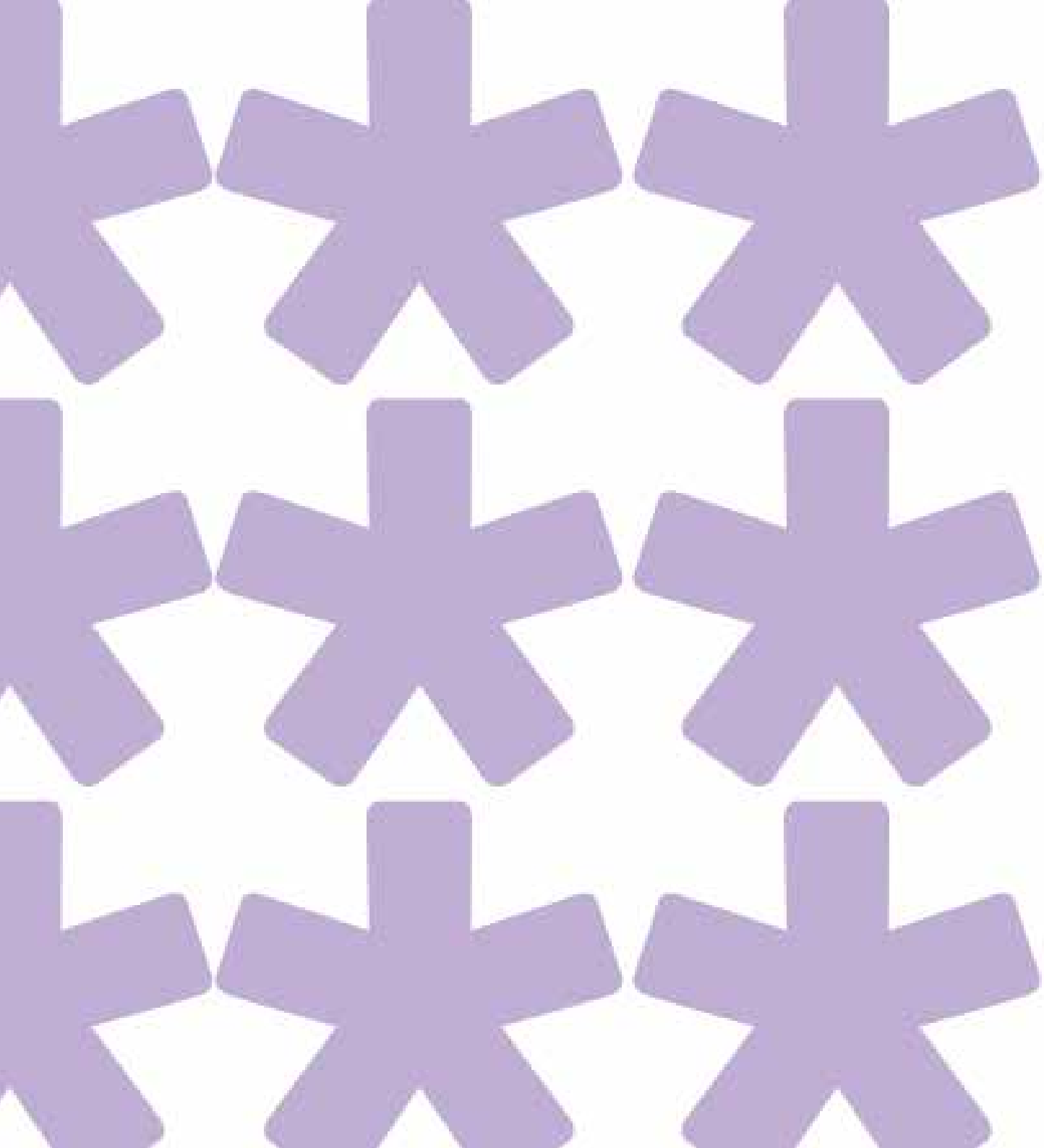
Umati Final Report

Sept 2012 - May 2013

Contents



*Background

Historically, election-monitoring efforts in Kenya have focused on scrutinizing actions of the political class so as to ensure free, fair and peaceful elections. However, the 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 avenues through which the public 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.[1] Sang has since been accused of crimes against humanity by the International Criminal Court due to his role in instigating mass violence through his utterances on his radio show.

Though, apart from Sang's, 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 Committee (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".[2] Notably, the Act mentions ethnic hatred to constitute racial, ethnic or national discrimination – and does not include hatred based on religion, gender, nationality, sexual preference, or any other group 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 realized 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 examines multi-lingual incidences of hate and dangerous speech in the Kenyan online space.

The goals of the Umati project were:

1. 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.

1 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.

2 National Cohesion and Integration Act 2008 s. 13.

2 To collect and monitor the occurrence of hate speech in the Kenyan online space.

3 To forward any distress calls the Umati team came across online, e.g. on Twitter and Facebook, to Uchaguzi (www.uchaguzi.com). Uchaguzi is a technology-based system that enables citizens to report and keep an eye on election-related events on the ground.

4 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 ran for nine months from September 2012 to end of 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, blog entries, videos and pictures. Apart from monitoring online content in English, a unique aspect of the Umati project was its focus on locally spoken vernacular languages. Online blogs, groups, pages and forums in Kikuyu, Luhya, Kalenjin, Luo, Kiswahili, Sheng/Slang and Somali were monitored.

From these platforms, hate speech incidents were collected and analysed based on a methodology that relied on definitions by Professor Susan Benesch of the American University. Benesch's introduces the term *dangerous speech,*[3] which is defined as a subset of hate speech with highest potential to catalyse violence. Her clear and detailed definition of dangerous speech, its constituents and effects, enabled the Umati project to build a process for collecting and analysing hate and dangerous speech found in the Kenyan online space between September 2012 and May 2013.

Key findings from the project, which are discussed at greater length in this report include:

- The occurrence of online hate speech cannot solely be relied on as a precursor to violence on the ground. Other factors might play a stronger role in determining violent or peaceful outcomes. Instead, it appears that online hate speech could be a window into the conversations Kenyans engage in offline, and thus offer a way to understand recurring issues that need to be addressed.

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

- 'KOT cuffing' contributed to the fact that only 3% of total hate speech comments collected by Umati originated on Twitter, while 90% were found on Facebook. We use this new term, 'KOT cuffing', to refer to a phenomenon observed on Twitter where tweets not acceptable by the status quo 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.

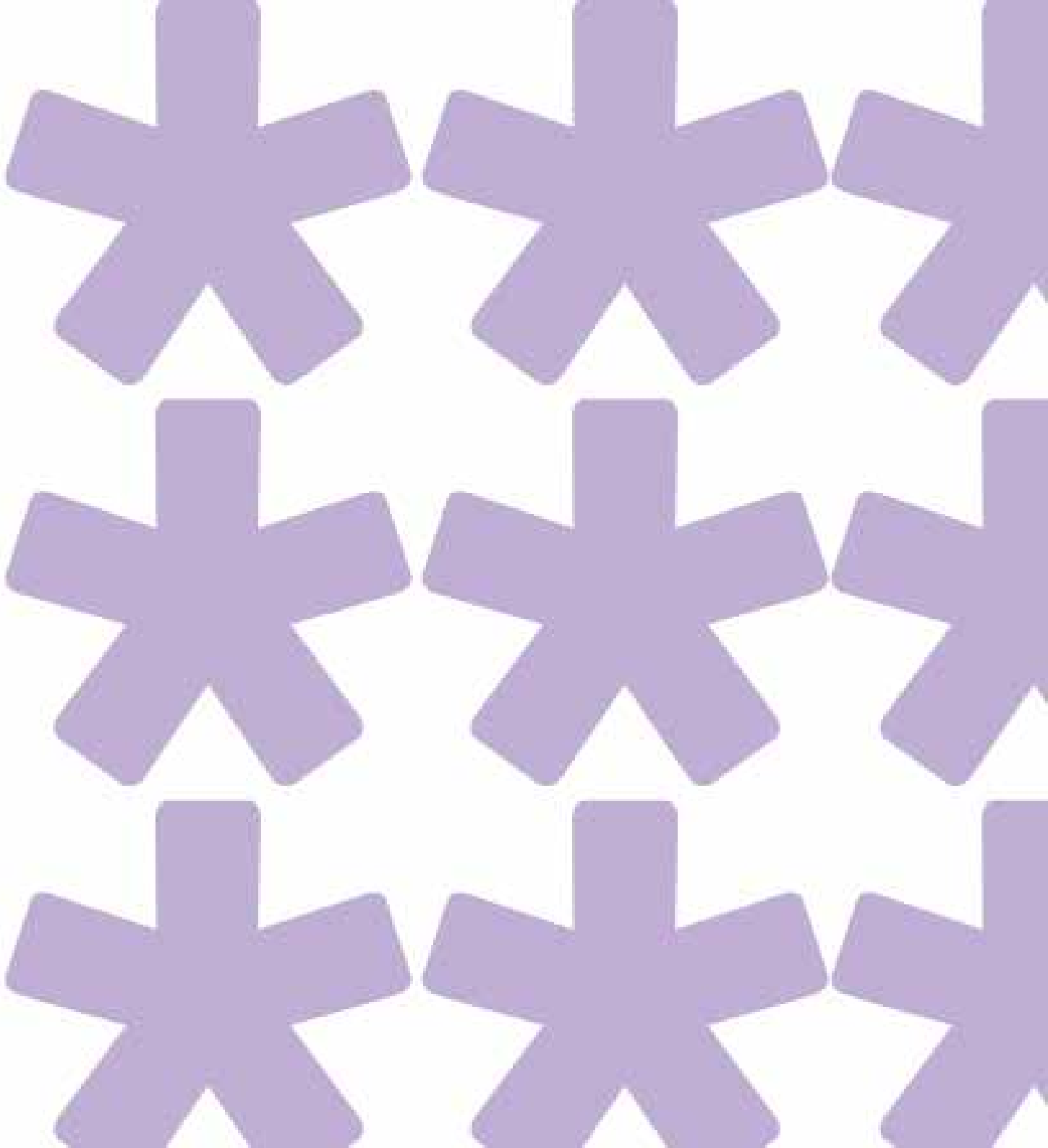
- There is a huge disparity between what the public perceives as hate speech and what the Umati project defines it as. From an exploratory survey conducted in May 2013, we found that the public perceives personal insults, propaganda and negative commentary about politicians as hate speech. The public's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

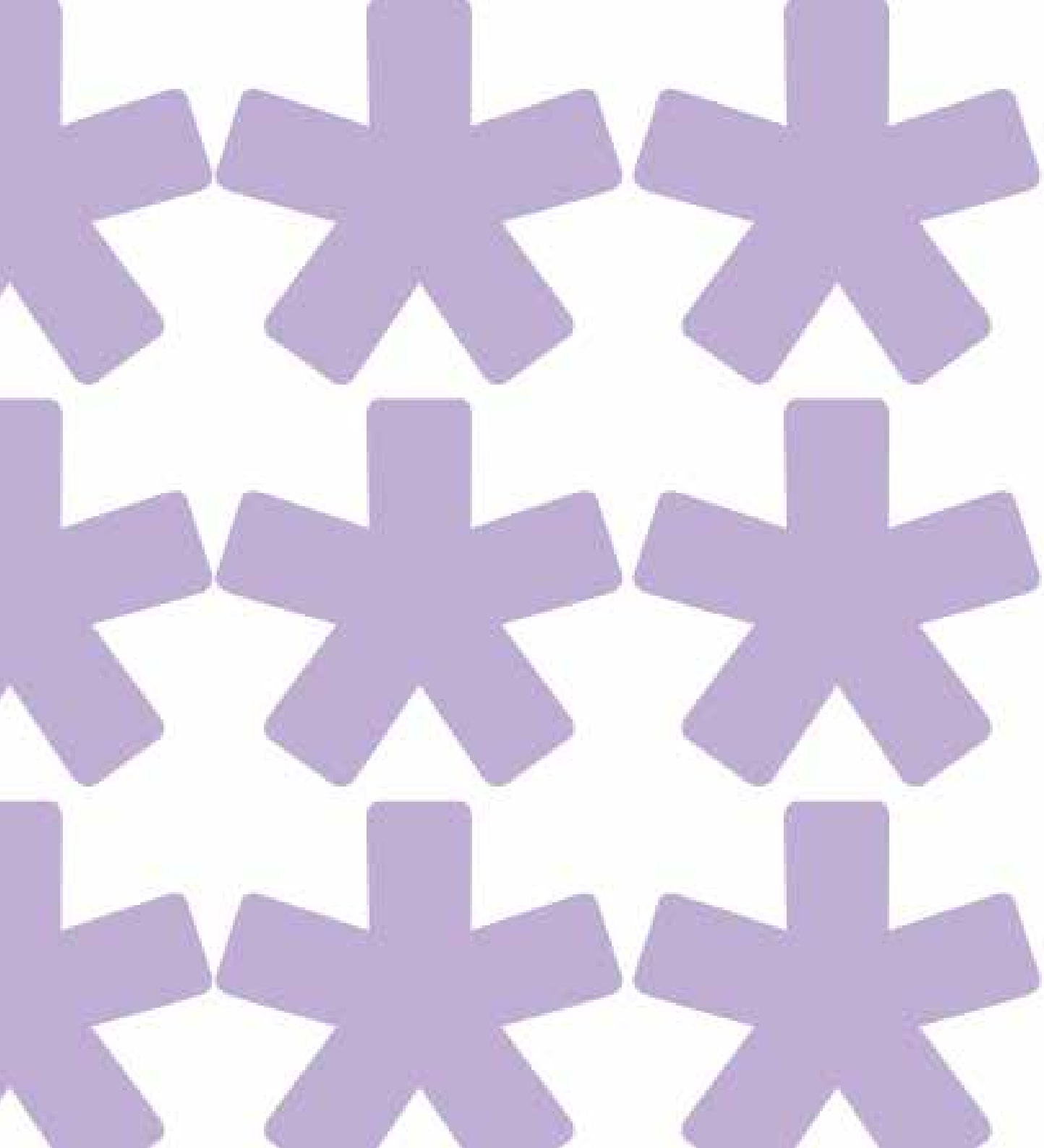
- Umati defines dangerous speech as a subset of hate speech that contains three out of the five possible calls to action, as defined by Benesch. Narrowing the definition of dangerous speech further was done in order to fit the Kenyan context. For example, stereotypical insults across tribes can amount to Benesch's definition of dangerous speech, however, applied to the Kenyan context, such stereotyping across tribes is usually largely perceived as harmless banter.

- Benesch's definition of dangerous speech was used by Umati to create a workable methodology that could be implemented in Kenya. Looking more closely at the elements of the Benesch Framework deemed relevant to the Kenyan context by the Umati project, could enable governing bodies, the general public, peace-building organisations, politicians, researchers, to accurately define and identify hate speech in Kenya, and take stronger measures against dangerous speech while protecting freedom of expression.

These findings, as well as the Umati methodology and challenges, are discussed further in the forthcoming sections.

3 S. Benesch, 'Dangerous Speech: A Proposal to Prevent Group Violence.' 12 January 2012. Retrieved from http://voicesthatpoison.org/proposed-guidelines-on-dangerous-speech/





*Literature Review

Review of Similar Projects

At the beginning of the Umati project, it was of great importance to develop a methodology that suited the Kenyan context, especially during elections. A review of the available literature on election monitoring processes and the identification and detection of online hate speech was therefore used to better understand the methodologies used in earlier election monitoring projects.

Our research revealed that crowdsourcing is a popular method for election monitoring.; the blog 'Stand up to Hate' (http://standuptohate.blogspot.com) encourages users to report offensive content on the site; the Canadian site 'Stop Racism and Hate Collective' (http://www.stopracism.ca) runs several initiatives aimed at combating online hate speech including, an online reporting form; training on how to report sites; and listings of sites and blogs that have been identified as offensive.[4]

Project iDitord in Armenia[5] went a step further and incorporated crowdsourcing with mapping, by asking contributors to send reports of election tampering to the Ushahidi mapping platform. Contributors sent these reports via a web-interface, mobile applications, Twitter and SMS. These reports were displayed geographically on a map depending on where they occurred.

Projects monitoring hate speech online include one by the French organisation MRAP (Mouvement contre le racisme et pour l'amitié entre les peuples).[6] The study looked into 2,000 URLs of forums, blogs, and social networking sites and various forms multimedia, including videos. The study revealed a sophisticated network of 'hate networks,' displaying how hate groups spread their ideologies across the Internet.

Another project by the Ukrainian organisation IHRPEX (The Institute of Human Rights and the Prevention of Xenophobia)[7] analysed the top 20 Ukrainian social and political websites and polled 623 users of these sites in order to assess their attitudes towards examples of hate speech on these websites. They defined hate speech as 'verbal offences, threats and displays of aggression to a certain social group or a certain person.' The study investigated key articles included on the sites and comments beneath these articles, and used the hate speech examples picked from these to query the respondents.

Additionally, a study by Warner and Hirschberg[8] developed an annotation scheme that they used to identify and categorise paragraphs that contained anti-semitic hate speech. The method they used involved human annotators who grouped hate paragraphs into 7 categories, depending on their content. A computer interface was created that allowed annotators to assign one or more of the seven labels to each paragraph.

Finally, the Sentinel Project (http://thesentinelproject.org) and Hatebase

4 British Institute of Human Rights, 'Mapping study on projects against hate speech online'. Young People Combating Hate Speech, DDCP-YD/CHS (2012) 2 , 2012, pp. 31-34.

5 P. Chadwick, 'Armenian Elections Monitoring: Crowdsourcing + Public Journalism + Mapping' in InterNews. Viewed on 3rd January 2013, http://innovation.internews.org/blogs/armenian-elections-monitoring-crowdsourcing-public-journalism-mapping..

6 http://www.mrap.fr, Viewed on 21st May 2013

7 The Institute of Human Rights and the Prevention of Xenophobia, 'The summary of the report: "Phenomenon of the cyber-hatred in the Ukrainian Internet space', in IHRPEX. 25 May 2011, viewed on 21st May 2013, [http://www.ihrpex.org/en/article/2086/the_summary_of_the_report_phenomenon_of_the_cyberhatred_in_the_ukrainian_internet_space],

8 W. Warner and J. Hirschberg, 'Detecting Hate Speech on the WorldWideWeb'. In Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012), Association for Computational Linguistics (2012), pp. 19-26.

(http://www.hatebase.org) collect crowd-sourced hate speech vocabulary and study these in order to create an early warning system to prevent genocide.

From the examples studied, none looked specifically at monitoring online hate speech during elections. Moreover, none of the examples collected, analysed and synthesized hate speech reports en masse through the use of contextualised methodologies that take into consideration unique differences among countries. The methodology that was thought ideal for Umati was one that would consider the dynamic and unique characteristics of the Kenyan online space, e.g., the multiple languages spoken online, the need for local monitors who understand not only the vernacular languages but the ethnically divided politics in Kenya, the lack of a workable definition of hate speech that suited the Kenyan context and budgetary limitations.

Umati thus created a new methodology to meet its needs, while at the same time applying best practices from the reviewed projects.

Deriving an actionable definition of harmful speech

The current definition of hate speech is defined broadly in Kenyan Law. Under Article 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." The Act mentions ethnic hatred only and leaves out other forms of hate that can be based on religion, gender, nationality, sexual preference or political affiliation.

Other Kenyan laws touch on hate speech in diverse ways. The 2010 Constitution notes that freedom of expression does not extend to hate speech, but fails however, to define the term. 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." The specific address to political polities inadvertently overlooks the general public as possible instigators of hate speech.

Given that Umati had to use a conclusive workable definition of hate speech in order to facilitate its identification and collection from the Internet, Umati adopted the more narrowly defined term, dangerous speech, which was coined by, Professor Susan Benesch of the American University. Benesch, who is an authority on hate speech as a precursor to violence in many countries, defines dangerous speech as speech that has a reasonable possibility of catalyzing violence.[9] It is important to note that the NCIC Act of 2008 and the Benesch Framework are complimentary and do not conflict, the Benesch Framework gives more guidance on how to identify hate speech that is particularly dangerous or inciteful.

Benesch developed a five-point analytical tool for identifying dangerous speech. These five factors[10] are:

- the speaker and his/her influence over an audience. For example, a political or religious leader has more influence over a crowd than a primary school teacher would.
- the susceptibility of the audience (the audience here being the listeners of the hate speech statement). E.g. are they in fear of the speaker? are they uneducated and misinformed and thus easily manipulated, are they marginalized, poor or desperate?
- the content in the speech that may be taken as inflammatory/offensive to the listener.
- the social and historical context of the speech
- the means of spreading the speech, including the language in which it is expressed.

9 S. Benesch, 'Dangerous Speech: A Proposal to Prevent Group Violence'. 23 February 2013. Viewed on 21st May 2013, http://voicesthatpoison.org/proposed-guidelines-on-dangerous-speech/

10 *ibid*

In other words, for a statement to amount to dangerous speech, one must consider not only the content of the statement, but also the speaker of the statement, his/her potential to move the audience to action against the targeted group, the susceptibility of the audience, the historical context and/or the underlying meaning of the statement, or code words used, and finally how the speech was disseminated, e.g, via radio, which has a wider reach than say in an email.

Note that all factors do not have to be present for a speech statement to amount to dangerous speech. For Umati, for example, the medium was already established, the Internet. It was also not possible for Umati to accurately establish the susceptibility of the audiences. Based on the 90-9-1 theory[11] that states that for each piece of content online, 9 people will reply/respond to it online, while 90 will read it but not ostensibly respond to it, we believed it would be difficult to gauge the exact influence of a speaker on his/her audience given that 90% of the potential audience will not respond to the statement reaction online.

Umati was challenged to develop a data collection methodology that could most accurately identify hate and dangerous speech online as per the Benesch definitions, while at the same time taking into consideration the pre-existing constraints.

Umati simplified the Benesch definition of dangerous speech into three points. The project held that dangerous speech is speech that:

11 B. McConnell, J. Huba, 'The 1% Rule: Charting citizen participation' in Church of the Customer Blog,11 May 2010, Retrieved on 10 July 2010, Viewed on 13th June 2013. http://web.archive.org/web/20100511081141/http://www.churchofthecustomer.com/blog/2006/05/charting_wiki_p.html

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

Dangerous speech is harmful speech that calls the audience to condone or take part in violent acts against a group of people.

It's important to note that an ugly or critical comment about an individual - a politician, for example - is not hate speech unless it targets that person as a member of a group. Hate speech is directed at a group, or at a person as part of a group: a tribe, religion, etc. During election periods, it is not uncommon for negative statements to be made against politicians and other influential personalities. This is a common part of the political process, as long as the statements do not constitute defamation, threats, hate speech or dangerous speech.

2 May contain one of the hallmarks/pillars of dangerous speech

From studying cases of violence that was exacerbated by inflammatory speech from a variety of countries and historical periods including Germany (1930s), Rwanda (1990s), Cote d'Ivoire (2002/3, 2011) and Kenya (2002, 2007/8), Benesch identified certain hallmarks often found in speech that led to violence.

Three hallmarks common in several dangerous speech statements are[12] :

- 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.

Each is discussed hereafter.

12 For further information and articles on the hallmarks and on Dangerous Speech generally, see www.voicesthatpoison.org

i. Compare a group of people with animals, insects or a derogatory term, especially in mother tongue

Before the 1994 genocide in Rwanda, the Hutus used the term "inyenzi" (cockroaches) to demean the Tutsis to less than human beings.[13] Psychological research suggests that it was easier for the Hutus to harm the Tutsis since they thought of them as mere insects.

In Kenya, atop using animal and insect names, our communities also have particular insults in vernacular language that are intended to demean certain groups.

Note that a speech statement can still be dangerous despite not having any of these three mentioned pillars of dangerous speech. The hallmarks serve as a diagnostic tool to identify some dangerous speech, since they are commonly (but not universally) found in it.

Also note the converse: a hallmark does not automatically make speech dangerous. As an example, if a mother tells her daughter to stop seeing a boy from another community, and calls the boy by the name of an animal, the speech is almost certainly not dangerous since the daughter will not react with violence against the boy or his community.

ii. Suggest that some people are spoiling the purity or integrity of the group

Of the five ethnic communities monitored by the Umati project, all are known to possess certain characteristics and/or perform certain socio-cultural activities that define them, e.g. Luos fish, Kikuyus run businesses, Kalenjins are pastoralists, Somalis trade and Luhyas are farmers/watchmen/ househelps.

13 N. Mitchell. 'First we call them insects: the prelude to horror', 26th April 2012. Viewed on 16 February 2013 http://www.abc.net.au/news/2012-04-26/mitchell-first-we-call-them-insects/3969192

However, in our research we found that commenters use these stereotypes negatively or state other negative stereotypes as truths against these communities. This category contains comments that use historically negative stereotypes to suggest that the targeted community is impure to the audience. Also, comments that we found from our research exhibited calls to remove these 'impure' groups from the society.

iii. Suggest that the audience faces a serious threat or violence from another group

Another indicator of a statement has the potential to promote violence is when the statement suggests that the audience should equip themselves because another group will attack them. Often, these comments are not based on truth but are instead intended to invoke fear in the audience so that they can defend themselves against the claimed imminent violence.

These statements often promote mistruths against the targeted community so that the audience can move to act against that community in the name of self-defense.

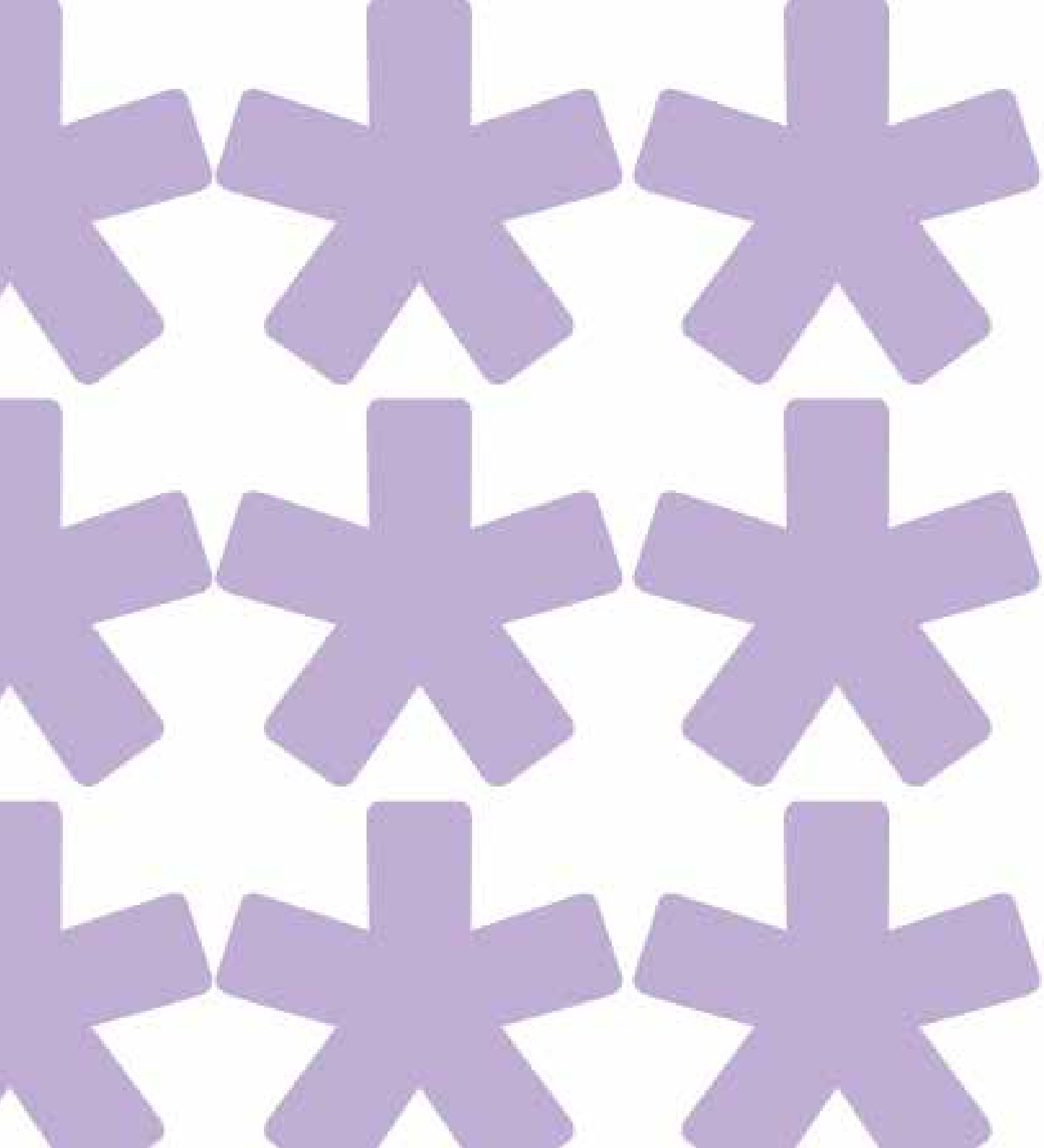
3 Contains a call to action

Dangerous speech often encourages the audience to condone or commit violent acts on the targeted group. The six calls to action common in dangerous speech are, calls to:

- discriminate,
- loot,
- riot,
- beat,
- forcefully evict, and
- kill.

Though this three-step process focuses on identifying dangerous speech by looking solely at the content of the speech act, it offers a pragmatic method that matches the reality of an online user, where they are often not aware or mindful of the speaker, the historical context and/or the influence the speaker has over the audience.

Nonetheless, the identification methodology used by the Umati team was more comprehensive in that it took into consideration at least three of the five dangerous speech factors found in the Benesch Framework.



*Methodology

Between September 2012 and May 2013, six human monitors scoured a collection of sites from the Kenyan online space for incidences of hate and dangerous speech. Apart from their ability to work independently and with minimal supervision, a key characteristic each monitor had to possess was a good command of English, Kiswahili and one of the vernacular languages studied by Umati. Seven languages were monitored; Kikuyu, Luhya, Kalenjin and Luo, representing the four largest ethnic groups in Kenya; [14] Swahili, the national language, and Sheng, an unofficial slang language; Somali, which is spoken by the largest immigrant community in Kenya; and English, which was monitored by all six.

For eight hours each day, each monitor manually scanned online platforms for incidences of hate and dangerous speech, recording them to online database through the use of a Google form. The Google form acted as a coding sheet with questions built based on the Benesch's definition of dangerous speech.

Appendix 1 displays the categorisation form, which contained questions asked around each incidence of hate speech collected by the monitors.

All hate speech incidents were translated by the monitors into English and then sorted into three hate speech subcategories - offensive speech, moderately dangerous speech and dangerous speech.

To enable the sorting of hate speech incidents into these three categories, a categorisation formula was devised that was dependent on two questions on the coding sheet:

i) On a scale of 1 to 3 with 1 being little influence and 3 being a lot of influence, how much influence does the speaker have on the audience? (code = N)

- 1 Little
- 2 Moderate
- 3 A lot of

ii) On a scale of 1 to 3, with 1 being barely inflammatory and 3 being extremely inflammatory, how inflammatory is the content of the text? (code = M)

- 1 Barely inflammatory
- 2 Moderately inflammatory
- 3 Extremely inflammatory

These two questions were aimed at gauging four factors from the Benesch framework;

- the speaker and his/her influence over the audience
- the susceptibility of the audience
- the degree of offensiveness contained in the content of the speech
- the social and historical context of the speech

14 Kenya National Bureau of Statistics 2009 Population Census

The answers to these two questions were dependent on five questions in the categorisation form. Below is part of a training manual the Umati monitors received, which explains how answers to the scale questions would be arrived at.

Categorisation Guide

Part A: Factors to be determined: SPEAKER/AUDIENCE/CONTEXT

A1. The speaker is

- a politician
- a journalist
- a blogger
- a public figure (includes media personalities)
- an elder/community leader
- an anonymous commenter
- an identifiable commenter

A2. Who is the audience most likely to react to this statement/article?

A3. The statement

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

Use the above questions to answer the first scale question below:

How much influence does the speaker have on the audience?

- 1 Little
- 2 Moderate
- 3 A lot of

Part B: Factor to be determined : CONTENT

B1. The text /article can be seen as encouraging the audience to

- Discriminate (can be 1 or 2) depending on B2)
- Riot
- Loot
- Forcefully evict
- Beat
- Kill
- None of the above

B2. Does the statement/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 another group
- None of the above

These two questions(B1 and B2) determine the second scale question below:

How inflammatory is the content of the text? (code: N)

- 1 Barely inflammatory
- 2
- 3 Extremely inflammatory

Guide to this question

1. In B1, if 'Discriminate', 'Riot' and/or 'Loot' is selected in B1, N can be 1 or 2.
2. In B1, anytime 'forcefully evict', 'beat' or 'kill' is selected N is 3.
3. In B2, 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, N is 3.

Finally, depending on the answers from the two scale questions (coded 'M' and 'N' respectively), a sorting formula was used that enabled the grouping of statements into the three hate speech sub-categories.

SORTING
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
Bucket 1 = Offensive Speech
Bucket 2 = Moderate dangerous speech
Bucket 3 = Dangerous speech

Understanding the three hate speech categories

1 Offensive speech

The primary intent for hate speech statements in this category was to insult a member of a certain group due to their belonging to the group, or insulting the entire group. Often, the speaker had little influence over the audience, the content of his/her speech was barely inflammatory, and generally the statement did not call upon the audience to commit a harmful action against the targeted group.

Statements in this category were therefore labeled "offensive" as they were aimed at verbally discriminating the targeted group and had low potential to spark violence.

However, it was noted that several statements that fell in this category contained strong insults or negative stereotypes, often in vernacular language, that encouraged the audience to hate the target group. From observing these words and statements, Umati noted that when used in other contexts, they did not amount to hate or dangerous speech. It is thus posited that, banning the use of 'offensive words' as an effort to curb dangerous speech, is not a viable way to mitigate hate and dangerous speech as these words in and of themselves may not be dangerous.

On the flipside, it was observed that if these words or statements were repeated by influential speakers and to more vulnerable crowds, they could very easily be dangerous statements, which, based on Benesch[15] have the highest potential to ignite violence.

2 Moderately Dangerous Speech

In this category, hate speech statements were made by speakers with little to moderate influence over their audience. The content of the statements had a mixed effect on the audience; to some, these statements could be viewed as inflammatory, while to others, they could be viewed as merely offensive.

The reasoning behind assigning statements to this category despite their subjective content, was based on a weighting that Umati created - the influence a speaker has over a crowd has a higher weighting than the content in the statement uttered by the speaker.

15 Benesch, op. cit.

3 Dangerous Speech

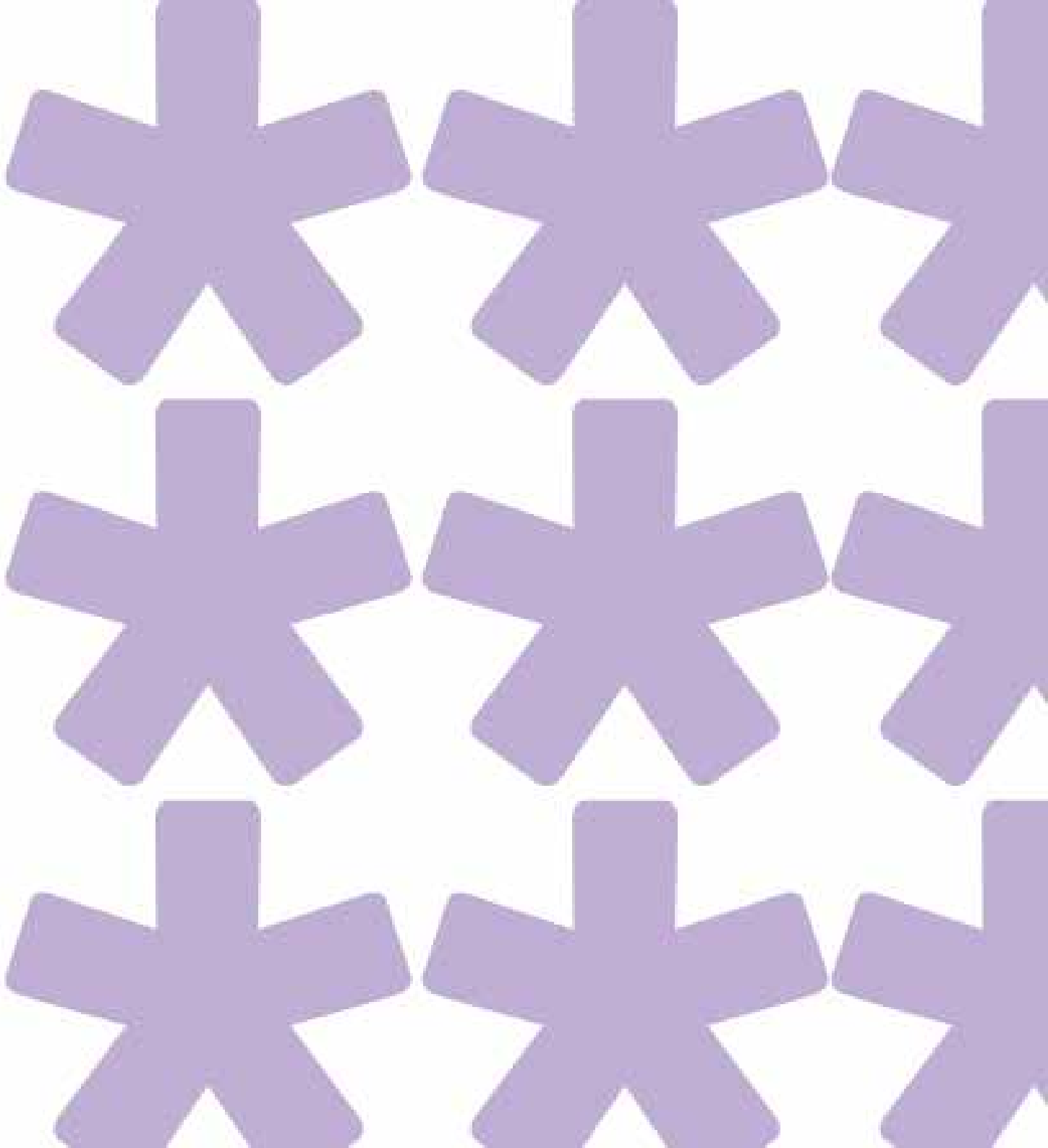
Statements in this category were made by speakers with a moderate to high influence over the crowd, were extremely inflammatory and had the highest potential to promote violence.

Umati further narrowed dangerous speech statements as those that contain clear or implied **calls to beat, forcefully evict** or **kill**. Benesch defines dangerous speech as speech that may contain any of the aforementioned five calls to action. However, Umati further narrowed this definition to speech that contains three of the most harmful calls to action. This was done in order to fit the Kenyan context where for example, stereotypical insults across tribes can amount to Benesch's definition of dangerous speech, however on the ground they are understood as merely offensive jokes. It was thus necessary to define the most harmful category of hate speech, as speech that contains clear calls to cause great harm to a group of people.

Analysis of category 3 data revealed that these statements provoke the audience to act violently by using the following tactics:

a) exacerbating fear in the audience and in this way encouraging them to protect themselves against the targeted group;
b) calling on the audience to seek revenge on the targeted group;
c) encouraging the audience to harm the targeted group based on often inaccurate beliefs or rumours that the speaker promotes as truth.

Comments in the Extremely Dangerous Speech category have the highest potential to catalyse violence, as they provide a plan of action that can be well understood and even acted upon by the intended audience.



*Findings

1 Online identities as a window into the state of the society

A hypothesis at the beginning of Umati, was that the project would collect only 5% – 10% of dangerous speech. This was based on certain perceptions and observations: 35.5% of the Kenyan population use the Internet.[16] Close to half of these Internet users are on Facebook. Extrapolating from the characteristics of Kenyan Facebook users as at the end of June 2012, it can be postulated that majority of internet users in Kenya are between the ages of 18 and 35, are ⅔ male and live majorly in urban areas.[17] Additionally, given Kenya's high literacy rate of 87%,[18] it can be posited further that majority of internet users have been exposed to multi-ethnic environments characteristic of most learning institutions.

Lastly, given that in both urban and rural contexts, the Kenyan society is predominantly conservative and thus averse to public displays of discontent, it was assumed that the behaviour of the Kenyan online user would be equally conservative, and that their most acerbic sentiments would not be expressed in the public online space.

16 Communications Commission of Kenya .(2012, June). Quarterly Sector Statistics Report Fourth Quarter Of The Financial Year 2011/12, (April-June 2012).Retrieved from www.cck.go.ke/.../SECTOR_STATISTICS_REPORT_Q3_11-12.pdf), Accessed on 21st May 2013.

17 Social Bakers .(2012).Kenya Facebook Statistics. Retrieved from http://www.socialbakers.com/facebook-statistics/kenya) Accessed on 21st May 2013

18 Index Mundi.(2013).Kenya literacy. Retrieved from(http://www.indexmundi.com/kenya/literacy.html)Accessed on 21st May 2013

Instead, the Umati Project found that a fourth of the 5,683 unique hate speech statements collected was considered dangerous speech. As earlier mentioned, dangerous speech is the most vitriolic category of hate speech as it contains a call to kill, to beat and/or to forcefully evict a particular group, or an individual because of their belonging to a particular group. Umati data shows that one out of every four hate speech comments in the Kenyan online space 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.

The table below shows the trend of the hate speech reports collected between October 2012 and May 2013.

Months	Offensive Speech	Moderately Dangerous Speech	Dangerous Speech	Total
Oct	34%	40%	26%	100%
Nov	21%	38%	42%	100%
Dec	22%	32%	46%	100%
Jan	27%	48%	26%	100%
Feb	29%	41%	30%	100%
Mar	37%	33%	30%	100%
Apr	51%	31%	18%	100%
May	47%	37%	16%	100%
Total	**37%**	**37%**	**26%**	**100%**

Figure 1 : Table showing total hate speech collected between October 2012 and May 2013, across the three hate speech sub-categories

Even more interesting is the dominance of identifiable commenters as the main actors of hate speech. For each hate speech report collected, Umati took note of the type of commenter of the speech. The possible types were:

- a blogger
- journalist
- politician
- public figure
- anonymous commenter
- elder/community leader or
- identifiable commenter

As shown in Table 2 below, 94% of hate speech actors were identifiable commenters. We defined an identifiable commenter as one who uses his/her real name, or pseudonym, when making hate and dangerous speech comments online. Pseudonyms were included in the bracket of identifiable commenters for three reasons. First, it was difficult for Umati to determine whether the actor's name is indeed a real or fake name.

Second, with additional resources, the real identity of an actor can be traced even when he/she uses a pseudonym, hence a pseudonym user can still be considered identifiable.

Finally, it was noted later in the project that online users often change their usernames so as to interact 'anonymously' on multiple Facebook pages. Nonetheless, there exists software that can be used to trace these fake usernames to the primary user account, making these hate speech commenters identifiable.

In the study of human activities on the Internet, online identity has often been a commonly recurring topic among the research community - Miller and Slater, Turkle and Sunde´n[19] are known pioneers in this research area.

19 D. Miller, & D. Slater, The Internet. An Ethnographic Approach. Oxford, Berg, 2000; S. Turkle, Life on the Screen: Identity in the Age of the Internet. Wiedenfeld & Nicolson, London, 1996; and J. Sunde´n, Material Virtualities: Approaching Online Textual Embodiment. Linko¨ping University, Linko¨ping, 2002 IN P.A. Aarsand, 'Frame switches and identity performances: Alternating between online and offline ', in Text & Talk 28-2, 2008, Walter de Gruyter , pp. 147-165.

Months / Speakers	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Grand Total
a blogger	5%	0%	3%	2%	1%	1%	1%	0%	1%
a journalist	1%	0%	1%	0%	0%	0%	0%	0%	0%
a politician	1%	3%	1%	1%	1%	1%	0%	0%	1%
a public figure	0%	0%	1%	0%	0%	0%	0%	0%	0%
an anonymous commenter	41%	1%	0%	0%	1%	1%	0%	0%	3%
an elder/ community leader	1%	0%	0%	0%	0%	0%	0%	0%	0%
an identifiable commenter	52%	95%	94%	96%	97%	98%	99%	100%	94%
Grand Total	100%	100%	100%	100%	100%	100%	100%	100%	100%

Figure 2 : Table showing speakers across the months

Dr. Aarsgard of the Linko¨ping University, Sweden, summarises their work by noting that,

> "several of these studies deal with the distinction of 'online/offline' and have shown how self-presentations in chat rooms, newsgroups, and games are closely related to other arenas and activities in people's lives."[20]

These studies suggest that a person's activities online are an extension of his/her activities offline, based on the actor's local experience and everyday life.[21] Therefore, what an actor says in a Facebook group, could very likely be the same thing he would say in a conversation at a bar.[22]

In relation to Umati data, the interconnected relationship between online and offline activity, indicates that instead of trying to draw out the impact online hate speech has on violence on the ground, it might be more accurate to view observable online activity as a **window** into the inaccessible offline activity of Kenyan internet users. Analysis of the topics recurring in Umati data gives a glimpse of the areas that affect the Kenyan society and what issues are in need of address. The recurring topics in the negative space that Umati observes give a broad indication of what socio-cultural issues continue to plague the Kenyan society.

The statement can be taken as offensive to:

- ☐ Luos
- ☐ Luhyas
- ☐ Kikuyus
- ☐ Kalenjins
- ☐ other tribe
- ☐ the Lower class
- ☐ the Upper class
- ☐ Christians
- ☐ Muslims
- ☐ Hindus
- ☐ other religion
- ☐ Asians
- ☐ Africans
- ☐ Whites
- ☐ Arabs
- ☐ political party members
- ☐ the Middle class
- ☐ politicians
- ☐ women
- ☐ Other:

As part of the data collection process, Umati took note of what group the hate speech actor intended to offend or victimize. The question below was answered for each hate speech report Umati collected, in order shed light on which groups were predominantly the subjects of hate speech online.

Putting these into broader groupings, hate speech statements discriminated people based on their:

- ethnicity (tribe)
- socio-economic class
- religion
- race
- political party affiliation
- government role i.e. whether politicians or not
- gender

From a sample of 703 dangerous speech statements, it was found that 88% discriminated the targeted group based on tribe, followed by 8% discrimination based on political party affiliation. The sample size was obtained by picking the top six groups that were targeted by dangerous speech statements, and regrouping them into the larger groupings mentioned above.

	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Grand Total
Tribe	19	52	27	55	71	147	150	114	635
particular political party members	5	35	14	11	3				68
Grand Total	24	87	41	66	74	147	150	114	703

Figure 3 : Table showing the top groups targeted with dangerous speech comments across the months.

20 P.A. Aarsand, 'Frame switches and identity performances: Alternating between online and offline ', in Text & Talk 28-2, 2008, Walter de Gruyter , pp. 147-165.

21 *ibid.*

22 *ibid.*

This data indicates that, in the Kenyan online space, a large majority of calls to kill, forcefully evict and beat are based on tribe (ethnicity) and political affiliation.

Furthermore, as shown figure 4 below, dangerous speech around these topics showed a sharp rise immediately before and after elections.

FIgure 5 compares all the hate speech sub-categories, and shows that dangerous speech was highest in March, when the 2013 elections were held. Offensive speech in the other hand was highest in April, after the elections, indicating a sharp rise in verbal conflicts online despite the peaceful elections.

The increase of hate speech after elections, offers a good window into what issues cause high levels of angst among the Kenyan public. In a proposed second phase of the project, Umati hopes to conduct deeper analysis of Umati data with a larger pool of experts, to reveal what these issues are. For example, finding out exactly what issues are raised in the statements against tribes. This can in turn offer an opportunity for civil societies and organisations to map out what issues to address before the next elections.

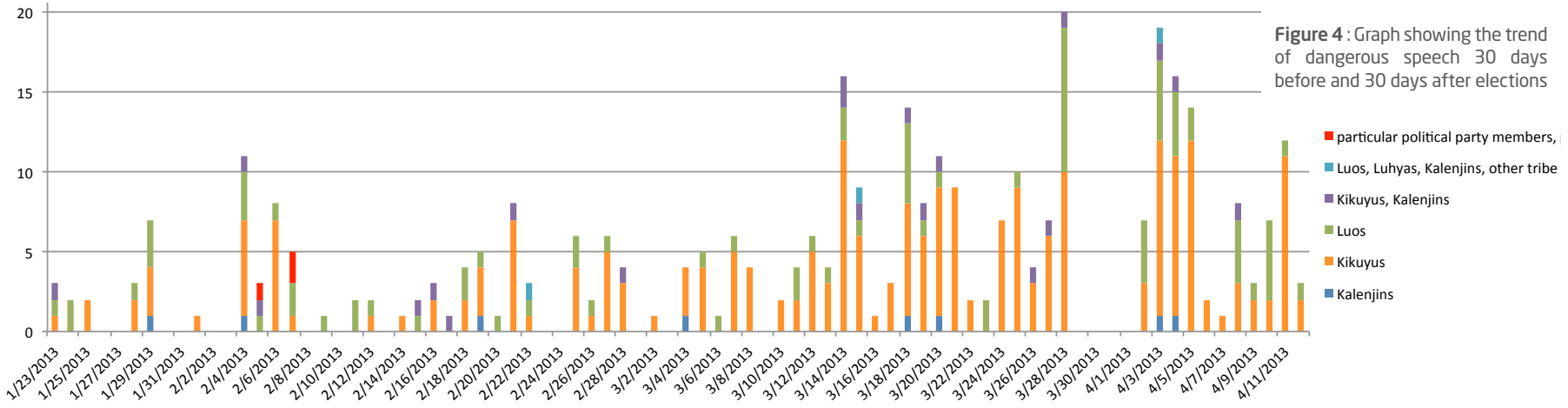


Figure 4 : Graph showing the trend of dangerous speech 30 days before and 30 days after elections

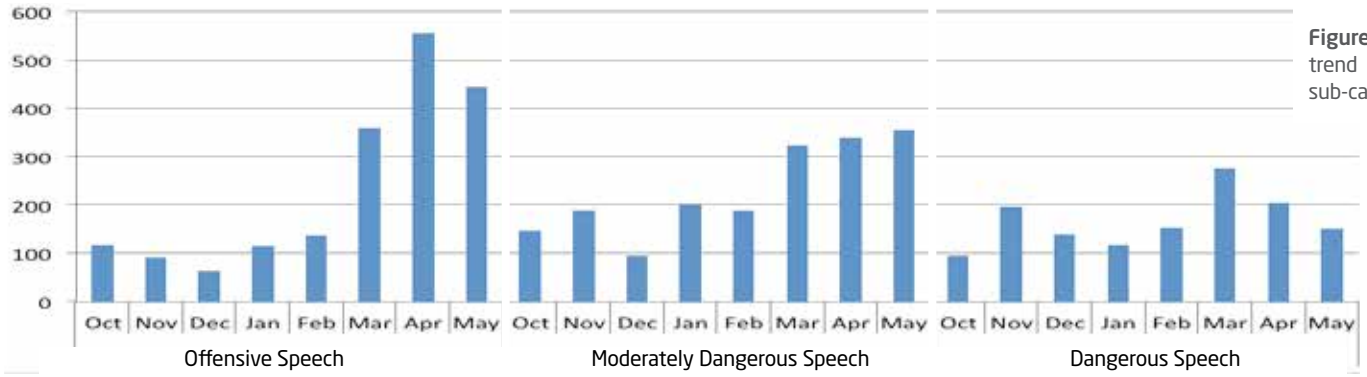


Figure 5 : Graph showing the trend of the three hate speech sub-categories across the months.

2 Influence of Platforms

Umati collected its data from four platforms: Facebook, Twitter, news sites and blogs. Both English and vernacular language platforms were monitored.

Despite monitoring Kikuyu, Luhya, Luo, Kalenjin and later Somali content online, few sites in these vernacular languages that fit the Umati research focus were found. This suggests most Kenyans online prefer to converse in English, Kiswahili, Sheng or Kenyan English slang.[23]

The Umati project also found that most of the data it collected originated from Facebook, as compared to Twitter, which is the second most popular social media site in Kenya. The table below shows the huge disparity between the two platforms.

Row labels	Facebook:Twitter(rounded)
Oct	n/a (100% Facebook)
Nov	n/a
Dec	37 : 1
Jan	n/a
Feb	n/a
Mar	48 : 1
Apr	n/a
May	n/a

Figure 6: Ratio of Facebook to Twitter incidences between October and May.

A number of reasons can explain this stark difference. Three emerge as the most pertinent to Umati:

23 Both Sheng and Slang are informal languages. They are a mixture of English, Kiswahili and mother tongue, differing primarily in influence; Sheng has more influence from Kiswahili and local languages, while Slang has stronger influence from English, both local and foreign.

Firstly, the Facebook architecture allows for conversations to be formed around a particular topic, and more so, to continue to exist around that topic. This is made possible through Facebook threads, groups and pages, all of which have independent lifespans. In contrast, on Twitter, topics have a lifespan that is dependent on popularity. Through the use of hash tags, the most discussed topics become 'trending topics' and consequently gain more exposure across Twitter. However, when a more popular topic arises, it stands the chance of overshadowing the current trending topic, thereby ending its lifespan. Unlike on Twitter, Facebook groups and pages allow for topics to exist independent of any activity on them. Meaning that, as long as the page/group is not closed, users can engage in them whenever they so desire, even after the topic becomes stale.

Secondly, Facebook threads, pages and groups allow online users who share similar opinion to collect around the topic at hand, and engage in discussions around it. Twitter's architecture also allows for users to form groups. However, these groups, known as lists on Twitter, are used only for reading tweets; a twitter user cannot post a tweet to a Twitter list.[24] Facebook thus effects continuous group collection and discussion much better than Twitter does.

Thirdly, the grouping of Facebook conversations into autonomous information domains allows for users to participate actively in these domains in parallel. In other words, a user can have a Facebook account where he posts his activities on his timeline, while at the same time engages in hate and dangerous speech conversations outside his timeline in other groups of his choosing. In Twitter, however, all the user's posts are contained in a singular information domain, and can be viewed by everyone that follows the user. It thus becomes difficult to post an offensive comment against a particular tribe for example, since all the user's posts are seen by everyone. In Facebook, the same user can post an offensive comment to a group that was formed to discriminate the particular tribe. He/she can thus populate his public timeline with harmless posts, and at the same time engage in hate speech in tribalist groups.

24 Twitter Help Center.(2013).Using Twitter Lists. Retrieved from https://support.twitter.com/articles/76460-using-twitter-lists). Accessed on 21st May 2013.

These three reasons enable the fourth, a term we have coined, KOT-cuffing.'[25] This comically coined term refers to an observed behaviour by Kenyans on Twitter (KOT), where tweets not acceptable to the status quo are shunned, and the author of the tweets, publicly ridiculed. The end result is that the 'offender' is forced to retract statements due to all crowd's feedback, and can even close his/her Twitter account altogether. The singular conversation stream architecture found on Twitter facilitates KOT cuffing since all posts are contained on a single timeline and can be viewed by all. It is hypothesized that the low number of hate speech reports on Twitter is a result of Kenyan social media users being more conscious of what they post on Twitter, in avoidance of the potential backlash from the crowd and public ridicule.

An example of KOT cuffing is by the account @gregoryivanovic, which on March 28, 2013, posted the following tweet,

> "I smell some KIKUYU'S stinking up the media waves. Those FUCKING KIKUYU sons of bitches. I wish I could find & KILL ONE!."

The tweet was reported to Uchaguzi 2013 and within 48 hours, the user had closed his account, likely because of backlash from other Twitter users appalled by his online behaviour. This incident is not unique; at least three such incidents of Twitter users closing their account or apologizing for their comments has come to Umati's attention in the months of February and March 2013.

On the other hand, the Facebook page "Not another Kikuyu President" was reported to Umati, NCIC, Uchaguzi and the @KenyanPolice[26] twitter account, yet the page still stands today[27].

25 Coined by Umati Research Lead, Kagonya Awori

26 It was found that this Twitter account does in fact not belong to the Kenya Police Service.

27 As of 21st June 2013.

3 Soft war

In contrast to the 2007 Kenyan elections, the 2013 election and post-election period has been regarded as peaceful, with relatively much fewer reported incidents of election related violence.[28] Despite the calm and largely peaceful elections however, local media and the National Cohesion and Integration Committee (NCIC) reported a shift in the violence expressed in the 2013 election cycle – the violence moved online.[29]

Umati data supported this; after the March 4th elections, there was a remarkable increase in hate speech statements against three particular ethnic groups; the Kikuyu (of which Uhuru Kenyatta, one of the primary presidential candidates at the time hails from); the Kalenjin (to which William Ruto, Kenyatta's running mate belongs); and the Luo (tribe of Raila Odinga, the second primary presidential candidate at the time).

This increase, particularly in hate based on tribe, was interesting to Umati in two ways. First, as mentioned earlier, Umati noted that most hate speech statements targeted groups based on political party or tribe. In March 2013, the large increase in hate speech statements was due to a sharp increase in hate speech statements that targeted tribes, with the top three tribes targeted being Kikuyu, Luo and Kalenjin. The graph below demonstrates this.

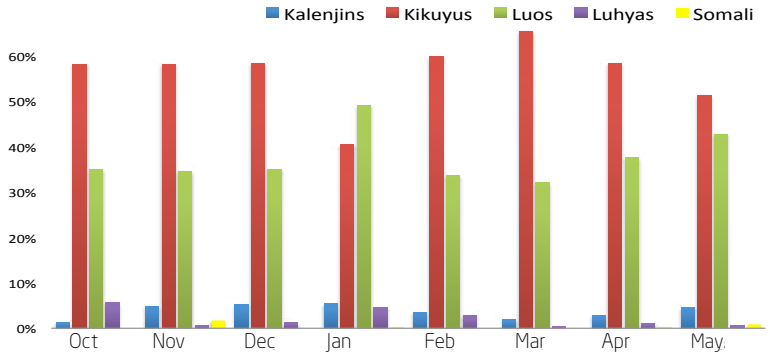


Figure 7: Graph showing the major tribes seen in hate speech between Oct 2012 and May 2013

28 International Crisis Group, 'Kenya After the Elections'. Crisis Group Africa Briefing N°94, 15 May 2013, p.3

29 T. Odula, 'Online war erupts in Kenya after peaceful vote', in Yahoo! News. 14 March 2013, viewed on 15 June 2013, http://news.yahoo.com/online-war-erupts-kenya-peaceful-vote-173648281.html

Secondly, as noted in the graph below, there was a sharp increase in the calls to kill and beat contained in the hate speech statements collected in March.

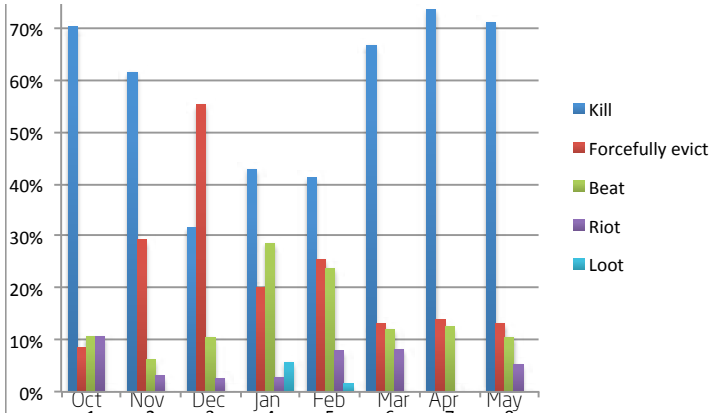


Figure 8 : Graph showing the calls to action across the months

Despite this clear increase in the volume and severity of hate speech incidences online during and after elections, there was little reported violence on the ground as compared to the 2007/8 post-election period. Notably, had the trend in Umati data been used in isolation to predict the possibility of post election violence, the prediction would have been that there would be violence, given the highly vitriolic data Umati came across in February and March (examples provided in Appendix II).

The peaceful election outcome suggests that there are overriding factors that can strongly contribute to the serenity of an election period. In their report,[30] the International Crisis Group posits that these overriding factors are: the public's lessons from the 2007/8 period, the rampant peace campaigns all over the country, the stern monitoring and self-censorship of traditional media, international pressure and increased government action in monitoring, security and prosecution.

30 International Crisis Group, 'Kenya After the Elections'. Crisis Group Africa Briefing N°94, 15 May 2013, p.1.

The question is, can these efforts be replicated? And if they can, should they be set as prerequisite for all Kenyan elections? Given the sharp increase in hate speech statements that Umati collected, were these efforts indeed effective or did they merely act as a medication to the symptoms, i.e. avoiding violence, and not the problem, treating the reasons for the violence?

In response to the soft war witnessed online, and the fulfillment of the project's goals, Umati relied on research by Benesch to provide the public with four possible ways to deal with hate and dangerous speech:

I. Silence the speaker
This is a role primarily played by governing bodies in which the speaker of dangerous speech is prosecuted. In January 2013, The Ministry of Information and Communication, through the Permanent Secretary Bitange Ndemo, announced that blogs found to contain hate speech would be closed and the perpetrators fined and/or jailed.[31] Similar prosecution would be given to those who spread hate messages via SMS and the media.

II. Limit the medium of dissemination of dangerous speech
This too has been implemented by governing bodies in Kenya, namely the Communications Commission of Kenya (CCK), who limited the means of spreading dangerous speech by monitoring SMSs and radio stations during the 2013 election period. However, the challenge here is that, like an amoeba, once one communication channel is cut, another quickly grows. Curbing hate speech on SMS and radio, may result in the hate speech actors moving their conversations to other unmonitored media such as Whatsapp groups, Facebook pages/groups, emails, and face-to-face meetings.

One way that the public can limit the means of dissemination online is by unfollowing or unfriending those believed to be hate speech propagators.

III. Empower the audience to be immune to incitement
The Umati team favours this approach as it gives power to the audience. By defining what type of speech is most harmful to a society, two goals are met:

- the mwananchi (citizen) is able to identify which comments/ statements are dangerous and is then able to react responsibly to these statements.
- by educating the public on exactly what kind of speech has harmful effects on the community, the public is then able to freely engage in the speech that is not harmful.

Umati engaged the public through outreach events that promoted education on the matter. Through an awareness campaign dubbed NipeUkweli (Swahili for 'Give me Truth'), Umati engaged bloggers, community radio, traditional media, and grassroots groups in order to debunk inflammatory rumours and myths, and thus reduce the possible effects of hate speech.

There are several ways an audience can choose to react to hate speech, the audience can:

- choose to ignore the statement;
- choose not to react to the statement;
- educate the speaker to engage in speech that is not dangerous;
- offer correct information so the others are encouraged to react peacefully.

IV. Discredit the speaker's utterances
One way to point out that a hate speech statement is false is by pointing out the truth.

Such responsible online activity was exemplified during the Mombasa violence that followed the death of Muslim cleric Sheikh Aboud Rogo, when inflammatory tweets were posted stating that a Mombasa church was demolished by Muslim bombings. A responsible social media user took a picture of the church (which was in fact standing and was not demolished) and posted it on Twitter stating, "stop the lies!".

Such dissemination of correct information has the potential to lead the audience to react peacefully to an utterance intended to incite them to violence.

31 P. Ngetich, 'Crackdown on hate speech starts today' on *The Star*, 9 Jan 2013. Viewed on 18 June 2013

4 The public's perception of hate speech

Based on the finding that most hate speech offenders were identifiable, Umati sought to investigate the public's understanding of hate speech.

A simple study was conducted where 12 participants were presented with 3 Facebook articles with corresponding comments from the public. Participants were asked to go through the articles and identify which of the comments constituted hate speech, and what severity rating (between 1 - low and 3 - high) they would assign to each hate speech statement.

Out of the 12 participants, 6 were from the public, 3 were Umati weekend monitors and 3 Umati weekday monitors. An additional objective of this test was to investigate whether there would be varied results between the three groups given their experience with Umati's definition and data collection methodology; the weekday monitors were very well trained and familiar with it, the weekend monitors were moderately trained and familiar with the methodology, while the public was not aware of it.

Results

First, from the 3 Facebook articles presented, the Umati monitors identified 44 statements as hate speech whereas the public identified 156 statements as hate speech. The table below shows this.

Category	Monitors' Total	Public Total
Offensive	24	55
Moderate	20	59
Dangerous	0	42
Total	**44**	**156**

Figure 9 : Quantitative comparison of the public's understanding of hate speech, and the Umati monitors' understanding

Further analysis was conducted to find out if statements were selected by both the monitors (weekday and weekend), and the public. Only 5% of the statements (10 out of the total 190) were picked by both the monitors and the public.

A possible explanation for this stark difference in selection is that there is a huge disparity between what the public perceives as hate speech and what the Umati project defines it as. From the data collected, the public's definition of hate speech includes personal insults, propaganda and negative commentary about a favoured politician/political party. The public's understanding of hate speech is also broader than the constitutional definition, which only takes into consideration discrimination on tribal lines.

Second, within the monitors, the weekday monitors picked 23 out of the 44 statements, while the weekend monitors picked 21 out of the 44 statements with the rankings shown below.

Category	Weekend Monitor Total	Weekday Monitor Total
Offensive	14	10
Moderate	7	13
Dangerous	0	0
Total	**21**	**23**

Figure 10 : Quantitative comparison of the Weekday and Weekend monitors' categorisation of hate speech statements

In the second exercise, where each respondent was required to assign a rating of 1 to 3 for each hate speech statement identified, the three weekend monitors rated the statements differently, while weekday monitors gave the same rating across the board. Results are displayed below.

Statement	Weekend Monitors Rating	Weekday Monitors Rating	Citizens Rating
Wakikuyu wametombwa na dogy huku kenya	1,1,1,1,2	2	3,3,3,2,2,2,
Why do you opt for uncut dogs instead of your cut men dicks		2	3,2,1
Deya ni kikuyu? Bloodsuckers of gwasi? Gwassi ni central? Kihii type.Thii ukarue.		2	2,1,1,1,
	1,2	1	3,1,2,2
the fuckin kikuyuz wth their fuckin president in a fuckin country...	2,2,1	2	3,2
Jaluos wil neva rule kenya! smely fish!	1	2	3,3,3,2
kikuyus why is it that ur minds revolves around luo's uncicum's dicks, yet ur cicum's dicks r so weak that u let dogs to fuck ur women 4r u. Shame on u wth ur small dicks kama peremende!		1	2
Wajaluo hapa wameamua kuambolish nyt runng n slypng wth maiti. Je huko	1	2	2,2
	1		1
Wakikiyu sio waschana kama sio dogy	2		2,3
I would like to say that these guys from lake side always want free things.i have lived with them n had to rebuke them atimes bcoz they want to b a parasite.looting and lazines is what occupy them..it is high time you started working			
hard and doing your own things			
ong'er frm the mountains r big fools			

Figure 11 :
The 10 Facebook comments that both the Umati monitors (weekend and weekday) and the public identified as hate speech.

The different categorisation of each statement is also compared across the three groups (ie weekend monitors, weekday monitors and the public)

From these observations, it can be posited that while the Umati methodology is teachable and replicable, extensive training, both informal and on-the-job, is necessary to standardise results. The more highly trained and experienced weekday monitors yielded better results compared to the weekend monitors who had been on the project for only three months.

Comparing the monitors and the public, the Umati methodology provided an easy, clear way to identify hate and dangerous speech.

On the other hand, it might also be argued that the definition used by Umati is too narrow, as suggested by the huge difference between the public's results and the monitors' results. However, the public's inclusion of personal insults and negative comments, as hate speech, is inaccurate based on what is contained in the Kenya Constitution 2010. According to Bill of Rights,

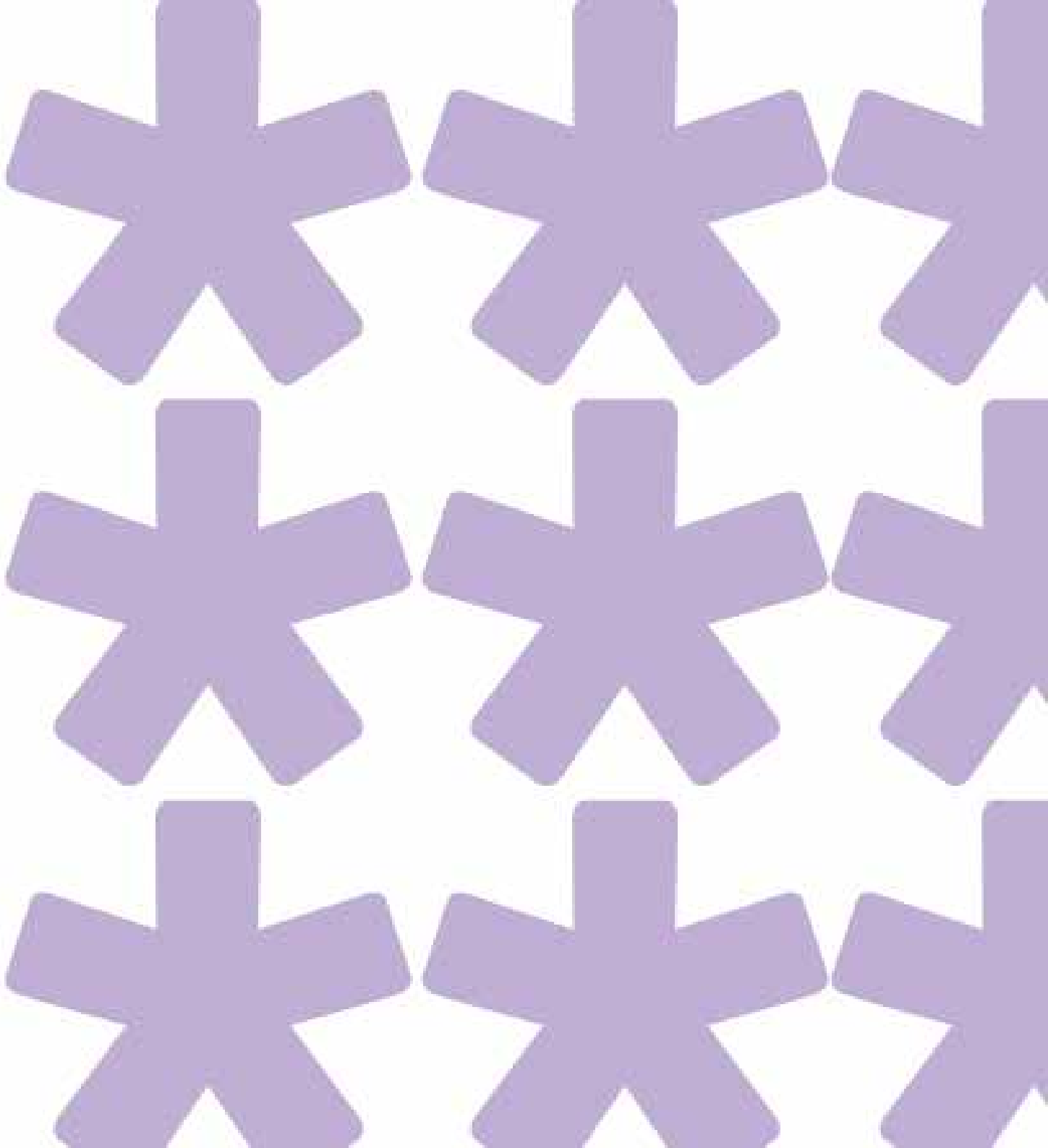
> "33.
> (2) The right to freedom of expression does not extend to—
> (a) propaganda for war;
> (b) incitement to violence;
> (c) hate speech; or
> (d) advocacy of hatred that—
> (i) constitutes ethnic incitement, vilification of others or incitement to cause harm; or
> (ii) is based on any ground of discrimination specified or contemplated in Article 27 (4).
> (3) In the exercise of the right to freedom of expression, every person shall respect the rights and reputation of others." [32]

The bill essentially states that incitement to violence, propaganda for war, and hate speech, and speech advocating for hate are not protected by the right to freedom of expression.

Nonetheless, insults and negative comments are protected by the freedom of expression. This means that an individual can critique the government or another individual without overstepping the boundaries of their freedom of expression.

More civic education and research should be done around public perceptions of hate speech.

32 Kenya Constitution 2010. Chapter Four, Ariicle 33.



*Challenges Faced In Umati

The benefits of human monitoring and detection processes are vast and are additionaly the best option for projects that involve highly contextualized information; as is the Umati project. However, use of human monitoring presents some challenges.

1 False alarms

One of the most important tasks in Umati is signal detection. This involves the monitors picking a target signal, in this case a hate speech statement, from the noise, in this case other online text.
There are four possible outcomes for signal detection tasks;

Hit: where the monitors correctly picked a hate speech statement from the online content they read.
Miss: where the monitors failed to pick hate speech statements from a piece of online text.
False alarm: where the statement picked by the monitors did not amount to hate speech.
Correct miss: where the monitors correctly failed to pick non-hate speech statements from a piece of online text they read.

Out of all the four outcomes , the most costly to the Umati project was false alarms. False alarms occurred throughout the project with the incidence reducing greatly in January after the monitors had to be retrained on the Umati methodology and process.
Additionally, the data already collected at that time had to be reread by each monitor in order to weed out incorrectly picked statements i.e. false alarms. It was found that several false alarms were insults against particular politicians. This further supported the finding that the definition of hate speech understood by the Kenyan public is broader than Umatis' definition. Between February and May, false alarms were still present in the Umati data due to the inclusion of the weekend monitors and general human error. It is posited that this has resulted in a 5% margin of error in the Umati data collected.

2 Fluctuations in productivity

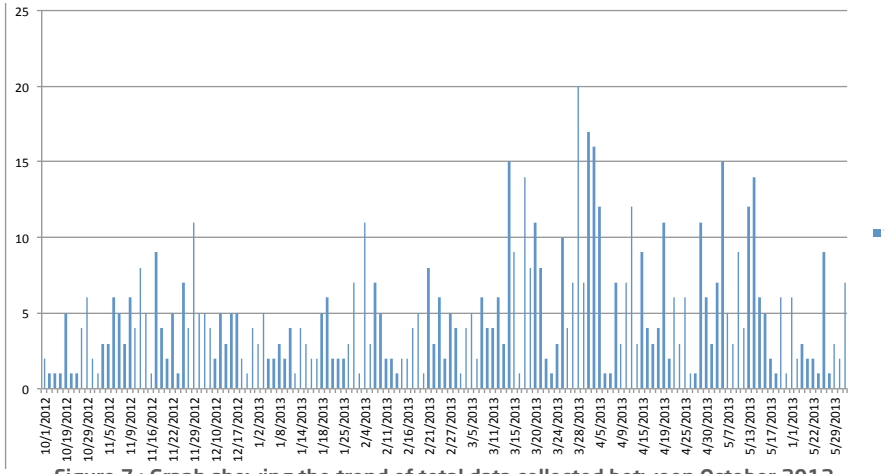


Figure 7 : Graph showing the trend of total data collected between October 2012 and May 2013.

As shown in the graph above, productivity levels fluctuated greatly across time. Factors that affected productivity include monitor fatigue, political and media events on the ground, introduction of new search tools, varying

productivity across monitors and varying team size i.e. when a monitor(s) was absent from work.

The challenge here is when the levels of productivity were due to monitor productivity. Though difficult to single out, it was noted the absence of particular monitors affected the total number of incidents collected. This affected the project during the analysis stage, when trying to establish what external factors directly contributed to the increase/decrease of hate speech online.

3 Task Dullness

Due to the inherently dull tasks performed in visual search and signal detection, 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 off every three months as long as the monitors made up the hours; and increasing work benefits such as remuneration, lunches and free counseling services.

4 Cost of Scalability

The high cost of scalability was realized when the weekend team was added to the project and when a team member had to be replaced. A significant amount of time had to be spent recruiting, training and retraining the new team members.

A financial cost was also incurred when the team size was increased to six, when a sixth monitor was added to monitor the Somali language. A new computer, desk, chair and pertinent software had to be purchased for the new team member.

5 Technology/multiple applications

Umati primarily uses the following software applications:

- Google Forms where the hate speech categorisation form sits.
- Microsoft Excel for analyzing the data and creating graphs.
- 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 on Facebook.
- Desktime (www.desktime.com) for the daily monitoring of each monitor's work hours.

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

Furthermore, a lot of time was lost when the same software had to be used across different operating system, for example, some Excel functionalities are supported in Windows but not in Apple.

The challenges Umati faced offered great oppurtunity for development of software that would meet the needs of online monitoring projects. One tool that can be used to meet some of the aforementioned challenges is SwiftRiver:

> "SwiftRiver[33] is a free and open source platform that helps people make sense of a lot of information in a short amount of time. [The platform consisits of] applications which combine natural language/ artificial intelligence process, data-mining for SMS and Twitter, and verification algorithms for different sources of information."

In the second phase of the project, Umati aims to build an online hate speech monitoring tool, based on the Swiftriver platform, that can solve the challenges it faced.

33 http://en.wikipedia.org/wiki/Ushahidi#cite_note-4

*Way Forward

The Umati project provided an opportunity to collect and study hate speech incidents from the Kenyan online space. This offered insight into the sentiment of the citizens especially around pivotal times like presidential elections. Findings garnered from the study include the divide that occurred online as opposed offline, aptly named the soft war; the influence of platforms on levels of hate speech; the role the media plays in driving conversations online and the public's definition of hate speech.

The four objectives of the Umati project, as detailed earlier, were to forward distress calls to the Uchaguzi platform; further education on the possible outcomes of hate speech to online and offline communities, and to develop a workable definition of hate speech that can be applied to the Kenyan context, and an accompanying methodology that can possibly be replicated to other countries. During the 9-month project, the Umati team was able to accomplish all of the above. We are especially excited about the possibility for scaling the developed definition and methodology to other cases. To this end, we are developing the second phase of the project building on lessons learned and findings from this study. Our interest in machine learning and human monitoring has influenced the design of the second phase of the Umati project, where an automated tool will be developed to 'learn' the Umati methodology and perform hate and dangerous speech monitoring over a longer period of time.

The second phase of the project will also entail working closely with civil society organisations who are equipped to deal with many of the socio-cultural issues that surfaced from the Umati data. A key concern here would be to ensure that peace-building and civic education efforts are not only concentrated around election periods but continue for longer to ensure a greater effect on the public.

Ultimately, Umati aims at having an effect far beyond the Kenya 2013 election cycle and far beyond Kenya. It is Umati's intention that the monitoring procedures systematized through this project will be employed by relevant organizations to suit various contexts.

*Umati Support Team

Daudi Were,
Project Director
As Ushahidi's Project Director, Daudi manages key partnerships and custom Ushahidi deployments for organizations ranging from large multinational organisations to grassroots NGOs. He was the point of contact for Umati and organized outreach events and linkages with the Uchaguzi platform.

Emmanuel Kala,
Technical Specialist
Emmanuel is the primary software developer on Ushahidi's SwiftRiver platform. He provided technical support on the Umati Project and collected user needs from the monitors' in order to improve the SwiftRiver platform for future use.

Rosemary Njeri
Finance Manager
Rosemary provided financial assistance on the Umati Project and monitored and interpreted cash flows. She managed the project's financial accounting, monitoring, and reporting.

James Ndiga
Project Assistant
As the Umati Project Assistant, James ran the outreach blogger roundtables and community forums in Mathare, Dandora, and Huruma. He also assisted with the media events held in the weeks leading up to the elections.

Professor Susan Benesch
Research Advisor
Susan Benesch is the Director of the Dangerous Speech Project, working to identify speech that is likely to catalyze violence, and to find the best policies to limit the force of such speech without curbing freedom of expression.

Juliana Rotich
Advisor
As Ushahidi's Executive Director, Juliana manages projects and aids in the development and testing of the Ushahidi platform. She provided advisory support on the Umati project.

Jessica Colaço
Advisor
Jessica provided senior expertise for the research design/research implementation. She provided expertise and technical inputs into analysis plans, coding, and reports.

Lillian Nduati
Outreach Advisor
As the iHub Media Lead, Lillian advised the Umati Project on outreach and engagement strategies and provided support for the NipeUkweli campaign and outreach events.

*Umati Core Team

Angela Crandall
Project Manager
Angela carried out project planning and scheduling, stakeholder management, task distribution and monitoring/supervision of their execution. She participated in designing the study instruments in line with the objectives, guided the analyst, and reviewed reports of findings.

Terry Musalia
Luhya Weekday Monitor
"The most memorable thing about the project is the fact that we engage in negative use of social media and dont even realize it sometimes, but now atleast I have learned how to handle myself."

Elvis Nyamolo
Luo Weekday Monitor
"Umati has enlightened me by giving me insight to social media and for an individual to be sensitive on information they post online and how it would impact to the readers of the information posted. "

Faith Morara
Kiswahili/Sheng Weekday Monitor
"Umati for me has been a roller coaster,i have learnt a lot but the most valuable lesson i got is we should all be responsible for that which we say,lets taste our own words before we spit them out!!"

Kagonya Awori
Research Lead/ Lead Analyst
With the support from the project manager, she oversaw all project functions and staff. This includes: scheduling and distributing workload; providing feedback to the technical specialist on the Uchaguzi and SwiftRiver platforms, and writing all Umati reports and presentations.

Anthony Metet
Kalenjin Weekday Monitor
"I worked for Ihub (Umati Project) for nine months and the experience was very exciting. The entire team was very cooperative and helpful."

Shamsa Abass
Somali Weekday Monitor
"I have learnt the different cultures,believes,causes of conflict and how to solve such problems."

Stephen Kamau
Kikuyu Weekday Monitor
"People should learn to treat each other as brothers/ sisters and have a spirit of embracing all ethnic communities despite their cultural backgrounds. Government will not always be there to prosecute all social media propagandists/bloggers rather each one of us should find a constructive activity that aims at building future of oneself and the nation."

*Umati Core Team

Joshua Obuya
Luo Weekend Monitor
"I learned that I am part of a society and I do not need to wait for someone else to come and improve my society. No matter how small a contribution to improving the society is, it contributes to making the big difference we might want."

Lydia cherotich
Kalenjin Weekend Monitor
"There are only two tribes in Kenya the poor and the rich. I saw this after the election period when Raila and the President Uhuru shook hands. Despite people in social media talking ill of other tribes, they remain friends, unlike the wananchi who talk ill about certain tribes instead of correcting someone as an individual."

Japheth Odonya
Luhya Weekend Monitor
"I [now] have a good understanding of the practical impact social media has on the society depending on how it is deployed; results can be progress or destruction. Hate multiplies - one person hates, another hates and everyone sees hate all over."

Abigael Wangui Gichuhi
Kikuyu Weekend Monitor
"One thing that stood out to me is the resilient and positive Kenyan spirit. A percentage of people out there were really tribal and baying for blood. This was however overshadowed by the numerous messages from persons preaching peace and love to the country."

Alex Orenge
Kiswahili/Sheng Weekend Monitor
"The project enabled me to improve on online content research skills, also I was able to observe how the social media has really influenced and promoted tribalism by use of hate and dangerous speech. "

Umati Final Report

Compiled and written by:
Kagonya Awori
Umati Research Lead
iHub Research

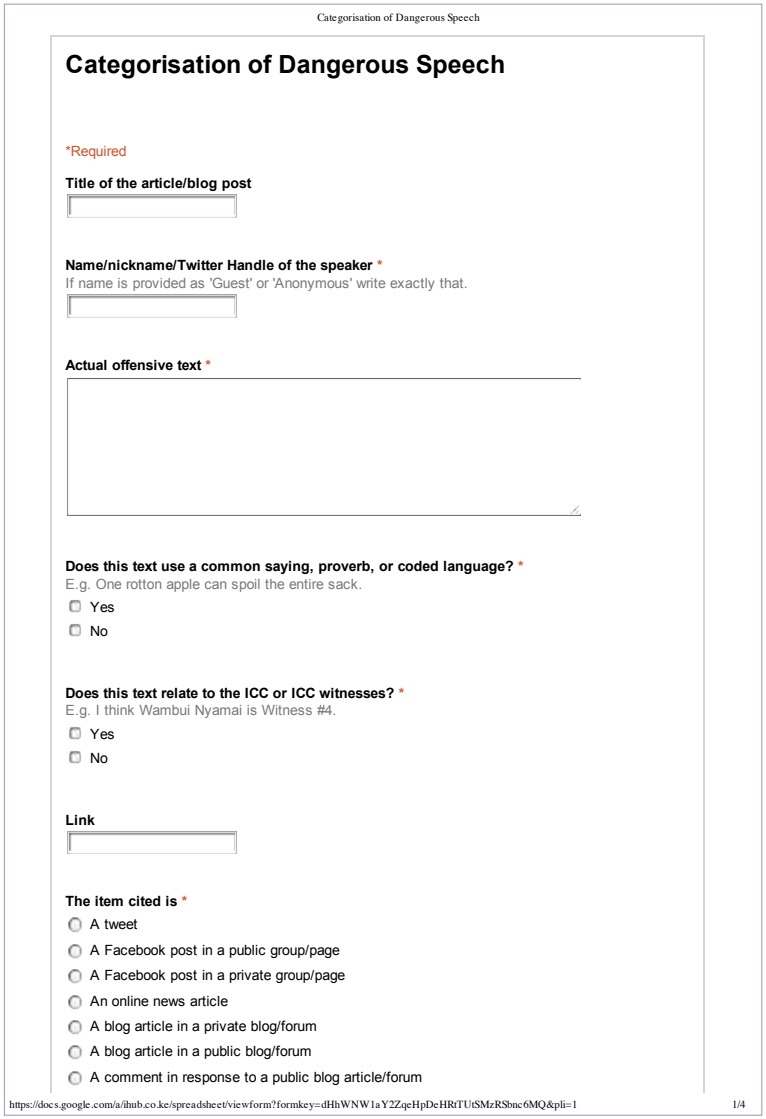
Reviewed by:
Dr. Gregory J.H. Deacon
Post Graduate Research Fellow
Oxford University

Edited by:
Angela Crandall
iHub Research Project Manager
iHub Research

June 2013

*Appendices

Appendix 1: Form used by Umati Monitors to collect and categorise data



Categorisation of Dangerous Speech

*Required

Title of the article/blog post

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

Actual offensive text *

Does this text use a common saying, proverb, or coded language? *
E.g. One rotton apple can spoil the entire sack.

- Yes
- No

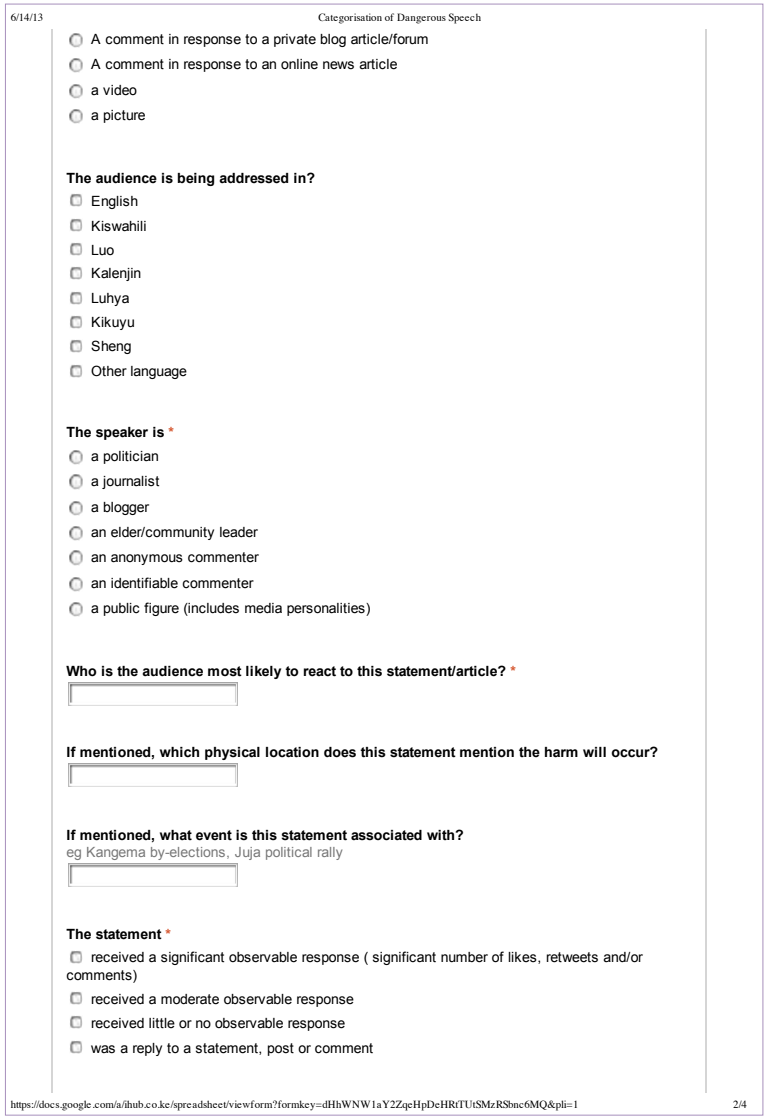
Does this text relate to the ICC or ICC witnesses? *
E.g. I think Wambui Nyamai is Witness #4.

- Yes
- No

Link

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 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 comment in response to an online news article
- a video
- a picture

The audience is being addressed in?

- English
- Kiswahili
- Luo
- Kalenjin
- Luhya
- Kikuyu
- Sheng
- Other language

The speaker is *

- a politician
- a journalist
- a blogger
- an elder/community leader
- an anonymous commenter
- an identifiable commenter
- a public figure (includes media personalities)

Who is the audience most likely to react to this statement/article? *

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

If mentioned, what event is this statement associated with?
eg Kangema by-elections, Juja political rally

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 response
- was a reply to a statement, post or comment

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

	1	2	3	
Little	○	○	○	A lot of

The text /article can be seen as encouraging the audience to *

- ☐ Discriminate
- ☐ Riot
- ☐ Loot
- ☐ Forcefully evict
- ☐ Beat
- ☐ Kill
- ☐ None of the above

Does the statement/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

How inflammatory is the content of the text? *

	1	2	3	
Barely inflamatory	○	○	○	Extremely inflamatory

The statement can be taken as offensive to

- ☐ Luos
- ☐ Luhyas
- ☐ Kikuyus
- ☐ Kalenjins
- ☐ other tribe
- ☐ the Lower class
- ☐ the Upper class
- ☐ Christians
- ☐ Muslims
- ☐ Hindus
- ☐ other religion
- ☐ Asians
- ☐ Africans
- ☐ Whites
- ☐ Arabs

- ☐ the Middle class
- ☐ politicians
- ☐ women
- ☐ Other:

Has this statement been used before and led to violence/ harm?

- ○ Yes
- ○ No
- ○ Dont know

☑ Send me a copy of my responses.

Submit

https://docs.google.com/a/ihub.co.ke/spreadsheet/viewform?formkey=dHhWNW1aY2ZqeHpDeHRtTUtSMzRSbnc6MQ&pli=1

Appendix 2: Some examples of dangerous speech collected in February and March 2013, that discriminated on tribe.

"Its time for us to leave manyatta and move to chalbi. That is what we did and that is what we will get. Plus some people are saying dont try a lorry. If you do its at your own risk to life and limb. Goodbye moyale. I came a poor fellow, made loads of money where the local fools cannot. Am rich am king. Fuck that fool called borana. We will now crush him for good. We are the world, they are feacal matter. But what to do between turbine and moyale? We will crush them. We are govt. Ukur is king. Konso burji is prince and rendille too. Kill all boranas."

"We,the Free Kenya Army ve declared war. We'll kill all kikuyus. Lets C hw u can rule us dead.Uhuru brought this to his pple."

"Kikuyus preparing to slaughter Kalenjins in ELDORET despite them being in the same coalition. This is happening in Eldoret now."

"Wakatwe vichwa hawa wabara chinja hao washenzi kabisa
("Behead all of them,non coastal residents,butcher them they are all fools")

"if its war,we wana give them they av never seen.last tym tulitumia machete and only MUNGIKI,THIS TYM (THAMAKA) WITH GUNS."

"Jaluo zote ziko bonde la ufa. Zitoke polepole. Coz lazima mtahiri na mabati ata mkose kuanza fujo. Iyo ni kitu imepangwa vizuri
("All luo's in the rift valley you better start moving out of there because you will be forcefully circumcised using iron sheets even if you don't cause chaos. this is something that has been well planned")

"Any County Commisioners in TURKANA, SOUTH EASTERN, NYANZA, WESTERN, NORTH EASTERN, COAST, NAROCK and NAIROBI, Waanze kujipanga, hatuwataki huku! Endeni CENTRAL mtawale na Wakikuyu! We don need you and we are ready to force you out of here! NA SI TAFADHALI
("Any County Commisioners in TURKANA, SOUTH EASTERN, NYANZA, WESTERN, NORTH EASTERN, COAST, NAROCK and NAIROBI,you better start arranging yourselves,we don't want you here,go to CENTRAL and lead with those kikuyu's We don need you and we are ready to force you out of here!and we are not PLEADING WITH YOU!")

Kumbe wameru wako na akili ndogo ivi. Funda kabisa. Hasira mingi na kukula miraa kama mambuzi ndio kazi.Meru's u r idiots. Kufieni mbali na izo kura zenu za Tharaka Nithi. Kubaff!!
("So meru's have small brains like that,donkeys!They are quick to get mad and all they know is chewing khat like goats.Meru's are idiots.you people should just die!")

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