DATA-EFFICIENT LEARNING FOR HEALTHCARE QUERIES IN LOW-RESOURCE AND CODE MIXED LANGUAGE SETTINGS

Stanslaus Mwongela, Jay Patel, Sathy Rajasekharan & Laura Wotton
Jacaranda Health
Nairobi, Kenya
{smwongela,jpatel,srajan,lwotton}@jacarandahealth.org

Mohamed Ahmed & Gilles Hacheme
Microsoft Africa Research Institute
Microsoft AI for Good
Nairobi, Kenya
{maxamed.axmed,ghacheme}@microsoft.com

Bernard Shibwabo & Julius Butime
School of Computer Science and Engineering Sciences
Strathmore Univesity
Nairobi, Kenya
{bshibwabo,jbutime}@strathmore.edu

ABSTRACT

The leading approaches in modern Natural Language Processing (NLP) are notoriously data-hungry. A good example is Transformer models, which achieve surging and state-of-the-art performance at the cost of big data. However, acquiring the big data needed is expensive and time-consuming for many application domains, limiting large adoption. Consequently, state-of-the-art NLP models perform poorly for low-resource languages such as African languages. Their performance is even worse when applied in sectors such as healthcare in low-resource settings. As a result, both academic and industrial communities are calling for more data-efficient models that use artificial learners but require less training data and less supervision. The current research aims to tackle these challenges by creating a data-efficient Transformer-based model for maternal queries intent detection in the settings of low-resource and code-mixed languages (Kiswahili, Sheng, and other local Kenyan Languages). Several experiments were carried out, including the use of pre-trained multilingual language models, language adaptive fine-tuning, supervised contrastive learning, and sample weighting in the loss function. The most efficient data-learner was obtained by Masked Language Model(MLM) adaptation on our unlabelled maternal queries and fine-tuning the adapted MLM checkpoint on our labelled training dataset. Sample weighting the loss function of the derived model increased the robustness and the overall performance of the model. The developed model was later deployed and is currently used to triage code-mixed maternal health queries at Jacaranda Health.

1 INTRODUCTION

The widespread adoption of digital and remote healthcare systems has revolutionized how patients access clinical information and support. 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 NLP today use Transformer architectures. 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).

Jacaranda Health is a non-profit organisation based in Kenya, Eswatini, and Ghana, whose mission is to improve maternal and newborn health outcomes in public health systems. The organisation 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 `PROMPTS` [1], 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 helpdesk 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 3000-4000 medical queries sent through the platform 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

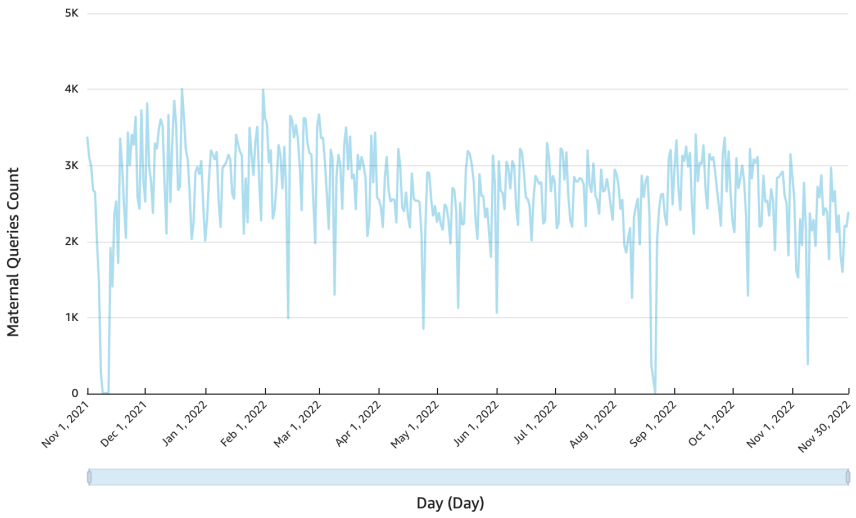

Figure 1: Trend of Maternal Health Queries Received via Prompts

The 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 well represented in existing state-of-the-art pre-trained NLP models. Additionally, the context of the text is maternal health discussions, which includes specialised vocabulary in local languages which is missing from standard NLP corpora. Furthermore, the maternal health queries distribution is highly imbalanced between different intents. $> 85\%$ of

[1] `https://www.jacarandahealth.org/prompts`

questions from mothers so far are regarding general queries on pregnancies. Thus, the goal of this research is to overcome these difficulties and create a model that can accurately predict and classify language in the context of maternal healthcare with limited data resources and small budgets for annotation.

In this work, we evaluate four strategies to build a data-efficient classifier for health queries intent detection at 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) and lastly language adaptive fine-tuning + downstream task fine-tuning with weighted loss. The last approach which involved first selecting a pre-trained Transformer multilingual model (XLM-Roberta) with Swahili checkpoint and adaptively fine-tuning it to our in-domain Jacaranda Health Maternal Queries worked best for our setting. In this paper, we presented a novel approach to fine-tuning pre-trained language models that addresses the challenges of class imbalance and limited labelled data in low-resource settings. Our approach, which involves adaptively fine-tuning on a large, unlabelled dataset and sample weighting in the loss function, showed significant improvement in performance compared to traditional fine-tuning methods. 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 class imbalance problem.

2 DATASET

Our dataset consists of queries posed by mothers to the PROMPTS system between July 2021 to July 2022 and is compliant with GDPR(General Data Protection Regulation) regulations of Kenya. The majority of the queries contained herein are unlabelled.

2.1 THE UNLABELLED DATASET

This dataset contains 939,819 questions sent through the PROMPTS SMS framework as asked by the mothers Jacaranda serves. These questions were asked between the months of July 2021 to July 2022. The unlabelled dataset includes original text as sent by the mother, AI-generated prediction, and Agent's response, among other attributes. To clean the data and prepare it for processing by our model, we first removed all patient identifying information, which included the phone number and mum's unique key identifier. Secondly, we dropped all the attributes, including the predicted label, to remain with only the original raw text. The training unlabelled dataset thus contained only one column, which was the original text. This original text contained low-resource language text, which was either in Swahili, Sheng (code mixed Swahili and English), and other local Kenyan languages.

2.2 THE LABELLED DATASET

The labelled dataset contains 107,714 questions asked by the mothers using our services and labelled by clinically trained personnel. The questions are in low-resource language text, which is either in Swahili, Sheng (code mixed Swahili and English), and other local Kenyan languages. The intent labels were 58 in total. The distribution of questions across the 58 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 Post Natal Care questions were classified as baby general. Figure 2 below illustrates the class distribution. The red bars denote danger signs intents.

3 METHODOLOGY

3.1 RESEARCH DESIGN

An experimental design approach was used in this work. Four experiments were proposed to solve the underlying problem.

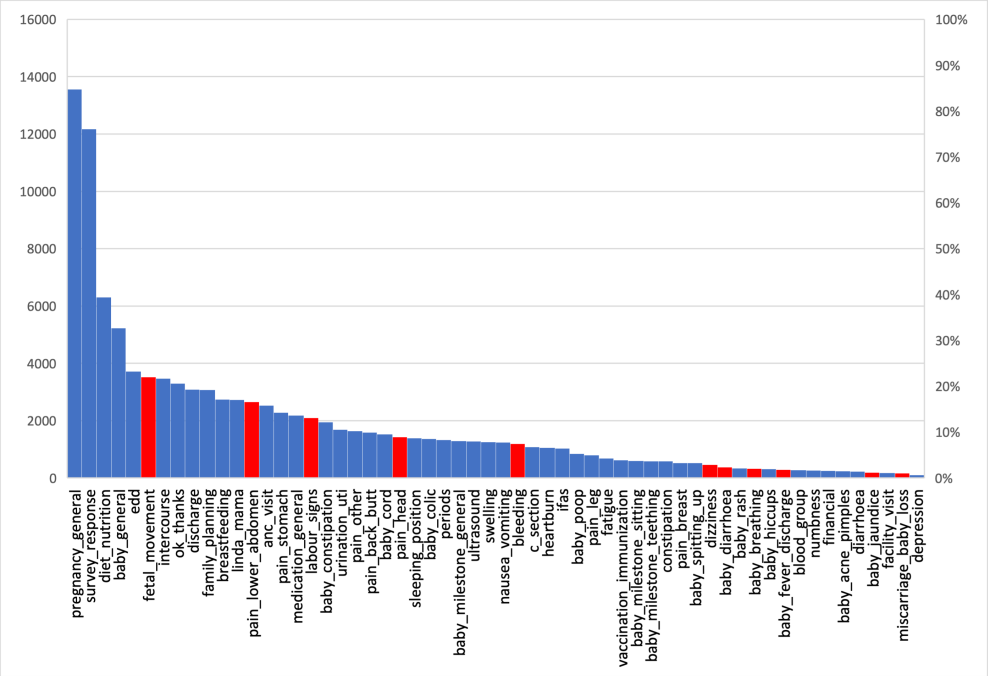

Figure 2: Labelled Data Class Distribution

3.2 Experiment 1: Fine-tuning using Pre-trained Transformer Multilingual Models

In this experiment, we selected 3 state-of-the-art Pre-trained Transformer Multilingual Models for a fine-tuning evaluation using the current 107k Jacaranda Health Labelled maternal queries training dataset. The selected models included XLM-Roberta-Large, MT5 and Afro-XLMR-Large.

3.2.1 Finetuning Using MT5EncoderModel

We created an experiment to fine-tune mT5 using our training dataset. The experiment was implemented using python 3.8, Pytorch, and Hugging Face libraries. MT5 is a multilingual variant of T5 that was pre-trained on a new Common Crawl-based dataset covering 101 languages (Xue et al., 2020). The dataset was first cleaned by dropping all the null values. The second step of cleaning involved dropping all other columns with the exception of the maternal query and the intent column. The dataset was then split to train and test using a stratified split-train-test strategy at a ratio of 80:20. The text rows and the intents were then tokenized using the T5 tokenizer and prepared for feeding into the model. Since the task was sequence classification, we chose the MT5EncoderModel and added a classification head on the CLS embedding. The finetuned MT5 achieved a balanced accuracy score of 0.61 and a weighted f1-score of 0.71.

3.2.2 Finetuning using XLM-Roberta-Large

XLM-ROBERTa is a multilingual version of RoBERTa. It is pre-trained on 2.5TB of filtered Common Crawl data containing 100 languages. It was introduced in the paper Unsupervised Cross-lingual Representation Learning at Scale by Conneau et al. (2019). The current model in use at Jacaranda Health is XLM-Roberta fine-tuned on the 107k+ labelled Jacaranda Health Care dataset by researchers from Penn State University and Jacaranda Health. The fine-tuned model was evaluated for precision, recall, F1 score, and balanced accuracy score The finetuned XLM-Roberta trained achieved a balanced accuracy score of 0.76 and a weighted f1 score of 0.76.

3.2.3 FINE-TUNING USING AFRO-XLMR LARGE

AfroXLMR-Large was created by MLM adaptation of XLM-R-large model on 17 African languages (Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Oromo, Naija, Kinyarwanda, Kirundi, Shona, Somali, Sesotho, Swahili, isiXhosa, Yoruba, and isiZulu) covering the major African language families and 3 high resource languages (Arabic, French, and English) Alabi et al. (2022). An experiment was created to finetune Afro-XLMR on our labelled training dataset. Similar data preprocessing and training were used as in the previous experiment. The finetuned XLM-Roberta achieved a balanced accuracy score of 0.75 and a weighted f1 score of 0.77.

3.3 EXPERIMENT 2: SUPERVISED CONTRASTIVE LEARNING USING XLM-ROBERTA

An experiment was created using a supervised contrastive learning objective, and an augmentation objective and implemented using a pre-trained multi-lingual Transformer model. The experiment is inspired by the paper, SupCL-Seq: Supervised Contrastive Learning for Downstream Optimized Sequence Representations by Sedghamiz et al. (2021). Given an anchor representation, a contrastive learner was trained to generate augmented altered views by altering the dropout mask probability in a standard Transformer architecture. A supervised contrastive loss was then utilized to maximize the model's capability of pulling together similar samples (e.g. anchors and their altered views) and pushing apart the samples belonging to the other classes. The loss function utilized was:

$$l_i^{\text{sup}} = \sum_{i \in \boldsymbol{I}} \frac{-1}{|P(\text{i})|}$$

$$\sum_{p \in P(\text{i})} log \sum_{b \in B(\text{i})} \frac{e^{sim(\tilde{x_i}, \tilde{x_p})/\tau}}{e^{sim(\tilde{x_i}, \tilde{x_b})/\tau}}$$

Where,$B(i) \equiv I\{i\}$ is the set of so-called negative pairs for the anchor $\tilde{x_i}$, τ is a temperature scaling parameter $sim(.)$ stands for any similarity function such as cosine similarity or the inner product.$P(i) \equiv p \in B(i) : \tilde{y_p} = \tilde{y_i}$is the positive pair set distinct from sample i and $|.|$ stands for cardinality.

After training for 10 epochs, the best checkpoint was extracted. This checkpoint had a supervised contrastive loss of 0.65. The checkpoint base model parameters were then frozen, and a sequence classification head was added on top. This was then used for inference and results obtained. With a contrastive loss of 0.65, the contrastively trained XLM-Roberta model achieved a Weighted F1 score of 0.75 and a balanced accuracy score of 0.72.

3.4 EXPERIMENT 3: LANGUAGE ADAPTIVE FINE-TUNED(LAFT) XLM-ROBERTA

Language adaptive fine-tuning (LAFT) involves fine-tuning a multilingual PLM on monolingual texts of a language using the pre-training objective. We extracted from the repository all the questions (900K+) mothers have asked and selected them without the labels. Using a Masked Language Objective, we adaptively fine-tuned an XLMRobertaMaskedLanguageModel on the Jacaranda Unlabelled Maternal health dataset. Masked Language Modelling is an example of auto-encoding language modelling (where an output is reconstructed from the corrupted input). In this experiment, we masked 15% of the words in each sentence question and trained our model to predict those masked words given the other words in the sentence question. By training using this objective, our model essentially learned certain (but not all) statistical properties of the word sequences for maternal health questions we receive from mums. An intrinsic evaluation strategy was used to evaluate our model. An intrinsic evaluation evaluates the quality of the Natural Language processing against some predetermined ground truth referred to as reference text (Hailu et al., 2020). Specifically, this work used the Perplexity (PPL) metric. This metric is used to quantify the uncertainty of our model in making predictions. A low perplexity only guarantees the confidence of the model, not accuracy.Given a tokenized sequence $x = (x_0, x_1, ..., x_t)$, perplexity is calculated as:

$$PPL(X) = exp\ \{- \sum\nolimits_{\text{i}}^{\text{t}} \log \text{p}\theta(\text{x}_0|\text{x}_{<\text{i}})\ \}$$

Table 1: Test results

Model	Precision	Recall	F1 Score	Balanced Accuracy Score
Experiment 1: Multilingual Models				
Finetuned MT5	0.73	0.71	0.71	0.60
Finetuned XLM-Roberta Large	0.76	0.76	0.76	0.73
Finetuned AFRO-XLMR	0.76	0.76	0.76	0.74
Experiment 2: Contrastive Learning				
Supervised Contrastive-XLMR	0.75	0.75	0.75	0.72
Experiment 3: Adaptive Finetuning				
Language Adaptive Fine tuned-XLMR	0.78	0.78	0.78	0.76
Experiment 4: Best Model Sample Weighting				
LAFT-XLMR+Sample Weighted Loss	0.78	0.78	0.78	0.80

Where; $log\ p\theta(x_0|x_{<\mathrm{i}})$ is the log-likelihood of the nth token conditioned on the preceding tokens $x_{<\mathrm{i}}$ according to our model. Our trained Jacaranda Maternal Masked Language Model achieved a perplexity score of 3.623, while the original XLMRobertaMaskedLanguageModel on evaluation on our unlabelled dataset had a perplexity score of 109.097. This means that our trained Masked Language Model significantly learned the statistical properties of the word sequences for maternal health questions we receive from mothers. The second step in the experiment involved fine-tuning our trained masked language model on downstream task (in our case, intent detection). The labelled Jacaranda Health dataset (107k) was fed to our masked language model with the aim of fine-tuning our masked language model to predict the different intents. The trained model achieved a balanced accuracy score of 0.76, and a weighted F1 score of 0.78.

3.5 Experiment 4: LAFT-XLMR with Sample Weighted Loss

We extracted the LAFT checkpoint in experiment 3 and used it to fine-tune on the labelled Jacaranda dataset.Similar steps as in experiment 3 were implemented with the modification of the loss to include sample weighting. The sample weighting strategy used was the Inverse of Number of Samples (INS). The samples were weighted as the inverse of the class frequency for the class they belonged to. This experiment led to the best outcomes , the balanced accuracy score improved to 0.80 while the weighted F1 score was at 0.78.

4 Results

The labelled stratified split-test dataset (21,000 medical queries) was used to evaluate the performance of the developed models. The evaluation was done across four performance metrics; weighted precision, weighted recall , weighted F1 score and weighted balanced accuracy score. Experiment 4 (LAFT-XLMR+Sample Weighted Loss)had the best outcomes. Table 1 shows the performance of the different models across the four metrics.

5 Discussion and Conclusions

This study sought to answer the question: how do we design a data-efficient Transformer model for the task of medical intent detection given: code mixed and low resource languages(Swahili,Sheng, Luo, Kikuyu and Luhya), highly imbalanced dataset, weak labelling and need for in domain cross-lingual transfer to other low resource languages? The model was developed to tackle the challenges

of imbalanced datasets, weak labelling, and the need for in-domain cross-lingual transfer. By implementing MLM adaptation and fine-tuning techniques, we achieved remarkable improvements in performance (+7 points over the baseline current model in use). Our results showcase the significance of taking into account the unique linguistic features and data constraints while developing machine learning models for medical intent detection. The MLM adaptation of our model led to learning better representations of the queries we receive from mothers enrolled in the PROMPTS framework. In essence,the trained MLM model acquired some certain statistical understanding of maternal health care language as used by mothers who send questions through the PROMPTS framework. The study emphasizes the effectiveness of sample weighting in resolving data imbalance. Sample weighting in the loss function improved the F1 score in 45 out of 58 classes, making it a more efficient and effective solution compared to under-sampling and over-sampling. Further, the study points out the possibilities for future improvements, such as incorporating active learning, reducing bias in the training dataset, exploring more contrastive learning approaches, and enhancing the model's cross-lingual transfer capabilities.

6 CONCLUSIONS

The current improvements to the model have improved its accuracy and efficiency at triaging mother's questions at scale - with a small technology team and at a 80% lower cost than our baseline model. Specifically, this has achieved:

- High performance at scale: Our NLP model triages 4,000+ messages daily in code-mixed Kiswahili English.
- Rapid detection/ referral of urgent cases: Mothers flagged by the helpdesk with a danger sign (eg. heavy bleeding) receive a response in < 1 minute before connecting with a human agent.
- Hospital follow-through; 90% of mothers flagged by the helpdesk report receipt of care.
- Data for decision-making: PROMPTS data (eg. feedback on care quality, Maternal Health issues) is informing how local governments direct limited resources.

The outcomes of this study hold the potential to propel the development of efficient solutions for maternal health care delivery and communication, especially in low-resource language settings, and provide a new approach for developing models for medical intent detection.

ACKNOWLEDGMENTS

We would like to acknowledge the team from Penn State University and the AI for Social Good Grant for the support in creating the initial model, Microsoft Africa Research Institute for the knowledge exchange and transfer sessions which led to the creation of the new model, and the clinical help desk team at Jacaranda Health for all the hours they put in generating annotations for the training data.

REFERENCES

Jesujoba O Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. Adapting pre-trained language models to african languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 4336–4349, 2022.

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 Guzmán, 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 Communications*, 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.