|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | INTERNATIONAL TELECOMMUNICATION UNION  **TELECOMMUNICATION STANDARDIZATION SECTOR**  STUDY PERIOD 2022-2024 | | FG-AI4H-S-033 | |
| **ITU-T Focus Group on AI for Health** | |
| **Original: English** | |
| **WG(s):** | | Plenary | Geneva, 3-5 July 2023 | |
| **DOCUMENT** | | | | |
| **Source:** | | Jacaranda Health | | |
| **Title:** | | Triaging High Volume Code Mixed Maternal Healthcare Queries in Low Resource Settings Using Natural Language Processing Techniques | | |
| **Contact:** | | Jay Patel Jacaranda Health Kenya | | Tel: +254704888680 E-mail: [jpatel@jacarandahealth.org](mailto:jpatel@jacarandahealth.org) |
| **Contact:** | | Stanslaus Mwongela Jacaranda Health Kenya | | Tel: +254719515156 E-mail: [smwongela@jacarandahealth.org](mailto:smwongela@jacarandahealth.org) |

|  |  |
| --- | --- |
| **Abstract:** | Maternal mortality remains a significant challenge in Kenya despite efforts to improve healthcare services. This study presents the innovative use of a SMS-based digital health platform called PROMPTS, developed by Jacaranda Health, a Non-Profit Organization based in Kenya, Ghana, and Eswatini. PROMPTS leverages Natural Language Processing (NLP) techniques to handle maternal healthcare queries and provide timely, accurate responses, addressing the pressing issues in low-resource settings.  PROMPTS reaches 2.2 million mothers in 1,110 Kenyan health facilities, offering free and easily accessible support through SMS. The platform empowers women with knowledge to make informed decisions about their health. Notably, PROMPTS is endorsed by government health systems, ensuring prompt response to all maternal inquiries regardless of volume, reducing the risk of vital questions being overlooked.  The study highlights PROMPTS' progress, challenges faced, and lessons learned. It discusses specialized datasets construction, effective triaging of queries with limited language resources and budgets, and handling long-tailed query distributions. Modeling approaches for extreme multi-class settings and high confusion are explored. Implementation and deployment of machine learning applications and ethical considerations in AI model creation are also examined.  This study showcases PROMPTS' transformative potential in reshaping health-seeking behaviors among Kenyan women, emphasizing NLP's critical role in improving maternal healthcare outcomes. It contributes to AI for healthcare in low-resource and domain-specific languages, emphasizing the importance of addressing data challenges and modeling complexities in such settings. |

# Introduction:

The wide-spread adoption of digital and remote healthcare systems has brought about a paradigm shift in the accessibility of clinical information and support for patients. Patients commonly interact with these systems by asking clinically-related questions, which, given the diverse nature of their conditions, demand nuanced responses (Bai et al., 2022).

If the objective is to guide patients down the most appropriate care pathway for their condition, the accuracy of these question-and-answer systems relies on the quality and expansiveness of the knowledge base (user questions, associated medical intents and associated responses) and on the intelligence of the Natural Language Processing (NLP) unit that processes this knowledge base (Wu et al., 2020).

The leading approaches in modern Natural Language processing are notoriously data hungry. A good example being Transformer models, which achieve surging and state-of-the-art performance at the cost of big data. The process of acquiring this big data is expensive and time-consuming for many application domains, which limits access and research. Existing research using these architectures occurs in the domain of high-resource languages, like English, Chinese, and Spanish, where training corpora are in abundance (Wongso et al., 2022). Scant NLP research exists around low resource and/or domain-specific languages (Litre et al., 2022). The healthcare domain in the settings of low-resource languages, for instance, African languages is one area where NLP research using these state-of-the-art deep learning models lags.

Jacaranda Health is a non-profit organization based in Kenya, Eswatini, and Ghana, whose mission is to improve maternal and newborn health outcomes in public health systems. The organization works with governments to deploy affordable, scalable solutions through public hospitals where underserved mothers receive care. To improve access to healthcare, Jacaranda has launched [PROMPTS,](https://www.jacarandahealth.org/prompts) a revolutionary free service SMS-based digital health platform designed to empower expecting and new mothers with the critical information they need to seek care at the right time and place. This cutting-edge solution fuses a series of messages aimed at influencing key behaviors associated with better health outcomes, such as prenatal care attendance, with a state-of-the-art NLP powered helpdesk. This help desk is capable of reading, responding to, and prioritizing incoming SMS queries from mothers in real-time, based on the urgency of their medical needs.

With up to 4000-5000 medical queries sent through PROMPTS every day, the NLP model plays a crucial role in ensuring that the response time for high-priority maternal queries and danger signs is optimized. Without it, a first-in-first-out approach would result in equal priority being given to all questions, leading to poor response times and potentially devastating consequences for maternal health. Today, PROMPTS reaches 2.2m mothers across 1,110+ Kenyan health facilities, with proven success in shifting health seeking behaviors linked with better health outcomes, including prenatal attendance and family planning uptake (2x). PROMPTS is one of the few digital health solutions actively endorsed by government health systems on such a large scale.

Overall, PROMPTS has made significant progress in transforming health-seeking behaviors among Kenyan women, offering hope in the battle against maternal mortality. In this work we share lessons we’ve learned in building datasets, triaging maternal health care queries in low resource settings (under resourced in-domain languages and small budgets), dealing with long-tailed distributions of maternal healthcare queries, modeling for extreme multi-class settings with high degrees of confusion, implementing and deploying ML apps and ethical considerations we observe in creating AI models.

# Problem Formulation

The Natural Language Processing (NLP) problem setting for Jacaranda Health Care involves training datasets in Swahili, Sheng (code-mixed), English, and other local Kenyan languages. These languages (Swahili, Sheng, and Kenyan local languages) are not adequately represented in existing state of the art pre-trained NLP models. Additionally, the context of the text is maternal health discussions, which includes specialized vocabulary in local languages which is missing from standard NLP corpora. Furthermore, the maternal health queries distribution is highly imbalanced, between different intents, for example > 85% of questions from mothers so far are regarding general queries on pregnancies. The degree of confusion (level of ambiguity or similarity between different labels) for some labels across the distribution of the queries is also high. Some labels are distinct and easily distinguishable from each other, while others have similarities or overlap in their characteristics, making them more prone to confusion.

# Research or Innovation:

In this work, we evaluate four strategies in designing a data-efficient classifier under the problem setting of Jacaranda Health. Specifically, we share the performance of models trained using; i) fine-tuning pre-trained multilingual models, ii) supervised contrastive learning using pre-trained multilingual models, iii) language adaptive fine-tuning with additional fine-tuning for the downstream task, iv) language adaptive fine-tuning + downstream task fine-tuning with weighted loss and lastly v) hierarchical clustering model adaptively fine tuned and sample weighted.

The last approach which involved first selecting a pre-trained transformer multilingual model (XLMRoberta) with Swahili checkpoint and adaptively fine-tuning it to our in-domain Jacaranda Health Maternal Queries worked and creating hierarchical models based on query clustering worked best for our settings. This work opens up new avenues for improving the performance of pre-trained models on smaller, task-specific datasets and highlights the importance of addressing the long-tailed distributions and degrees of confusion in multi-class settings.

# Dataset:

Our dataset consists of queries posed by mothers to the PROMPTS system between July 2021 to July 2022. This dataset contains 939,819 unlabelled questions sent through the PROMPTS SMS framework as asked by the mothers Jacaranda serves. The questions are in low-resource language text, which is either in Swahili, Sheng (code mixed Swahili and English), and other local Kenyan languages, and were asked between the months of July 2021 to July 2022.

The labeled dataset contained 162,525 questions asked by the mothers using our services

and labeled by clinically trained personnel. The intent labels were 176 in total. The distribution of questions across the 176 intents was highly imbalanced.

The queries’ distribution of the annotated data (training data) reflected the actual distribution of the queries in the real-world setting they were collected (mothers’ queries distribution as received via the PROMPTS pipeline). Furthermore, some of the labels were very weak, for instance, “baby\_general” and “pregnancy\_general”. Noisy Antenatal Care (ANC) questions are mostly classified as pregnancy\_general while noisy Postnatal Care questions were classified as baby\_general.

# Technical Details:

We conducted four experiments to address the problem we specified. These experiments included fine tuning pre-trained multi-lingual transformer models, supervised contrastive learning, language adaptive finetuning using masked language modeling objective and hierarchical clustering.

## Fine Tuning using Pre-Trained Language Models.

In this experiment, we aimed to evaluate the performance of three state-of-the-art pre-trained transformer multilingual models: XLM-Roberta-Large, MT5, and Afro-XLMR-Large. The evaluation was conducted through fine-tuning using the Jacaranda Health Labeled maternal queries training dataset, which consisted of 107,000 instances.

The first model, MT5, is a multilingual variant of T5 that was pre-trained on a large dataset covering 101 languages. The dataset was cleaned by removing null values and irrelevant columns, leaving only the maternal query and the intent column. The dataset was then split into train and test sets using a stratified split strategy with an 80:20 ratio. The text rows and intents were tokenized using the T5 tokenizer, and the MT5EncoderModel was chosen for fine-tuning. A classification head was added on top of the CLS embedding. The fine-tuned MT5 model achieved a balanced accuracy score of 0.61 and a weighted F1 score of 0.71.

The second model, XLM-Roberta-Large, is a multilingual version of RoBERTa pre-trained on a large amount of Common Crawl data in 100 languages. The model was fine-tuned on the Jacaranda Health Care dataset, resulting in a fine-tuned model evaluated for precision, recall, F1 score, and balanced accuracy score. The fine-tuned XLM-Roberta-Large achieved a balanced accuracy score of 0.76 and a weighted F1 score of 0.76.

The third model, Afro-XLMR-Large, was created by adapting XLM-Roberta-Large using MLM on 17 African languages and 3 high-resource languages. The same data preprocessing and training approach as the previous experiments were applied to finetune the Afro-XLMR-Large model. The fine-tuned model achieved a balanced accuracy score of 0.75 and a weighted F1 score of 0.77.

Overall, these experiments demonstrated the performance of different pre-trained transformer multilingual models in the task of fine-tuning for maternal query classification. The results varied across the models, with XLM-Roberta-Large achieving the highest scores in terms of accuracy and F1 score.

## Supervised Contrastive Learning.

In this experiment, we utilized a supervised contrastive learning objective and an augmentation objective, implementing them with a pre-trained multi-lingual transformer model. The inspiration for this experiment came from a paper called "SupCL-Seq: Supervised Contrastive Learning for Downstream Optimized Sequence Representations" by Sedghamiz et al. (2021).

The experiment involved training a contrastive learner to generate augmented altered views by modifying the dropout mask probability in a standard Transformer architecture. This alteration process created different views of the original anchor representation. The goal was to train the model to bring together similar samples (anchors and their altered views) and separate samples belonging to other classes. To achieve this, a supervised contrastive loss was employed.

The training process involved training the model for 10 epochs, and the best checkpoint was selected based on a supervised contrastive loss of 0.65. The checkpoint represented the model's parameters at that stage. Subsequently, the base model parameters of the checkpoint were frozen, and a sequence classification head was added on top of the model. This modified model was then used for inference to evaluate its performance.

The contrastively trained XLM-Roberta model, with a contrastive loss of 0.65, achieved a Weighted F1 score of 0.75 and a balanced accuracy score of 0.72. These metrics indicate the model's performance in terms of its ability to classify sequences accurately. A Weighted F1 score of 0.75 suggests that the model achieves a good balance between precision and recall across the different classes, while a balanced accuracy score of 0.72 indicates its overall accuracy in predicting class labels.

Overall, this experiment demonstrates the effectiveness of supervised contrastive learning in improving the representation capabilities of a pre-trained multi-lingual transformer model. By leveraging contrastive learning and augmentation techniques, we were able to enhance the model's ability to discriminate between different samples, leading to improved performance in sequence classification tasks.

## Language Adaptive Fine Tuning (LAFT)

Language Adaptive Fine-Tuning (LAFT) is a process that involves fine-tuning a multilingual Pre-trained Language Model (PLM) on monolingual texts of a specific language using the pre-training objective. In this experiment, we focused on maternal health questions and utilized the Jacaranda Unlabelled Maternal Health dataset, which contained over 900,000 questions from mothers.

To adaptively fine-tune the model, we employed the Masked Language Objective, which is a form of auto-encoding language modeling. For each sentence question, 15% of the words were masked, and the model was trained to predict these masked words based on the remaining words in the sentence. By training the model using this objective, it learned certain statistical properties of the word sequences in maternal health questions.

We employed an intrinsic evaluation strategy to assess the quality of the trained model. Intrinsic evaluation compares the model's output to a predetermined reference text, known as the ground truth. Perplexity (PPL) metric was used to measure the uncertainty of the model's predictions. A low perplexity indicates higher confidence but does not necessarily guarantee accuracy. The trained Jacaranda Maternal Masked Language Model achieved a perplexity score of 3.623, while the original XLM Roberta Masked Language Model scored 109.097 on evaluation with the unlabelled dataset. This suggests that the trained model successfully learned the statistical properties of word sequences in maternal health questions.

The second step of the experiment involved fine-tuning the trained masked language model on a downstream task, specifically intent detection. The labeled Jacaranda Health dataset, which consisted of 107,000 examples, was used to fine-tune the model and predict different intents. The trained model achieved a balanced accuracy score of 0.76 and a weighted F1 score of 0.78. These metrics indicate the model's performance in accurately classifying different intents related to maternal health.

In summary, the experiment successfully applied Language Adaptive Fine-Tuning (LAFT) to a multilingual Pre-trained Language Model (PLM) for maternal health questions. By training the model using the Masked Language Objective and fine-tuning it on intent detection, we achieved promising results. The trained model exhibited a strong understanding of the statistical properties of maternal health question sequences, and it demonstrated good performance in accurately classifying different intents.

## LAFT + Hierarchical Clustering + Sample Weighting

We investigated different clustering techniques to identify common intents that could be grouped together at higher levels of the hierarchy and divided into more specific intents at lower levels. To achieve this, we calculated the degrees of confusion by creating a matrix that examined incorrect predictions for a given label and the labels it was frequently confused with. We also explored the computation of inter and intra-class distances using cosine similarity and Euclidean distance. Additionally, we examined the K-nearest neighbors (KNN) algorithm, using embeddings from the CLS token and the Faiss library for computing vector similarity. However, we found that extracting the embeddings was computationally expensive and time-consuming compared to using the incorrect predictions matrix to search for clusters.

By clustering the data, we were able to determine ways to collapse labels into higher-level root clusters and further break them down into more specific sub-levels. For example, all queries related to different types of pain, such as leg pain, back pain, stomach pain, butt pain, muscle joint pain, and breast pain, were initially clustered together under the broader category of "pain." We then trained another model to identify the specific type of pain within this cluster.

In total, our root model consisted of 58 intents, with 5 larger categories (Pain Cluster, Pregnancy General Cluster, Baby General Cluster, Diet Nutrition Cluster, and Pain Head Cluster). When broken down further, the total number of intents reached 176. Each model was trained separately, and the outputs of the root model were directed to the corresponding cluster model at level 2 of the hierarchy based on the predicted cluster. We trained the hierarchical model using the LAFT checkpoint, and we applied sample weighting to individual clusters.

The root model, with its 58 intents, achieved a precision of 87%, recall of 84%, an F1 score of 85%, and a balanced accuracy score of 88%. The "pregnancy\_general" intent showed a significant improvement, with the F1 score increasing from 0.69 to 0.83. Similarly, the "baby\_general" intent demonstrated a significant improvement of 24.56% compared to the best experiments in section 5.1, where XLM-Roberta was fine-tuned.

The pregnancy general model, consisting of 36 intents, achieved a precision of 87%, recall of 83%, F1 score of 84%, and balanced accuracy score of 85% when tested on 6643 previously unseen questions.The baby general model, comprising 40 intents, achieved a precision of 81%, recall of 78%, F1 score of 78%, and balanced accuracy score of 79% on a test set of 6083 unseen questions. The pain model, with 7 intents, achieved a precision of 91%, recall of 88%, F1 score of 88%, and balanced accuracy score of 86% when evaluated on 2211 unseen test questions. The diet nutrition model, encompassing 33 intents, achieved a precision of 90%, recall of 88%, F1 score of 88%, and balanced accuracy score of 88% on 3,745 previously unseen test questions. The pain\_head model, consisting of 2 intents, achieved a precision of 95%, recall of 94%, F1 score of 94%, and balanced accuracy score of 94% on a test set comprising 3,009 unseen questions.

# Implementation and Deployment:

The developed model has been deployed through our PROMPTS pipeline. The pipeline uses a micro-services architecture and within it we have 3 services; RapidPro Service, NLP Service, Ticketing Service. The pipeline is deployed on AWS (Amazon Web Services) and we use managed service (App Runner) to create instances of each service.

The different services communicate with each other using pub-sub messaging and we use Redis as the channel which is available on Amazon Elastic ache. Publish-subscribe messaging, or pub/sub messaging, is an asynchronous communication model that makes it easy for us to build highly functional and architecturally complex applications in the cloud. Pub/sub messaging provides instant event notifications for our decoupled components (our 3 microservices). Each service is connected to AWS Relational Database Storage for persistent storage of data using Object-Relational Mapping (ORM Framework) which provides a way to map database tables to classes, database rows to objects, and database operations to method calls, abstracting away the low-level SQL (Structured Query Language) queries. Our enrolled mothers interact with the pipeline through SMS which is routed through SMS gateway to the RapidPro Service. The messages are then published to the ticketing service and NLP service using pub-sub messaging.

Our NLP service uses the Amazon SageMaker Hugging Face Inference Toolkit. This is an open-source library for serving Transformers models on Amazon SageMaker. It provides default pre-processing, prediction, and post-processing for certain Transformers models and tasks. Since we want control on our pre-processing and post-processing, we override some of the default methods of the HuggingFaceHandlerService. Our NLP models are saved in S3 buckets and we create inference endpoints for pre-processing and post-processing. We also create our S3 bucket endpoint which creates a custom hugging face model class using the files zipped in the S3 bucket. We use this class to create and deploy our SageMaker endpoint. After the above setup, we dockerize the application and push it to the ECR (Amazon Elastic Container Registry) repository, where the main duty of our script is to deploy our docker image on EC2 and creating the prediction endpoint which we’ll invoke in the lambda function. AWS Lambda is a serverless, event-driven compute service that lets us run code for virtually any type of application or backend service without provisioning or managing servers. The central piece of our architecture is our Lambda function which is the bridge between our machine-learning model and our user-facing web application. The Lambda function receives text data (question) from the user, pre-processes it, passes it to our SageMaker endpoint, and then returns the predictions back to the user in a readable format. The client sends a post request that invokes the prediction endpoint in the lambda function, taking the input from the event and giving a response of the intent which is sent back to the client via the api-gateway.

The maternal query predicted intent and priority are then published to the ticketing service where the queries are queued according to the priority and assigned to our clinical help-desk for response. The responses are published through the RapidPro Service and routed to the mother via SMS gateway.

# Ethical and Regulatory Considerations:

All data collected from mothers through PROMPTS is private to Jacaranda Health, and we do not share this data with any external parties for any reason. In line with Kenya Data Protection Act (2019), we obtain consent to use the data via a consent message that the mother receives after enrolling to the platform. We do not collect any personally identifiable information except for phone numbers (which we also mask in our storage locations), and do not know our users’ names, identification numbers, addresses, or any other personal demographic data. All data is stored on secure cloud servers, with security measures in place to limit access and reduce risk of data exposure. As we expand into Ghana, and other SSA countries, we will continue to ensure compliance with national data regulations as a standard of practice.

# Collaboration and Partnerships:

This project depends on our deep ties to our 20 local government partners, who provide the broad/ sustainable distribution channel for PROMPTS to reach mothers across public facilities. PROMPTS has been endorsed for national scale by Kenya’s Ministry of Health, and we’ll ensure this journey is well-supported by local governments by providing reliable, real-time data to help them spend limited resources with greatest impact. We look for technical support from.

**– Google Country Offices**: We see exciting potential for on-the-ground collaboration (eg. language-specific ML support) with Google’s AI Center in Accra and Product Development Centre in Nairobi.

**– NVIDIA**: NVIDIA provides access to GPUs and ML toolkits, which power our ML models. We hope to leverage their technical knowledge around Large Language Models.

**– Microsoft Africa Research Institute (MARI)**: We currently consult with the principal researcher at MARI to improve our NLP model and will continue to seek their support as it evolves.

# Evaluation and Impact Assessment:

Our models are evaluated using both internal and external validation measures. The internal validation establishes the performance of our model using four metrics: precision, recall, F1 score and balanced accuracy score. The current model has an average balanced accuracy score of 88%, F1- Score of 85%, recall of 84% and precision of 87%. As for the impact, our A/B tests indicate that mums on PROMPTS are 3.51 times more likely to seek care for those danger signs than those who are not on PROMPTS and are 1.8 times more likely to take up family planning and long lasting contraceptives.

# Conclusion:

In conclusion, Jacaranda Health, a Non-Profit Organization based in Kenya, Ghana, and Eswatini, has developed PROMPTS, an SMS-based digital health platform, to address the high maternal mortality rates in Kenya. PROMPTS utilizes Natural Language Processing (NLP) techniques to handle maternal healthcare queries in low-resource settings. It has reached 2.2 million mothers across 1,110 Kenyan health facilities, providing free and easily accessible SMS-based support.

To improve the performance of NLP models for maternal health queries, we conducted several experiments. We evaluated the performance of pre-trained multilingual models through fine-tuning and contrastive learning. We also employed Language Adaptive Fine-Tuning (LAFT) to train models specifically on maternal health questions, achieving good results in intent detection. Additionally, they implemented a hierarchical clustering approach to group similar intents and achieve more accurate classifications.

The developed models achieved promising results, on the balanced accuracy scores and weighting F1 Score. By addressing the challenges of imbalanced distributions, degrees of confusion, and low-resource languages, our innovative approaches have significantly improved the triaging of high-volume maternal healthcare queries.

The implemented models have been deployed through our PROMPTS pipeline, using a micro-services architecture. The platform has demonstrated the potential to transform health-seeking behaviors among Kenyan women, empowering them with knowledge and guidance to make informed decisions about their health and the well-being of their newborns. These efforts offer hope in the battle against maternal mortality and serve as a valuable example of leveraging AI and NLP technologies for improving healthcare services in low-resource settings.

# References:

Guirong Bai, Shizhu He, Kang Liu, and Jun Zhao. Incremental intent detection for medical domain with contrast replay networks. In *Findings of the Association for Computational Linguistics: ACL 2022,* pp. 3549–3556, 2022.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm´an, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv*:1911.02116, 2019.

Tulu Tilahun Hailu, Junqing Yu, Tessfu Geteye Fantaye, et al. Intrinsic and extrinsic automatic evaluation strategies for paraphrase generation systems. *Journal of Computer and Communication*s, 8(02):1, 2020.

Gabriela Litre, Fabrice Hirsch, Patrick Caron, Alexander Andrason, Nathalie Bonnardel, Valerie Fointiat, Wilhelmina Onyothi Nekoto, Jade Abbott, Cristiana Dobre, Juliana Dalboni, et al. Participatory detection of language barriers towards multilingual sustainability (ies) in africa. *Sustainability*, 14(13):8133, 2022.

Hooman Sedghamiz, Shivam Raval, Enrico Santus, Tuka Alhanai, and Mohammad Ghassemi. Supcl-seq: Supervised contrastive learning for downstream optimized sequence representations. *arXiv preprint arXiv*:2109.07424, 2021.

Wilson Wongso, Henry Lucky, and Derwin Suhartono. Pre-trained transformer-based language models for sundanese. *Journal of Big Data*, 9(1):1–17, 2022.

Chaochen Wu, Guan Luo, Chao Guo, Yin Ren, Anni Zheng, and Cheng Yang. An attention-based multi-task model for named entity recognition and intent analysis of chinese online medical questions. *Journal of Biomedical Informatics*, 108:103511, 2020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv*:2010.11934, 2020.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_