

Group Charlie

Multi-Lingual SMS

Andrew, Chihang, Daria, Swaraj and Tanvi

University of Cambridge

March 4, 2015

Background

- ▶ Africa's Voices: Opinion polling done by broadcasting a variety of questions from several radio stations, where answers were received as SMS messages.

Background

- ▶ Africa's Voices: Opinion polling done by broadcasting a variety of questions from several radio stations, where answers were received as SMS messages.
- ▶ Main problem: The SMS messages were in different languages and sometimes included irrelevant information, such as details of the sender.

Background

- ▶ Africa's Voices: Opinion polling done by broadcasting a variety of questions from several radio stations, where answers were received as SMS messages.
- ▶ Main problem: The SMS messages were in different languages and sometimes included irrelevant information, such as details of the sender.
- ▶ Sub-problem: Machine translation is not feasible for minority languages with little economic value. As a result of this, translation was taken to be an expensive task.

Background

- ▶ Africa's Voices: Opinion polling done by broadcasting a variety of questions from several radio stations, where answers were received as SMS messages.
- ▶ Main problem: The SMS messages were in different languages and sometimes included irrelevant information, such as details of the sender.
- ▶ Sub-problem: Machine translation is not feasible for minority languages with little economic value. As a result of this, translation was taken to be an expensive task.
- ▶ Problem we're trying to solve: allow for easy detection of opinion trends in such data, despite it being in several languages.

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.
- ▶ This is a tool for scientists — it should have a friendly and interactive UI while not getting in the way.

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.
- ▶ This is a tool for scientists — it should have a friendly and interactive UI while not getting in the way.
- ▶ Due to translation being expensive, we need to allow for the researchers to determine more important words/topics for which they can prioritise translation.

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.
- ▶ This is a tool for scientists — it should have a friendly and interactive UI while not getting in the way.
- ▶ Due to translation being expensive, we need to allow for the researchers to determine more important words/topics for which they can prioritise translation.

Additional requirements set along the way:

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.
- ▶ This is a tool for scientists — it should have a friendly and interactive UI while not getting in the way.
- ▶ Due to translation being expensive, we need to allow for the researchers to determine more important words/topics for which they can prioritise translation.

Additional requirements set along the way:

- ▶ Allow for identification of the language that a message was written in.

Our goals

- ▶ To create a visual browser that will allow international policy teams to observe trends.
- ▶ This is a tool for scientists — it should have a friendly and interactive UI while not getting in the way.
- ▶ Due to translation being expensive, we need to allow for the researchers to determine more important words/topics for which they can prioritise translation.

Additional requirements set along the way:

- ▶ Allow for identification of the language that a message was written in.
- ▶ Allow for 'cleaning up' of the data by removing names, txt-speak and other slang.

Overview of the project

The project has been implemented as a web application.

Overview of the project

The project has been implemented as a web application.
User input:

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
- ▶ Categories and respective keywords associated with them

Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
- ▶ Categories and respective keywords associated with them

Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
- ▶ Categories and respective keywords associated with them
Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

- ▶ Word cloud

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
 - ▶ Categories and respective keywords associated with them
- Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

- ▶ Word cloud
- ▶ Pie chart

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
 - ▶ Categories and respective keywords associated with them
- Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

- ▶ Word cloud
- ▶ Pie chart
- ▶ Bar graph

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
 - ▶ Categories and respective keywords associated with them
- Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

- ▶ Word cloud
- ▶ Pie chart
- ▶ Bar graph
- ▶ Table view

Overview of the project

The project has been implemented as a web application.

User input:

- ▶ CSV file containing relevant columns (country, radio station, message)
 - ▶ Categories and respective keywords associated with them
- Format — **Cat1**: KeyA, KeyB; **Cat2**: KeyC, KeyD, KeyE

Viewing 'modes' implemented:

- ▶ Word cloud
- ▶ Pie chart
- ▶ Bar graph
- ▶ Table view

A data manipulation page has also been added.

Word Cloud

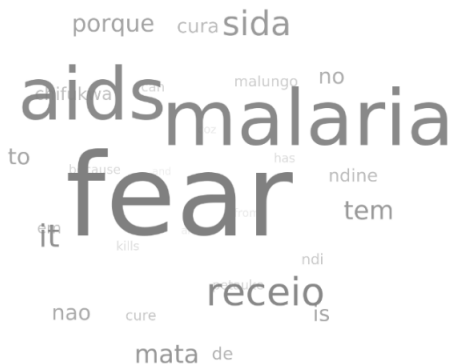


Figure: Word cloud — useful for viewing major keywords/concerns

Pie Chart

Overview of Opinions Regarding: Which disease do you fear more - AIDS or malaria?

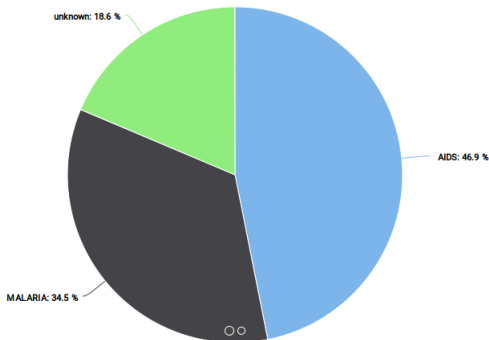


Figure: Pie chart — useful for viewing major opinions

Bar Graph

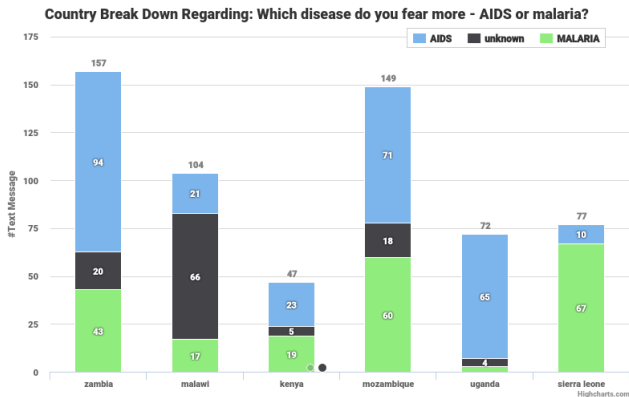


Figure: Bar graph — allows for breakdown by country

Table View

Show **100** entries

Search:

Country ▲	RStation ⚙	SMS	Opinion ⚙
kenya	nam lolwe	Fear! BOTH. Reasons, MALARIA, hehekei na mtu, AIDS, Yawa Ndilo..., Kinda kod yath chupa mopong, Asetho. Odhiambo Ogolo, kanyakoo- Asembo.	AIDS
kenya	nam lolwe	Jael kod Agangu to me i fear both ,First Malaria can kill within one day and second Aids has NO cure , Am Abenjo Kenna Ja Asembo Mabinju.	AIDS
kenya	nam lolwe	IFEAR AYAKI KOMAKI TO GIKO EN MANA THOO KAPONI INGENGIAN RV JACK OGAA WUOD NYORE.	unknown
kenya	nam lolwe	I fear both Malaria and HIV/AIDS MOSNA JAE AN OKELLO JAKABBLACK	AIDS
kenya	nam lolwe	Fear both malarial & ADS nikech kagimaki to gimoni tiyo tiji kaka pile kendo itiyi gi pesa mangeny eyor thieth and etc, to tho tochuno nyikech nosigi, an JALANGO OTIENO afuayo ne all Lolwe damu.	MALARIA
kenya	nam lolwe	Fear both malarial & ADS nikech kagimaki to gimoni tiyo tiji kaka pile kendo itiyi gi pesa mangeny eyor thieth and etc,	MALARIA
kenya	nam lolwe	FEAR;CHOLERA IT KILLS FASTER,WHEN IT HAPPENS IN THE VILLAGE IT BRINGS ALOT OF WORRY & HURRY,GEORGE WUOD CHIEF FRM RARIEDA.	unknown
kenya	nam lolwe	FIYAR AN JOHANA FUNDI KOMBE KIBUYE OKIMI RAS WAGOTICH GI FUNDE TE RAILA HALL KA	unknown
kenya	nam lolwe	Fear Hiv and aids since it has no cure	AIDS
kenya	nam lolwe	FEARE AIDS AN ATOTI OGOLA JAMANDAS FROM KAMARIGA ALUORE NE KECH OCHAMO DEL SPIDI MODHIERA NONO. OMIYO ALUORE MATHOTH SANA	AIDS
kenya	nam lolwe	Fired,HIV AIDS,because this is a disease that is no cured,also many pple knew that a persion with aid has already die.Abele from,gul bar kwamudhi.	AIDS
kenya	nam lolwe	I fear AIDS because it has no cure	AIDS
kenya	nam lolwe	Fear AIDS. It has no cure	AIDS

Figure: Table view — used to view individual SMS messages

Data Manipulation

Modifications Save Changes Visualise Download File

Show entries

Search:

	Country	RStation	Modified SMS	Original	Language	Opinion
<input type="checkbox"/>	zambia	breeze	<input type="text" value="fear aids malaria non of the two is better we have to fear both louis sinda"/>	fear aids malaria non of the two is better we have to fear both louis sinda	<input type="text" value="English"/>	<input type="text" value="AIDS"/>
<input type="checkbox"/>	zambia	breeze	<input type="text" value="fear accident bcz kils fast daniel phiri magazini"/>	fear accident bcz kils fast daniel phiri magazini	<input type="text" value="English"/>	<input type="text" value="unknown"/>
<input type="checkbox"/>	zambia	breeze	<input type="text" value="wopani maleya chikwa hiv nichadya chamuthu chapa paziko portpher sinda"/>	wopani maleya chikwa hiv nichadya chamuthu chapa paziko portpher sinda	<input type="text" value="Nyanja"/>	<input type="text" value="unknown"/>
<input type="checkbox"/>	malawi	mudzi wathu	<input type="text" value="ayi adzipiliza chifukwa nkhani ndi ya voti"/>	ayi adzipiliza chifukwa nkhani ndi ya voti	<input type="text" value="Chichewa"/>	<input type="text" value="unknown"/>
<input type="checkbox"/>	malawi	mudzi wathu	<input type="text" value="vote yanga atha kuimanso nkhani ndi vote basi onyamata omwa mbuna"/>	vote yanga atha kuimanso nkhani ndi vote basi onyamata omwa mbuna	<input type="text" value="Chichewa"/>	<input type="text" value="unknown"/>
<input type="checkbox"/>	malawi	mudzi wathu	<input type="text" value="kawili nikokwana kuwopa ukawalala palast allen"/>	kawili nikokwana kuwopa ukawalala palast allen	<input type="text" value="Chichewa"/>	<input type="text" value="unknown"/>

Figure: Data manipulation — used to modify/correct data

Technical details

Language module:

- ▶ Language identification for such languages is expensive, similar to machine translation.

Technical details

Language module:

- ▶ Language identification for such languages is expensive, similar to machine translation.
- ▶ Our solution: do frequency analysis. Works out much better in the long run.

Technical details

Language module:

- ▶ Language identification for such languages is expensive, similar to machine translation.
- ▶ Our solution: do frequency analysis. Works out much better in the long run.

Other details:

Technical details

Language module:

- ▶ Language identification for such languages is expensive, similar to machine translation.
- ▶ Our solution: do frequency analysis. Works out much better in the long run.

Other details:

- ▶ Backend: Django

Technical details

Language module:

- ▶ Language identification for such languages is expensive, similar to machine translation.
- ▶ Our solution: do frequency analysis. Works out much better in the long run.

Other details:

- ▶ Backend: Django
- ▶ Frontend: Bootstrap

Thanks for listening!