UMATI

Building an Intelligent Umati Monitor

*iHub Research

DISCOVERY . KNOWLEDGE . SHARING

July 2015

Contents

Background of the project[1]

Historically, election-monitoring efforts in Kenya have focused on scrutinising actions of the political class so as to ensure free, fair and peaceful elections. However, Kenya's 2007 elections, which escalated to the worst post-election violence in Kenya's history, greatly demonstrated the public's ability to mount conflict during an election period. Hate speech was noted as one of the key vehicles that promoted the 2007-8 Post Election Violence (PEV), in which over 1,200 people were killed and over 600,000 displaced from their homes.

A key example of a hate speech act from the 2007-8 Post-Election Violence period is by radio presenter Joshua Arap Sang, who through his morning show on Kass FM - a local radio station that broadcasts in the vernacular Kalenjin language- used code to communicate to his listeners where and when to commit attacks on the rival political party supporters.[2] Sang has since been accused of crimes against humanity by the International Criminal Court due to his alleged role in instigating mass violence through his utterances on his radio show.

Though there are few documented cases of hate speech that resulted in violence during the 2007-8 election period, the Kenyan government, through the National Cohesion and Integration Commission (NCIC), has since greatly increased its monitoring and prosecution of hate speech. This in turn resulted in an increased demand from the general public, peace-building organisations, politicians and government officials for how to define, identify, report and mitigate hate speech, especially given the vague definition present in the National Cohesion and Integration Act (NCIC) of 2008.
Under Section 13 of the National Cohesion and Integration Act of 2008, a person who uses speech (including words, programs, images or plays) that is

> "threatening, abusive or insulting or involves the use of threatening, abusive or insulting words or behaviour commits an offence if such person intends thereby to stir up ethnic hatred, or having regard to all the circumstances, ethnic hatred is likely to be stirred up".[3]

Notably, the Act mentions ethnic hatred to constitute racial, ethnic or national discrimination, but does not include hatred based on religion, gender, nationality, sexual preference, or any other category.

Other Kenyan laws also touch on hate speech: the 2010 Constitution notes that freedom of expression does not extend to hate speech, but does not define that term; while the Kenya's Code of Conduct for political parties (attached to the Political Parties Act) forbids parties to "advocate hatred that constitutes ethnic incitement, vilification of others or incitement to cause harm."

In response to the realised negative potential of hate speech and its contentious definition, Ushahidi teamed up with iHub Research to create Umati (Swahili for "crowd"), a media monitoring project that collects and analyses multilingual incidents of hate and dangerous speech from the Kenyan online space.

[1] Largely adapted from the Umati I Final Report Umati 2013 report (download link: http://research.ihub.co.ke/uploads/2013/june/1372415606___936.pdf)

[2] The Hague Academic Coalition, 'Joshua Arap Sang' in The Hague Justice Portal. Viewed on 10th June 2013, http://www.haguejusticeportal.net/index.php?id=12477.

[3] National Cohesion and Integration Act 2008 s. 13.

The Umati Project

Umati was launched in October 2012, six months before the Kenya general elections (held on March 4, 2013) and consists of two phases. Umati I was primarily an online monitoring project that collected and analysed hate and dangerous speech statements from the Kenyan online space around the 2013 general elections.

Umati II was initiated in July 2013 and is ongoing. Umati II's main goal is to build an intelligent tool that can perform the duties of Umati I, for projects outside Kenya, and beyond election periods.

The following section talks more about Umati I and II.

Umati I

Umati I ran for nine months, between September 2012 and May 2013. The project monitored particular blogs, forums, online newspapers and Facebook and Twitter content generated by Kenyans. Online content that was monitored includes tweets, status updates, comments, posts, forums, blogs, videos and pictures.

Apart from monitoring online content in English, a unique aspect of the Umati project was its focus on Kenya's ethnic languages. Kenya has over 42 tribes, each with a distinct and well developed language. Seven languages were monitored; Kikuyu, Luhya, Kalenjin and Luo, representing the four largest tribes in Kenya[4]; Swahili, the national language, and Sheng, an unofficial slang language; Somali, which is spoken by the largest immigrant community in Kenya; and English.

Umati I had four goals:

- To propose both a workable definition of hate speech and a contextualised methodology for online hate speech tracking, that can be replicated locally and in other countries.
- To collect and monitor the occurrence of dangerous speech in the Kenyan online space.
- To forward any distress calls the Umati team came across online to Uchaguzi (www.uchaguzi.com). Uchaguzi is a technology-based system for citizen reporting during elections. It further distributes information to relevant partners so that they may action them, e.g. distress calls to the Police.
- To further education on the possible outcomes of hate speech, so as to promote civil communication and interaction in both online and offline spaces.

Umati categorised hate speech incidents based on a framework by Professor Susan Benesch of American University. Benesch introduces the term dangerous speechwhich is defined as "speech that has a special potential to catalyse violence"[5] .

Benesch's breakdown of dangerous speech, its constituents and effects, enabled the Umati project to build a workable methodology for collecting and analysing hate and dangerous speech.

Findings from Umati I[6]

- 1 out of every 4 hate speech comments collected by Umati between October 2012 and May 2013, was a call to kill, beat or forcefully evict another group, or a person because of their belonging to a group. Yet, on the ground, the 2013 general elections were generally peaceful. The occurrence of online hate speech can therefore *not* be *solely* relied upon as a precursor to violence on the ground. Instead, online hate speech can serve as a glimpse of the vitriolic conversations Kenyans engage in offline, and thus surface existing tensions in the society.

- Most Kenyans prefer to converse in English, Kiswahili, Sheng or Kenyan English slang when online. There were very few hate speech statements purely in ethnic languages.

- KOT cuffing (Kenyans On Twitter cuffing) contributed to Umati

[4] Kenya National Bureau of Statistics 2009 Population Census

[5] http://www.dangerousspeech.org/guidelines

[6] Umati 2013 report (download link: http://research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf)

collecting only 3% of total hate speech comments from Twitter, while 90% were found on Facebook. Coined by Kagonya Awori, the Umati I project lead, 'KOT cuffing' refers to a phenomenon observed on Twitter, where tweets not acceptable to KOT are openly shunned, and the author of the tweets publicly ridiculed. The end result is that the 'offender' is forced to retract statements or even close his/her Twitter account altogether. Other factors contributed to more posts being collected from Facebook. These are discussed at length in the Umati 2013 report.

- There is a huge disparity between what the public perceives as hate speech and what the Umati project defines it as. From an exploratory survey conducted in May 2013, we found that the public perceives personal insults, propaganda and negative commentary about politicians as hate speech. The public's understanding of hate speech is also broader than the current constitutional definition, which only takes into consideration discrimination along tribal lines.

- Umati further categorised dangerous speech into three groups: offensive speech, moderately dangerous speech and extremely dangerous speech. Introducing a spectrum to the definition of dangerous speech was done in order to fit the Kenyan context.

Umati II

Umati II looks at employing Machine Learning (ML) and Natural Language Processing (NLP) techniques to detect, collect, select, and sort hate and dangerous speech from the Kenyan online space. The goal is to automate aspects of Umati I's process in order to increase the breadth and applicability of online hate speech monitoring.

Umati II was motivated in part, by the technical challenges experienced in Umati I. For one, the Umati monitors had to navigate between four to five software applications at a time, in order to collect and save hate speech statements. The software applications included:

- A Google Form which constitutes the Umati Categorisation form. (Discussed further in Section 3)
- Internet browsers for viewing multiple online pages simultaneously
- HyperTexts (www.hypertexts.no) for monitoring Facebook pages
- Open Status Search (www.openstatussearch.com/) for deep searches of Facebook status updates.
- Twitter Fall and Topsy for searching for and analysing tweets.

Additionally, Microsoft Excel was used to categorise and analyse Umati data, and Desktime (www.desktime.com) was used to track the daily productivity of each Umati monitor.

Use of multiple applications reduced monitors' productivity as they had to juggle between multiple browser windows, search tools and the Google Form to complete the collection of each hate speech statement.

Another issue was the high cost of scalability. A significant amount of time had to be spent recruiting, training and retraining the new team members, e.g. when a weekend team was added, or replacing a monitor. A financial cost was also incurred when the team size increased; a new iMac computer, desk, chair and pertinent software had to be purchased for each new team member.

Finally, due to the inherently dull tasks performed in Umati (that is visual search and signal detection), and the vitriolic nature of the speech being monitored, it was critical to maintain high morale in the monitors throughout the project period. This was effected by involving monitors in other more active events that included Umati media events, talks and roundtable discussions; granting up to three days leave every three months; and increasing work benefits such as remuneration, lunches and free counselling services. While these accommodations were fruitful, they did not solve drops in monitors' productivity, false alarms and correct misses in data collection, or having to replace or retrain monitors after 3 - 5 months on the job.

These and other reasons led to Umati II's goals:

1. To refine the Umati methodology developed in Phase I.
2. To increase scalability of the project through automation.
3. To test the Umati methodology in situations outside of elections, and in other countries, in order to improve its contextual applicability and global deployment.
4. To explore non-punitive, citizen-centred approaches for reducing dangerous speech online.

This report focuses mainly on Umati II's first two goals. We seek to augment the Umati project from a manual to an automated process through Machine Learning and Natural Language Processing techniques.

This report will discuss the Umati Methodology, and thus lay the foundation for the Intelligent Umati Monitor we are building. We will then outline the components of the Intelligent Tool we seek to build, and our current status in the process.

Umati Methodology

Umati I primarily utilised a manual data collection and analysis process. For eight hours a day, each monitor manually scanned online platforms for incidents of hate and dangerous speech. Each statement was then pasted into an online form where several questions had to be answered about it, and then later sorted in one of the three speech categories.

Figure 1 illustrates the Umati I process:

1 Manually scan online spaces

Selection criteria

1. Discrimates a group of people or a person because of their belonging to a group, and;
2. Contains one of the hallmarks of dangerous speech.

2 Remove the noise

OR

1. Discrimates a group of people or a person because of their belonging to a group, and;
2. Contains a call to violent action.

Paste the statement into the Umati Categorisation Form, and provide more info about the statement 3

4 All statements are stored on a shared database

5 Group statements according to Umati Categorisation Formula

N1 + M1 =
Bucket 1
N1 + M2 =
Bucket 1
N1 + M3 =
Bucket 2
N2 + M1 =
Bucket 2
N2 + M2 =
Bucket 2
N2 + M3 =
Bucket 3
N3 + M1 =
Bucket 3
N3 + M2 =
Bucket 3

6 Make sense of the data

7 Disseminate outputs

The above manual process was used in Umati I and collected over 7,000 hate and dangerous speech statements over 8 months. The main goal of Umati II is to build an intelligent tool that can augment data collection, categorisation, analysis and dissemination.

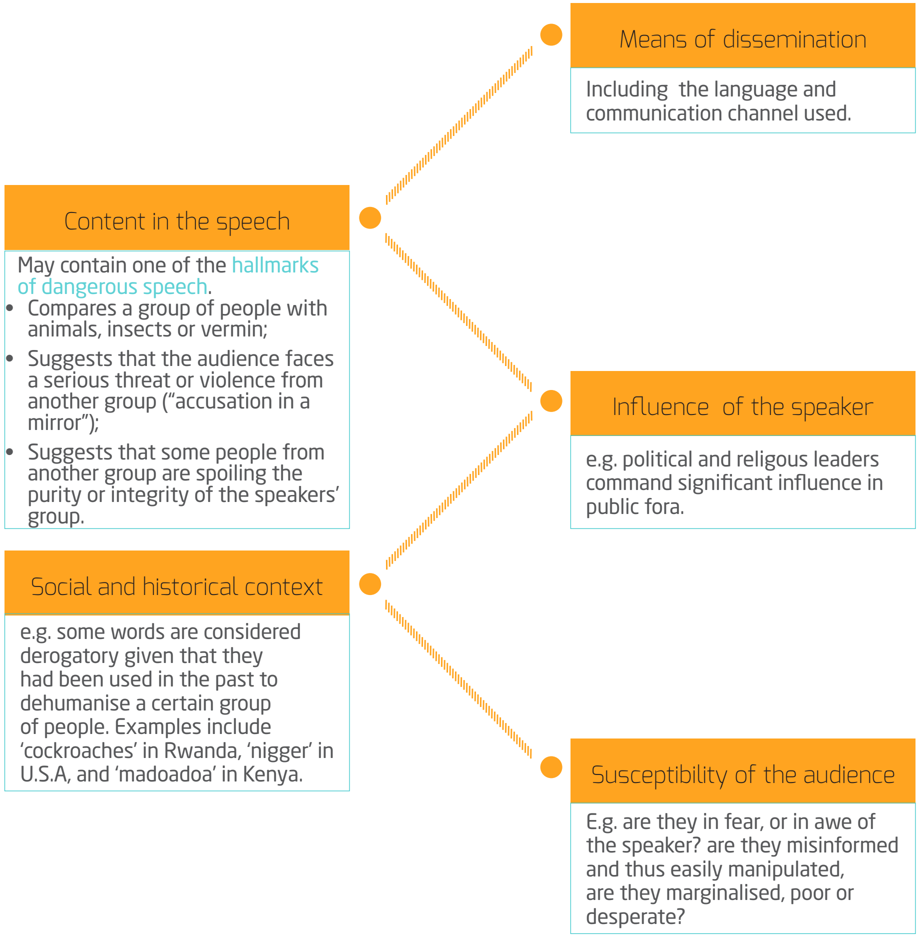
The following sections expound on the Umati Form, the Umati Categorisation Formula, and the tools that constitute the Umati Intelligent Monitor.

The Umati Categorisation Form

Background of the Umati Categorisation Form

Umati developed a form to capture meta-data about each hate speech statement collected by the monitors. The meta data is based on the five components of Dangerous Speech devised by Benesch[7].

According to Benesch, in order to determine that a speech statement is dangerous, one needs to examine not only the content of the speech act, but all of the following five criteria:



Means of dissemination

Including the language and communication channel used.

Content in the speech

May contain one of the hallmarks of dangerous speech.

- Compares a group of people with animals, insects or vermin;
- Suggests that the audience faces a serious threat or violence from another group ("accusation in a mirror");
- Suggests that some people from another group are spoiling the purity or integrity of the speakers' group.

Influence of the speaker

e.g. political and religous leaders command significant influence in public fora.

Social and historical context

e.g. some words are considered derogatory given that they had been used in the past to dehumanise a certain group of people. Examples include 'cockroaches' in Rwanda, 'nigger' in U.S.A, and 'madoadoa' in Kenya.

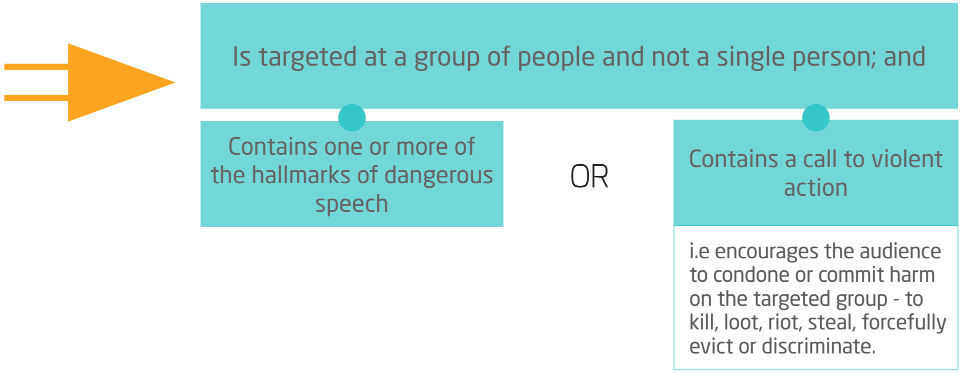
Susceptibility of the audience

E.g. are they in fear, or in awe of the speaker? are they misinformed and thus easily manipulated, are they marginalised, poor or desperate?

To note here is that these five criteria play a variable role in catalysing an audience to violence. For example, we posited that a significant number of those participating in dangerous speech lived in urban, muliti-ethnic areas and were educated. Therefore, the factors that catalysed them to violence may not be that they were misinformed or in fear of the speaker, but perhaps that the social and historical context of the speech statement provoked them to violence.

Also note that violence need not be on the ground for it to constitute violence; generating, condoning and disseminating dangerous speech online can also be deemed as participating in violence.

Based on the five criteria, we came up with the **selection criteria** below, that could be used by both Umati (as outlined in Step 2 in the Umati process above) and the public to easily identify dangerous speech.
A dangerous speech statement :



Is targeted at a group of people and not a single person; and

Contains one or more of the hallmarks of dangerous speech

OR

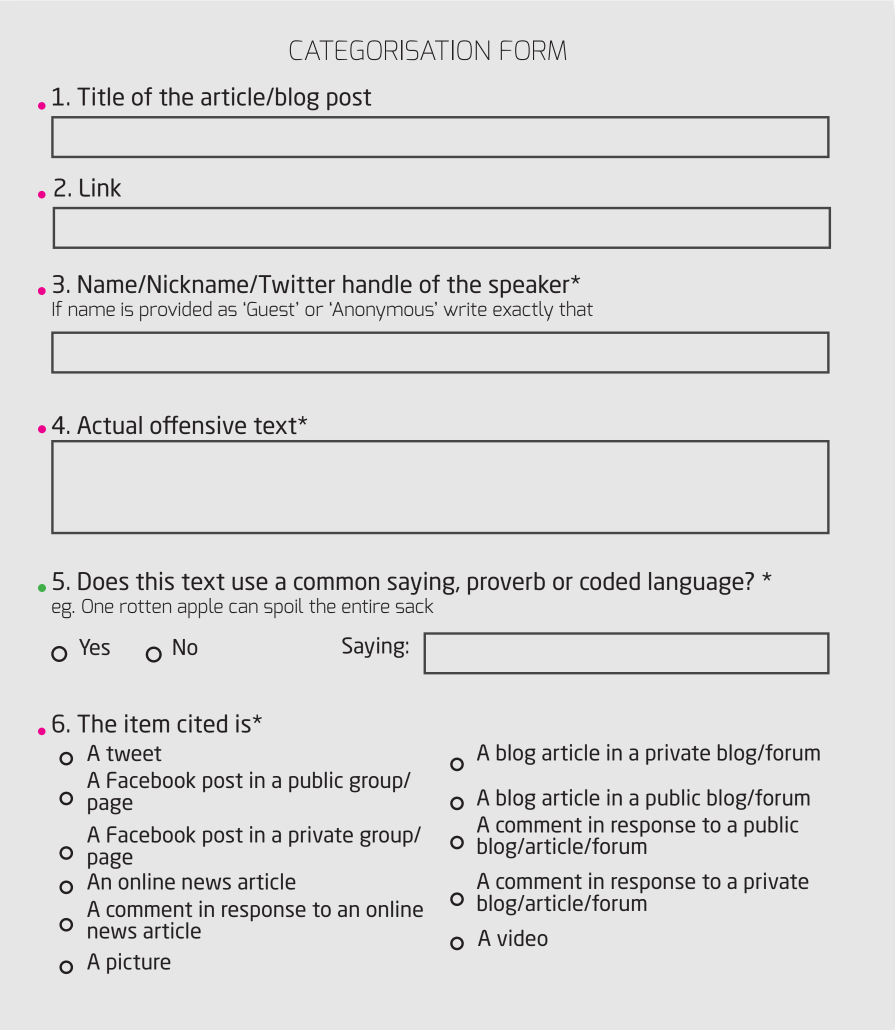
Contains a call to violent action

i.e encourages the audience to condone or commit harm on the targeted group - to kill, loot, riot, steal, forcefully evict or discriminate.

[7] S. Benesch, 'Dangerous Speech: A Proposal to Prevent Group Violence'. 23 February 2013. Viewed on 21st May 2013, http://www.dangerousspeech.org/guidelines

Umati Categorisation Form

Once a potentially dangerous speech statement was identified, Umati monitors used this form to collect information about the statement. The Umati Categorisation Form was designed based on Benesch's five factors, and further customised to collect more granular details that were relevant to the Kenyan context at the time.



CATEGORISATION FORM

1. Title of the article/blog post

2. Link

3. Name/Nickname/Twitter handle of the speaker*
If name is provided as 'Guest' or 'Anonymous' write exactly that

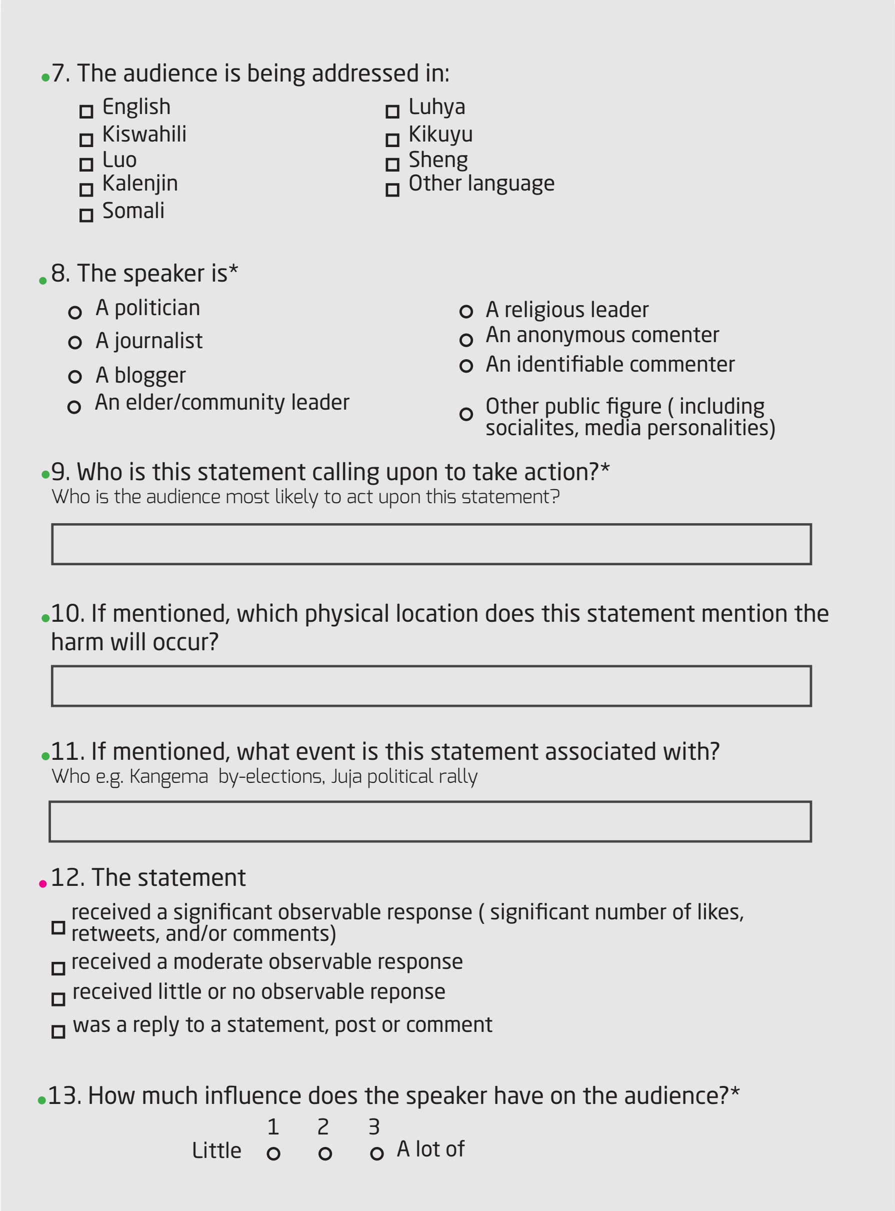
4. Actual offensive text*

5. Does this text use a common saying, proverb or coded language? *
eg. One rotten apple can spoil the entire sack

○ Yes ○ No Saying:

6. The item cited is*

○ A tweet
○ A Facebook post in a public group/page
○ A Facebook post in a private group/page
○ An online news article
○ A comment in response to an online news article
○ A picture
○ A blog article in a private blog/forum
○ A blog article in a public blog/forum
○ A comment in response to a public blog/article/forum
○ A comment in response to a private blog/article/forum
○ A video



7. The audience is being addressed in:

□ English
□ Kiswahili
□ Luo
□ Kalenjin
□ Somali
□ Luhya
□ Kikuyu
□ Sheng
□ Other language

8. The speaker is*

○ A politician
○ A journalist
○ A blogger
○ An elder/community leader
○ A religious leader
○ An anonymous comenter
○ An identifiable commenter
○ Other public figure (including socialites, media personalities)

9. Who is this statement calling upon to take action?*
Who is the audience most likely to act upon this statement?

10. If mentioned, which physical location does this statement mention the harm will occur?

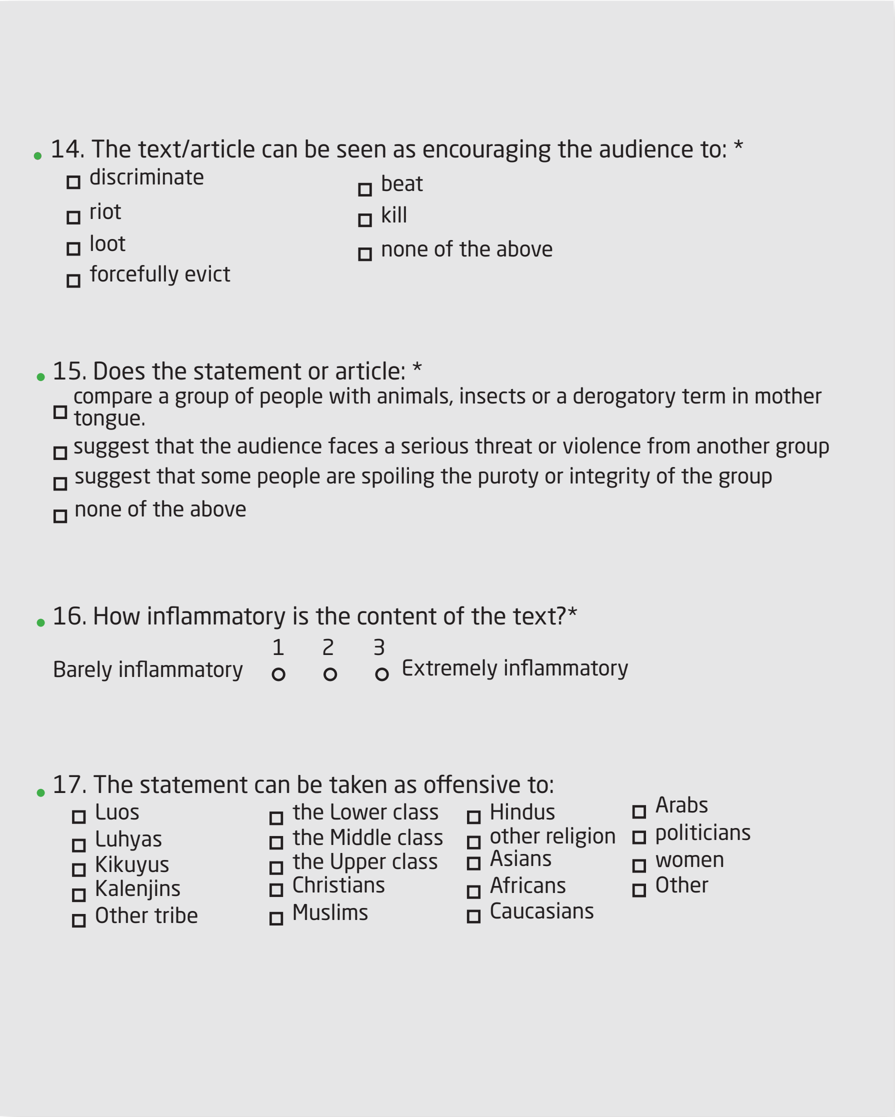
11. If mentioned, what event is this statement associated with?
Who e.g. Kangema by-elections, Juja political rally

12. The statement

□ received a significant observable response (significant number of likes, retweets, and/or comments)
□ received a moderate observable response
□ received little or no observable reponse
□ was a reply to a statement, post or comment

13. How much influence does the speaker have on the audience?*

	1	2	3	
Little	○	○	○	A lot of



● 14. The text/article can be seen as encouraging the audience to: *

□ discriminate
□ riot
□ loot
□ forcefully evict
□ beat
□ kill
□ none of the above

● 15. Does the statement or article: *

□ compare a group of people with animals, insects or a derogatory term in mother tongue.
□ suggest that the audience faces a serious threat or violence from another group
□ suggest that some people are spoiling the puroty or integrity of the group
□ none of the above

● 16. How inflammatory is the content of the text?*

	1	2	3	
Barely inflammatory	○	○	○	Extremely inflammatory

● 17. The statement can be taken as offensive to:

□ Luos
□ Luhyas
□ Kikuyus
□ Kalenjins
□ Other tribe
□ the Lower class
□ the Middle class
□ the Upper class
□ Christians
□ Muslims
□ Hindus
□ other religion
□ Asians
□ Africans
□ Caucasians
□ Arabs
□ politicians
□ women
□ Other

● In Umati II, we noted that answers to these questions can be collected automatically by the Intelligent Umati Monitor as opposed to the human monitor. We discuss this further in our Next Steps section.

● These questions still have to be answered by the human monitor. We discuss this further in our Next Steps section.

Questions were added to the form or revised to suit the evolving needs of the project. We have also revised it further in this report.

Albeit a lengthy form to fill for each hate speech statement collected, the varied questions allowed for the collection of richer meta-data, that proved useful for varied levels of analyses later in the project.

Hate vs Dangerous Speech

We note the contentious use and definition of the terms 'Hate speech' and 'Dangerous speech'. This project defines the latter as a subset of the former. Consequently, Umati focuses on collecting and analysing dangerous speech given that it is the subset of hate speech with the highest potential to catalyse violence.

However, in common parlance, the term used to refer to vitriolic speech is hate speech. Thus, while Umati is strictly a dangerous speech monitoring project, we sometimes refer to it as a 'hate speech monitoring project' or 'hate *and* dangerous speech monitoring project' in order to resonate with the public.

Categorising Collected Data

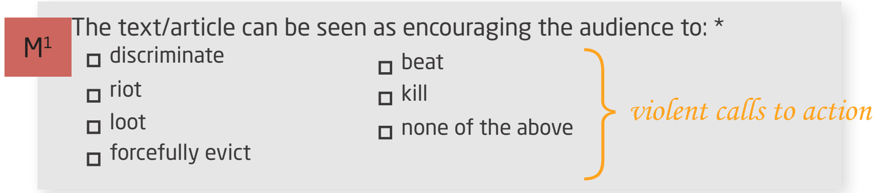
Umati further categorised the dangerous speech statement it collected into three groups: offensive speech, moderately dangerous speech and extremely dangerous speech. We introduced this extra level of categorisation in order to fit the Benesch framework into the Kenyan context.

Through Umati I, we determined that hate/dangerous speech cannot be viewed as a simple dichotomy, i.e. that a statement is either dangerous speech or not dangerous. Instead, it would be more befitting to view incitement as falling over a spectrum. Thus, the categorisation of a speech act is determined not only by the content of the speech act, but also on the influence the speaker has over the audience, and and by gauging whether the speech act was understood as a call to violence. Consequently, the influence and how inflammatory a speech act is, varies. We thus applied weightings to these variables and used these to categorise dangerous speech statements along the spectrum. We picked three points of the spectrum as the three dangerous speech categories, from least to most acerbic : offensive speech, moderately dangerous speech and extremely dangerous speech.

In practice, we formulated an algorithm from some of the questions on the Umati Form in order to categorise speech incidents into the three buckets. The questions we added to the form were aimed at measuring the influence of the speaker on the audience, the susceptibility of the audience and how inflammatory the content of the statement is.

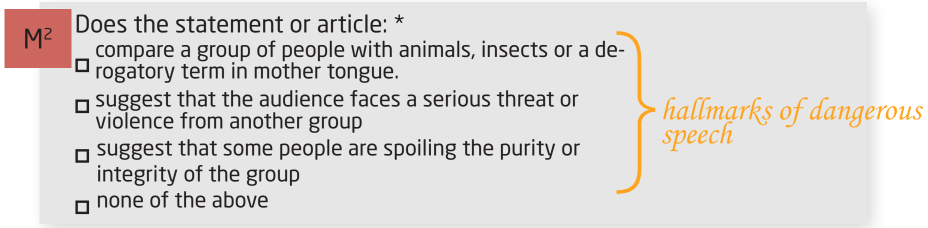
In the next section we elaborate on which questions we used from the form, and how we used them to come up with the Umati algorithm.

Firstly, to determine the how inflammatory the content of the speech act is, we used these two questions, M^1 and M^2 below, to determine which violent calls to action the speech statement contained, if any, and which hallmarks of dangerous speech it contained, if any.



M^1 The text/article can be seen as encouraging the audience to: *

□ discriminate
□ riot
□ loot
□ forcefully evict
□ beat
□ kill
□ none of the above

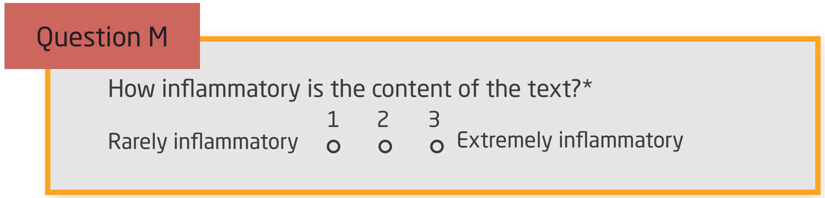


M^2 Does the statement or article: *

□ compare a group of people with animals, insects or a derogatory term in mother tongue.
□ suggest that the audience faces a serious threat or violence from another group
□ suggest that some people are spoiling the purity or integrity of the group
□ none of the above

The two questions assign different weightings to the calls to action and hallmarks e.g. calls to kill, beat and forcefully evict have a higher weighting than calls to discriminate. The guide below explains how we assigned the weightings to the question **M** based on M^1 and M^2.

1. In M^1, if ‘Discriminate’, M can be 1 or 2.
2. In M^1, if ‘Riot’ and/or ‘Loot’, M can be 2 or 3.
3. In M^1, anytime ‘forcefully evict’, ‘beat’ or ‘kill’ is selected, M is 3.
4. In M^2, If “Suggest that the audience faces a serious threat or violence from another group “ and/or “Suggest that some people are spoiling the purity or integrity of another group “ are selected, M is 3.



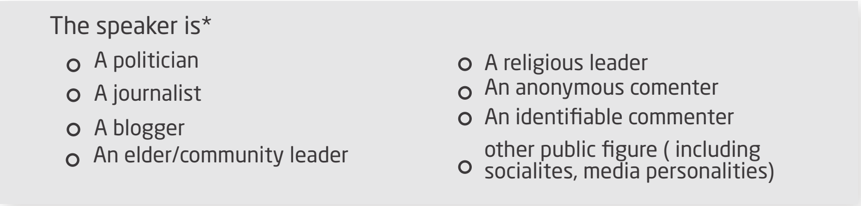
Question M

How inflammatory is the content of the text?*

Rarely inflammatory 1 ○ 2 ○ 3 ○ Extremely inflammatory

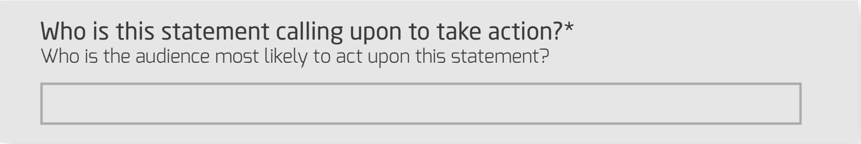
The answer to this question above is coded as M. Thus if 1 is selected, M=1 and consequently the M = M1. If 2 is selected, M= M2.

Then, to measure the influence of the speaker and the susceptibility of the audience, these three questions were used.



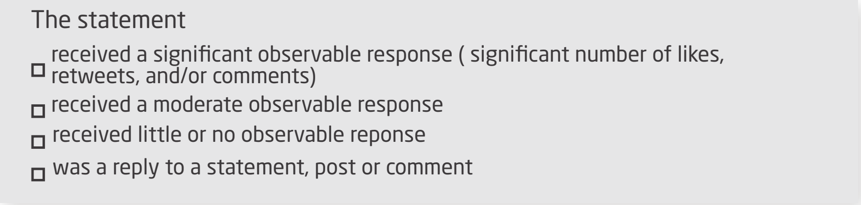
The speaker is*

○ A politician
○ A journalist
○ A blogger
○ An elder/community leader
○ A religious leader
○ An anonymous comenter
○ An identifiable commenter
○ other public figure (including socialites, media personalities)



Who is this statement calling upon to take action?*

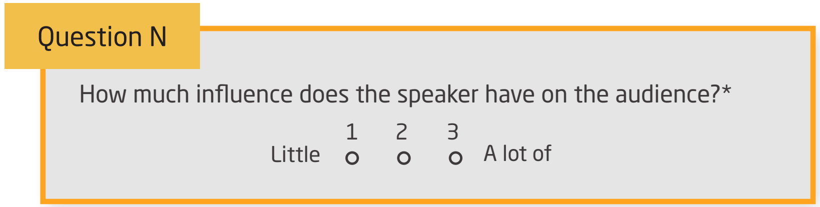
Who is the audience most likely to act upon this statement?



The statement

□ received a significant observable response (significant number of likes, retweets, and/or comments)
□ received a moderate observable response
□ received little or no observable reponse
□ was a reply to a statement, post or comment

The answers to these three questions enabled the Umati monitor to measure influence through the question below. We coded this question **N**.



Question N

How much influence does the speaker have on the audience?*

	1	2	3	
Little	○	○	○	A lot of

If the influence is little, then N=1 and N = N1.
If 'A lot of', then N=3 and N = N3.

Finally, depending on the answers from the questions M and N, the following sorting formula is applied to the speech statements.

SORTING FORMULA
M1 + N1 = Bucket 1
M1 + N2 = Bucket 1
M1 +N3 = Bucket 2
M2 + N1 = Bucket 2
M2 + N2 = Bucket 2
M2 + N3 = Bucket 3
M3 + N1 = Bucket 3
M3 + N2 = Bucket 3
M3 + N3 = Bucket 3

HATE SPEECH CATEGORIES[1]
Bucket 1 = Offensive Speech
Bucket 2 = Moderately Dangerous speech
Bucket 3 = Extremely Dangerous speech

All statements are then grouped into the three dangerous speech categories. This allows for richer data analysis and generation of information outputs that are relevant to Umati's partners and stakeholders.

By answering questions M and N, we have attempted to quantitatively consider three of the five criteria for dangerous speech when categorising dangerous speech along the spectrum. The three criteria we considered are : the content of the speech act (Question M), the social and historical context in the content of the speech act (Question M) and the influence the speaker has on the online audience (Question N).

[1] In the Umati I report (http://research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf) we defined these categories as Offensive, Moderately Dangerous and Dangerous Speech leading to some ambiguity between the sub-category and the category. We have thus revised these categories into Offensive, Moderately Dangerous and Extremely Dangerous Speech.

The Intelligent Umati Monitor

As stated earlier, the main goal of Umati II is to augment the Umati project from a manual to an automated process through Machine Learning and Natural Language Processing techniques.

The diagram below illustrates the tools Umati II seeks to build, and how they will augment the manual Umati I process we have discussed in the previous section of this report.

1 Scan online spaces

Tool: Trawler
Goal: To continually search through a specified list of online pages, user accounts, groups and forums.
Current Process: We have built a tool that integrates data collection from Facebook, Twitter and Disqus.
The intention is that the search runs continually, and thus able to capture data during unexpected events e.g riots and clashes.
Current Status: In use
Future Work: To collect data from other sources.

2 Remove the noise

Tool: Sieve
Goal: To reduce the collected statements to only potentially dangerous speech statements.
Tasks: Apply the selection criteria to sieve out the noise and thus remain with potentially dangerous speech statements.
Current Process:
The sieve uses a Regex tool to identify which collected statements contain potentially discriminatory words against groups of people, eg against a particular tribe, gender, race etc.
Current Status: Incomplete
Future Work: To improve the accuracy of the Regex tool.

3 Paste the statement into the Umati Categorisation Form, and provide more info about the statement

Tool: Tagger
Goal: To collect meta-data (as per the Categorisation Form) for each dangerous speech statement selected in Step 2.
Tasks: This is one of the most complex steps for automation. Some meta-data is easy to collect, e.g. date, location and speaker name, while other meta-data will result from additional sub-processes, a robust Machine Learning algorithm and primarily, human input.
As illustrated on the Categorisation Form, the pink dots will be tagged by the tool, while the green dots will be tagged by human monitors.
These tags will also be used to improve the Sieve's accuracy. Tagging by human monitors is mandatory in this step, and also necessary for the training the Sieve.
Current Status: Only able to tag data from Twitter as either true or false.
Future Work: Reduce both the human monitor's load and dullness of the tagging task.

All statements are stored on shared database 4

Tool: Online Database
Goal: Provide authorised access to authenticated Umati staff and partners.
Current process: Use current off-the-shelf platforms to store data on the cloud.
Current Status: In use

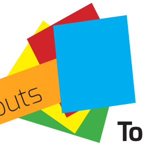
N1 + M1 =
Bucket 1
N1 + M2 =
Bucket 1
N1 + M3 =
Bucket 2
N2 + M1 =
Bucket 2
N2 + M2 =
Bucket 2
N2 + M3 =
Bucket 3
N3 + M1 =
Bucket 3
N3 + M2 =
Bucket 3

5 Group statements according to Umati Categorisation Formula

Tool: Bucket
Goal: Categorise data into the three dangerous speech buckets.
Process: The Categorisation formula will be applied by the tool to sort data into the buckets.
Current Status: Incomplete

Make sense of the data 6

Tool: Report Generator
Goal: Execute data analysis techniques in order to glean meaningful information.
Current Process: Applied Sentiment Analysis techniques, specifically through an external API known as Indico.io. This has enabled us to track Twitter for the general sentiment around events on the ground e.g. reactions to the gunning down of controversial cleric, Makaburi[1] and the attack in Mpeketoni[2].
Current Status: In use
Future Work: To accommodate other outputs e.g. reports, graphs and models.



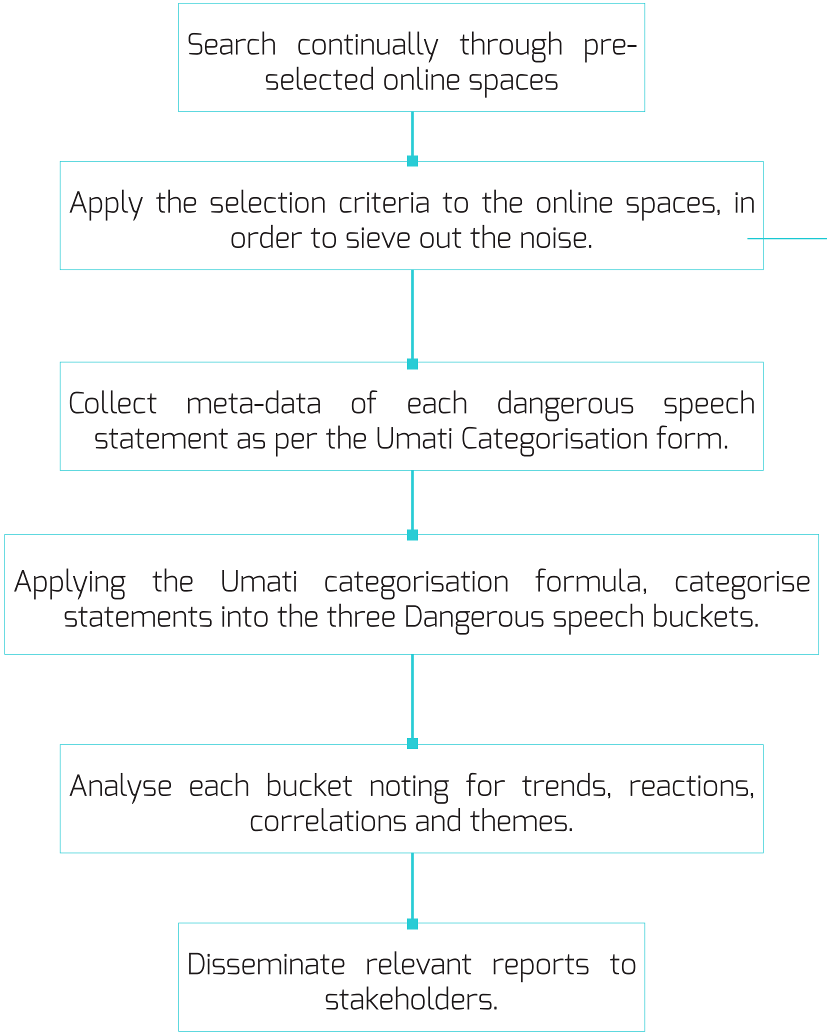
7 Disseminate outputs

Tool: Messenger
Goal: Disseminate relevant information on Umati to stakeholders and the public, including sensitising the public on dangerous speech.
Current Process: In Umati I, the Uchaguzi platform and email groups were used to connect with partners and relay calls for help in a timely manner.
Currently, reports are made publicly available on the iHub website and through media forums to relevant stakeholders.
Status: In use

[1] http://www.bbc.com/news/world-africa-26958455

[2] http://www.bbc.com/news/world-africa-27862510

We illustrate this same process as an algorithm in the flowchart below.



Search continually through pre-selected online spaces

Apply the selection criteria to the online spaces, in order to sieve out the noise.

Collect meta-data of each dangerous speech statement as per the Umati Categorisation form.

Applying the Umati categorisation formula, categorise statements into the three Dangerous speech buckets.

Analyse each bucket noting for trends, reactions, correlations and themes.

Disseminate relevant reports to stakeholders.

The selection criteria consists of three components.
The statement:

1. is targeted at a group of people and not a single person
2. contains a call to violent action
3. contains one or more of the hallmarks of dangerous speech

However, we foresee that our proposed Intelligent Umati Monitor will be limited in being able to determine whether the statement contains one of the hallmarks of dangerous speech.

At this stage of the process therefore, a human monitor is needed to determine whether the statements collected do have the hallmarks. The Intelligent Umati Monitor will thus sieve dangerous speech from predefined online locations, based on criteria 1 and 2 above.

Next Steps

The Umati project, in this current phase and with generous funding from the MacArthur Foundation, will run until December 2015. During this time, our focus is on building the Intelligent Umati Monitor and streamlining the analysis and synthesis of the data we collect.

We anticipate that the main benefit of the Intelligent Umati Monitor is that it will augment the manual dangerous speech monitoring process we used in Umati I, to an extent that it will be cheaper, more robust and more effective to run online monitoring over longer periods of time. It is also intended that this tool will be used by other countries and organisations, including our existing partners, to further augment their projects.

We are also keen on using the Intelligent Umati Monitor to address particular challenges we faced in Umati I, a key one being the unavoidable data collection inconsistencies that plague projects that rely extensively on human monitoring. A specific challenge is in gauging the influence of the speaker and how inflammatory a speech content is, as per our Umati process, and in determining whether a speech statement contains any of the hallmarks of dangerous speech as described on page 9. In Umati I, we noted that these inconsistencies led to the incorrect categorisation of speech statements. For example, we found several statements with calls to forcefully evict, in the Offensive Speech bucket. Correcting this required us to comb through the data again, and attempt to fix the inconsistencies. This exercise was costly to the project and made us aware that a large margin of error lies in the application of the Umati process. Collecting, categorising and analysing dangerous speech is more an art than an exact science. Nonetheless, the Umati Intelligent Monitor is seen as a solution to this challenge.

What is important to state here however, is that the Intelligent Monitor cannot replace the human monitor, but will instead assist the human monitor in decision making. In other words, the Intelligent Monitor will primarily assist in collecting potentially dangerous speech statements, collecting the meta-data about these statements, and provide this information to the human monitor, who will then use this information to classify and analyse the speech statements.The human-tool collaboration is especially evident in the Tagger tool we described earlier. The Intelligent tool will collect some of the necessary meta-data, those marked with pink dots on the Umati Form, and the human monitor will generate the rest. For example, the social and historical contexts of a country are best understood by those who live in the country and understand existing nuances, norms and dynamics. Thus, the Intelligent Umati Monitor may, at this point, not be able to discern whether a speech statement is discriminatory yet it does not contain any acerbic words. Answers to such questions (with green dots on the Umati form) will be added by human monitors.

The Intelligent Umati Monitor is therefore being built to support the human monitor in collecting, monitoring and analysing dangerous speech.

Another component of Umati is that it is part of a larger process. While the Umati project focuses on collecting, monitoring and analysing dangerous speech, the outputs from the project demand consideration as they are disseminated to various stakeholders. Therefore, apart from the technical aspects of the project, it remains important to engage partners in analysing and acting upon the outputs of Umati. For example, Umati partnered with Uchaguzi during the Kenya 2013 General Election, in order to forward all calls of help it came across online, to respective security bodies.

Currently, we periodically share snapshot analyses with various stakeholders, including the public via news articles, as we continue to analyse the massive data sets collected through the automated process. This has proven important to engage the public in discourse around the different ebbs and flows of dangerous speech as events take place.

As the project scales to different countries, we are also gaining massive insights on the applicability of the methodology around elections. The project is currently running in Nigeria, in conjunction with CITAD (Center for Information and Technology Development], who are conflict and peace building practitioners who saw the importance of monitoring online speech to inform their interventions. The Umati methodology has also been adopted to monitor online discussions around the 2015 elections in Ethiopia. Interest in adopting the methodology continues to be expressed by practitioners and researchers in different countries. In Kenya, we are currently assessing the relevance of the methodology and dangerous speech monitoring beyond the election period as other events such as unfortunate terror attacks (and their reactions online) unfold.

In conclusion, we aim to complete the Intelligent Umati Monitor by the close of the year. The tools created will be made available in Open Source, and the code created thus far is available on Github (https://github.com/iHub/UmatiCodebase). As for the data we collect, while it will not be available to the public, it will be open for use by data scientists, social scientists and other practitioners, for discourse analysis beyond dangerous speech monitoring. Access to and use of Umati data by external partners will be bound by our honour code.

We will also devise a toolkit with step by step considerations for replicating Umati, and we intend that this will receive input from other deployments of online speech monitoring around the world.

In that vein, we invite partners to collaborate with us in building the Intelligent Umati Monitor or analysing the Umati process. If interested in collaborating with us, do email us on umati@ihub.co.ke for more information.

The latest outputs from the Umati project are available on the iHub website: www.ihub.co.ke/umati.

UMATI

*iHub Research

DISCOVERY . KNOWLEDGE . SHARING

Written and compiled by:
Kagonya Awori - Umati I Lead, Umati Advisor

Contributors:
Nanjira Sambuli - Umati Lead
Sidney Ochieng' - Umati Data analyst/programmer
Chris Orwa - Umati Data analyst/programmer

Interim Research Lead
Leo Mutuku

For more information on this project, contact

Umati Project Team
iHub Research
Nairobi, Kenya

umati@ihub.co.ke
www.research.ihub.co.ke | Twitter: @iHubResearch