USING MACHINE LEARNING TO ANALYSE RADIO CONTENT IN UGANDA

Opportunities for Sustainable Development and Humanitarian Action

September 2017

GLOBAL PULSE

PULSE LAB KAMPALA

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

This publication would not have been possible without the leadership, commitment and investment of many colleagues from Pulse Lab Kampala. UN Global Pulse would like to thank the Embassy of Sweden in Uganda, alongside the Government of the Netherlands and the Government of Denmark for supporting its network of Pulse Labs and the activities under this project. Special thanks also goes to the United Nations Development Operations Coordination Office (DOCO) for their support of the Radio Content Analysis Tool.

This report was authored by John Quinn and Paula Hidalgo-Sanchis from the Pulse Lab Kampala office. The pilot studies detailed in this publication represent the endless work of the team at Pulse Lab Kampala. The development of technology to enable these studies has been possible thanks to collaboration with Thomas Niesler, Raghav Menon and Armin Saeb from the Department of Electrical and Electronic Engineering at Stellenbosch University of South Africa.

UN Global Pulse wishes to also thank colleagues from the network of Pulse Labs for their inputs into the creation of this publication. Appreciation also goes to: Ellen Lust from the University of Gothenburg, Steven Goldfinch from UNDP Uganda, Eddie Mukooyo from the Ministry of Health in Uganda, colleagues of the National Emergency Coordination and Operations Center of Uganda (NECOC) from the Office of the Prime Minister of Uganda and the NGO Save The Children Uganda for their contributions.

EXECUTIVE SUMMARY

In September 2015 Member States of the United Nations adopted, by consensus, the 2030 Agenda for Sustainable Development. The 17 Sustainable Development Goals (SDGs) contained in the 2030 Agenda constitute a transformative plan for people, planet, prosperity, partnerships and peace. A data revolution was recognized as an enabler of the SDGs not only to monitor progress but to inclusively engage stakeholders at all levels to advance evidence-based policies and programmes to reach the most vulnerable.

In a world of increasing interconnectivity, radio remains a primary source of information for communities in many parts of the world, including Uganda. Radio reaches large groups of people in real-time and is often a medium for community discussions on subjects like healthcare, education, the provision of services and even politics.

There is a wealth of data that can be extracted from public radio conversations and these data can be parsed to support sustainable development and humanitarian efforts. Insights about the spread of infectious diseases, or the way people move during a disaster, or how they perceive healthcare campaigns or access to jobs and education, can be derived from radio talk.

This report outlines the methodology and processes of the Radio Content Analysis Tool, a prototype developed by Pulse Lab Kampala to analyse public radio content in Uganda and explore its value for informing development of UN projects and programmes on the ground. It distills the technology behind the creation of the Radio Content Analysis Tool and presents the lessons learned along the way.

The report also details the results of several pilot studies that were conducted together with partners from the Government, UN agencies and academia to understand the validity and value of unfiltered public radio discussions for development.

Hypothesis: Given the prominence of radio as a forum for public discussion, could this information be analysed to bring complete, timely and contextually rich insights that development practitioners can use?

By sampling different indigenous languages, types of broadcasters, and locations within Uganda, the pilot studies assess the potential uses of radio talk across five topics: perceptions towards refugees in Uganda, the impact of small-scale disasters on livelihoods, perceptions around the delivery of healthcare services, understanding the spread of infectious diseases, and monitoring the effectiveness of awareness raising radio campaigns.

The processes and lessons detailed in this report are intended to serve as examples and inspiration for using radio talk and data analytics to inform decision-making processes in development and humanitarian scenarios, in contexts where other sources of data may be missing or insufficient.

INTRODUCTION

In Uganda, where most of the population lives in rural areas, radio is a vibrant platform for public discussions, information sharing and news. Talk shows and phone-ins are popular ways for people to voice their needs, concerns and opinions. According to the 2014 Population and Housing Census released by the Uganda Bureau of Statistics[1], 55% of households reported radio as their main source of information, followed by word of mouth, 20% of households, and internet, only 7% of households.

Part of the reason radio has remained relevant is the way that new technologies have been incorporated into how radio stations interact with their listeners. Access to communication technologies has been changing the way that African radio stations operate, improving information gathering and engagement with communities[2]. The prevalence of mobile phones is a particularly important factor; with phone-in discussion programmes community members can become correspondents, commentators and critics[3].

The Uganda Communications Commission (UCC) reports 216 FM radio stations registered as of 2014, broadcasting from 299 locations around the country (see Figure 1). They are a mixture of commercial, community and religious-based organizations, though these categories are not mutually exclusive. For example, it is common for radio stations to have a commercial aspect, running adverts in order to make them financially sustainable. Approximately 7.5 million words are estimated to be spoken on these radio stations on a daily basis[4], around 80 times the amount of language content in the daily national newspapers. Because phone-ins and listener participation are a major part of radio programming, they provide a substantial source of information about public perceptions.

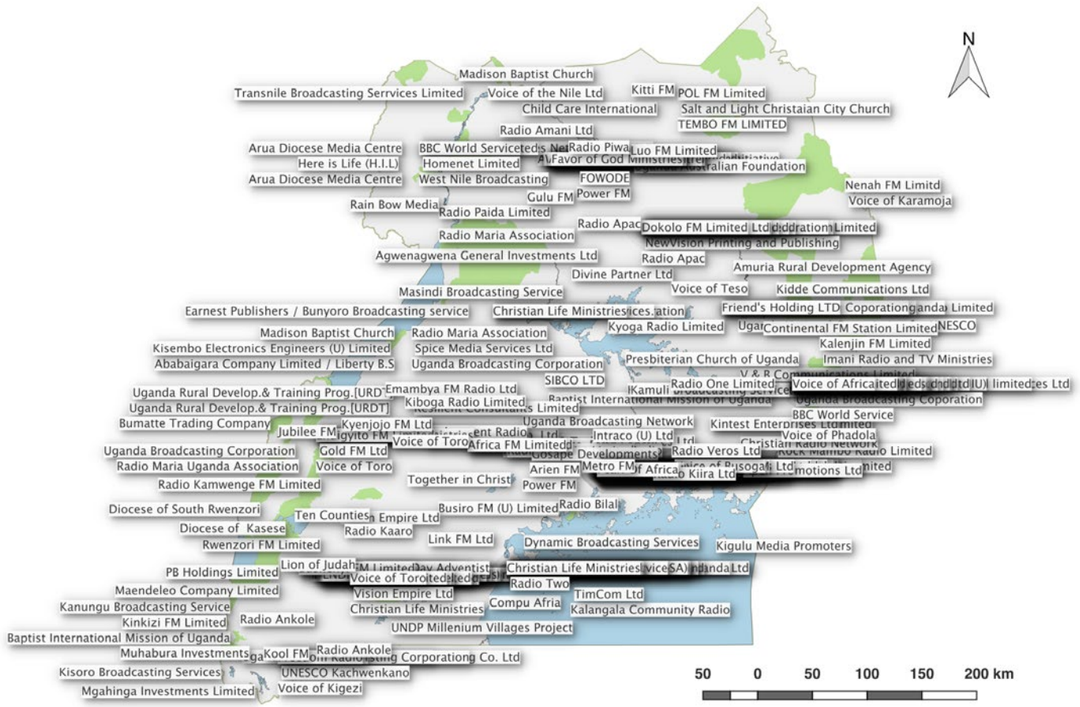

Figure 1: Locations of radio stations in Uganda (source data: Uganda Communications Commission, 2014).

There is a wealth of data that can be extracted from public radio conversations and these data can be parsed to support sustainable development and humanitarian efforts. Insights about the spread of infectious diseases, or the way people move during a disaster, or how they perceive healthcare campaigns or access to jobs and education, can be derived from radio talk.

Given the potential of radio data, there are a number of efforts[5] in place to use public radio content to advance the SDGs. Ongoing initiatives in Africa are promoting public debates on radio and asking audiences to voice their opinions on different topics. While these are important examples of the ways in which the perceptions of radio audiences can be systematically analysed, there are drawbacks to this type of approach. The approach is difficult to scale, requiring configuration for each individual radio station using them. It is also to be noted that answers depend on the type of questions listeners are asked, as well as the way in which the moderators frame the discussion. The overall manner in which information is collected is similar to the logic of a survey or focus group: it is a stimulus-response model in which the answers obtained might depend heavily on the choice of question asked. Therefore, the overall significance of information obtained, and its possible impact on the SDGs, may not be fully accomplished.

The radio analysis project carried out by Pulse Lab Kampala takes a different approach to harnessing the potential of public radio content. Rather than prompt the audience for feedback, the Lab developed a Radio Content Analysis Tool that passively filters public radio discussions in order to identify conversations on particular topics of relevance to the SDGs and humanitarian action. This approach offers certain advantages in that it could be scaled up to a wide range of radio stations and could reduce certain biases during the analysis of the data.

To assess the validity and value of the Radio Content Analysis Tool, Pulse Lab Kampala is working with the Government and UN partners on a number of pilot studies to understand what type of information can be obtained from radio talk in Uganda and how it might be used to advance and inform the SDGs.

THE TECHNOLOGY BEHIND THE RADIO CONTENT ANALYSIS TOOL

The Radio Content Analysis Tool was developed as part of a project conducted in collaboration with the Stellenbosch University in South Africa. The tool works by converting public discussions that take place on radio into text. Once converted, the text can be searched for topics of interest.

The first step in radio mining analysis is to physically receive radio transmissions. The project used a portable hardware setup composed of Raspberry Pi computers linked to an FM antenna with RTL-SDR radio dongles. The equipment was deployed in the central and northern regions of Uganda, namely capital city Kampala and Gulu. The predominant languages in the two cities are Luganda and Acholi respectively. The two locations were strategically chosen based on the number of radio stations they host and their socio-economic conditions. Kampala is the country's capital and has 59 radio stations, accounting for 27% of the national total.

[1] http://www.ubos.org/2016/03/24/census-2014-final-results/

[2] Nassanga, Goretti Linda, Linje Manyozo, and Claudia Lopes. "ICTs and radio in Africa: How the uptake of ICT has influenced the newsroom culture among community radio journalists." *Telematics and Informatics* 30.3 (2013): 258-266.

[3] Girard, Bruce. "Community Radio, New Technologies and Policy." *Fighting Poverty: Utilising Community Radio in a Digital Age* (2008).

[4] Pulse Lab Kampala estimation based on 216 stations broadcasting 18 hours per day, with talk 25% of the time, at 130 words per minute.

[5] For example, Africa's Voices (http://www.africasvoices.org) conducts radio debates and ask listeners about their opinions on the debates via surveys conducted with SMS or instant messages. TracFM (http://tracfm.org/) also uses SMS to ask listeners for their opinions.

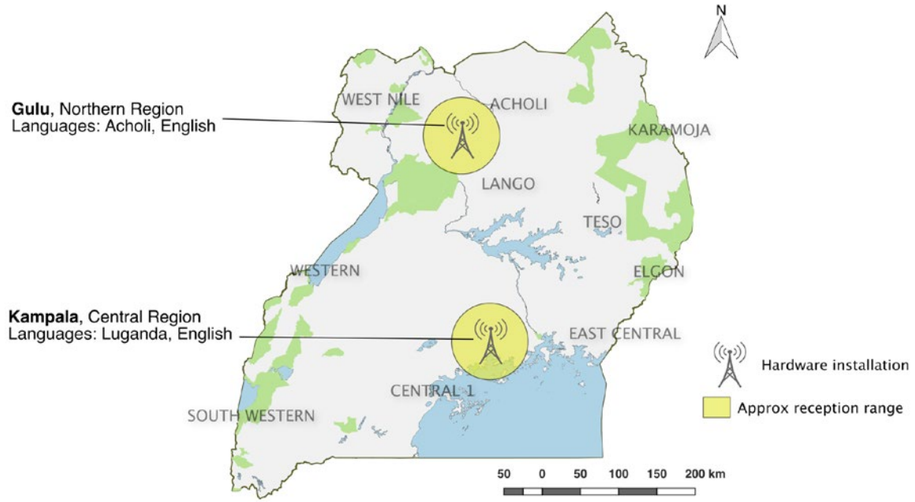

Figure 2: Locations where radio hardware was deployed.

Gulu is the largest population centre in Northern Uganda, and is of particular interest for SDG achievement due to high poverty rates, post-conflict reconstruction efforts, vulnerability to a range of hazards, and a large-scale refugee settlement programme. There are 15 radio stations that broadcast from Gulu, roughly 7% of the national total.

To weed out "radio noise," like for example music, Pulse Lab Kampala then developed a speech-recognition software to identify speech vs nonspeech broadcasts and delete the latter. On a daily basis, the software can capture hundreds of hours of public radio content in the form of raw data.

The project worked to convert this large and unstructured dataset, containing both relevant and irrelevant data, into smaller, structured datasets of categorized text for topics relevant to development. An overview of the process is detailed in Figure 3.

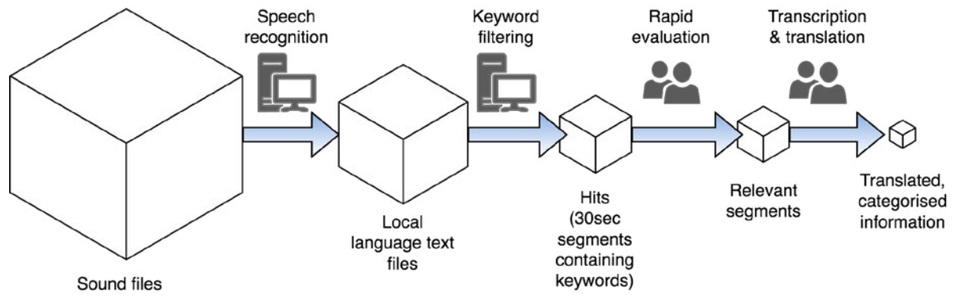

Figure 3: Overall radio mining process for converting a large quantity of unstructured audio data into a small quantity of relevant, structured information.

The overall radio mining process, has two automated software stages and two human analysis stages. This semi-automated approach allows a relatively small team of analysts to process a large quantity of audio recordings.

1. Automated software stage

The first automated software stage is *automatic speech recognition* (ASR) where the software extracts all the words recognized in the received audio and converts them into text. The software built as part of the project is able to recognize a vocabulary of 34,415 different Luganda words and 14,381 Acholi words. The sequences of recognized words can contain errors, due to variations in speakers, pronunciation, or background noise. Around 50 out of 100 words are correctly recognized in Luganda and 60 out of 100 words are correctly recognized in Acholi. The accuracy depends on factors such as how slowly and clearly the speech is pronounced. For example, speech is more easily recognized during news broadcasts, where presenters are trained to articulate clearly, and is less recognized during call-ins, where the quality of the audio is poorer and the speech is a rapid conversational style.

The second automated software stage filters the broadcasts to look for keywords and keyphrases relevant to pre-defined topics. The filtering can take place even in the presence of speech recognition errors, due to the fact that speakers usually utter several words and phrases related to a particular topic of interest when engaging in radio talk. However, refining the keyword logic is a time-consuming process, as there are many ways in which people can express opinions on topics of interest, and sometimes, even if a term usually has the connotation pertinent to a particular topic, it may be used in a phrase that is not relevant.

For each of the pilot studies described in this report, the keyword filters take the form of three lists: required keywords/keyphrases, optional keywords/keyphrases and keywords/keyphrases to exclude. For instance, when filtering for discussions on healthcare, the software highlights the broadcast if any of the Luganda words *omusawo* (doctor) or *eddagala* (medicine) are spoken. Other terms such as *obulamu* (health/life) are optional, in that they are less specific to the topic at hand, but where they appear they constitute additional evidence that the topic of interest is being discussed. The exclusion list also helps reduce the number of irrelevant matches; for example *okulya obulamu* (literally "eating life") simply means to enjoy oneself, so the health filter rejects any discussion containing the term.

The output of the automated stage is a daily list of filter hits, ranked by relevance, which is calculated using the number of matching keywords. Because of the complexity of some of these languages and the imperfection of the automatic speech recognition, the filters do not necessarily capture every discussion of relevance, and not all discussions matching a filter are necessarily relevant. It is also to be noted that the choice of keyword logic is one potential source of bias in the overall analysis.

2. Human analysis stage

Analysts will examine original radio audio clips rich in relevant keywords to discard false detections of topics and extract further semantic detail. The analysis is done in two stages: First, a rapid assessment is conducted, where analysts listen to an extract of the original audio clips for which there was a cluster of matching filter rules. They tag the recording as either a *false match* (if the keywords were wrongly detected), as *not relevant* (if the keywords were correctly detected, but in a context that was not of interest, e.g. "…there has been a flood of accusations…"), or otherwise as *relevant*. Some of the *not relevant* tags are usually identified when the keywords are mentioned in adverts or the lyrics of songs. For those audio clips identified as *relevant*, literal translations into English are provided. Providing literal translations is a key step to ensure the analysis captures the voices of the people while avoiding personal interpretations or summaries that could distort the data.

Second, the analysts annotate the *relevant* fragments by categories (e.g. disaster.flood, health.disease-outbreak.cholera). These categories are either pre-identified depending on the relevant research angle or driven by a corpus analysis, a posteriori, and added later on. The type of speaker–either a member of the public, a news anchor, a government official etc.– is also recorded. The results are incorporated in a structured, searchable database.

At the time this report was written, the analysis team had manually reviewed 80,166 filter matches. A large part of the analysis efforts consisted of identifying those filter matches in which the keywords were being used in the sense intended in the topic filter. Overall, the team found that around 96% of filter matches were not relevant, including over 18,000 filter matches (23% of the total) where keywords were spoken as part of an advert. From the relevant matches, 873 were fully transcribed, translated and categorized, and they represent the basis for the pilot studies detailed in the report.

TALK RADIO AS A SOURCE OF BIG DATA - OPPORTUNITIES AND CHALLENGES

The project sought to analyse public radio discussions at large scale and for less known languages to help provide useful insights to inform on SDG-related topics. In order to explore the potential use of this data, the project looked to first understand its relevance.

The first question the analysis asked was: *What data in a 24/7 public radio flow can be used to monitor and help achieve the SDGs?* After analysing the data available, radio content was divided into:

a) **Irrelevant broadcasts:** content related to sports, music and celebrity news;
b) **Relevant broadcasts:** local news reports, shows with district officials as guest speakers and *call in* shows where citizens express their opinions on a variety of topics and where they share information on local events (see Table 1).

Type of information		Reported by...
Reports (local facts and realities)	News reports	Part of the regular programme on local radio stations
	Testimonials	Individuals who have witnessed events or heard about them and participate in a talk show
		Representatives of grassroots organisations invited to shows by local radio stations
Opinions		Individuals expressing perceptions, concerns or hopes through participation in talk shows
		Representatives of grassroots organisations invited by local radio stations to participate in shows
		Local officials invited to shows by local radio stations

Table 1: Categories of public radio discussions in Uganda found to be of most relevance

Second, the project explored the biases in the data. All data sources are biased in one way or another, but the existence of biases does not mean the data cannot be used. However, understanding which biases the data possess is important in order to draw the appropriate conclusions. In the analysis of public radio content, the project identified the following biases:

- Talk shows were identified as the main source of *relevant content*, in particular individuals' testimonies and opinions. However, the analysis revealed that men participate in talk shows much more often than women. Also, middle-aged and elderly men seem to participate in talk shows more often than young people.

- Unless speakers mention a particular affiliation, radio content does not reveal the socio-economic background of talk show participants, or their affiliation with grassroots or other types of organisations. Hence, the types of weighting methods that are commonly used to adjust for bias in survey and questionnaire data cannot be used for radio data.
- For talk shows that do not call their listeners back, there is a cost incurred by the person calling in. This can deter the poorest members of a community from contributing to a discussion.
- Different radio stations have varying types of management and profile of presenters and listeners: a commercial radio station operating from Kampala provides a different view of events compared to a church-based community radio station operating in a rural area, for example.
- Talk shows are usually on a particular theme, or are framed in a particular way by the presenter. While this bias is mitigated by analysing the discussion of many talk shows from many different radio stations, in general the discussion is not entirely spontaneous and there can be prompting to some extent.
- Reports or allegations made on the radio are not necessarily true, for example where the speaker is motivated by sensationalist or political objectives.
- Some topics are less likely to be discussed on the radio, for example because of social stigma or fear of retribution, adding to selection bias.

Once the biases are understood and subsequently an analysis that takes them into consideration has been designed, public radio content can be analysed for insights useful for the SDGs.

A COMPENDIUM OF PILOT STUDIES

To validate and understand the value of radio data to inform on issues pertinent to sustainable development and humanitarian action, Pulse Lab Kampala worked on five pilot studies together with partners from various UN agencies.

This section summarizes the findings and lessons learned from the studies and presents preliminary observations on the usefulness of radio data to inform on specific topics of interest: (i) perceptions of host communities towards refugees, (ii) understanding the impact of small-scale natural disasters, (iii) perceptions regarding local governance and the quality of public health service delivery, (iv) understanding the spread of infectious diseases, and (v) gauging the effectiveness of behavioural change campaigns.

While in some cases, the analysis showed radio data can provide a wealth of relevant information, in others the data was found to be less relevant. In those cases, radio data could be coupled with other traditional and new sources of data for better results, or a different type of big data–mobile or social media data– could be used for the analysis.

UNDERSTANDING PERCEPTIONS TO A REFUGEE INFLUX THROUGH ANALYSIS OF LOCAL RADIO CONTENT

An outbreak of conflict in July 2016 caused thousands of South Sudanese to flee to neighbouring countries, especially Uganda, where the number registered in 2016 reached over 600,000 people, mainly women and children. Due to the protracted nature of the crisis in South Sudan, the UN Country Team continues to consider options that enable both refugees and host communities to build resilience and improve self-reliance. Moreover, the South Sudanese influx presents increasing social, economic and environmental pressures on host communities that, unless addressed through innovative and targeted support, could result in conflict and instability. In response to the UN needs, Pulse Lab Kampala was tasked by the UN in Uganda with unearthing the attitudes and intentions of host communities towards refugees. The Lab used automated speech-to-text methodologies to analyse local radio content to provide insights around the refugee influx in Uganda.

Methodology

The analysis was conducted from July 2016 to February 2017. Filters to detect conversations regarding refugees were created and the translation team annotated all mentions of refugees, including issues of acceptance, health or increasing social tensions. To be able to identify the level of concern of host communities towards refugees, the project conducted an inductive and manual tagging of relevant keywords. This classification would not have been possible using the automated semantic analysis given the high variety of well-being issues discussed (e.g. car accidents, health service delivery, livelihoods) and the complexity of sentiments expressed.

Results

Findings showed a high degree of acceptance of refugees among Ugandans at the time when the influx of refugees started. However, some concerns were expressed about the spread of infectious diseases and an increase in pregnancies. Ugandans also raised the question of the right of refugees to agricultural land and implications on forests of more people requiring wood for shelter and firewood for cooking. The figure below represents the main topics and volume of conversations over a one-month period from mid-July to mid-August 2016.

Figure 4. Topics related to the South Sudan refugee crisis during the early stages of the influx in July 2016. Volume of discussions is represented by the size of each of the bubbles, while topics of discussions are represented through the different colours.

A further analysis of radio content from November to December 2016 revealed new issues related to the refugee crisis.

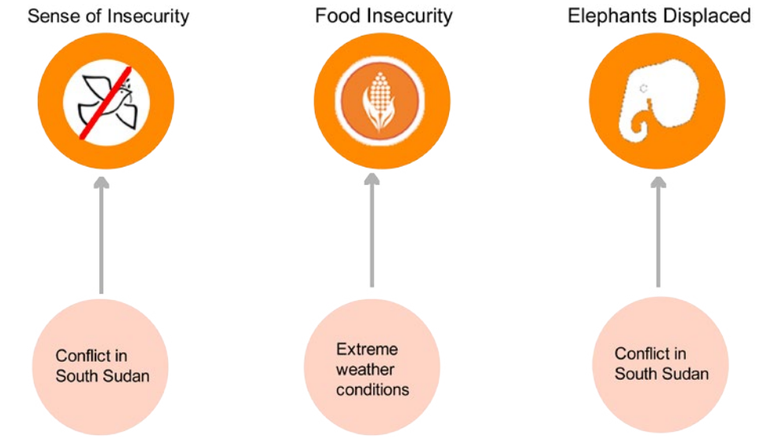

Figure 5. Events related to the unfolding refugee crisis (lower), which are reflected in topics of radio discussions extracted by Radio Content Analysis Tool (upper).

The study also revealed discussions and rumours of a cholera outbreak in refugee settlements in Northern Uganda two weeks before the outbreak was officially announced by the Government in August 2016.

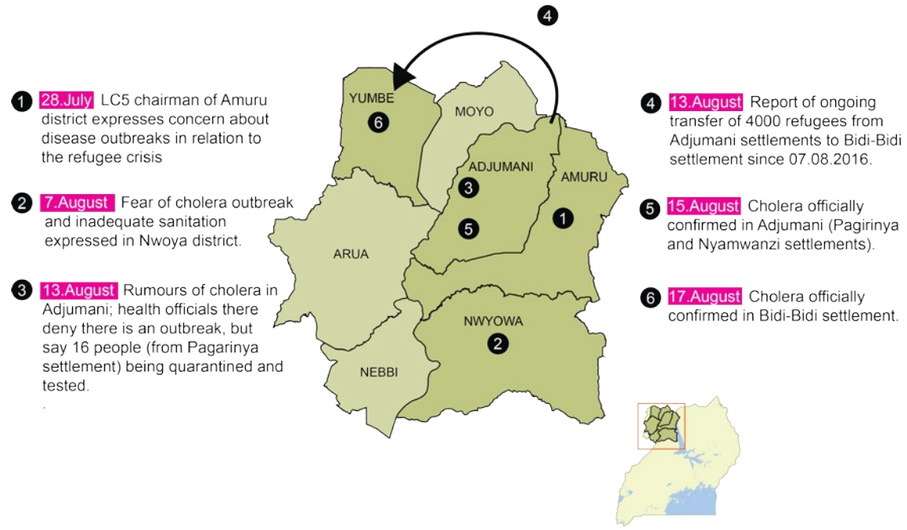

Figure 6. Timeline of events and radio discussion related to cholera outbreak in Northern Uganda refugee settlements in August 2016.

This effort demonstrated that speech-to-text processing (by automated dictation software), combined with natural language processing (with software that can interpret text), may be used to close an information gap across the digital divide by generating machine-readable big data about refugees in rural areas where up-to-date information of any kind is seldom available. The data could be used by humanitarian organizations and government institutions to monitor the response on the ground and inform decision-making processes. This type of data could also be used together with other sources, like data from social media, to feed early warning systems on conflict prevention.

DOCUMENTING THE IMPACT OF SMALL-SCALE, LOCAL DISASTERS WITH RADIO DATA

Instruments currently used by the Ugandan authorities[6] to collect disaster-related data do not always contain complete information about the effects and impact of small-scale, local disasters. Most of the small damages caused by natural disasters are not reported by media, in newspapers or government reported mechanisms and fail to be included in disaster risk management efforts.

First hand testimonials on how natural hazards impact communities are common in public radio discussions, especially at the time when the community is being affected.

Pulse Lab Kampala worked with the United Nations Development Programme (UNDP) and the National Emergency Coordination and Operations Centre (NECOC) in the Office of the Prime Minister to test whether useful data can be extracted from talk radio on the occurrence and impact of small scale disasters on households, communities and small and medium-sized enterprises. The project used the Radio Content Analysis Tool to collect and analyse information on the location, extent of damage and type of disaster.

Methodology
The analysis was conducted on public radio content from October 2016 to January 2017. Typical local disasters were tagged to fit the categories of the management information system used by the Ugandan authorities and provide information about the location (local level), type of natural disaster (flood, drought, lightning) and quantitative losses.

Results
In Northern Uganda, a large number of discussions related to the damages and losses resulting from natural disasters–such as drought or lightning–and the coping mechanisms of communities.

> *The heavy rains with hail stones have destroyed 13 buildings in Kampeka and Ssemuto sub-counties that are in Nakaseke district, which has left hundreds and hundreds without homes and has affected them as well. The rain that fell so much has also destroyed many roofs of buildings and among them are 3 school buildings that were left without roofs. Two health centers as well were destroyed completely. In addition, things planted in the garden [crops] about 100 acres in total [were destroyed] in the two sub-counties to the extent that you look at them with pity. The schools that have been destroyed are Semuto secondary school, Tuzunkere primary school and Erina memorial primary school. Kikandwa health center two and Lukumumbi Health center two are among the hospitals that have been destroyed."*
>
> (August 2016/ Media /Northern region)

> *"Rain that fell with hailstones that were very strong destroyed a lot of crops in over 20 hectares of land in Atiak Amuru district within two days only."*
>
> (August 2016/ Media /Northern region)

[6] DesInventar is a disaster management information system adopted by 94 countries for the systematic collection, documentation and analysis of data about losses caused by disasters associated with natural hazards.

Findings also revealed complaints from local leaders who correlated the influx of refugees from South Sudan with the displacement of elephants from their natural habitat.

> *"About 400 elephants and so on are just seated every day [daily] in the sub counties of Lapono and Paimol so foods like millet, sorghum and whatnot, the elephants are doing what, spoiling them all."*
>
> (October 2016, Northern region)

Reports and testimonials from the central region of Uganda were found to contain less relevant information, which could be attributed to reports coming from more national officials and media outlets rather than grassroots and communities. Only 36% of reports mentioned location of the disaster and 18% talked about the quantitative losses associated with the event.

The analysis yielded useful insights on the extent of small scale disasters, as shown in the snippets above, as well as the impact on communities and their coping strategies. The difference in the large amount of information gathered from Northern Uganda, compared to the relatively little information from the central region, may mean radio data analysis efforts could be most useful in areas where radio is the main source of information.

UNDERSTANDING PERCEPTIONS ON THE QUALITY OF HEALTH SERVICE DELIVERY

The objective of this pilot study was to evaluate the availability of public radio content to analyse perceptions on the quality of health service delivery that could be used to inform local governance processes. The study was conducted together with the Ministry of Health of Uganda, and with guidance from the Department of Political Science at Gothenburg University.

Methodology
The analysis was conducted from October 2016 to January 2017 and filtered radio talk that captured discussions related to a variety of issues pertinent to healthcare service delivery. To ensure a data-driven and inductive approach rather than base an analysis on pre-conceived topics of what is important for communities, the project used the entire corpus of data extracted from radio discussions to identify recurrent semantic categories.

Four categories were identified as part of the analysis (see Figure 7) and radio content was manually tagged into each category to allow for a deeper understanding of the context and to lower the error rate in categorization.

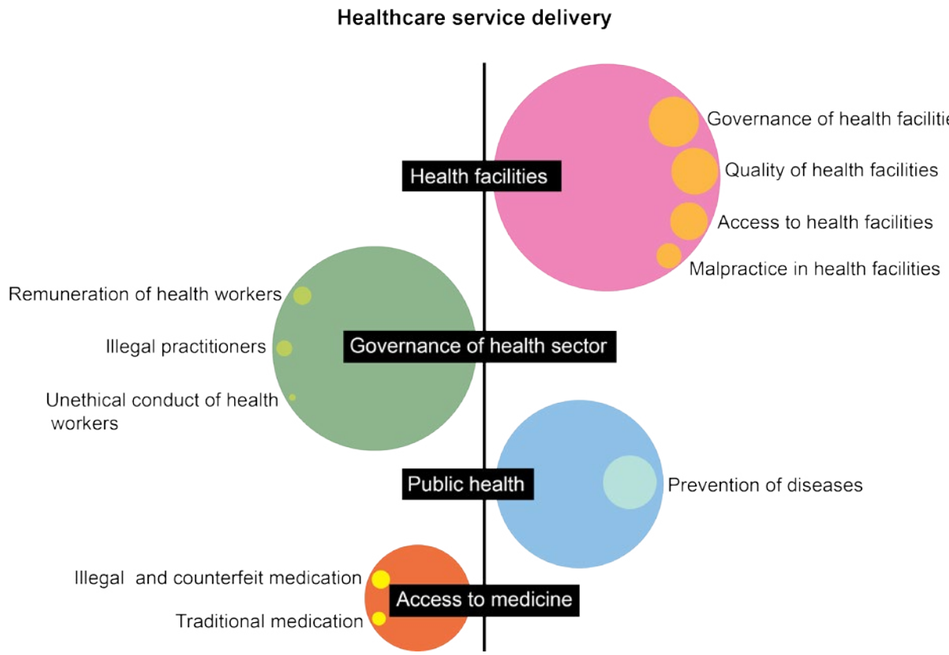

Figure 7. Topics of radio discussion identified related to healthcare service delivery.

Results
The analysis showed a large volume of ongoing conversations around the delivery of health services. The bulk of conversations described complaints and reports around the poor quality of facilities and the difficulties people encounter when trying to access medicine. A number of discussions brought up issues pertinent to disease prevention strategies as well as the use of illegal or counterfeit medication to treat patients. Figure 7 describes the volume of discussions around each of the four categories.

Both in the central and northern regions of Uganda, findings showed a large volume of discussions around the topic of *governance of the health sector*, with people commenting on issues related to the remuneration of health workers, instances of illegal practices or the unethical conduct of some of the health workers.

> *"When you go to the hospital they tell us 'go to the clinic of so and so'…. Most of the doctors now own clinics, and if possible the government should prohibit doctors working in government hospitals from having clinics. There is a lot of medicine theft by the doctors and they take the medicine to their private clinics."*
>
> (August 2016/ Grassroots /Northern region)

> *"Sometimes you go to a hospital and a doctor tells you; "I would have treated you but you voted badly".*
>
> (September 2016/ Media /Central region)

The initial findings show that real-time information from radio discussions can provide a glimpse into how people perceive public health services, and what they deem to be issues of concern. These insights can be used to evaluate and tailor the quality of health services.

ANALYSING RADIO DATA FOR INSIGHTS INTO MALARIA OUTBREAKS

Malaria is a leading cause of death in Uganda, accounting for over 27% of deaths annually[7]. Statistics show that Uganda has the world's highest malaria incidence, with a rate of 478 cases per 1,000 people every year. The Acholi sub-region still bears the harshest brunt of the disease, with the prevalence soaring every year.

Pulse Lab Kampala set out to test whether radio talk can provide information on the incidence of malaria and the cases that are being reported to inform the implementation of malaria eradication programmes.

The project analysed radio data for the period October 2016 to February 2017. Filters were set up to screen for words related to malaria, including its translation in local dialects– "omusujja" in Luganda (literally "fever", Luganda having no specific word for "malaria").

However, findings proved mostly inconclusive with only few malaria-related discussions being identified. Although the project did not yield the expected results, it highlighted some potential shortcomings of radio data analysis, revealing that certain topics of interest may not be widely discussed over radio, or that they may be harder to filter.

[7] http://health.go.ug/content/malaria-bulletin-2016

MONITORING RADIO CAMPAIGNS IN REAL TIME TO GAUGE THEIR EFFECTIVENESS

Radio campaigns are popular ways to reach remote villages in countries like Uganda, and development organizations use them to implement behavioural change campaigns. Issues addressed range from water and sanitation, to gender violence, to HIV/ AIDS campaigns.

However, monitoring and evaluating the results of these campaigns is challenging and it usually requires people on the ground surveying or interviewing communities where the radio stations are broadcasting the campaign. This is labour-intensive, particularly for campaigns which are long-running on multiple stations or in inaccessible areas.

Pulse Lab Kampala worked with Save the Children on a study to monitor the implementation of the Every Last Child campaign. Save the Children, an international non-governmental organization that promotes children's rights, is running the Every Last Child global campaign to promote better access to life-saving healthcare and quality education for all children. As part of global efforts, in July 2016 the organization launched a radio and video campaign in Uganda, to reach children in remote areas of the country.

Pulse Lab Kampala used the Radio Content Analysis Tool to monitor the frequency with which the campaign is promoted, with particular focus on a song that aired as part of the campaign.

Methodology
Software was trained to recognise the song, and applied to the recording stream for Gulu (Mega FM and Radio King), and Kampala (Capital FM). The accuracy of the automatic detection was tested during an evaluation period where a human analyst listened to radio broadcasts to manually identify the number of times the song was played, which was then compared to the number of times it was picked up by the software.

Station	Total plays	Number detected	Number missed	False alarms
Mega FM (Gulu)	30	21	9	0
Capital FM (Kampala)	4	2	2	0

Results of the automated campaign detection.

Results
The detection rate of the automatic software was 68% with zero false alarm rate, suggesting that campaign detection using the radio tool can be an effective monitoring and evaluation method. However, the tool could be improved to ensure better accuracy of reception. For example, missed detections could be caused by weak reception of radio stations. This is an issue particularly with the radio mining setting where a single antenna is used to receive multiple radio stations; it can be difficult to place the antenna in such a way that all stations are clearly received. Investing in radio infrastructure could improve the results.

DATA PRIVACY AND DATA PROTECTION

The complexities of many data innovation projects can put fundamental human rights, including the right to privacy at risk, sometimes in unexpected ways. Although radio is a public forum of discussion, UN Global Pulse employs data privacy and data protection[8] measures to ensure that individuals are not exposed in a way that could harm them.

Pulse Lab Kampala used specific tools, like data anonymization, restricting access to the data during project implementation and destroying the data once the project was concluded, to protect and respect the privacy of individuals and communities. Although the Radio Content Analysis Tool analysed data from public radio broadcasts, projects did not identify opinions of individuals, but rather used data analytics to understand trends pertinent to the SDGs at local or regional level.

Below are some measures UN Global Pulse implements to ensure the protection of individuals' rights, which are useful when analysing data from radio discussions.

Limited access and keywords/phrases
The data generated from projects is accessible only to UN Global Pulse and kept on a secure server. Additionally, only project-relevant speech segments obtained after filtering are made available. For example, a project that aims to understand the effects of natural disasters will identify and aggregate discussions regarding the early onset of a drought. 'Scrap data' that is not relevant will not be kept.

Anonymization of voices
Since keywords and phrases need to be understood in their context, it is useful to be able to listen to the pertinent segments where they appear. To maintain anonymity, one solution is to modify the sound in order to anonymise it. In the pilot studies detailed in this report, UN Global Pulse used the transcribed text rather than the audio recordings to filter for topics of interest. Any audio extracts used as examples were first vetted to ensure the content of the discussions did not pose privacy risks.

Restricting access to the data
People phoning into radio are sometimes asked to provide their name and where they are calling from. Also, at times, their voice could be used to easily identify them. UN Global Pulse does not release any of the raw data it collects. Instead, Pulse Lab Kampala produced visualizations for exploration that serve as an analysis based on aggregated findings.

Saving recordings
Recordings were initially saved for the development of the Radio Content Analysis Tool to train the speech recognition software for previously unavailable languages and dialects. Once a pilot project using radio data is completed, recordings that do not contain relevant data are discarded.

[8] UN Global Pulse Data Privacy and Data Protection Principles, UN Global Pulse, 'Data Innovation Risk Assessment Tool,' 2016

CONCLUSIONS

Radio remains a rich source of information, especially in the developing world, where phone-ins and radio programmes are the most common way for people to publicly voice their opinions, concerns and ideas.

The use of aggregated radio data holds potential to inform on sustainable development and humanitarian efforts. The research and pilot studies conducted with UN partners have demonstrated proof of concept, and there is an opportunity to create large-scale projects that will provide valuable insights into a wide range of SGD-related goals and indicators.

While there is myriad information to be parsed from radio talk, the various biases that come with analysing this data should be carefully considered. Complementing these insights with information from other sources, including traditional data collection methods, could provide more granular, accurate information in some instances.

Findings of the pilot studies have provided a basis for UN Global Pulse and partners to understand the type of information that can be gauged from radio talk in Uganda and how it might be used to advance the SDGs. Table 2 provides a comparison of opportunities and challenges of radio data with regard to other sources of information:

Information source	**Enabling attributes of radio data**	**Potential roadblocks of radio data**
Surveys and focus group discussions	Sample of population whose voices are analysed is larger (population scale, rather than limited sample size); Not restricted to answers given to particular questions, “unfiltered”; Can be carried out on an ongoing basis, rather than snapshots, and hence useful for emerging issues.	Unknown demographics of contributors; Possibility that a particular topic of interest is not widely discussed; More difficult to include native speakers of different languages and dialects.
Analysis of social media data	Wider demographic range in contexts such as Uganda, particularly for poor households; Spoken discussions often go into more depth than is typical in social media posts.	Complexity in developing speech-to-text technology; Results may be skewed due to speech-recognition errors (the system mis-recognize certain words or phrases).

Table 2: Relative advantages and potential roadblocks of public radio analysis in comparison with other methods of extracting information.

The development of technology to analyse African vernacular languages continues and Pulse Lab Kampala is currently testing a second generation prototype with two additional Ugandan languages.

Going forth, UN Global Pulse and partners plan to develop a large-scale demonstration project using radio content to inform sustainable development and humanitarian action in Uganda. Engagement is ongoing to define the scope of the new project that will be launched in 2018. The areas of analysis will have a common focus to support the Leave No One Behind agenda.

How to cite this publication:
UN Global Pulse, ‘Using machine learning to analyse radio talk in Uganda,’ 2017


GLOBAL
PULSE


PULSE
LAB KAMPALA