steps for data preparation, model selection, training, and evaluation to ensure the effectiveness of the Swahili Extractive Question Answering system while leveraging the KenSwQuAD dataset as a benchmark for evaluating the model's performance.

## Data Preparation

The process of preparing the Kencorpus Swahili Question Answering (QA) dataset involves several crucial steps to ensure high-quality data for training and evaluation of natural language processing (NLP) models. In our case we used domain-specific dataset. Question and answer pairs were provided in .csv file format and the context text are in labeled .txt file where the answer to every question a segment of text is, or span, from the corresponding reading passage (Wanjawa et al., 2023). The Haystack annotation tool is employed in this process to facilitate the creation of a structured and annotated dataset that can be used for training a Bert model on an extractive question and answer downstream task. The Haystack tool allows for precise marking of answer spans, making it easy to extract the required information. We then exported the data in SQUAD format and saved it in a JSON file. We then split the data into training, validation, and test sets(“Annotating {Data},” n.d.).

## Model Selection: Bert

We introduce BERT and its detailed implementation in this section. There are two steps in our framework: pre-training and fine-tuning. We applied Bert to solve this problem using Hugging Face Library(*Hugging {Face} – {The} {AI} Community Building the Future.*, n.d.). For Question Answering Bert takes 2 parameters input question and the text which contains the answer as a packed sequence. It will produce a 768-dim vector output corresponding to each token. In downstream tasks like Question-Answering we will have 2 linear layers - one for start position and another for end position(start token classifier and end token classifier). We have separate weights for each of them. During finetuning they are trained together. During inference for every token in the text, we feed its final embedding into the start token classifier as well as end token classifier. For each token internally a dot product occurs with a start token vector and produces logits corresponding to that token. Similarly for the end token classifier as well. Thus, the model will produce start logits and end logits corresponding to all the input tokens.



*Figure1: Start token classification*

A diagram of a diagram

Description automatically generated

*Figure 2: end token classification*

The use of Transformers has become common, and our implementation is almost identical to the original, we will omit an exhaustive background description of the model architecture. In this work, we denote the number of layers (i.e., Transformer blocks) as L, the hidden size as H, and the number of self-attention heads as A. We primarily report results on two model sizes: BERTBASE (L=12, H=768, A=12, Total Parameters=110M) and BERTLARGE (L=24, H=1024, A=16, Total Parameters=340M) (*Hugging {Face} – {The} {AI} Community Building the Future.*, n.d.).

We ran a simple inference pipeline on the three Bert based models to see how the input and prediction looks like on all the models. This was to determine the best Bert based model to fine tune on our downstream task. The models used were Swahili-question-answer-latest-cased which is a Bert based model trained on Swahili text only, Bert-base-multilingual-cased which is trained on 104 languages(Schlinger, 2019), distilbert-base-cased-distilled-squad from the hugging face library. We ran an example set in question, answer and context format and they had different confidence score: innocent-charles/Swahili-question-answer-latest-cased(from the hugging face library)0.0016, Bert-base-multilingual-cased 0.0018, distilbert-base-cased-distilled-squad 0.77. The implemented simple inference pipeline effectively leveraged the strengths of multiple BERT-based models. By selecting the prediction with the highest confidence score, the pipeline improved the overall reliability of the outputs.

IV. RESULTS

In our experiments, we use Kencorpus Swahili Question Answering Dataset (KenSwQuAD) question-answering dataset for fine-tuning and evaluation(Wanjawa et al., 2023). We used 230 data points for training, 50 data points for testing and 30 data points for validation created using haystack annotation tool. We tested distilbert on a downstream task of extractive question and answer. According to (Kenton et al., 2019) Bert input is a combination of 3 inputs: Word piece embedding, The token embeddings and Positional embeddings were generated by model itself. We passed Segmentation embedding where it took 0 for all tokens related to question and 1 for all tokens related to Context Positional embedding and segmentation embedding. After tokenization the input was in the format [CLS] question [SEP] context [SEP]. We then added suitable padding and the converted word ids and mapped with embedding matrix to generate Embedding vector and corresponding positional embedding vector. We used the pretrained "distilbert-base-cased" weights which were not fine-tuned. We used two metrics for evaluation exact match(EM) and F1 score. For each question-answer pair if the characters of the models prediction exactly match with the characters of true answer, then EM=1 else 0. F1 score depends on precision and recall (Blagec et al., 2022). We had a very poor score as expected(exact\_match:2.9 and F1 score:11.2). During the fine-tuning, the parameters are optimized using the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of =2e-5and ϵ parameter of 1e-8. The batch size was set to 2 with 2 epochs. The results were 6. 45 exact match and 35.4 F1 score. We believe that better performance will be recorded if more data points and computing resources are used to fine tune the model.

V. CONCLUSION

The research on DistilBERT for question answering tasks, particularly in extractive question answering systems, has provided valuable insights into the model's performance. While the model may exhibit slightly lower accuracy compared to other state-of-the-art models like BERT, RoBERTa, and ALBERT, its efficiency and effectiveness in generating answers were notable. Further exploration of DistilBERT's capabilities in domain-specific question answering tasks to assess its performance in specialized contexts. Additionally, conducting comparative studies with other transformer models and evaluating DistilBERT's adaptability to different languages and domains could provide deeper insights into its versatility and effectiveness.

# REFERENCES

Ahmed, H. K., & Hussein, J. A. (2021). Design and Implementation of a Chatbot for Kurdish Language Speakers Using Chatfuel Platform. *Kurdistan Journal of Applied Research*. https://doi.org/10.24017/science.2020.2.10

Annotating {Data}. (n.d.). In *Haystack Documentation*. Retrieved July 4, 2024, from https://docs.haystack.deepset.ai/v1.25/docs/annotation

Antoun, W., Baly, F., & Hajj, H. (2020). *AraBERT: Transformer-based Model for Arabic Language Understanding*. *May*. http://arxiv.org/abs/2003.00104

Artetxe, M., Ruder, S., & Yogatama, D. (2020). *On the Cross-Lingual Transferability of Monolingual Representations*. https://doi.org/10.18653/v1/2020.acl-main.421

Blagec, K., Dorffner, G., Moradi, M., Ott, S., & Samwald, M. (2022). A global analysis of metrics used for measuring performance in natural language processing. *NLP-Power 2022 - 1st Workshop on Efficient Benchmarking in NLP, Proceedings of the Workshop*, *3*, 52–63. https://doi.org/10.18653/v1/2022.nlppower-1.6

Chaybouti, S. (2021). *EfficientQA : A RoBERTa Based Phrase-Indexed Question-Answering System*. https://doi.org/10.48550/arxiv.2101.02157

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). *Unsupervised Cross-Lingual Representation Learning at Scale*. https://doi.org/10.18653/v1/2020.acl-main.747

Cui, Y., Che, W., Liu, T., Qin, B., Wang, S., & Hu, G. (2019). *Cross-Lingual Machine Reading Comprehension*. https://doi.org/10.18653/v1/d19-1169

Debnath, A., Rajabi, N., Alam, F. F., & Anastasopoulos, A. (2021). *Towards More Equitable Question Answering Systems: How Much More Data Do You Need?* https://doi.org/10.18653/v1/2021.acl-short.79

Farea, A. (2024). Experimental Design of Extractive Question-Answering Systems: Influence of Error Scores and Answer Length. *Journal of Artificial Intelligence Research*. https://doi.org/10.1613/jair.1.15642

*Hugging {Face} – {The} {AI} community building the future.* (n.d.). Retrieved August 4, 2024, from https://huggingface.co/

Kenton, M. C., Kristina, L., & Devlin, J. (2019). *BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding*. 4171–4186.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–15.

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N., & Nutanong, S. (2021). *WangchanBERTa: Pretraining transformer-based Thai Language Models*. http://arxiv.org/abs/2101.09635

Martin, G., Mswahili, M. E., Jeong, Y.-S., & Young-Seob, J. (2022). *SwahBERT: Language Model of Swahili*. https://doi.org/10.18653/v1/2022.naacl-main.23

Masua, B., & Masasi, N. (2020). Enhancing text pre-processing for Swahili language: Datasets for common Swahili stop-words, slangs and typos with equivalent proper words. *Data in Brief*, *33*. https://doi.org/10.1016/j.dib.2020.106517

Nassenstein, N., & Dimmendaal, G. J. (2020). Bunia Swahili and Emblematic Language Use. *Journal of Language Contact*. https://doi.org/10.1163/19552629-01203008

Nayak, A., Timmapathini, H., Ponnalagu, K., & Venkoparao, V. G. (2020). *Domain Adaptation Challenges of BERT in Tokenization and Sub-Word Representations of Out-of-Vocabulary Words*. https://doi.org/10.18653/v1/2020.insights-1.1

Ojokoh, B. A., Igbe, T., Afolabi, B., & Daramola, O. (2023). A Graph Model With Integrated Pattern and Query-Based Technique for Extracting Answer to Questions in Community Question Answering System. *Social Network Analysis and Mining*. https://doi.org/10.1007/s13278-023-01046-3

ÖZKURT, C. (2024). *Comparative Analysis of State-of-the-Art Q\\&amp;amp;A Models: BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD V2 Dataset*. https://doi.org/10.21203/rs.3.rs-3956898/v1

Rajpurkar, P., Jia, R., & Liang, P. (2018). *Know What You Don’t Know: Unanswerable Questions for SQuAD*. https://doi.org/10.18653/v1/p18-2124

Sai Sharath, J., & Banafsheh, R. (2020). Question Answering over Knowledge Base using Language Model Embeddings. *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN48605.2020.9206698

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT , a distilled version of BERT : smaller , faster , cheaper and lighter*. 2–6.

Schlinger, E. (2019). *How multilingual is Multilingual BERT?*

Shidaganti, G. (2024). Exploratory Analysis on the Natural Language Processing Models for Task Specific Purposes. *Bulletin of Electrical Engineering and Informatics*. https://doi.org/10.11591/eei.v13i2.6360

V, S. (2023). The DistilBERT Model: A Promising Approach to Improve Machine Reading Comprehension Models. *International Journal on Recent and Innovation Trends in Computing and Communication*. https://doi.org/10.17762/ijritcc.v11i8.7957

Wanjawa, B. W., Wanzare, L. D. A., Indede, F., McOnyango, O., Muchemi, L., & Ombui, E. (2023). KenSwQuAD - A Question Answering Dataset for Swahili Low-resource Language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(4), 1–17. https://doi.org/10.1145/3578553

Xu, M.-A., & Khanna, R. (2020). *Importance of the Single-Span Task Formulation to Extractive Question-Answering*. https://doi.org/10.5121/csit.2020.101809

Zhang, Y., Nie, P., Geng, X., Ramamurthy, A., Song, L., & Jiang, D. (2020). *DC-BERT: Decoupling Question and Document for Efficient Contextual Encoding*. https://doi.org/10.48550/arxiv.2002.12591