COMPLETION REPORT

Exploring the use of Big Data in Somalia Peacebuilding

Pulse Lab Kampala/UN Global Pulse
July 2018

Photo credit: Zohra Bensemra / Reuters: https://www.pri.org

GLOBAL PULSE

PULSE LAB KAMPALA

CONTENTS

LIST OF ANNEXES

ABOUT THE PILOT PROJECT

With the generous support of the United Nations Peacebuilding and Support Office (PBSO), UN Global Pulse and partners piloted and evaluated technology prototypes to gauge public perceptions to support the ongoing peace and state building processes in Somalia without security risk exposure to UN personnel. The experimental prototypes rely on big data analytics and artificial intelligence to explore the value of analysing public discussions on social media (Facebook) and radio to support the work of the United Nations and the Federal Government in Somalia. The main project results include:

RADIO CONTENT ANALYSIS TOOLKIT: the prototype mines voices from public radio discussions through automated speech recognition developed specifically for Somali language.

SOCIAL MEDIA ANALYSIS TOOLKIT: the technology tracks topics of relevance in Facebook (FB) by using keywords that contain predefined Somali words.

'QATALOG' ANALYSIS TOOL: the user-friendly tool allows UN personnel and partners with basic computer skills to access and analyse big data from public FB and radio discussions.

ANALYSIS OF FB AND RADIO DISCUSSIONS: in addition to the different analysis done to frame big data specifically for the Somali context, 4 sets of analysis results were produced on public discussions from FB and radio.

A brochure on the pilot project is available in Annex A.

OUTCOME PROGRESS

Stakeholder engagement has been ensured from the initial steps of project implementation. Discussions, brainstorming sessions, two workshops in Mogadishu and working sessions in Kampala have provided the opportunity to all stakeholders to co-design the pilot initiative in consensus.

Roles and modalities of engagement among partners for project implementation were agreed upon. A Big Data Advisory Committee (BDAC) was formed to guide and support end-to-end project implementation. The partners included in the BDAC were: UN Assistance Mission in Somalia (UNSOM), UN Peacebuilding Commission (PBSO), UN Department of Political Affairs (DPA) / Somalia Team, UN Development Programme in Somalia and Uganda, UN Resident Coordinator Office (RCO) in Somalia and UN Global Pulse.

Senior management was informed on the project, including a briefing to the UN Country Team in Somalia and two briefings to the Special Representative of the UN Secretary General in Somalia in January and November 2017.

One-to-one sessions were organized between project analysts and UN colleagues in Somalia, for sharing views on the technology prototypes under development and to explore how these could support the work of the teams in the country.

The following package has been prepared on the engagement process:

Annex B. List of consultations and brainstorming sessions
Annex C. Inputs from stakeholders in brainstorming sessions
Annex D. Terms of reference of BDAC

The agreed expected deliverables of the pilot project were:

- Social media (FB) text analytics toolkit for Somali Arabic developed. The toolkit will include the software programme and a technical guide.
- Prototype for analysis of radio content in Somali Arabic language developed. The toolkit will include the software programme and a technical guide.
- Report and dashboard produced with pilot Big Data analysis to inform peace and state building process in Somalia.

Important milestones were achieved in the development of the technology prototypes. Especially relevant is the advancement with the radio prototype that is a worldwide innovation. Also relevant is the development of Qatalog, a tool for the analysis of FB and radio content. Additionally, analysis results were produced to probe the value of the analysis.

The success of the pilot initiative encouraged UN Global Pulse to start planning the next phase-scaled up of the pilot.

OUPUT PROGRESS

ACHIEVEMENTS ON THE DEVELOPMENT OF THE TECHNOLOGY PACKAGES - RADIO

A documentation package has been prepared on the technology prototype, including:

Annex E. Brochure on the radio content analysis toolkit
Annex F. Technical guide on the radio toolkit
Annex H. Overview of challenges with radio hardware and software

Overview of the radio content analysis toolkit

The raw data that is received from public radio stations is very large and the radio equipment streams hundreds of hours of public content every day. With the machine filtering, the initially large and unstructured datasets become a smaller dataset.

The application reduces 3,906 weekly hours of public radio content to 653 by first filtering out music, and only targeting speech. Then, a second filter is applied to target radio programmes relevant for analysis, reducing content to 251 weekly hours. A final filter is applied to target relevant conversations on a specific topic.

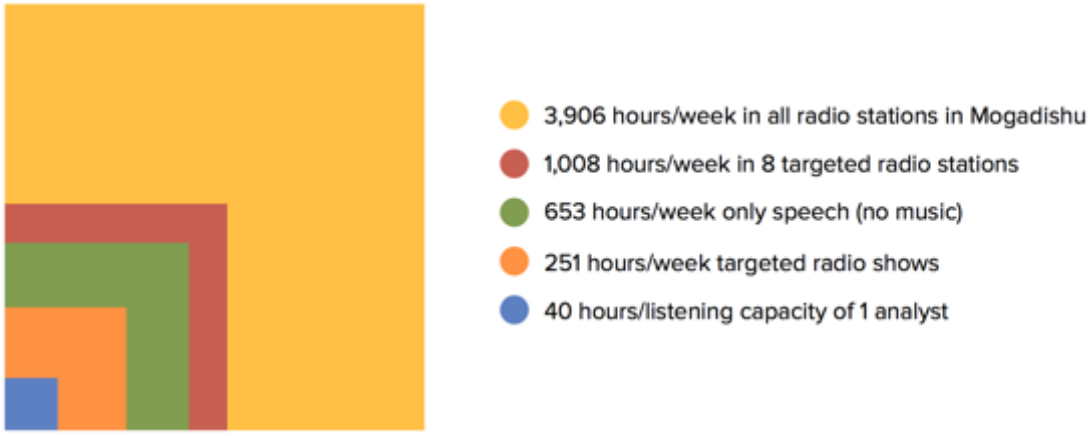


Illustration of the radio content filtering process in Mogadishu

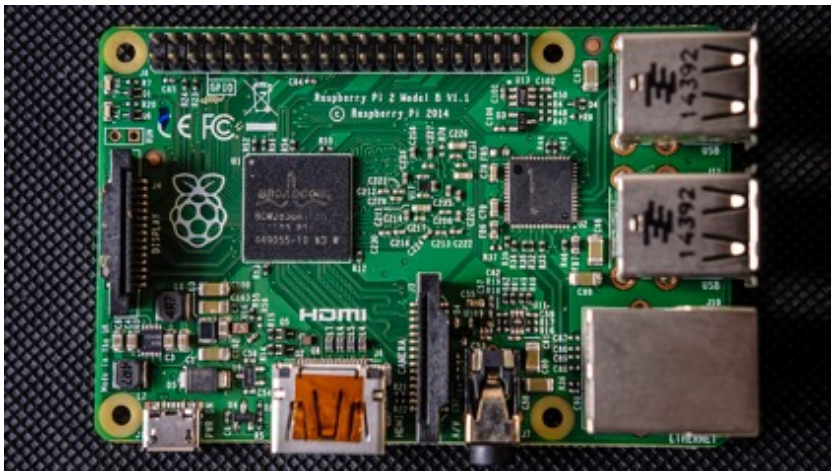
Radio hardware

The IT equipment to capture and analyse radio content was deployed and installed in a first location in Mogadishu at the UN MIA compound in January 2017. Several challenges were experienced (described in documentation package) and in consequence, in May 2018, it was moved to a new location, with the following enhancements:

- Secured space, with limited number of staff that has access;
- Indoor space and enhanced housing of equipment, protecting it from corrosion;
- Reliable internet connection;
- Stable power through an uninterruptible power source (UPS);
- Improved radio signal due to better position of antenna;

- Increased space to host additional hardware, with capacity to host hardware to capture up to 12 radio stations.

The system was set up to receive radio content flow streaming from 10 radio stations in Mogadishu area. The radio stations broadcast mostly in Somali language. With enhancements in the equipment, an average of 1,200 radio clips of 5 minutes duration can be uploaded every day to a cloud server, totalling approximately 6GB.



Piece of hardware for radio content analysis, called Raspberry-pi.

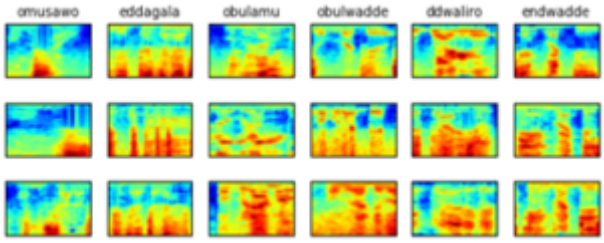
Radio software

The Somali speech recognition software is able to process audio files from broadcast radio and recognises words from a vocabulary of 3,099 Somali terms, with a world-level accuracy of approximately 43% on radio recordings from Mogadishu.

To develop it, Pulse Lab Kampala worked with Stellenbosch University of South Africa. Two post docs worked full time on the radio prototype directed by the Head of the Department (who supported the project on a pro bono basis).

This work built on the working prototype developed by Pulse Lab Kampala Luganda and Acholi languages of Uganda.

Prior to the first version of the software, a proof of concept was done with another African language (Lugada), to test the neural network method.



Examples of signal processing-based representations of keyword recordings used as input for the new speech recognition software, for keywords in Ugandan languages. Similar methods to the "OK Google" keyword detector on Android were used.

Then, training data was created to develop the system for Somali language; it consisted of:

a) 2 hours and 10 minutes of detailed (with specifications and annotations) transcription of Somali audio content into text.
b) Approximately 100,000 worlds of Somali text materials identified to build a wordlist and a basic language model.
c) Phonetic rules dictionary for Somali language transformed into software code, to enhance the language model.

Software was also developed to:

- Filter out music content - the software identifies and filters out audio clips with more than 70% of music content;
- Browse radio broadcasts from Mogadishu according to times/frequencies that have been identified as relevant for analysis, i.e. which are known to include topical discussion and phone-ins. This software is named 'Goldie' and it automatically distributes the computing resources so that higher priority radio programmes can be analysed with greater computing power than lower-priority radio stations.

In addition, developments were made to integrate all software pieces in the radio content analysis flow. This allowed the automatic targeting of relevant audio clips after audio data is streamed by the hardware.

ACHIEVEMENTS ON THE DEVELOPMENT OF THE TECHNOLOGY PACKAGES – FB

A documentation package has been prepared on the technology prototype, including:

Annex I. Brochure on the social media analysis toolkit
Annex J. Technical guide on the social media analysis toolkit

Overview of the FB analysis toolkit

The technology prototype allows for the analysis of public FB content in Somali, extracting relevant comments and posts made on public pages. There are two ways a user can interact with FB: through the online platform and via the application programming interface (API). The prototype developed uses the API to access the FB database. The software converts information on the platform into structured sets of data that are ready for analysis.

The top 5 functionalities of the software developed are to:

- Target specific public FB pages and extract comments on these pages;
- Target FB comments posted over specific periods of time;
- Target FB comments using specific keywords;

- Tag FB comments according to analytical needs (including gender);
- Identify trending topics on FB.

Particularly innovative is the functionality to identify trending topics. The software looks for words mentioned in FB the past 48 h that did not appear as much in previous weeks, identifying topics people may be talking about now that they were not discussing before.

ACHIEVEMENTS ON THE DEVELOPMENT OF THE TECHNOLOGY PACKAGES – QATALOG ANALYSIS TOOL

Qatalog tool and a video with user guidelines about it can be accessed in the following links:
https://Qatalog.unglobalpulse.net/
https://drive.google.com/open?id=1tdCgUUZNQ4jPM9bsyE7eMqexARl4vqal

A documentation package has been prepared on the tool including:

Annex K. Qatalog benefit analysis
Annex L. Qatalog feednack from one-to-one sessions
Annex M. Qatalog enhancements suggested by users

Considering the feedback from stakeholders during brainstorming sessions and meetings of the BDAC, UN Global Pulse proposed to build a platform to facilitate the analysis of big data content in Somali by any user with basic computer skills. The objective was to develop a tool where analysts could perform the analysis of FB and radio content by themselves. This development is additional to the deliverables initially agreed for the project.

Functionalities of the tool were defined in constant consultation with potential users from the UN in Somalia. Two UN Global Pulse analysts met on a weekly basis with colleagues in Nairobi and Kampala and collected their feedback in an automated way using Google Forms. Once a usable version of the tool was developed, two colleagues from the UN Assistance Mission in Somalia (UNSOM) tested it and provided their feedback during a working sessions in Kampala.

Overview of Qatalog

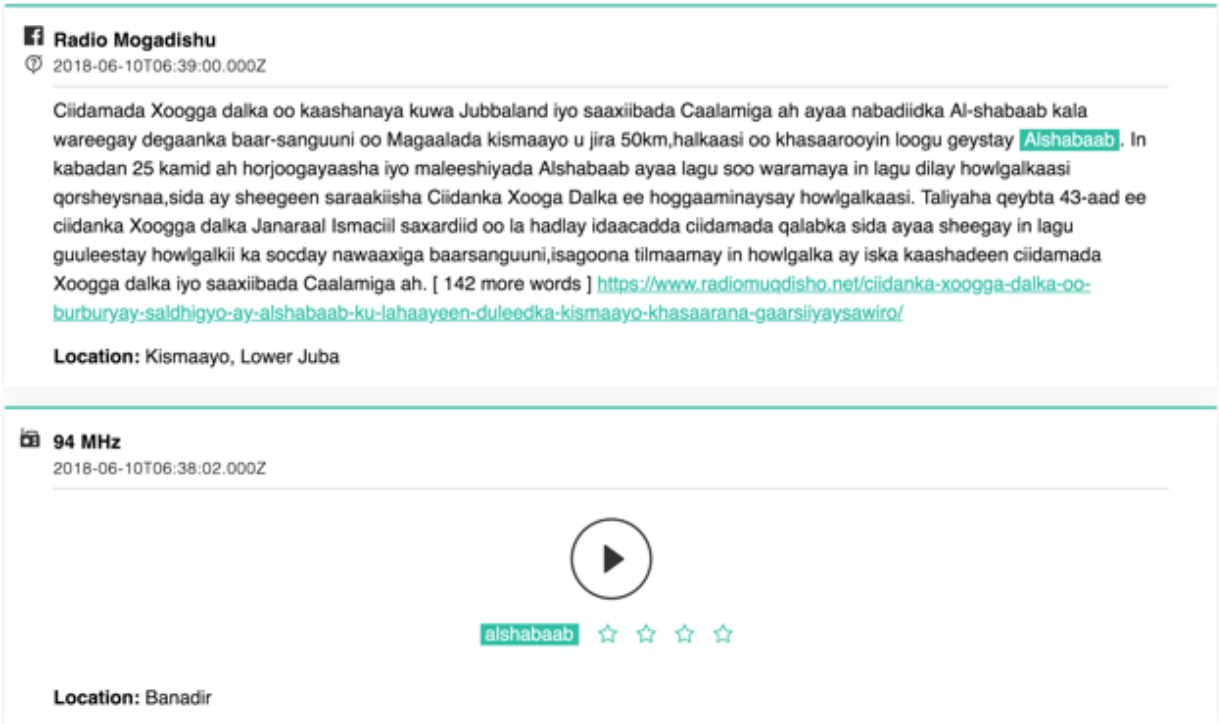
The technology prototypes developed for the real-time analysis of radio and FB content in Somali language were integrated in an easy-to-access tool. The platform can be used to gauge Somali public perceptions, opinions and reports without risk exposure to personnel.

The top 5 functionalities of the tool developed are to:

- Analysis of FB and radio content: users can specify keywords and a time frame of their interest to retrieve relevant content, and then assign user-defined tags to do

in-depth analysis of the content and generate statistics.

- Automatic translation of comments from Somali to English: an integration with Google Translate provides rough text translations.
- Dynamic identification of trending topics in FB: supports identifying popular topics discussed and guides users on the selection of keywords.
- Automatic geotagging of FB posts: it assigns a location tag to comments containing the name of a Somali location.
- Export raw data to a spreadsheet.



Screenshot of FB comment and radio clip in Qatalog

ACHIEVEMENTS ON DATA ANALYSIS

A documentation package has been prepared on the analysis performed, including:

Annex O. Analysis on the use of media in Somalia
Annex P. Evaluation of biases and potentialities of radio and FB analysis
Annex Q. List of radio stations and methodology to select them
Annex R. Schedule of radio stations
Annex S. List of FB pages and methodology to select them
Annex T. Keywords per topic of analysis

Prior to launching the social media and radio pilots UN Global Pulse assessed the types and frequency of media used in Somalia and the biases and potentialities of radio and FB. This was done to frame big data specifically for the Somali context.

RADIO

A complete list of radio stations that can be captured with reasonable quality in Mogadishu was compiled. 8 radio stations were targeted for analysis after a screening process that consisted of listening throughout several broadcasting hours to map program profiles of several radio stations. The stations with call-in programs and shows discussing political, socio-economic and cultural issues were identified as relevant.

A classification of radio programmes was generated after manually listening to almost all broadcasting hours of each of the 8 selected radio stations in Mogadishu. The classification was made into "Golden" (including concerns from local population), "Silver" (including studio discussion), "Bronze" (including news) or "Black" (not relevant for analysis).

Somali analysts did a qualitative analysis of radio content for each of the topics analysed throughout the project. A total of 7,500 clips were evaluated.

FB

A total of 2,500 public FB groups were targeted for analysis. The targeting was done with a software identifying keywords related to locations in Somalia (for example Mogadishu or Baidoa) included in the group name. In addition, hundreds of groups were selected manually and included in the selection for analysis.

Somali keywords were defined for each of the topic analysed in the project. For their definition, an expert on SDG16 and an expert in African languages guided the work of two full time Somali analysts.

DATA ANALYSIS RESULTS

A microsite with an overview of analysis results was created and is accessible here:
https://peacebuilding.unglobalpulse.net/Somalia

At the workshops in Mogadishu in January and November 2017 participants identified the "data gap" that the Big Data analysis aimed to address. The topics of analysis were identified by participants and guided the exploratory works conducted on public FB and radio content.

First set of analysis

First Big Data analysis results were shared during the second meeting of the Big Data Project Advisory Group. Results included:

- Example of quantitative and qualitative analysis from public FB posts.

- Example of qualitative analysis from public radio content (extracts from news and phone in).

Quantities of public FB posts on Somali pages mentioning AMISOM. The spike in messages on 2017-02-09 is in reaction to the decision of the president to use national bodyguards rather than AMISOM personnel, eliciting a largely positive reaction.

Second set of analysis

A second set of Big Data analysis results were shared with BDAC. The analysis was conducted around 3 topics namely AMISOM, drought and Al-Shabaab. Results included:

a) Quantitative and qualitative analysis of FB discussions. A total of 7,000 FB messages were analysed, of which 45% were translated to English.
b) Qualitative analysis of relevant radio content on the topics.

Third set of analysis

The analysis was conducted around 4 topics namely AMISOM, drought, Al-Shabaab and corruption/accountability. Results were shared with partners at the working session at Mogadishu in November 2017, and included:

a) Data mining results overview for FB and radio.
b) Trends analysis of FB discussions related to key events and real-time examples of radio clips from the last bomb attacks that took place in Mogadishu on October 2017.
c) Examples of in-depth analysis performed for the four topics under study, including a categorization of topics being discussed by the Somali population.

Fragments of public discussions on FB and radio after terrorist attack in Mogadishu on October 24th, 2017.

Fourth set of analysis

The fourth analysis was produced on corruption. The topic was suggested by the Special Representative of the Secretary General (SRSG) in Somalia, who also stated that would be a relevant topic for the Federal Government of Somalia (FGS). It included:

a) Analysis of influences in FB on the topic of corruption.
b) Qualitative analysis of radio content with examples of audio clips on the topic of corruption.

SCALING UP

Feedback from users on Qatalog tool and analysis results was very positive and UN Global Pulse is defining new functionalities for the tool with users. The Food and Agriculture Organisation (FAO), the International Organisation for Migration (IOM), the UN Development Program (UNDP), UNICEF, the UN Population Fund (UNFPA) and the United Nations Support Office in Somalia (UNSOS) are some of the agencies in Somalia already interested in using the tools developed with the pilot initiative.

Next steps involve adding functionalities to Qatalog and performing more in-depth case studies on topics to continue exploring the use of Big Data for peacekeeping. Engagement with the Federal Government of Somalia (FGS) will be pursued further during the scaling up phase.

A brochure on scaling up is available in Annex G.

LESSONS LEARNT

Output	Lesson learnt
Engagement with stakeholders	Project implementation can benefit with a project member working to support daily interact with project stakeholders.
Radio content analysis toolkit	*Hardware* o High deterioration due to proximity to the sea. The antenna was replaced after 1 year and anti-corrosion spray was used on a regular basis. o Continues power breakdowns affected the flow of analysis. Troubleshooting was needed at least once per month to ensure continuous data streaming. o International procuring needed for all hardware (no local availability). *Software* - Large number of phonemes are included in Somali language and some of them (like aspirated consonants), are not used in other African languages. Speech from other languages was used as training data for the Somali software, and since some of the phonemes are not shared, accuracy of the software was lower than expected.
FB analysis toolkit	o Frequent adaptations of the software were needed due to changes in the Application Programming Interface (API) by FB. o Deletion of comments by FB or users after some time might affect streaming analysis.
Data analysis	All Al-Shabaab owned radio stations stream outside of Mogadishu.

BOTTLENECKS OVERCAME

Bottleneck	Description
Delayed access to project funds	The project was signed on September 2016, funds were received by UNDP Somalia by mid-November 2016 and UNDP Somalia approved the budget in Atlas by late December 2016, as a consequence, fund were accessible for project implementation by January 2017.
Delayed recruitment of personnel	With the approval of TORs for project personnel by the BDAC in late January 2017, the recruitment process of personnel key to the implementation started. The process was completed in 5 months. To support this process Pulse Lab Kampala conducted 5 field mission of a week duration to Nairobi. To overcome further delays in hiring a translation company, UN Global Pulse consulted UN offices in Nairobi for over 1 month until a Long-Term Agreement (LTA) was found and taped into.
Challenges with radio hardware	The first installation of equipment (at UNSOM's premises) carried out in January 2017 allowed the reception of 3 radio stations and a total of 300 - 500 audio files per day each containing 5 minutes of speech.

Bottleneck	Description
	After the equipment was disconnected without notice to Pulse Lab Kampala, a mission to relocate and reinstall the equipment (at UNDP's premises) took place. The second installation site supports improved data uploads with approximately 700 audio files per day from 8 radio stations. The equipment was placed by UN Global Pulse in the roof of UNDP's conference room at MIA as no other space has been allocated in an office or server room by any partner. The location was not ideal and caused further delays for the following reasons: i) outdoors and exposed to dust and high humidity levels, what causes hardware malfunction; ii) unsecured so equipment could be misplaced; iii) not large enough to store additional IT equipment needed for functional improvement (as a screen monitor or UPS); iv) difficult to access for troubleshooting. In July 2017, basic IT part of the equipment – Raspberry Pi malfunctioned, and Global Pulse could not secure support on the ground to address this issue causing additional delays in the project implementation. The Researcher-Somali based in Mogadishu was trained with basic IT skills to provide support for basic issues related to the equipment. Further delays were caused when, in December 12th 2017, the equipment was disconnected from electricity after the power cable was unplugged at UNDP´s conference room. Its re-connection was only possible 7 days after, when technical support was received. The power disruption also caused the misconfiguration of the equipment (4 SD memory cards became corrupted). An official request was made in December 2017 to UNDP Somalia for the allocation of office space in the MIA compound in Mogadishu, to host the equipment. No support was secured. It was only in May 2018, when the equipment was relocated in a secured and indoor location thanks to the support of the United Nations Support Office in Somalia (UNSOS).
Delayed allocation of office space for Somali analyst	The Researcher - Somali based in Mogadishu joined the team in early August 2017. Space was only temporarily allocated in March 2018.