Detecting Urgency in Multilingual Medical SMS in Kenya

Narshion Ngao[◇] **Zeyu Wang**[†] **Lawrence Nderu**[◇]
Tobias Mwalili[◇] **Tal August**[†] **Keshet Ronen**[†]
[†]University of Washington, Seattle, WA, USA
[◇]Jomo Kenyatta University of Agriculture & Technology, Nairobi, Kenya

Abstract

Access to mobile phones in many low- and middle-income countries has increased exponentially over the last 20 years, providing an opportunity to connect patients with healthcare interventions through mobile phones (known as mobile health). A barrier to large-scale implementation of interactive mobile health interventions is the human effort needed to manage participant messages. In this study, we explore the use of natural language processing to improve healthcare workers' management of messages from pregnant and postpartum women in Kenya. Using multilingual, low-resource language text messages from the Mobile solutions for Women and Children's health (Mobile WACh NEO) study, we developed models to assess urgency of incoming messages. We evaluated models using a novel approach that focuses on clinical usefulness in either triaging or prioritizing messages. Our best-performing models did not reach the threshold for clinical usefulness we set, but have the potential to improve nurse workflow and responsiveness to urgent messages.

1 Introduction

In many low- and middle-income countries, access to healthcare is limited and unaffordable. Interactive short message service (SMS) communication with healthcare workers has shown great potential to promote access to care in such contexts by providing remote information and support (Hall et al., 2015; Rono et al., 2021).

One such system is the Mobile solutions for Women and Children's health (Mobile WACh) platform, an interactive semi-automated platform designed to connect pregnant and postpartum women to healthcare workers through SMS (Perrier et al., 2015; Unger et al., 2019, 2018; Harrington et al., 2019; Kinuthia et al., 2021; Ronen et al., 2021). Studies using this platform have reported significant impacts on health outcomes like breastfeeding and postpartum contraception (Unger et al., 2018; Harrington et al., 2019).

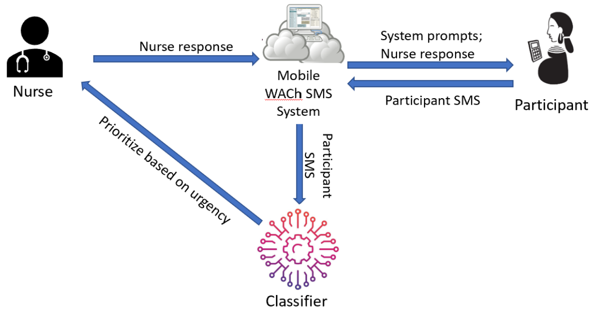

Figure 1: Task Definition Workflow

While promising, a major limitation of mobile health interventions is the human effort required to manage messages. Nurses involved in the Mobile WACh platform received hundreds of messages per day (Unger et al., 2019). Many of these messages did not require immediate responses; however, nurses couldn't distinguish between urgent and non-urgent messages without reading them.

Natural language processing (NLP) can potentially be used to automatically triage and prioritize incoming messages. This may significantly improve worker efficiency and improve reliability of the healthcare system in low-resource settings (Rono et al., 2021; Barron et al., 2017). In recent years, researchers have used NLP for content analysis of incoming messages in digital health interventions (Schwab-Reese et al., 2019; Klimis et al., 2021) and for analyzing messages in mental health discussions (Zhang and Danescu-Niculescu-Mizil, 2020; Althoff et al., 2016).

This paper explores the possibility of a clinically useful model to detect urgent participant messages in an interactive mobile health system in Kenya. Our study focuses on a dataset drawn from the Mobile WACh NEO studies (Unger et al., 2019; Ronen et al., 2021). The dataset contains real-world, informal messages in multiple low-resource languages

Proceedings of the AACL-IJCNLP 2022 Student Research Workshop, pages 68–75
November 20, 2022. ©2022 Association for Computational Linguistics

(Swahili, Luo and Sheng) and English. We explore methods for handling the unique challenges presented in our dataset, including additional pretraining (Gururangan et al., 2020) and adding prior conversation context. We tested several approaches to classifying urgency in labeled participant messages, with a focus on classifiers that could have clinical utility in improving healthcare worker workflow. Models based on mBERT can achieve performance levels close to our threshold of clinical usefulness, suggesting such systems could be useful to healthcare workers in the future. We discuss our findings and next steps for integrating such NLP models into real-world systems that could significantly improve global healthcare delivery.

2 Task Definition

We aim to introduce models that classify messages based on urgency as a way of prioritizing or triaging messages for nurses. Specifically, given a message, our task is to identify if the message requires immediate nurse attention (urgent) or can be looked at later (non-urgent). Figure 1 summarizes the Mobile WACh system and NLP task. Because of the sensitivity of all participant messages in this context, our task is not intended to replace nurses by filtering participant messages or generating responses.

3 Dataset

Our data consists of messages, from the Mobile WACh NEO pilot (messages sent between 05-12-2017 and 20-02-2019) (Unger et al., 2019) and Mobile WACh NEO RCT (messages sent between 09-09-2020 and 04-05-2022) studies (Ronen et al., 2021). Messages were exchanged between pregnant/postpartum women, nurses and the automated Mobile WACh system. The Mobile WACh NEO pilot dataset consists of a total of 58,834 messages that were exchanged between 800 participants, the automated system and 2 nurses. The Mobile WACh NEO RCT had a total of 161,735 messages that were exchanged between 1,724 participants, the automated system and 12 nurses. Therefore, the combined dataset consisted of 220,560 messages from 2,523 participants and 14 nurses (after cleaning). Automated messages were sent to participants weekly during pregnancy until 38 weeks gestation, then 2 messages daily for the first 2 weeks after delivery, and then every 2 days for 6 weeks follow-up post delivery. Participants could send messages to the system at any time. Nurses in the study manually replied to participant messages. These nurses had the same training and qualification as nurses in the public health facilities, however, they were employed by the study and did not have routine care provision responsibilities outside of study. A total of 112,220 (50.9%) messages were sent by the Mobile WACh system, 65,572 (29.7%) by participants and 42,768 (19.4%) by nurses (Table 1). Automated system messages were sent in English, Swahili (a Bantu language) or Luo (a Nilotic language) based on each participant's preference. Participant messages were sent in the participant's language of choice; about half (50.4%) were in English, 36.8% were in Swahili, 5.4% were in Luo, 4.5% were code-switched, and 2.9% were in a slang fusion known as Sheng (Table 2). To clean the dataset of any identifiable information, we removed standard salutation, and any location, nurse, or participant names. Automated messages used to validate participant registration in the SMS system were also removed. The total number of messages described here were the final dataset after the cleaning exercise.

Table 1: Messages By Source. Around a third of participant messages were less than 10 characters, suggesting many participant messages were short and depend on previous message context for detecting urgency.

Sent By	Total Messages	Messages with less than 10 characters	Mean number of characters in a message (std)
nurse	42768 (19.4%)	2500	97.9 (103.5)
participant	65572 (29.7%)	19769	36.5 (39.8)
system	112220 (50.9%)	0	257.3 (102.7)

The dataset we present here is typical of how language is used in Kenya (Bosire, 2006; Mondal et al., 2021). For instance, Swahili words used by participants in Nairobi may have different connotations from the same word in standard Swahili or Swahili used in Western Kenya. It is worth noting that this dataset also contains languages (Sheng and Luo) not commonly included in training for multilingual transformer-based models like mBERT (Devlin et al., 2019). Table 2 illustrates the breakdown of participant messages by language.

3.1 Urgency Labelling

Two nurses at the study clinics labelled a total of 11,129 messages from 772 participants. Of these, 30 participants were selected from the Mo-

Table 2: Labelled Participant Messages by language. While the majority of messages are in English, Swahili and Luo make up more than 40% of the total messages. Note that the total number of labelled messages was 11129.

Language	Total Messages	Percentage
english	5646	50.7%
swahili	3893	35.0%
sheng	572	5.1%
luo	566	5.1%
Code-Switched	452	4.1%
TOTAL	65572	100%

bile WACh NEO pilot study and had a total of 1,477 messages. The remaining 742 participants were from the Mobile WACh RCT study. Nurses labelled urgency based on how quickly a given participant message should be replied to by a nurse: 1) immediately, 2) within 2 hours, 3) before end of work day 4) by tomorrow 5) no need to reply. Nurses were instructed to use information from prior messages to inform assessment of the urgency of a given participant message. A sampled agreement between two raters had a Cohen Kappa score of 0.75, indicating high agreement. From the labelled data, we split the 5 urgency categories into a binary label of urgent (categories 1 and 2) and not urgent (categories 3, 4 and 5). The distribution of urgency labels was imbalanced (2,383 out of 11,129 were labelled as urgent, 21.4% of all labelled messages). This data represents the reality that in the context of the Mobile WACh studies, most messages received were not urgent. Because we are interested in a model than can eventually be useful in this real-world context, we leave the data imbalanced. Finally, the data was split into training 7,790 (70%), test 2,337(20%), and validation 1,002(10%) sets, having been stratified by label.

4 Classifying Urgency

We defined the task of predicting message urgency as a binary classification task. We tested two modeling approaches. Our first approach was a penalized logistic regression (penalty: l2, maximum iterations: 570) with bigram lexical features as input. The bigram features were extracted from uni- and bi-grams from the messages using Scikit-learn's count vectorizer (Pedregosa et al., 2011). In our second approach, we evaluated a fine-tuned multilingual BERT model (mBERT) (Devlin et al., 2019). mBERT was pretrained in 104 languages including English and Swahili. Linear models have been used as a baseline in mobile health classification tasks (Losada et al., 2020), and mBERT is a strong multi-lingual text classification model.

4.1 Adding Context

We observed that many participant messages are short (Table 1) and messages like "okay", "thank you", "no", and "yes" can have different meanings depending on the context of the conversation. Past work has found that including prior message context when analyzing SMS messages can be helpful for understanding conversation trajectory (Althoff et al., 2016) and appropriate responses (Zhang and Danescu-Niculescu-Mizil, 2020). We took inspiration from this work and evaluated whether adding preceding message context to participant messages would improve model performance.

We represented context by prepending the message preceding a participant message. We developed two versions of the dataset: one in which each participant message was prepended with the preceding system message (system context) and one in which each message was prepended with the preceding nurse message, or, in the event there was no nurse message, then the most recent system message (nurse context). Example messages are displayed in Table 3. We compared results for both the logistic regression with bigram features and mBERT using these approaches.

4.2 Additional Pretraining

It has been shown that additional in-domain and task-adaptive pretraining can improve model performance in a variety of settings (Gururangan et al., 2020). Since our dataset differs from the languages and domains used to pretrain mBERT, we reasoned this may be particularly impactful in our task.

We explored two versions of pretraining. In the first approach, we pretrained on all 49,786 participant messages that were not in the test or validation sets (this included both labelled and unlabelled data). Similar to our approach for fine-tuning data, we tested pretraining with participant messages that were prepended with system messages or nurse messages. In the second approach, we used the 11,129 labeled (both urgent and non-urgent) participant messages that were also prepended with system messages or nurse messages (Table 3). Note that in this second approach, we did not include the

Table 3: Sample messages with contexts

System Message	Nurse Message	Participant Message	Urgency Label
Make sure you come in for antenatal care even at the end of your pregnancy We check for any problems and help you prepare a birth plan Do you have any questions or concerns Are you feeling the baby move often	Am glad your OK have a nice day	You To	0
We are checking to see how you are doing How is your bleeding Do you have any pain in your lower abdomen Any fevers Please let us know if you feel unwell	Are you still having the headache	yeah	1
Regular strong stomach pains are a sign of labour If you feel this strong tightening regularly pains leaking of fluid or bleeding go to the facility Do you feel any contractions Do you have any concerns	Hello That is fine Please avoid strenuous activities at this point in your pregnancy	Its OK I willthanks for your concerns	0
Newborns sleep a lot but wont stay asleep for more than 24 hours at a time You may still be up several times at night to change feed and comfort your baby Take naps with your baby and try and interact with your baby during the day and keep things dark and quiet at night How is the baby sleeping	Hello there is no problem with topping up for the baby if the baby not satisfied Where are you getting the milk to top up What do you mean by yellow skin and which treatment is this for yellow skin that you are referring to	My baby had jodesyellow skin colour and was put on photo therapynow am asking can the baby suffer from the same problem a gain the Normal skin colour of the baby turning to yellow…	1

labels, only the text of the messages, for pretraining. We explore this method for mBERT.

We present the models in which pretraining and fine-tuning data are matched in terms of the context used (i.e., system, nurse, or no context) since we observed that this led to the largest increases in performance. We pretrained with masked language modelling with 15% of the text masked and used a batch size of 4, with a maximum input sequence size of 512. During fine tuning the models, we used default hyper parameters apart from batch size which was 16. The default parameters can be found at (huggingface.co).

5 Evaluation

While F1 score is a common classification evaluation metric, clinically useful systems may not require both high recall and high precision to improve healthcare worker workflow. We visualize the trade offs of models' precision and recall with precision-recall curves. We can use these PR curves to pick potential models that could be clinically useful by defining two regions in the graph: a triage region and a prioritize region.

A precision-recall curve shows the trade offs of models' precision and recall across a range of classification thresholds. This allows us to visualization of trade-offs of the results of the system between high precision and high recall. We can use models' precision-recall curves to visualization of trade-offs of high precision and high recall and pick potential models that could be clinically useful by defining two regions in the graph: one region for a triage model and the other for a prioritize model (see Figure 2).

We defined three potential model use cases and their evaluation criteria: 1) triage, 2) prioritize, or 3) combination. Most machine learning models for binary classification output a real number between 0 and 1 and use 0.5 as the default threshold for classification. While in most scenarios this threshold is sufficient, in our case it is helpful to examine the model performance with a range of thresholds which might be better suited for different scenarios (e.g., triaging or prioritizing messages).

Below we define the evaluations for these regions (§5.2 & 5.1) and their combination (§5.3).

5.1 Triage

An ideal triage model is aimed at reducing the number of messages that the healthcare staff need to read by ruling out messages that do not indicate urgency. A triage model needs to be able to reduce message volume enough to justify its implementation costs (e.g., debugging or training nurses to use the system) while also ensuring a minimal number of false negatives. Within the Mobile WACh studies, we choose the threshold of 30%. This means that the model should assign non-urgent (negative) to at least 30% of the messages while maintaining near-perfect recall. Knowing the number of samples in the dataset and the number of actual positive labels, we can get the relationship between precision and recall:

$$precision_{triage} \geq \frac{recall_{triage} \cdot actualpositives}{datasize \cdot 70\%}$$

We can take a high value for recall$_{triage}$ (95% in our case) and calculate a threshold for precision$_{triage}$. This creates a region in the precision-recall graph that a triage model's

Table 4: Performance of bigram and mBERT models with varying context. Bolded text indicates best-performing model in terms of F1.

Model	Pre Data	Pre Context	FT Context	Precision	Recall	F1
Bigrams	-	-	none	51	20	29
Bigrams	-	-	system	58	29	39
Bigrams	-	-	nurse	59	29	39
mBERT	-	-	none	46	34	39
mBERT	-	-	system	50	27	35
mBERT	-	-	nurse	52	38	44
mBERT	labelled	system	system	50	32	39
mBERT	**labelled**	**nurse**	**nurse**	**50**	**45**	**47**
mBERT	unlabelled	system	system	49	39	44
mBERT	unlabelled	nurse	nurse	48	38	42

precision-recall curve crosses (Figure 2).

5.2 Prioritize

An ideal prioritize model should identify urgent messages that should be replied to more quickly than other messages. This approach is helpful when the healthcare staff need guidance on which messages to read first. Since all the messages will eventually be reviewed, the focus is not on reducing false negatives, but false positives, since this will determine the trust of the healthcare staff in the system. This model should have a high precision and maintain a significant number of positive cases. We decide on a threshold of 10% here. This means the model should predict a message as urgent at least 10% of the time while maintaining a near-perfect precision. Similar to the triage region, we can calculate the relationship between precision and recall for a prioritize model:

$$recall_{prioritize} \geq \frac{precision_{prioritize} \cdot datasize \cdot 10\%}{actualpositives}$$

We can take a high precision$_{prioritize}$ (95% in our case) and calculate a threshold for recall$_{prioritize}$. This creates a region on the graph that a prioritize model's precision-recall curve crosses (Figure 2).

5.3 Combination

A combination model is one that is able to meet the targets of both triage and prioritize models. The model should have a high F1 score. When a model's precision-recall curve cuts across the overlapping region between the triage and prioritize regions, the model is a combination model.

6 Results

Table 4 summarizes the performance of our models. The best performing bigram model used participant messages with nurse context, outperforming the bigram model with system context by one precision point. The mBERT model using nurse context achieved an F1 score of 44, with 52 precision and 38 recall. While the precision is worse than the bigram model, the recall is higher by 8 points. The models incorporating nurse context messages performed better than system context. The effect of incorporating context on the performance of mBERT models is summarized in Table 5.

Table 5: Effect on mBERT model performance of prepending messages with context and additional pretraining. Pretraining here refers to pretraining on the labelled data with matched context (i.e., System + pretraining is pretraining on participant messages prepended with system messages, using labeled data only). Baseline model was with no pretraining and no context added to the messages.

Metric	Baseline	System	Nurse	Nurse + pretraining	System + pretraining
Precision	46	50 (+4)	52 (+6)	50 (+4)	50 (+4)
Recall	34	27 (-7)	38 (+4)	45 (+11)	32 (-2)
F1	39	35 (-4)	44 (+5)	47 (+8)	39 (+0)

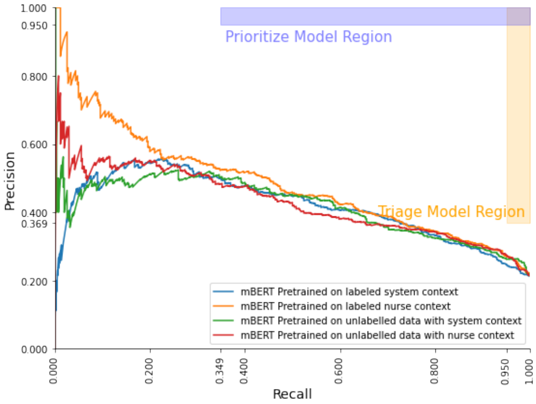

Figure 2: Performance of mBERT models with additional pre-training on varying data type

6.1 Additional Pretraining

Recall improved when the models were pretrained with nurse context messages. For system context models, only pretraining with unlabelled data increased recall. Table 5 details these results. The highest performing model was mBERT with nurse pretraining on labelled data. Figure 2 presents precision-recall curves for the above pretrained models. We found that the model pretrained and finetuned on labelled nurse context messages was the best model overall, though it did not pass into either the prioritize or triage region.

7 Related Work

NLP techniques for information extraction have been used in several SMS based mHealth applications. For example, Gupta et al. (2020) developed a virtual assistant health coach using text messages in English to help patients set physical activity goals. Lowres et al. (2020) developed NLP models to triage incoming English SMS text messages to reduce the burden of healthcare worker review. Fewer studies have applied these methods to low-resource languages and multilingual datasets. The MomConnect program from South Africa's National Department of Health is one such application where Engelhard et al. (2018) used multilingual data from the project to perform a feasibility study on triaging incoming messages of pregnant clients. Using data from the same program, Daniel et al. (2019) created an automated multilingual digital helpdesk service. Daniel et al. (2019) reported the challenges of this dataset as being multilingual, in low-resource languages, and with high prevalence of code-switching, spelling errors and abbreviations. Like the MomConnect data, Mobile WACh messages are in multiple low-resource languages, with code-switching, misspellings and abbreviations.

8 Discussion & Conclusion

Consistent with prior literature (Gururangan et al., 2020), our results showed that performing additional pretraining boosts performance. Our evaluations show that our modeling approaches have the potential to support healthcare workers in a unique low-resource and multilingual setting, though more work must be done to have the models achieve clinical usefulness based on our measures. Moving forward, to improve performance of these models, future studies could look into how to optimize the models when the dataset is skewed for non-urgent messages (as is the case currently). Another approach would be to explore models explicitly trained on the languages in our dataset, for example, models trained on Swahili datasets or code-switched languages from East Africa (Ogueji et al., 2021) and (flax community). We also plan to validate our highest-performing models with healthcare workers and implement a model in a pilot context similar to the Mobile WACh SMS system.

9 Ethical Considerations

The context of this study requires careful attention to preserving patient anonymity and the potential for unforeseen consequences. The Mobile WACh NEO pilot and RCT studies were approved by our institution's ethics and review board. All participants provided written informed consent for participation in the studies, including participation in the SMS intervention and use of data for secondary analyses. All patient data were made anonymous for our analyses. Because of the sensitive nature of the messages, the dataset will not be made publicly available, though researchers are welcome to contact the Mobile WACh study for anonymized data.

Deploying any system for triaging or prioritizing patient messages also must be piloted in real-world settings. While our analyses suggest that such systems are possible, ensuring that patient messages are not mislabeled is paramount. A single urgent message mislabelled could be catastrophic for a patient. Our evaluations aim to capture such considerations, but additional safegaurds are necessary. For example, all messages should be reviewed by nurses within a day regardless of model predictions, or patients should have a way of overriding model predictions if they have an urgent issue. We are excited to pilot models in real-world settings to see how they can support mobile health interventions.

References

Tim Althoff, Kevin Clark, and Jure Leskovec. 2016. Large-scale analysis of counseling conversations: An application of natural language processing to mental health. *Transactions of the Association for Computational Linguistics*, 4:463–476.

Peter Barron, Joanne Peter, Amnesty E. LeFevre, Jane Sebidi, Marcha Bekker, Robert Allen, Annie Neo Parsons, Peter Benjamin, and Yogan Pillay. 2017. Mobile health messaging service and helpdesk for south african mothers (momconnect): history, successes and challenges. *BMJ Glob Health*.

Mokaya Bosire. 2006. Hybrid languages: The case of sheng.

J. E. Daniel, Willie Brink, Ryan Eloff, and Charles Copley. 2019. Towards automating healthcare question answering in a noisy multilingual low-resource setting. In *ACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages

4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matt Engelhard, Charles Copley, Jacqui Watson, Yogan Pillay, Peter Barron, and Amnesty E Lefevre. 2018. Optimising mhealth helpdesk responsiveness in south africa: towards automated message triage. *BMJ Global Health*, 3.

flax community. Gpt2 swahili.

Itika Gupta, Barbara Maria Di Eugenio, Brian D. Ziebart, Aiswarya Baiju, Bing Liu, Ben S. Gerber, Lisa Kay Sharp, Nadia Nabulsi, and Mary H. Smart. 2020. Human-human health coaching via text messages: Corpus, annotation, and analysis. In *SIGDIAL*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Amanda K. Hall, Heather J. Cole-Lewis, and Jay M. Bernhardt. 2015. Mobile text messaging for health: a systematic review of reviews. *Annual review of public health*, 36:393–415.

Elizabeth K. Harrington, Alison L. Drake, Daniel Matemo, Keshet Ronen, Alfred Osoti, Grace C John-Stewart, John Kinuthia, and Jennifer A. Unger. 2019. An mhealth sms intervention on postpartum contraceptive use among women and couples in kenya: A randomized controlled trial. *American Journal of Public Health*, 109:934–941.

huggingface.co. multilingual bert training arguments.

John Kinuthia, Keshet Ronen, Jennifer A. Unger, Wenwen Jiang, Daniel Matemo, Trevor Perrier, Lusi Osborn, Bhavna H Chohan, Alison L. Drake, Barbra A. Richardson, and Grace C John-Stewart. 2021. Sms messaging to improve retention and viral suppression in prevention of mother-to-child hiv transmission (pmtct) programs in kenya: A 3-arm randomized clinical trial. *PLoS Medicine*, 18.

Harry Klimis, Joel Nothman, Di Lu, Chao Sun, N Wah Cheung, Julie Redfern, Aravinda Thiagalingam, and Clara K Chow. 2021. Text message analysis using machine learning to assess predictors of engagement with mobile health chronic disease prevention programs: Content analysis. *JMIR Mhealth Uhealth*, 9(11):e27779.

David E. Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 272–287, Cham. Springer International Publishing.

Nicole Lowres, Andrew Duckworth, Julie Redfern, Aravinda Thiagalingam, and Clara K. Chow. 2020. Use of a machine learning program to correctly triage incoming text messaging replies from a cardiovascular text–based secondary prevention program: Feasibility study. *JMIR mHealth and uHealth*, 8.

Ishani Mondal, Kalika Bali, Mohit Jain, Monojit Choudhury, Ashish Sharma, Evans Gitau, Jacki O'Neill, Kagonya Awori, and Sarah Njeri Gitau. 2021. A linguistic annotation framework to study interactions in multilingual healthcare conversational forums. *Proceedings of The Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

T Perrier, N Dell, Brian DeRenzi, Anderson R, John Kinuthia, and Jennifer A. Unger. 2015. Engaging pregnant women in kenya with a hybrid computer-human sms communication system. proceedings of the 33rd annual acm conference on human factors in computing systems. *ACM Press*, CHI 15.

Keshet Ronen, Esther M Choo, Brenda Wandika, Jenna I Udren, Lusi Osborn, Peninah Kithao, Anna B Hedstrom, Millicent Masinde, Manasi Kumar, Dalton C. Wamalwa, Barbra A. Richardson, John Kinuthia, and Jennifer A. Unger. 2021. Evaluation of a two-way sms messaging strategy to reduce neonatal mortality: rationale, design and methods of the mobile wach neo randomised controlled trial in kenya. *BMJ Open*, 11.

Hillary Rono, Andrew Bastawrous, David Macleod, Ronald Mamboleo, Cosmas Bunywera, Emmanuel Wanjala, Stephen Gichuhi, and Matthew J. Burton. 2021. Effectiveness of an mhealth system on access to eye health services in kenya: a cluster-randomised controlled trial. *The Lancet. Digital Health*, 3:e414 – e424.

Laura Marie Schwab-Reese, Nitya Kanuri, and Scottye J. Cash. 2019. Child maltreatment disclosure to a text messaging–based crisis service: Content analysis. *JMIR mHealth and uHealth*, 7.

J Unger, Keshet Ronen, Trevor Perrier, Brian DeRenzi, Jennifer A. Slyker, AL Drake, Danstan O Mogaka, John Kinuthia, and Gc. John-Stewart. 2018. Short message service communication improves exclusive breastfeeding and early postpartum contraception in a low- to middle-income country setting: a randomised trial. *BJOG: An International Journal of Obstetrics & Gynaecology*, 125:1620 – 1629.

Jennifer Unger, Brenda Wandika, Keshet Ronen, Claire Rothschild, Jay Shih, Dalton Wamalwa, Wangui Muthigani, Maneesh Batra, John kinuthia, and Grace John-Stewart. 2019. Mobile wach neo: Engagement of pregnant and postpartum women with a two-way sms service to improve neonatal outcomes.

Justine Zhang and Cristian Danescu-Niculescu-Mizil. 2020. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of ACL*.