umati

MONITORING ONLINE DANGEROUS SPEECH IN KENYA

January - November 2013

By Nanjira Sambuli, Faith Morara, & Christine Mahihu
Edited by Elizabeth Rensor & Angela Okune
Support by Natasha Wissanji, Leo Mutuku, Patrick Costello & Chris Orwa
Design by Zack Adell


umati

Umati Report

Acknowledgements

January - November 2013

Thank you to Ushahidi, Internews, PACT, and Chemonics/KTI for their mentorship and funding support; Professor Susan Benesch for her support and feedback; Wambui Kamiru, who helped develop the foundational work for the project; and Kagonya Awori, for her significant contribution to the Umati body of work.

Umati Report

EXECUTIVE SUMMARY

January - November 2013

The Umati project emerged out of concern that mobile and digital technologies may have played a catalysing role in the Kenyan 2007/08 post-election violence. The project seeks to better understand the use of dangerous speech in the Kenyan online space by monitoring particular blogs, forums, online newspapers, Facebook and Twitter, the two most popular social networks in the country. Online content monitored includes tweets, status updates and subsequent comments, posts, and blog entries.

In order to understand the changes in online inflammatory speech used over time, the Umati project developed a contextualized methodology for identifying, collecting, and categorizing inflammatory speech in the Kenyan online space. To categorize hate speech, the Umati project uses Susan Benesch's definition of dangerous speech, that is, speech that has the potential to catalyse collective violence. The key variables of the five-part Benesch framework uses a speaker's influence, audience receptiveness, speech content being understood as a call to action, the social and historical context of the speech and the medium of dissemination. The framework enabled the Umati project to develop a methodology for the collection and analysis of online hate speech. We developed the categorization spectrum of offensive speech, moderately dangerous or extremely dangerous speech especially based on the perceived speaker's level of influence and the content as perceived to be a call to action.
The project's key findings in 2013 were:

1. **Dangerous speech captured was predominantly based on ethnicity and religious affiliation, and much online hate speech comes in reaction to events that transpire or are witnessed offline.**
2. **Online hate speech disseminators largely identify themselves with a real or fake name and use languages widely understood in Kenya (English, Swahili, and Sheng).**
3. **Over 90% of all online inflammatory speech captured by Umati was on Facebook, making it the highest source of such content.**

Running from September 2012 to date, the Umati project has created the largest database of hate speech incidents in any one country (over 7,000 incidents). The project is now in its second phase, automating, where applicable, the online monitoring process in order to enable the methodology to be more replicable locally and in other countries.

Though instances of online hate speech catalysing events offline have not yet been well-established, we believe that the project's findings offer a window of insight into the state of Kenyan society. From this we conclude that the root causes of hate speech—both online and offline—should be investigated and addressed. Monitoring, in and of itself, is not a complete solution.

1

Umati Report

INTRODUCTION

January - November 2013

The term hate speech does not have a universally agreed-upon definition. It includes, but is not limited to speech that advocates for or encourages violent acts against a specific group or creates a climate of hate or prejudice, that could in turn, encourage the committing of hate crimes. In this context, speech can include any form of expression, including images, film and music. It is important to keep in mind that a hate comment about an individual does not necessarily constitute hate speech, unless it targets the individual as part of a group.

Kenyan laws make mention of hate speech. 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." Notably, the Act mentions ethnic hatred to constitute hatred against a group of persons defined by reference to colour, race, nationality (including citizenship) or ethnic or national origins– and does not include hatred based on religion, gender, sexual preference, or any other group category. The 2010 Constitution notes that freedom of expression does not extend to hate speech, but does not define that term, while 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."

Incendiary remarks by politicians and other notable public figures such as musicians (through lyrics) have been noted to incite violence in Kenya's historical past, specifically around election periods, with a culmination noted during the 2007 election period and its aftermath. Efforts to monitor hate speech usage have been in place through undertakings by Kenyan civil society as well as police authorities, but the migration of hate speech online remained neither monitored nor analyzed. The migration of hate speech online can be attributed to the significant increase in Internet penetration in Kenya and consequent social media adoption. Umati (Swahili for 'crowd') thus emerged out of concern that mobile and digital technologies may have played a catalyzing role in Kenyan 2007/08 post-election violence and the inadequate assessment of dissemination of potentially harmful speech online.

A strong evidence base linking online content and offline actions has not yet been established. While it is almost impossible to prove causation between a speech act and violence, we believe speech acts make a contribution to the 'piling up' of discourse around dehumanization of a group of people. As Benesch explains,[1] you cannot say that one person smoking one cigarette means many will die of lung cancer. But if many people smoke many cigarettes, then it is highly likely that a lot more of them will die of lung cancer. In the same way, a build-up of dangerous speech can prime a group towards collective violence by shifting the rhetoric towards condoning of violence. Online rhetoric can also contribute to this 'priming of the pump' and condoning of violence. Users online, for a variety of reasons, are interacting and sharing ideas and opinions. Even after they leave the online world, those ideas continue to shape their behaviors and interactions offline. Therefore, in order to understand the online interactions and comments shared during an election period, the Umati project was launched in October 2012, six months before the Kenyan general elections (March 4, 2013). The Umati project exists in two distinct phases: the first phase of the project established the following initial goals:

1. *To better understand the type of speech most harmful to Kenyan society, by monitoring speech disseminated online.*
2. *To forward calls for help to Uchaguzi, a technology-based system that enabled citizens to report and keep an eye on election-related events on the ground.*[2]
3. *To define a process for online hate speech tracking that could be replicated elsewhere.*
4. *To further civic education on dangerous speech, as observed online, so that Kenyans are more responsible in their communication and interactions with people from different backgrounds.*

The second phase of the Umati project (July 2013 to January 2016) further aims:

1. *To improve the Umati methodology developed in phase I and increase the system's scalability through automation where applicable.*
2. *To test the Umati methodology in additional countries ahead of national elections in order to improve and increase global applicability of the methodology.*
3. *To explore non-punitive, citizen-centered approaches for reducing dangerous speech online.*

1 *S. Benesch, personal communications, January 13, 2014.*

2 *Uchaguzi was an election-specific deployment by Ushahidi and other stakeholders that saw collaboration between citizens, election observers, humanitarian response agencies, civil society, community-based organisations, law enforcement agencies, and digital humanitarians to monitor elections. For more information, see www.uchaguzi.co.ke.*

Historical Background of Hate Speech in Kenya

Kenya has a history of hate speech in the political arena. Ethnic clashes preceding and following the 1992 elections were largely fuelled by politicians' negative verbal campaigns. The vitriolic campaigning used ethnic stereotypes to reinforce and rally support while simultaneously spreading suspicion against other ethnic groups. The multi-ethnic Rift Valley Province was most affected by the clashes, as members of the Luo, Bukusu and Kikuyu communities were victims of what have since been recognized as well-coordinated attacks.[3]

Intense, ugly and insulting rhetoric aimed at people based on ethnic affiliation formed the basis of hate speech in Kenya in the 1990s. For example, one minister at the time, in an effort to portray members of the Kikuyu community as untrustworthy of the country's leadership, described them as "ugly, with brown teeth and jigger-infested feet."[4] He further portrayed them as greedy and selfish, a recurrent stereotype that has been a basis for offensive speech against the community to this day. Around the same time, President Moi also referred to the Luo community as a cheap and easy-to-buy people.[5]

The 2005 Constitutional referendum was also overtaken by the ethnic rivalries perpetuated by hate speech. Despite the 'government of unity' that had replaced President Moi's 24-year rule, grievances of broken promises overwhelmed any objective discussions. Meanwhile, the newly liberated media, with vernacular stations for each ethnic group, became a new medium for the propagation of hate speech, and reduced the question of supporting the draft constitution to an ethnic matter. During this time, the Kenya National Commission on Human Rights (KNCHR) and the Kenya Human Rights Commission (KHRC) publicized and published utterances of hate speech by politicians during the campaigns. This was in a bid to 'name and shame', as well as spark debate on the impact of ethnic biases and speech in politics.[6] After their hate speech monitoring exercise during this period, the KNCHR recommended that Parliament should enact legislation, as a matter of national urgency, to bolster existing laws that monitored and restricted hate speech.

3 *Kiai, M. (2010). Speech, Power and Violence: Hate Speech and the Political Crisis in Kenya [PDF]. Retrieved from http://www.ushmm.org/m/pdfs/20100423-speech-power-violence-kiai.pdf*

4 *Ibid.*

5 *Ibid.*

6 *Kenya National Commission on Human Rights, Kenya Human Rights Commission. Behaving Badly Report [PDF]. (2006). Retrieved from http://www.knchr.org/Portals/0/CivilAndPoliticalReports/BehavingBadly.pdf*

The climax in the evolution of political hate speech in Kenya came in 2007/2008. KNCHR monitored offline hate speech incidents during the 2007 election campaign period, prior to the outbreak of violence. They noted that in the months preceding and following the general election in December 2007, there was a spike in hate speech used by ordinary citizens who primarily leveraged mobile-phone SMS/text messaging to spread hate and inciteful messages. Supporters of various political parties also distributed leaflets with messages that encouraged "hate passions" and violence. The KNCHR report also noted the failure of the Kenyan Parliament to enact the proposed legislation against hate speech which would have criminalized the use of such language.[7] (The proposed legislation was different from the NCI 2008 Act with its hate speech provision).

Disputed election results announced on December 30th, 2007 catalyzed post-election violence that left 1,200 people dead and approximately 300,000 displaced.[8] The Commission of Inquiry into the Post-Election Violence (CIPEV) found that use of hate speech and incendiary remarks by politicians, FM local media stations and the public played a role in instigating the violence by poisoning an already tense political environment.[9]

In response to the gravity of the post-election violence, a National Accord and Reconciliation Agreement[10] was signed in 2008, following a dialogue between the two main parties entangled in the disputed election outcome, that sought to provide a peaceful solution to the political impasse and violence that had engulfed the country. The National Cohesion and Integration Act and Commission draw their existence from the dialogue outcome. The NCI Act defines and criminalizes offences of hate speech, racial and ethnic contempt[11] with a number of hate speech offences having been taken to court under this law,[12] while the Commission was formed to address the long term issues and undertake reforms as stipulated in the National Accord.[13] The International Criminal Court has also brought charges against the current President Uhuru Kenyatta and his Deputy William Ruto, as well as a journalist, Joshua Sang' for the alleged roles they played in inciting the violence.

Since 2008, the growth of online and social media has created a new space for the dissemination of hate speech.

7 *Kenya National Commission on Human Rights. (2007). Still Behaving Badly: Second Periodic Report of the Election-Monitoring Project[PDF]. Retrieved from http://cbrayton.files.wordpress.com/2008/01/election_report.pdf*

8 *Kiai, M. (2010). Speech, Power and Violence: Hate Speech and the Political Crisis in Kenya.*

9 *Commission of Inquiry into Post Election Violence Final Report [PDF]. (2008). Retrieved from http://reliefweb.int/sites/reliefweb.int/files/resources/15A00F569813F4D549257607001F459D-Full_Report.pdf*

10 *Mzalendo. Text of the National Accord and Reconciliation Act[Blog Post]. (2008, March 8). Retrieved from http://www.mzalendo.com/blog/2008/03/08/text-of-the-national-accord-and-reconciliation-act/*

11 *National Cohesion and Integration Act 2008. s.13 and s.62*

12 *iLaw Kenya. Challenges of Prosecuting Hate Speech-Related Offences [Blog Post]. (2013, May 29). Retrieved from http://ilaw-kenya.com/kenya/index.php/blawg/item/26-challenges-of-prosecuting-hate-speech-related-offences*

13 *Agenda 4: http://reliefweb.int/sites/reliefweb.int/files/resources/Background-Note.pdf*

Online Hate Speech & Umati

Since the submarine fibre optic cables landed on Kenyan shores in 2009, Internet penetration has been on a steep increase.[14] Greater access to affordable Internet, especially through the use of smart and feature phones,[15] has seen increased use of social media in the country. Such platforms offer new spaces for people to express their feelings, especially during times of heightened anxiety such as election periods. With over 2 million active[16] Kenyan Facebook users in April 2013 (an estimated 19.2% of the country's online population) and over 2.48 million geo-located tweets generated in Kenya in the 4th quarter of 2011, it is clear that social media is heavily used by Kenyan and continues to grow in popularity. As of April 2013, the number of active Kenyan users on Facebook grew by more than 58,400 within 6 months. Nonetheless, preparations by the government and NGOs to monitor hate speech during the 2013 general elections in Kenya did not include a plan to monitor online hate speech. The Kenya Police, for instance, were only prepared for physical monitoring of hate speech, such as using recorders in political rallies.[17]

In the build up to the 2007 Kenyan elections, mediums of propagating hate speech were generally limited to broadcast media transmissions, print media, SMS and email. Anecdotal evidence from the 2007/08 post-election violence suggested that online spaces such as forums and blogs were used to plan and incite violence and hate. However, at that time, no system existed to track such data.

14 *Communications Commission of Kenya. (2013). Quarterly Sector Statistics Report First Quarter of The Financial Year 2013/14 (Jul-Sept 2013) [PDF]. Retrieved from http://www.cck.go.ke/resc/downloads/Sector_Statistics_Report_Q1_201314.pdf*

15 *Ibid.*

16 *The number of people who have been active on Facebook during a 30-day-period.*

17 *Kaberia, J., Musau,N. (2013, May 3). Kenyan Authorities in the Dock over Hate Speech[Blog Post]. Retrieved from http://iwpr.net/report-news/kenyan-authorities-dock-over-hate-speech*

In the build up to the 2007 Kenyan elections, mediums of propagating hate speech were generally limited to broadcast media transmissions, print media, SMS and email. Anecdotal evidence from the 2007/08 post-election violence suggested that online spaces such as forums and blogs were used to plan and incite violence and hate. However, at that time, no system existed to track such data.

New media have diversified the audiences that engage in online communication, as seen in Umati's findings. Because these online spaces are a new medium for disseminating hate speech, their influence on the actions of the audience has yet to be observed. One possible result is the creation of a vicious cycle as audiences convene around hateful content, converse in a self-selected group, and form new ideas or support their original biases with the hateful beliefs of others. However, it is also possible to create a virtuous cycle as new media spaces act as an alternative source of information that neutralizes the negative impacts of offline hate speech.

With this in mind, iHub and Ushahidi[18] teams therefore decided to develop a systematic process of collection and categorization of inflammatory speech online in order to better understand the inflammatory conversations taking place online prior to, and during the 2013 elections. It should be noted that it is not the goal of Umati to find and prosecute the perpetrators of hate speech. Umati is a civil society project, not a legal or policing body.

18 Ushahidi is a non-profit tech company that began during the Kenyan post-election violence as a way to visualize citizen-generated information from the ground. Ushahidi specializes in developing free and open source software for information collection, visualization and interactive mapping. For more information, see http://www.ushahidi.com.

M

Umati Report

METHODOLOGY

January - November 2013

The Umati project adopted a definition of harmful speech that takes into consideration forms of hate speech beyond those based on ethnicity and race as provisioned in Kenyan law.[19] As noted in the background, Umati began months to the Kenyan General Elections with the aim to systematically capture and understand the conversations occurring in online Kenyan public spaces. In particular, we were interested in speech and conversations that could potentially incite and promote violence. In order to collect such speech, we needed to first adequately identify it, especially in the Kenyan context. This process led us to discover Susan Benesch's Dangerous Speech Guidelines.

Benesch defines 'dangerous speech' as speech with the potential to catalyse mass collective violence.[20] Benesch's Dangerous Speech Framework offers the following key variables for identifying dangerous speech:

THE BENESCH DANGEROUS SPEECH FRAMEWORK

SPEAKERS' INFLUENCE	AUDIENCE RECEPTIVENESS	SPEECH CONTENT	MEDIUM OF DISSEMINATION	SOCIAL/HISTORICAL SPEECH CONTEXT
eg a political, cultural or religious leader will likely have influence over a crowd	subject to incitement by a speaker	content that may be taken as inflammatory by the audience and understood as a call to violence	this includes language used and the medium for dissemination	eg previous clashes or competition between groups that can make them more prone to incitement

NB: not all variables must be present for speech to qualify as dangerous speech.

19 *National Cohesion and Integration Act 2008 s. 13*

20 *Benesch, S. (2012, January 12).Dangerous Speech: A Proposal to Prevent Group Violence. Retrieved from http://voicesthatpoison.org/guidelines/*

21 *Ibid.*

The Umati team built on the Benesch Framework to form a practical identification method that incorporated several variables from it. Specifically, the project found that the following 3 components of the Benesch Framework were the most relevant for the identification of online hate speech in Kenya:

FACTORS FOR IDENTIFYING ONLINE INFLAMMATORY SPEECH IN KENYA (from the Benesch Framework):

TARGETS A GROUP OF PEOPLE

- dangerous speech towards a group can occur across different lines, including religion, tribe/ethnicity, gender, sexuality, political affiliation and race
- NB: a hate comment about an individual is not necessarily hate speech UNLESS it targets the individual as part of a group

MAY CONTAIN ONE HALLMARK OF DANGEROUS SPEECH

- Comparing a group of people with animals, insects or vermin
- Suggesting that the audience faces a serious threat or violence from another group, specifically the same group that is a target of the inflammatory speech ("accusation in a mirror")
- Suggesting that some people from another group are spoiling the purity or integrity of the speakers' group

CONTAINS A CALL TO ACTION

- encourages a particular audience to commit acts of violence towards a group of people.
- These can include calls to kill, beat/injure, loot, riot, and forcefully evict.

Note that a causal link is almost impossible to draw between dangerous speech and on-the-ground violence, with many factors contributing to bringing about a physical violent act. However, speech has the capacity to catalyse or inflame violence. Actors are still legally and morally responsible if they commit violence in response to incitement or dangerous speech.

We found estimating speech 'dangerousness' to be somewhere between an art and a science. Umati is essentially trying to draw lines in a continuum of speech where often it is quite hard to draw these lines. Not only does the spectrum of hate speech change over time, the online space is also constantly in flux. Thus the development of our Umati methodology continues to be a work in progress.

Monitoring Process and Tools

Though initially Umati was to use an open source software, SwiftRiver, due to special customizations and needs of the Umati process that could not be built quickly enough for Umati Phase I, the software was unable to be used. Additionally, we realized that the nuanced insights necessary to accurately review local vernacular languages required heavy human input that a computer could not replicate at the time (because there was no previous database of mother tongue text corpus).[22]

Umati Phase I relied on a manual process of collecting and categorizing online hate speech. Human input proved necessary for accurately reviewing local vernacular languages and local vocabulary, which in turn allowed the creation of an inflammatory speech[23] database. Between October 2012 and November 2013, up to eleven monitors[24] scanned a collection of online sites: forums, blogs, the comments sections of online mainstream media sites, and social networks.

The monitors hired were selected for their knowledge of the particular vernacular languages monitored as well as their general understanding of how online platforms work. They monitored seven languages: English and Kiswahili (Kenya's official languages); Kikuyu, Luhya, Kalenjin, and Luo (representing the four largest ethnic groups in Kenya); Sheng;[25] and Somali (spoken by the largest immigrant community, as well as Somali-Kenyans). In the second phase, monitoring of Somali and Luhya languages has been discontinued, due to the incredibly low incidents of hate speech found during Phase I. Our data also indicated that most hate speech was propagated in English, Swahili and Sheng, with very few noted incidents in pure vernacular languages. Additional research is needed to investigate potential use of linguistic 'code-switching'[26] between the various languages in one conversation.

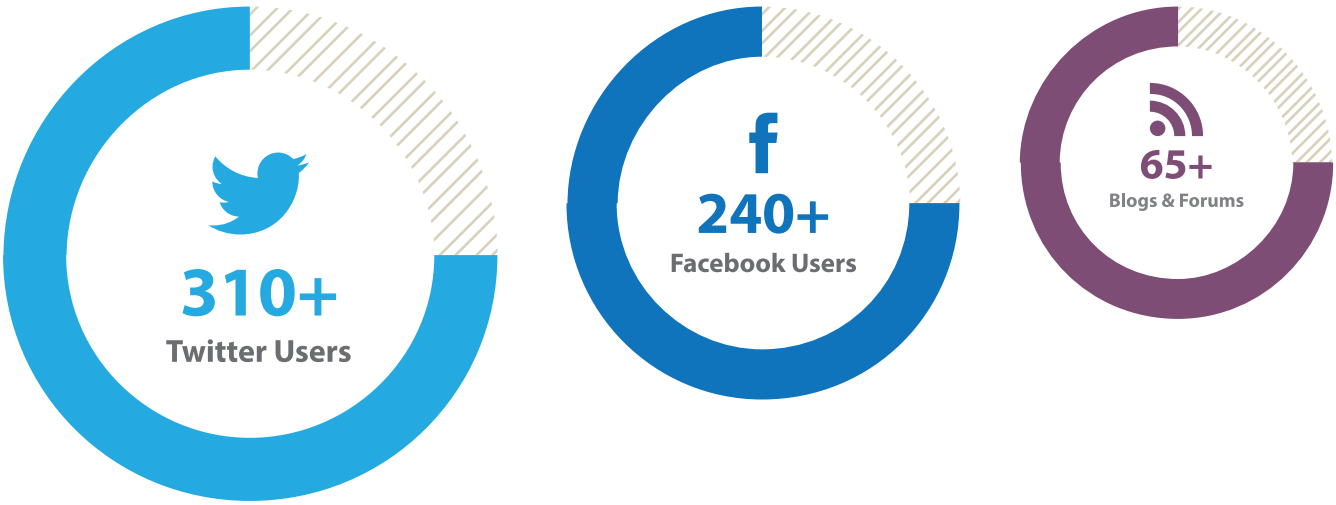
The monitors have continued to use and augment a list of online content sources initially developed during the set-up of the Umati project. At the start of Phase II, the monitors reviewed and updated the source list. As of November 2013, the source list covered 65+ blogs and forums; 240+ Facebook users, groups and pages; 310+ Twitter users; all major online Kenyan newspapers; and YouTube channels for the five main Kenyan media houses.

23 Inflammatory speech is used to refer to all hate speech categories: from offensive to the dangerous speech subset.

24 From October 2012 – January 2013, five monitors (covering all languages except for Somali) monitored Monday through Friday from 8:30 AM to 5:30 PM. From February 2014 through April 2014, in addition to the weekday monitoring, an additional 6-person monitoring team (original five languages plus Somali) was hired to increase monitoring activities to also include the weekends. From May 2014 onwards, the monitoring returned to weekdays only, using the original five-person team.

25 Sheng is a pidgin language incorporating Swahili, local languages, and English and largely used by youth in Nairobi.

26 Code switching is when a speaker alternates between two or more languages in the context of a single conversation, often to convey a thought or say something in secret.



The monitors referred to the source list daily and visited all major sites on a daily basis. Over the course of the 2013 data collection, after it became apparent that hate speech was more readily found on Facebook, the monitors focused more of their attention to data collection on Facebook. Also, the monitors had trouble following the conversations occurring on Twitter for a variety of reasons, especially due to the real-time characteristic of the platform and lack of persistence of conversations over time. The monitors worked from 8:30 am to 5:30 pm from Monday through Friday and Monday through Sunday during the months of January through April 2013. With a more automated collection tool, we hope it will prove easier to capture such real-time conversations and improve the Umati Twitter data collection process.

The monitors took advantage of several online applications to assist in monitoring social media platforms. These included Topsy,[27] Twitterfall,[28] and Trendsmap[29] for Twitter. These search engines give real-time insights into online conversations and enable one to monitor how content is being shared, who is sharing it, the key influencers and sentiments over time, by use of keywords and hashtags.[30] Trendsmap also gives a detailed view of current trends on Twitter with the help of Google Maps to depict the geographical location of each trend. For Facebook, the monitors leveraged the Open Status Search tool[31] that allows keyword searches for public Facebook conversations.

Since July 2013, the Umati team has begun work on incorporating more automation in the data collection process, through Machine Learning and Natural Language Processing techniques, where applicable. If these techniques can be used to teach a computing machine to accurately capture the speech acts and context nuances, it will increase the efficiency and scalability of the Umati project going forward. Several tools are still being assessed for the online automation process.

27 http://topsy.com

28 http://www.twitterfall.com

29 http://trendsmap.com

30 A # symbol used in Twitter conversations to categorize messages and mark keywords and topics in a tweet.

31 http://openstatussearch.com

Categorization Process

Each morning, monitor manually scanned through the online platforms for incidents of hate and dangerous speech, recording the speech acts they perceived to be hateful in an online database through the use of a Google Form. The form was comprised of a set of questions that the monitor answered for each incidence of hate speech. In this process, all hate speech statements were translated to English and sorted into three categories. The categories, in ascending order of severity, are:

Category One: Offensive Speech

- Hate speech comments in this category are mainly insults to a particular group.
- Often, the speaker has little influence over the audience and the content is barely inflammatory, with no calls to action.
- Most statements in this category are discriminatory and have very low prospects of causing violence on the ground.

Category Two: Moderately Dangerous Speech

- comments are moderately inflammatory and made by speakers with little to moderate influence.
- Audiences may react differently; to some, these comments may be inflammatory, while to others, they are not.

Category Three: Extremely Dangerous Speech

- made by speakers with moderate to high influence over a particular online audience.
- extremely dangerous statements with a high potential to inspire violence.
- Such comments are usually calls to action (calls to beat, kill, and forcefully evict), stated as truths or orders.

The full categorization formula, including the data entry form, is included in Appendix A.

Reporting to Uchaguzi

When imminent threats of violence were found during the election period, the Umati team extracted the relevant information and forwarded it by email to a listserv of specific people from donor agencies, Umati partners, and Uchaguzi key decision makers. The information was then forwarded to an on-the-ground early warning team for verification and/or action. This process was triggered five times from January 2013 to April 2013 and on-the-ground teams verified and mobilized based on the information passed to them.

The Uchaguzi team also had access to all of Umati's backend data. However, a notable challenge for integration between the Umati online hate speech monitoring and Uchaguzi election-watch portal and process was that the threshold level or 'when to act' on hate speech was hard for authorities to determine. Future deployments of Umati/Uchaguzi should think through an explicit threshold for when action should be taken based on hate speech instances online.

F

Umati Report

FINDINGS

January - November 2013

This section discusses the key findings from the Umati project in 2013 overall. Note that an earlier report[32] dives into findings specifically from the election period. We first highlight key events in Kenya that may have inspired use of dangerous speech online throughout the year, followed by a breakdown of the three key findings, which are:

1. Dangerous speech captured was predominantly based on ethnicity and religious af filiation, and much online hate speech comes in reaction to events that transpire or are witnessed offline.
2. Online hate speech disseminators largely identify themselves with a real or fake name and use languages widely understood in Kenya (English, Swahili, and Sheng).
3. Over 90% of all online inflammatory speech captured by Umati was on Facebook, making it the highest source of such content.

Key Kenyan events in 2013

Looking at the total volume of hate speech instances observed over the span of the project (January to November 2013), there is a noticeable peak during the election period (March - April 2013).

While this may be partly explained by the increased monitoring (weekends were monitored from February 2013 until April 2013), the data shows an increase in overall collected data, even when data collected on the weekends was excluded.[33] When classified into the three categories of hate speech defined by the Umati project (offensive, moderately dangerous, and extremely dangerous), the data showed that each category also peaked during the election period. The post-election period saw a steady decline in comments, although this is exaggerated in the period between June and July, during which, the project took a two-month hiatus. Figure 1 displays the quantities of each category of comment over the year.

32 See the Umati final report from Phase I available at http://www.research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf.

33 See page 10 of Umati March report, available at http://www.research.ihub.co.ke/uploads/2013/april/1365508815_819_823.pdf.

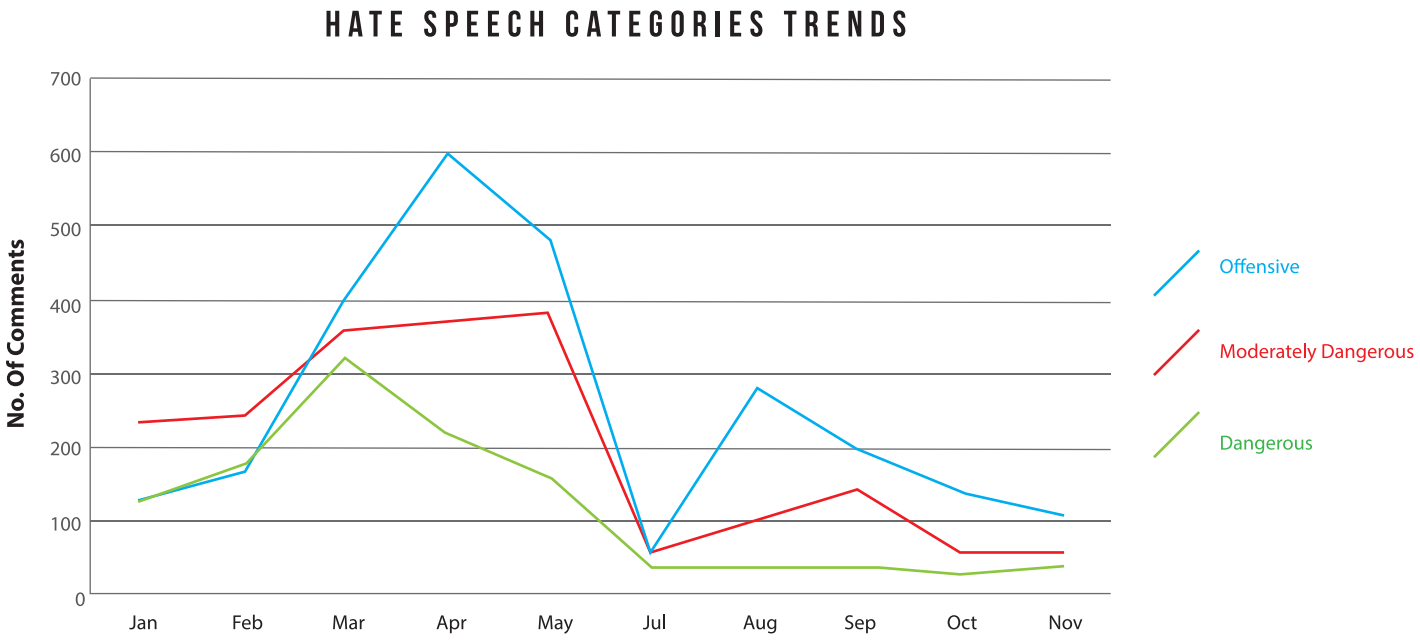
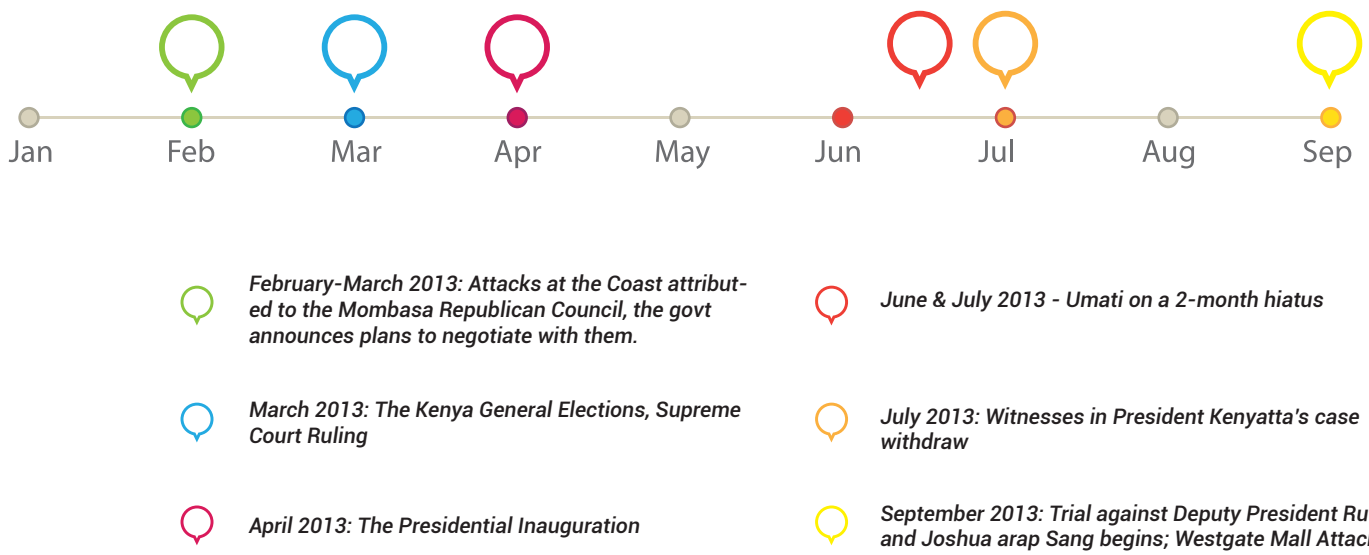


Figure 1: *Timeline of hate speech comments by category throughout 2013. N = 5718*



Besides the elections on March 4th, several other influential events were observed over the course of the year:

The International Criminal Court cases

The ICC cases against current President Uhuru Kenyatta, Deputy President William Ruto, and a journalist, Joshua Sang, spurred online discussion and sometimes hate speech in 2013.

Umati data collected throughout 2013 suggests that hate speech related to the ICC cases was not as widespread before the trials began as it has been since. In May, the Kenyan government wrote to the United Nations Security Council asking them to halt the case against President Kenyatta.[34] At the African Union Summit in the same month, conversations focused heavily on the narrative that the ICC was targeting African leaders.[35] These events elicited reactions, some of which account for the spike in the offensive speech category as seen in Figure 2 below. William Ruto's and Joshua Sang's trials began on September 10, at which point Umati noted a surge in hate and dangerous speech related to the cases through to October (see Figure 2). The push to have the President's case deferred for a year, and its eventual postponement,[36] contributed to much of the conversation and subsequent hate speech in the month of November.

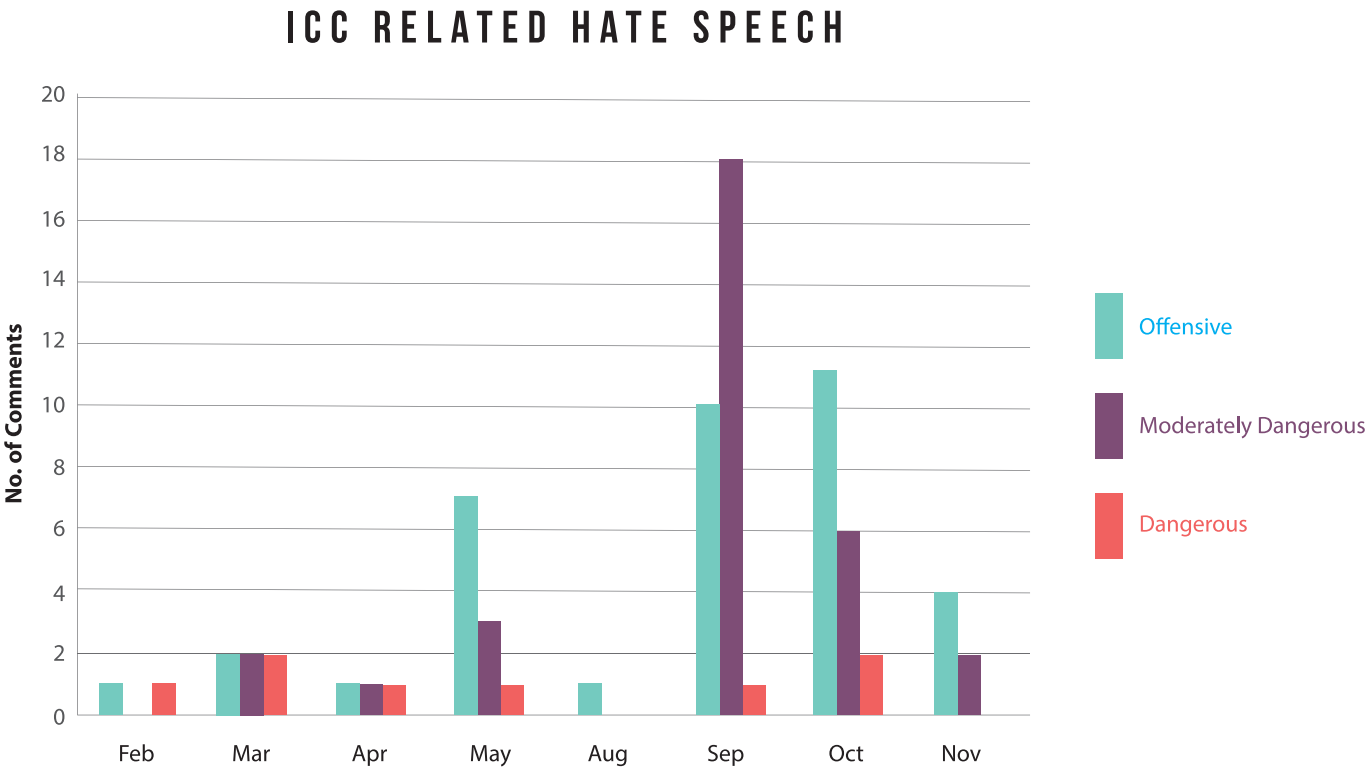


Figure 2: *Hate speech related to the ICC cases throughout 2013. N = 76*

34 Kenya asks UN to halt ICC charges against Kenyatta. (2013, May 9). Retrieved from http://bbc.co.uk

35 African Union accuses ICC of 'hunting Africans'. (2013, May 27). Retrieved from http://bbc.co.uk

36 Uhuru Kenyatta's trial at the ICC moved to February 5. (2013,October 13). Retrieved from http://nation.co.ke

The ICC cases appear to have ignited ethnic tensions online due to the nature of the cases, which focus on the post-election violence of 2007/08. Witness testimonies before the Court triggered suspicions and allegations between the Kikuyu and Kalenjin tribes as the President is Kikuyu and most witnesses testifying against him are Kalenjin, while the Deputy President is Kalenjin and most witnesses testifying against him are Kikuyu. The onset of the trials also saw members of the Kikuyu and Kalenjin community accuse the Luo community of betrayal for their supposed support of the international proceedings. This resulted in a significant surge in hate speech online, as members from these ethnic groups engaged in fierce online exchanges. In addition, the first witness' identity was leaked online to prove that the witness belonged to a certain ethnic community. At least one mainstream media journalist shared the leaked identity information on their Twitter page. In response, the court issued a stern warning to journalists, bloggers, and social media users.[37] The framing of the ICC cases in mainstream media also fanned the flames of online hate speech, as headlines hinted at this sense of ethnic betrayal at The Hague.[38] Findings related to the ethnic breakdown of hate speech comments are discussed later in this section.

The International Criminal Court cases

Kenya has increasingly come under terror-related attacks throughout the country since the start of the Kenya Defence Forces' military operation in neighbouring Somalia. Several of these attacks have been considered to be retaliations by the Al Shabaab militant group. In April, for instance, unknown gunmen attacked a hotel in Garissa, a town in North Eastern Kenya, killing six people and injuring others.[39] This led to online hate speech against Muslims and Kenyan Somalis, stereotyping members of both groups as terrorists, accounting for the spike noted in that month (see Figure 3 below).

The Nairobi Westgate Mall attack on September 21, 2013 was the most severe attack by Al Shabaab in Kenya in recent years. All three categories of hate speech spiked in September, although offensive and moderately dangerous increased dramatically, and dangerous speech only increased slightly.

37 *Maliti, T. (2013, September 18). Judges Caution That It Is An Offense To Reveal A Witness' Identity. The International Criminal Court Kenya Monitor. Retrieved from http://www.icckenya.org/2013/09/judges-caution-that-it-is-an-offense-to-reveal-a-witness-identity/*

38 *Sambuli, N., Morara, F. 2013. Online Monitoring of the ICC and Devolution Processes in Kenya: July - September 2013. iHub Research. Retrieved from http://www.ihub.co.ke/ihubresearch/jb_ICCDevolutionProcessesReportpdf2013-11-1-09-42-07.pdf*

39 *Six killed in gun attack in Garissa. (2013, April 18). Retrieved from http://nation.co.ke*

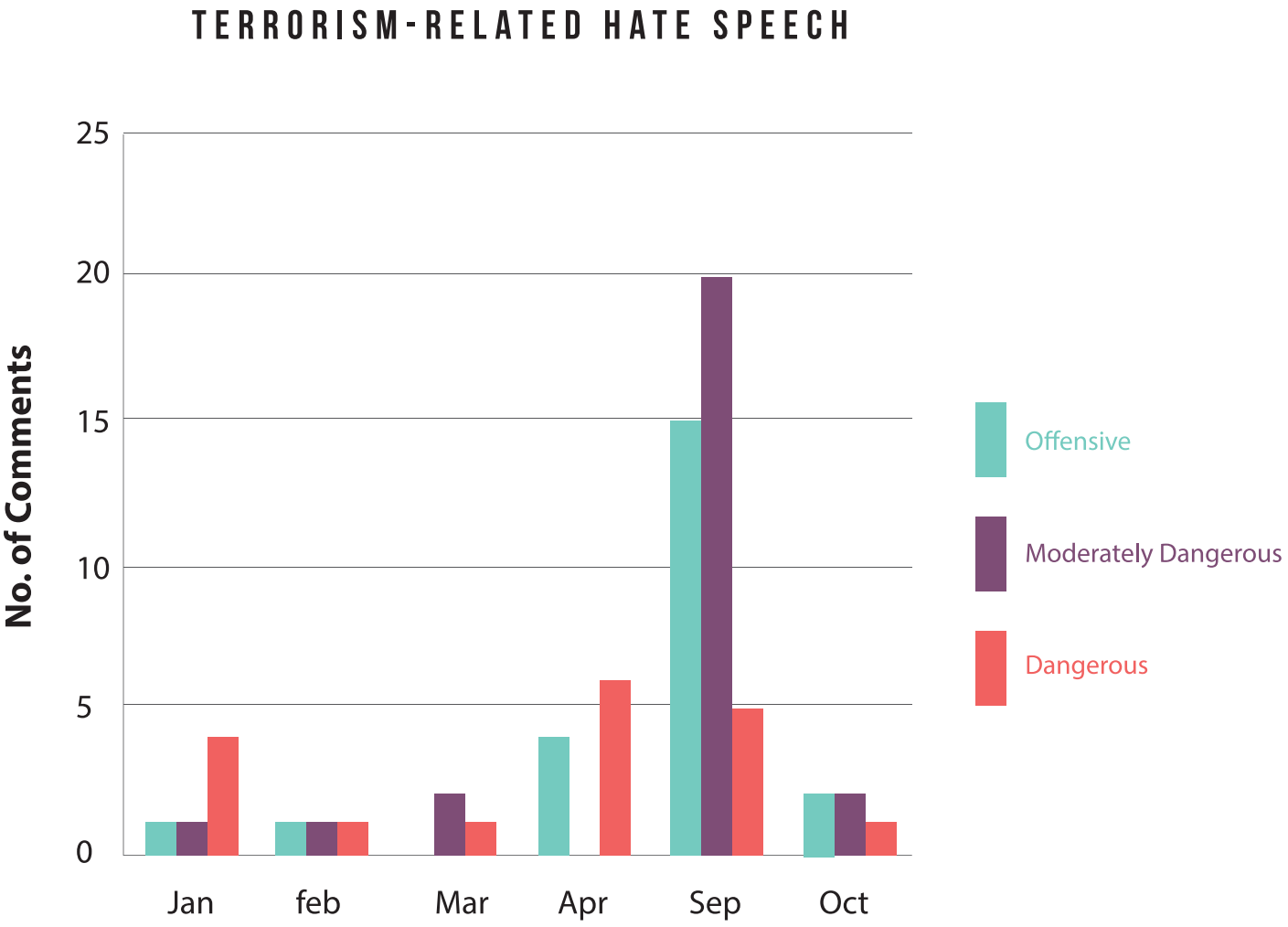


Figure 3: *Terrorism-related hate speech reactions throughout 2013. N = 67*

A list of additional key Kenyan events in 2013 that may have influenced online hate speech is included in Appendix B.

1. Ethnicity and Religion: two recurring themes in 2013 Kenyan hate speech comments

a. Ethnicity-based hate speech

Umati has noted that much online hate speech comes in reaction to events that transpire or are witnessed offline. In 2013, one of the key events that took place in Kenya was the March 2013 General Election that was closely contested and whose outcome was challenged through a petition filed by presidential contender Raila Odinga at the Supreme Court. Although the election was mostly peaceful, politicians and their supporters still used ethnicity-based arguments to win votes. This ethnic campaigning further entrenched divisions across the country, a phenomenon visible in the hate speech comments during the same time.

During the election period, hate speech mostly targeted ethnic groups, with the Kikuyus, Luos and Kalenjins, three of the dominant ethnic groups in the political arena consistently being the most mentioned. The combination of Kikuyus and Kalenjins in hate speech could be attributed to the ruling coalition's ethnic makeup: the President is Kikuyu and his Deputy is Kalenjin. This coalition has been referred to as the 'tyranny of numbers' because these are the two most populous ethnic groups in Kenya and therefore will always form a majority if they create a coalition.

There was a gradual increase in hate speech against the three aforementioned ethnic groups in the months leading to the elections. Hate speech against Kikuyus and Kalenjins as a joint target, and political party members, decreased slightly after the elections. March and April saw the biggest spike of hate speech against Kikuyus. Hate speech targeted towards Luos spiked in April after the election petition and the Supreme Court's ruling. Presidential contender Raila Odinga, of the Luo community, filed and lost the election petition case contesting the election results.

There is no clear trend among other ethnic groups as the intensity of speech directed at each, fluctuated over the year (see Figure 4). Hate speech targeted at political parties could be viewed as an extension of hate speech targeted at members of ethnic groups because a primary identity point, by politicians and citizens, is one's ethnicity. This is particularly the case with comments that reference the two main political coalitions in the country: the Jubilee Coalition, which forms the current government, and the Coalition for Reforms and Democracy (CORD), the minority party in Parliament. Hate speech that targets the Jubilee Coalition often includes references to members of the Kikuyu and Kalenjin communities. Similarly, comments that target CORD usually include references predominantly made to members of the Luo community.

Throughout 2013, with an exception in the month of September, Kikuyus are the dominant group at whom hate speech is targeted, followed by Luos.

b. Religion-based hate speech

After the Westgate Mall attack on September 21, there was a surge in hate speech targeting Muslims. Al-Shabaab, the Islamic militant group, claimed responsibility for the attack, citing payback for the Kenyan military's involvement in Somalia.

This led to an online outburst claiming that all Muslims are terrorists, and other hate speech following similar reasoning. Other terror-related attacks throughout the year and after the Westgate Mall attack have also contributed to hate speech against Muslims.

2. Online hate speech disseminators largely identify themselves with a real or fake name and use widely understood languages.

a. Leading disseminators of hate speech

We defined speakers as 'identifiable' when they used their real name, or a pseudonym when making hate and dangerous speech comments online. We coded speakers by whether or not they had an online identity associated with their account since this entails a traceable history of online activity from a user profile. Without any online identity, one cannot have a sustained 'online relationship' with other users.

Identifiable commenters were the highest drivers of hate speech online for all three categories - offensive, moderately dangerous and dangerous speech. They accounted for 96% of all hate speech and 97% of the dangerous speech subset collected in 2013. Figure 5 breaks down the total comments in each of the three categories by the type of commenter. The number of identifiable commenters during the election period (January - April) was higher than in later months of the year, as shown in Figure 6. This could be attributed to the registration and commenting policies enforced by blog and tabloid sites, which we saw become more lax after the elections passed peacefully in March. In particular, one tabloid news site where dangerous speech was regularly found, dropped its registration policy in May and allowed users to comment anonymously again.

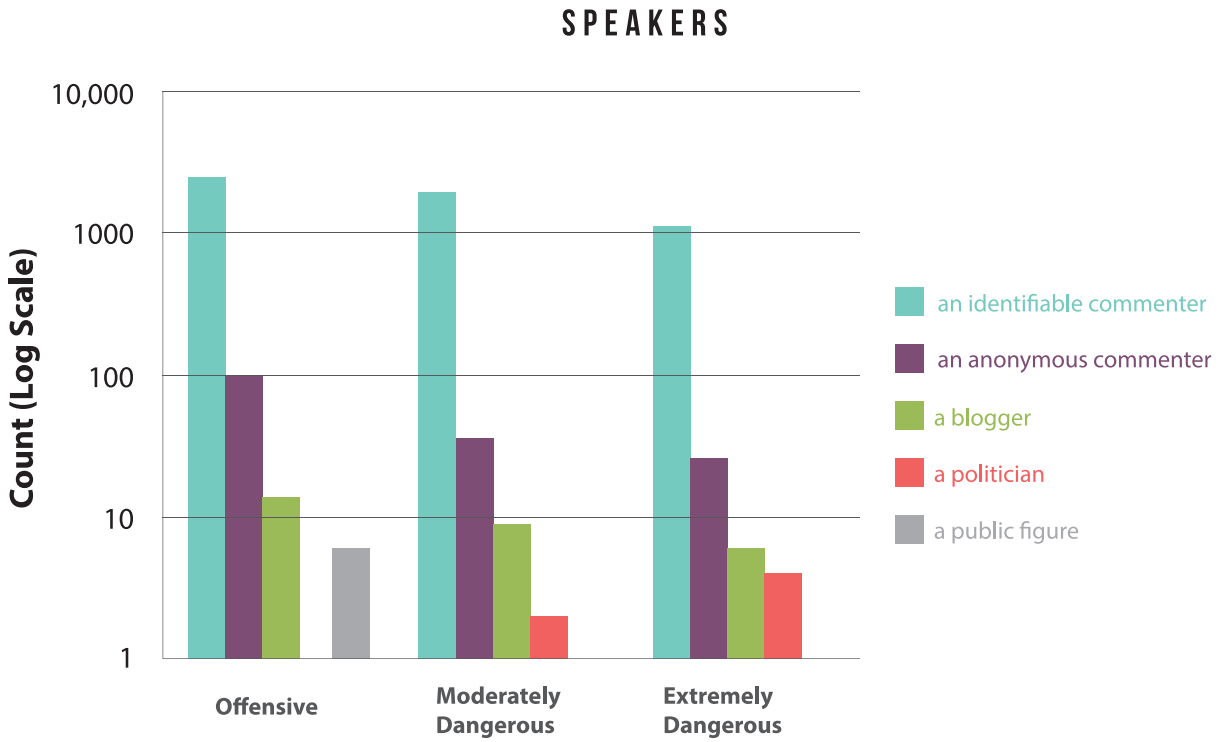


Figure 5: *Types of commenters for each hate speech category n = 5714*

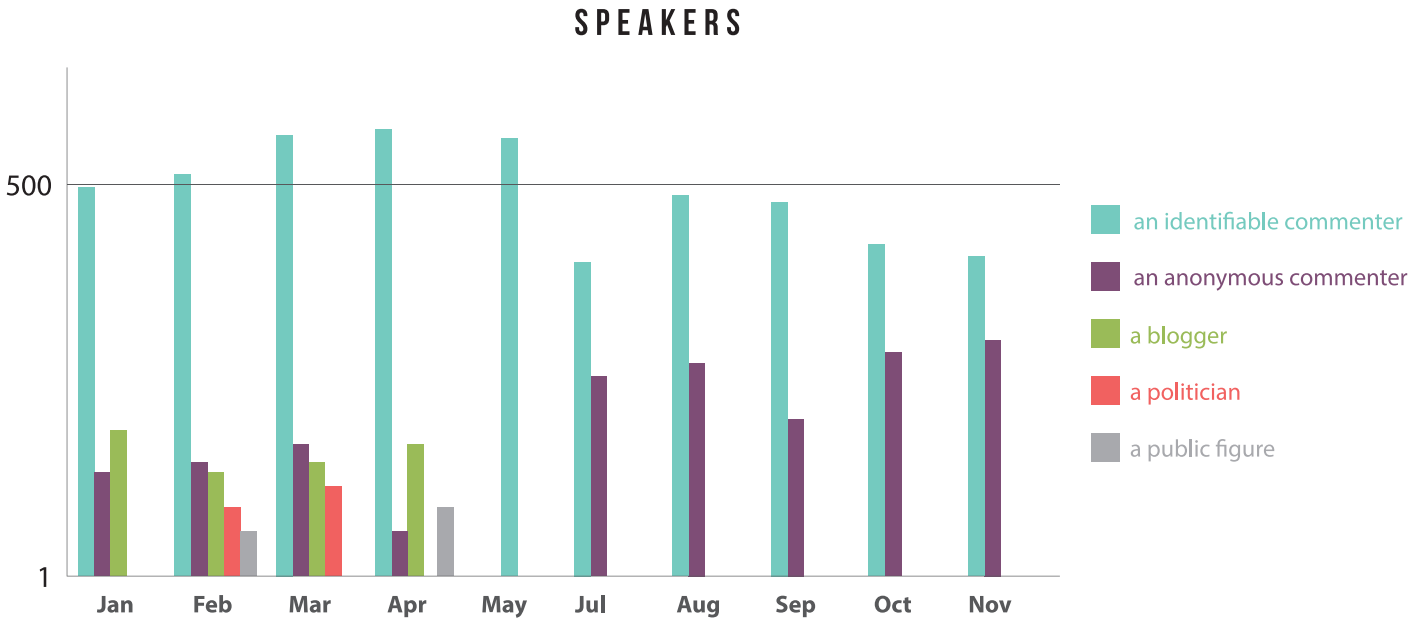


Figure 6: *Hate speech propagators per speaker category throughout 2013. N= 5714*

As the country moved past the election period, the number of comments by identifiable commenters declined. Nonetheless, this category of commenters remains consistently higher than anonymous commenters throughout the year.

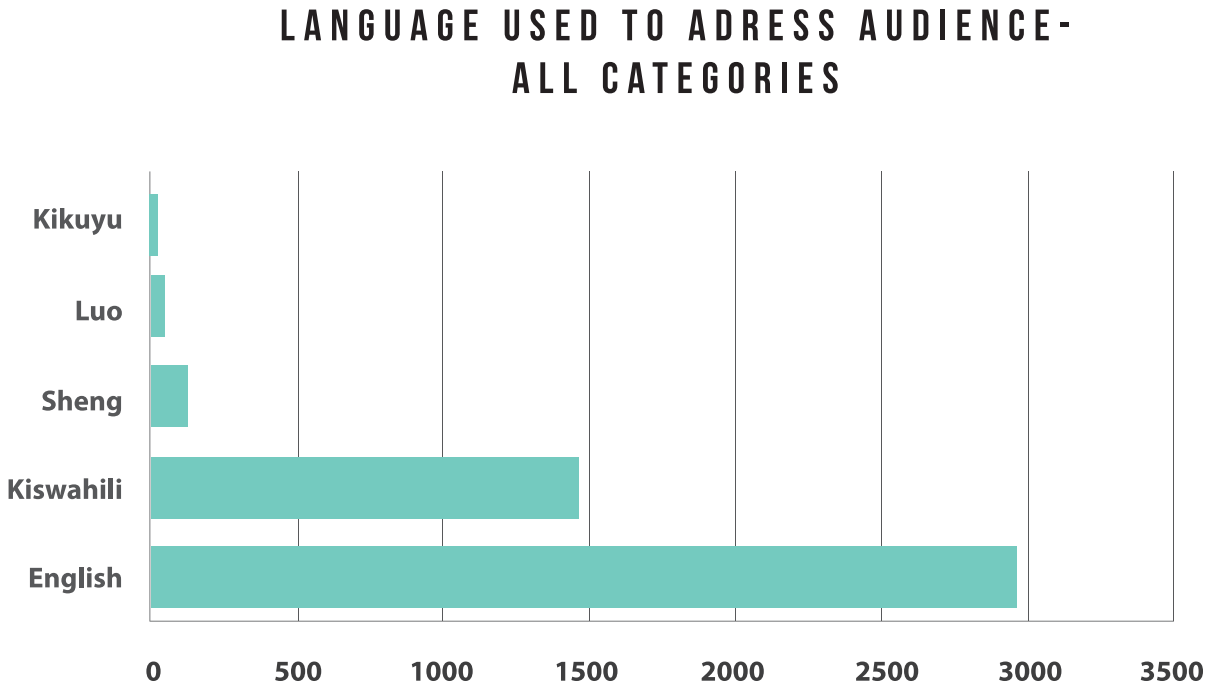
Identifiable speakers do not necessarily carry much influence online and in general, received few observable responses to their comments online. It is possible that such commenters believe their comments will not generate any serious consequences due to their lack of influence, and therefore are not concerned with being linked to these comments. This sense of 'online impunity' could be further attributed to the fact that there have been no cases in Kenya where online hate speech propagators have been successfully prosecuted. In March 2013, two renowned Kenyan bloggers were charged with stirring ethnic hatred through their social media pages . However, while these charges received some initial publicity, public attention waned when the court took no subsequent action. The NCIC holds that the cases are still underway but any progress on the cases has not been made available to the public. In another effort to address hate speech, one of the country's leading TV stations, NTV, started a campaign to name and shame 'tribal extremists.' The campaign ran during the period leading to the 2013 elections, but it was ultimately too short-lived to shift behaviours and perceptions. Furthermore, the fact that very few politicians, among several known to have employed inflammatory speech, have been prosecuted could embolden hate speech disseminators online, giving them the sense of being safe because these noticeable figures have not been pursued either.

Politicians and other public figures (including bloggers), are distinguished from the identifiable commenters; they have more influence, although their contribution to online hate speech is smaller in quantity. In general, the comments of these public and influential speakers take the form of a reaction to a news or blog article, with few instances of freestanding posts, especially after the election period, as seen in figure 6 above. In one instance, hateful comments by politician Johnstone Muthama (now Machakos County Senator) were uploaded online and reported on in blogs, which stirred further hateful speech. The politician's notoriety resulted in the creation of a Facebook page that raises awareness on the impact of his speech *(see Appendix C).*

a. Leading disseminators of hate speech

The most common languages for expressing hate speech imply that those who engage in online hate speech are either most comfortable in the national languages (English and Swahili) or perhaps intend for their speech to be widely understood (as opposed to use of coded or tribal language). English, Swahili and Sheng emerged as the three most common languages (respectively) used in online hate speech.

These are the languages that reach the greatest audiences in Kenya; English and Kiswahili are the country's official languages, and Kiswahili is also the national language. Figure 7 shows the breakdown of comments by language. We hope to conduct further research to investigate potential use of linguistic 'code-switching' between the various languages in one conversation.



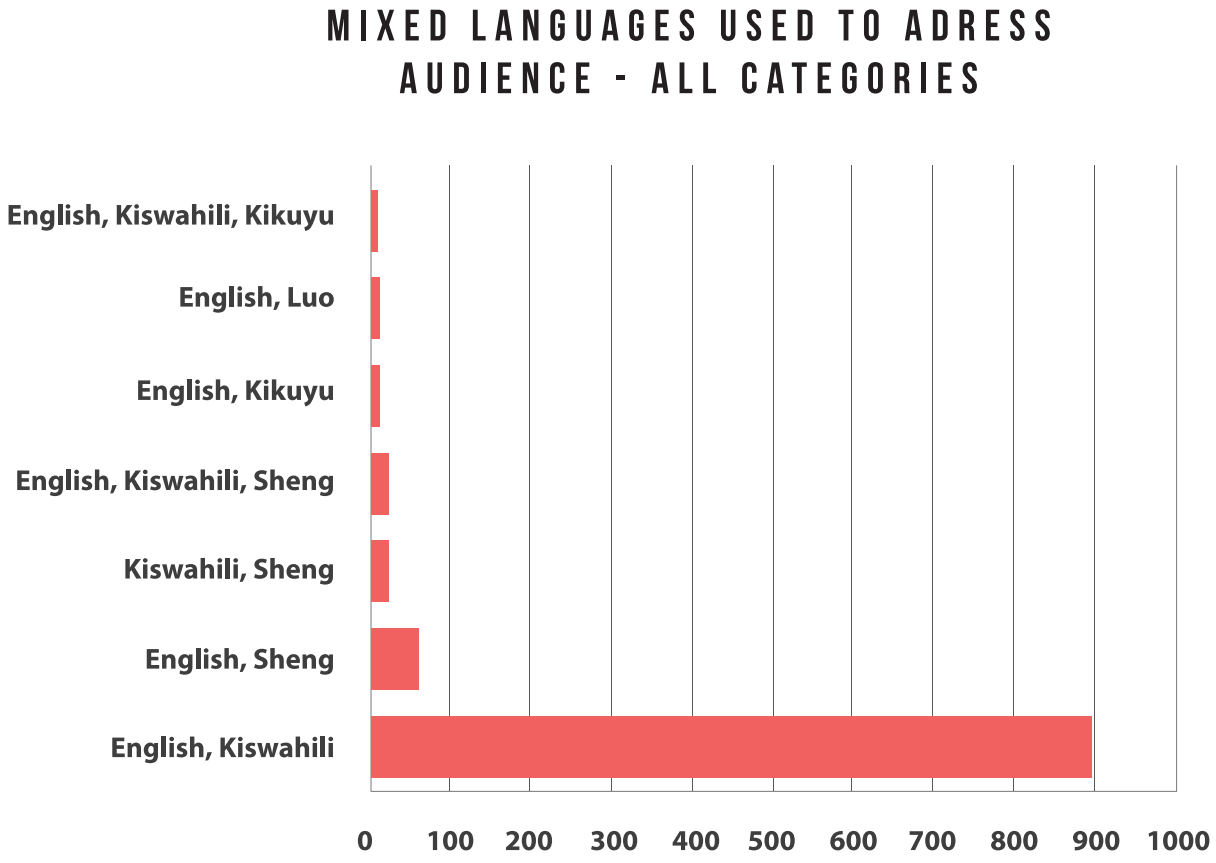


Figure 7a and b: *Languages used to disseminate online hate speech. Na = 4610; Nb= 1042*

3. Online Platforms

Over 90% of all online inflammatory speech captured by Umati was on Facebook, making it the highest source of such content. The online platforms monitored by the Umati project included Facebook, Twitter, online blogs and forums, comments sections of Kenyan mainstream media online newspapers, and YouTube channels for the five main Kenyan media houses.

Facebook, the most popular social media platform in Kenya, was the predominant source of online hate speech captured by Umati. As Figure 8 shows, 90% of all hate speech collected came from public Facebook pages and groups.

Cite Location	Count of the items cited
A Facebook Post in a public group/page	89.75%
A comment in response to an online news article	4.62%
A comment in response to a public blog article/forum	2.41%
A facebook post in a private group/page	0.86%
A comment in response to a private blog article/forum	0.75%
A blog article in a public blog/page	0.51%
A tweet	0.59%
A video on Youtube	0.28%
An online news article	0.12%
A blog article in a private blog/forum	0.05%
A picture from facebook	0.05%
Grand Total	100.00%

Figure 8: *Online platforms on which hate speech statements were found throughout 2013 (n=5717). A private blog is where one needs to sign in to be a member in order to comment on a given post, i.e. one cannot comment unless the owner approves membership. A public blog does not require signing in for commenters.*

Comments posted on online news articles, blog articles and forums were the next most common sources for hate speech. Figure 8 shows the frequency of hate speech on blogs, newspaper sites, and forums throughout the year. The comments section of online news articles harboured the second most incidents of dangerous speech. Although Twitter is the second most popular social networking site in Kenya, very few occurrences of hate speech were collected off it (see Figure 9 for a comparison of Facebook and Twitter).

It is good to qualify that over time, as noted in the methodology section, as the monitors began finding more dangerous speech on Facebook, they also focused more of their efforts on collecting data off of Facebook. In order to strengthen our finding that most of the dangerous speech in the Kenyan online space was found on Facebook, we are in the process of revisiting the Twitter data from that period (Sept 2012 - Nov 2013) using automated data collection by keywords and filters. Our findings from this revisiting of the Twitter data will be published in future reports.

Month	Facebook	Twitter	Facebook:Twitter Ratio
JAN	97.0%	3.0%	32 : 1
FEB	99.4%	0.6%	169 : 1
MAR	98.7%	1.3%	74 : 1
APR	99.6%	0.4%	280 : 1
MAY	100%	0.0%	NA
JUL	100%	0.0%	NA
AUG	100%	0.0%	NA
SEP	99.7%	0.3%	326 : 1
OCT	99.4%	0.6%	170 : 1
NOV	100%	0.0%	NA

Figure 9: *Comparison of hate speech found on Facebook and Twitter in 2013. N = 5214*

Why was Facebook the dominant site for Kenyan online hate speech?

Estimates of Internet penetration in Kenya stand at approximately 47% of the total population, or 19.1 million. Among these users, Facebook is the most commonly used social media platform with latest estimates at 2,015,600 Facebook users as of April 2013. This number also represents a relatively wide socio-economic group, with media reports and findings from fieldwork suggesting that Facebook use and growth is occurring in both urban and rural Kenyan settings. Thus, it is to be expected that Facebook, with a broader socio-economic group of Kenyan users could have wider spread and use for hate speech dissemination online.

Facebook communities are formed around groups and pages, allowing users who share similar opinions and ideas to congregate and engage in discussions around a topic. Members can participate whenever they desire because topics do not necessarily expire after a period of time. Thus, conversations can persist over time, with users returning to the same conversations on Facebook to check back on what others have said and continue to add to the discussion. Similarly, in the comments section of a blog post, readers can engage as long as the moderator/author leaves the post open for such activity. Twitter, by contrast, is better suited for real-time news dissemination and discussions; all shared in 140 characters per tweet. Due to the real-time nature of most Twitter feeds, new topics tend to overshadow those trending previously. It is therefore easier to revert to a hate speech-based conversation or topic on Facebook (as well as online forums). We have also found that users revert back to hate-based conversations on Facebook long after they initially began. Further inquiry around this behavior of going back to such pages should be conducted.

Finally, Facebook also allows for covert and overt behaviour among the users; a user is able to control the privacy of their personal profile and engage in public or private discussions on pages or in a group. Thus, someone could refrain from disseminating hate speech on their personal page (which is visible to the public, or all of their friends), and instead, only engage in hate speech in the perceived safety of a group or page that is open to and used by those with similar opinions. Twitter, in contrast, is designed so that a single user's posts are contained in one domain and are all viewable by the public or at least the user's followers. In either case, it is not possible to designate some tweets as private or restricted and others as public, unless a user creates multiple accounts to disseminate different content. So, the setup of Facebook is more amenable to those who might want to post hate speech that is not visible to the public.

Cuffing in the online sphere

There may also be another reason that hate speech was curtailed on Twitter. From the Umati data, we noted that hate speech in the Kenyan twittersphere had been subjected to "KoT (Kenyans on Twitter) cuffing, where tweets considered unacceptable by the status quo were openly shunned, and the author of the tweets, publicly ridiculed.

A particularly interesting example of 'KoT cuffing' was noted during the election period. One twitter user posted a series of tweets that were perceived as unacceptable by other tweeters. He asserted his impunity in several instances, including when the Attorney General tweeted at him, warning that his speech could land him in trouble. KoT cuffing responses ranged from demanding apologies, to forwarding his account activity to authorities such as the NCIC, and insulting him. In response to the cuffing, he denied having posted the offensive comments.

Below are some examples of his offensive tweets from March 2013:
*"I can die for Uhuru Kenyatta. And..... I can kill for Uhuru kenyatta.**Don't sue me**#patriotism"*

*"Even if you cc NCIC, i'll say, come baby come and arrest me. Uhuru will get me out.... Yay! *Shoot me now*"*

This second tweet stirred more cuffing reactions than the first one, including a response from the Attorney General's twitter account:
"Your tweets might get you in trouble. Think before you leap."

His reaction to the Attorney General is indicative of a sense of impunity, due to the lack of consequences for more influential people that propagate hate speech. He responded:

"Sorry Mr. But with all respect, why do you hunt me, a useless kenyan but spare big politicians who insight?"

And you, @agmuigai, with all the wisdom you think you have can't follow me and my baseless tweet leaving guys like muthama out there.

While this user is still active on Twitter, he has since toned down on inflammatory tweets. Umati is interested in working together with the tech and human rights communities to better understand such online self-regulating behavior and how it may be encouraged.

Censoring an entire account or page is another approach (though not one we advocate for) to dealing with hate speech on social media. For example, during the Westgate Mall attack, several Twitter accounts that claimed to be linked to the Al Shabaab were suspended for posting direct threats to Kenyans. Given the global attention that the attack warranted, it is probable that Twitter was monitoring accounts connected to Al Shabaab for any violations of Twitter rules.

It has also been noted that some Facebook pages that were active in disseminating or inciting hate speech through posts eliciting comments become dormant over time, that is, have no recent updates or comments from users. This is especially the case with pages and groups that were rather active during the election period, and some during the Westgate Mall attack. This could indicate that groups and pages are formed around events of interest, either before they occur (in the case where this can be established, e.g. elections) or those set up when an event unfolds, and that once the event is past, users move on to other pages/groups of interest. This kind of migration is interesting, and future Umati work will attempt to establish whether patterns around shifts of conversations and online hate speech disseminators can be established and used for predictive modeling.

Other examples of dealing with hate and dangerous speech observed in the Kenyan online sphere are discussed in the Umati Phase I final report.

C

Umati Report

CONCLUSION

January - November 2013

As noted in the introduction, ethnicity has been a primary lens through which political, economic and social issues are viewed and reacted to in Kenya. However, as different events transpired through 2013, most notably the Nairobi Westgate Mall attack, religious affiliation appears to have become a new frontier for Kenyan online discriminatory and hate speech. We have also found that much of the online hate speech tends to be a reaction to an offline event and in some instances, a reaction to the framing of an event by news organizations as well as blogs and other influential persons disseminating information online. These are both ongoing areas of inquiry as we track the characteristics of online conversations changing over time.

We have also found that online speakers are identifiable, either through use of their real name, pseudonyms, or through a traceable history of online activity. The language used to disseminate hate and dangerous speech is widely understood in the country, though a few incidents of coded language, known to have been used in past historical contexts were noted during the election period. Given the database of incidents captured throughout the project, and through continued monitoring, we plan to study how particular speech references come to, over time, be understood as a call to action.

Thirdly, we found that much of the dangerous speech in the Kenyan online space is found on Facebook, and assessed the social network information-sharing structures for both Facebook and Twitter to better make sense of this finding. Hate speech collected off Facebook has also been found to predominantly be reactions to events that take place offline, that are then reported about on traditional media (print, broadcast and online), on blogs as well as comments or reactions to events as they unfold in real time. We will be conducting further content analysis to assess the characteristics of hate speech found on Facebook, as a function of being either an original post, a comment, page title, and more. In our automated collection phase, we are re-scraping data for comparative analysis with the data collected by human monitors in 2013, and will pick out more metadata that can better help us in our automated online hate speech monitoring.

We are yet to find instances of online hate speech catalysing events offline. As 'netizens' congregate and converse online, forming networks around issues of interest, the possibility of organizing offline reactions to online conversations is likely. Authorities, while acknowledging and endorsing online media through adoption (as has been the case with various arms of the Kenyan government) are yet to appreciate these findings and address them effectively. The immediate risk, as seen in the local and continental legislative process trends, could lead to the infringement or rolling back of freedoms of the Internet and expression that facilitate the space where both good speech and hate speech are conducted on the web. As part of our third objective in Phase II of Umati, we will be exploring how to reduce online dangerous speech through online-based and offline civic engagement efforts, by engaging as many stakeholders as possible on matters pertaining freedoms of speech and expression, and how these are understood and exercised. While we are primarily looking at online methods, we are also building on experience from conducting a short campaign called Nipe Ukweli (Swahili for "Give me truth"), conducted in the weeks leading up to the elections both online and offline (realizing that there are limits to the influence of online content). Worried about the increasingly high levels of Dangerous Speech online in the months leading up to the elections, Umati launched the Nipe Ukweli campaign to explore non-governmental ways to reduce dangerous speech. See Appendix D for some of the lessons learned from the campaign.

Recommendations

Curtailing online hate speech should not infringe on freedoms of expression, including online freedoms. It is important to establish frameworks at a policy level that assist in distinguishing between speech that should be sanctioned and that which is protected under the freedom of expression.

This is especially crucial as authorities adjust to the new online avenues of self-expression. The root causes of hate speech—both online and offline—should be investigated and addressed. We acknowledge that monitoring, in and of itself, is not a complete solution. Lastly, citizens and government authorities alike must also be equipped with the knowledge to recognize hate speech and appreciate the impact it has on society. The NCIC, in its role and mandate of promoting harmony, cohesion and peaceful coexistence/integration should facilitate forums and outreach for citizens to address the core issues that are manifested through inflammatory and inciteful speech. Findings such as Umati's could be a starting point to shape future initiatives.

Impact

Over the course of 2013, the Umati project has developed the largest database of hate speech from one country to date (7,000+ incidents). The project garnered incredible attention both locally and internationally, with over 24 different news articles written about Umati from media houses including Al Jazeera, The Guardian, Nation Media, and Reuters. In addition to wide media coverage, the Umati project was able to gain traction in academic, non-governmental, and governmental circles. Our research findings and reports were cited in several publications such as the International Crisis Group report on Kenya's 2013 elections.

The Umati project is now globally recognized for its flagship work around online hate speech monitoring. With new support for Umati Phase II, we will be able to continue to define more clearly the boundaries of online freedom of expression. Over the next two years, Umati Phase II will also entail automating, where possible, the monitoring process in Kenya through Machine Learning (ML) and Natural Language Processing (NLP) in order to improve the productivity and lower the cost of running the project in other country contexts.

Umati Report

APPENDIX A
METHODOLOGY

January - November 2013

Categorization formula

To enable the sorting of the hate speech into the three aforementioned categories, a categorisation formula was devised that was mainly dependent on two questions on the coding sheet/categorisation form.

The two questions are:

a) How much influence does the speaker have on the audience?

This question was on a scale of 1 to 3, with 1 being little influence and 3 being a lot of influence (code N)

1-Little influence
2-Moderate influence
3-A lot of influence

b) How inflammatory is the content of the text?

Again on a scale of 1 to 3, with 1 being barely inflammatory and 3 being extremely inflammatory (code M)

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

The two questions were aimed at gauging four main factors from the Benesch framework:

- *The speaker and their influence over the audience.*
- *The susceptibility of the audience*
- *How offensive is the content of the speech.*
- *The social and historical context of the speech.*

The answers to these two questions above were dependent on other five questions on the form and these were:

A1 The speaker is

- *A politician*
- *A journalist*
- *A blogger*
- *A public figure (this also 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 comments, retweets, likes, shares)*
- *Received a moderate observable response*
- *Received no observable response*
- *Was a reply to a statement, post, or comment*

With the above set of questions, the monitors were then able to answer the first scale of the question relating to the speaker's influence listed above (code N).

The second set of questions would then be used to determine the content of the speech

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

- *Discriminate*
- *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 audiences 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*

The two sets of questions above then determined the second most important question, on the inflammatory scale of the content, also listed above (code M).

Finally, based on the answers from the two scale questions, (code N and M), a formula was used that enabled the grouping of the statements into the three hate speech categories:

Sorting

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
N3+M3=Bucket 3

Hate Speech Categories

Bucket 1: *Offensive speech*
Bucket 2: *Moderate dangerous speech*
Bucket 3: *Dangerous speech.*

New categorization entries were introduced at various stages as monitors adjusted the methodology and categorization process. The name of the speaker was tracked to establish if there were notorious and repeat hate speakers. Monitors also included the date the entry was recorded, the speaker's location, if identifiable, and whether any coded language (such as proverbs or 'hidden' sayings) was used to propagate hate speech.

Categorization Form Used by Umati Monitors

Title of the article/blog post

Original date the article/post was put up
Day/Month/Year

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

What country/location is the speaker from?
Optional (only if mentioned)

***Actual offensive text** **

English translation of actual offensive text
Only if original content is not in English

Any additional comment on the actual offensive text

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

- [ ] Yes
- [ ] No

Does this text relate to the devolution of government? *
E.g. Why are all elected officials for the county of Kisumu are Kikuyus!

- [ ] Yes
- [ ] No

Link

Appendix A

***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 from youtube
- ○ a video from media house
- ○ a picture from Facebook

The audience is being addressed in?

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

Appendix A

***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

Appendix A

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/Injure
- ☐ 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
- [ ] political party members
- [ ] 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.***

B

Umati Report

APPENDIX B

2013 EVENTS THAT MAY HAVE INFLUENCED ONLINE HATE SPEECH

January - November 2013

Several key events in 2013 could be related to increased frequency of dangerous speech. Such events are:

- *General elections and by-elections thereafter.*
- *The contentious bills passed by the parliament (these include: Media Bill and the Kenya Information and Communications (Amendment) Bill, Matrimonial Property bill, VAT bill, Retirement Bill, bill seeking to give Members of Parliament powers to set their own salaries).*
- *Devolution issues (push for referendum on county funds, county bills, ethnic rivalry in coun ties).*
- *ICC trials at The Hague.*
- *Repatriation of Somali refugees from Kenya back to their country.*
- *Deaths of politicians and religious leaders all over the country.*
- *Bestiality cases.*
- *The Nairobi Westgate Mall attack.*
- *Politicians' behavior. (Nairobi Governor, Evans Kidero slapping the Nairobi Women's Rep resentative Rachel Shebesh, Nairobi Senator Mike Sonko and Ms. Shebesh's alleged affair and subsequent viral photos).*
- *Kenya Government's announcement on shutting down refugee camps in Northern Kenya and repatriation of Somali refugees.*

In relation to the Umati data, the relationship between the online and offline activity indicates that instead of trying to draw out the impact that online speech has on violence on the ground, it might be more accurate to view observable online activity as a small window into the offline sentiments of Kenyans.

C

Umati Report

APPENDIX C

EXAMPLES OF HATE SPEECH

January - November 2013

Machakos County Senator Johnstone Muthama's hateful comments, just after the elections, were uploaded on Youtube, and stirred further hateful reactions on Facebook pages and on the video's comments section.

His statements, in Kiswahili, allude to stealing of votes by the majority party, and further claims that Kenya's president is the opposition party's leader. Though he made no mention of tribes, just leaders, his statements were inflammatory and inciteful. He was yet again implicated in hate speech use when he expressed his disdain on the Governor's appointment of an Asian to the Machakos County Executive Committee. The NCIC called for his arrest after he failed to heed summons over hate speech accusations. In responding to the NCIC, he said he would not be intimidated by use of the term 'hate speech' to silence opposition. Facebook users took notice and a page titled 'Kenyans Against Johnstone Muthama's Hate Speech' was set up to demand an apology from the Senator.

Johnstone Muthama's Hate Speech Comments on YouTube video, with English translation:

"Niliiwambia wezi wamepanga kuiba, lakini nawaambia hivi wakithubutu watakiona cha mtema kuni, watakiona cha mtema kuni, ni wezi ambao hawana aibu, waliiba mwaka wa 2007 mwaka huu hatutakubali, hawa watu walitoa wapi kura? Waliiba...haturudi nyuma mpaka uingie ikulu, tunasema rais wa kenya ni Raila."

("I told you thieves planned to steal. I'm telling you now if they dare, they will regret. They are thieves with no shame; they stole in 2007 and they will not be allowed this year. Where did they get their votes? They stole… we are not going back until you get in state house. We are saying the President of Kenya is Raila.")

D

Umati Report

APPENDIX D

IHUB RESEARCH'S NIPE UKWELI ('GIVE ME TRUTH') CAMPAIGN

January - November 2013

Reposted from an iHub blog post written by James Ndiga (April 25, 2013) available at: http://www.ihub.co.ke/blog/2013/04/the-responsibility-of-working-with-communities/.

In line with our third objective of Phase I looking at ways of furthering civic education on dangerous speech so that Kenyans are more responsible in their communication and interactions with people from differing backgrounds, we came up with an initiative in February 2013 under Umati called "NipeUkweli" or "Give me truth." NipeUkweli explored non-governmental ways to reduce dangerous speech. In the weeks leading up to the Kenyan election on March 4th, NipeUkweli worked to educate citizens both online and offline about:

- *Different categories of hate speech;*
- *How to identify dangerous hate speech;*
- *Frameworks of dangerous speech;*
- *Contents of dangerous speech;*
- *Ways of combating dangerous speech;*
- *How to report such incidence of dangerous speech through Uchaguzi.*

With little time remaining to the elections, NipeUkweli had to devise a quick and effective means of disseminating useful information about Dangerous Speech and ensuring that the target audiences (communities potentially more "vulnerable" to incitement) were reached. We therefore embarked on having community forums and using community radio stations, e.g. Koch FM based in Korogocho, and Safari Africa Radio, to reach out to grass-roots communities in urban Nairobi slum areas.

We also identified violence-prone areas, namely Mathare, Kariobangi, Kamukunji, and Dandora, where we also held various forums. These were some of the sites worst hit by post-election violence in 2007-2008.

It was interesting relating with the people in these areas—from church leaders and the elderly, to hostile youths. All seem to have politically matured since 2007 and were talking of peace, love and unity. This was evident in the way the community members recited community pledges together, singing the national anthem and reciting the Lord's Prayer.

After the completion of the Elections, we decided to conduct feedback forums to understand the community's experiences during elections and on their use of the Uchaguzi platform. "Mwalimu James Umerudi! Karibu sana (Teacher James, You've come back! Welcome!)."

The feedback forums revealed that citizens were glad that the Uchaguzi technology system and NipeUkweli had not failed them. We heard positive feedback that most of their complaints sent through the system were addressed and dealt with. We heard many praises and then an old man (mzee) stood and said "Tumewasifa sana na tuko na imani na nyinyi, kama mliweza kujumuisha haya mashirika yote yafanye kazi na nyinyi wakati wa Uchaguzi mbona msiyashughulikie mawaswala ya security, gender violence, rape, burglary na mengineo ambayo yatatusumbua kwa miaka ingine tano ama mtangoja hadi 2018 wakati wa Uchaguzi mje tena. (We have all praises for you and we are hopeful, we are glad you managed to partner with all these big organization to successfully monitor elections, why don't you now work on resolving day-to-day issues in the slums that include: rape, gender violence, burglary, child abuse among others. Or will you wait until 2018 when we are expected to have our next general election and come talk about Uchaguzi and NipeUkweli platforms again?)."

This blunt question revealed the desire and need for longer-term systems to be in place. Instead of only focusing on election-related violence, we learned through the feedback forums that communities want to use similar technology-based systems to deal with other issues, especially around general security in the slum areas. This was great proof of concept for a technology-based system to address the issues faced in Kenyan communities. But the question that naturally follows is about the mandate of technology organizations and our competence and capacity for addressing community needs and demands.

As much as we empathize with what the citizens need, it cannot be our job, for example, as a Dangerous Speech research project to create a tech system for acting on community reports on their daily security issues. This feeling of helplessness and "it's too big for us to do alone", seems to echo GSMA's recent SMS Code of Conduct, which highlights, "do not launch a…service unless you have the ability (and capacity/resources) to act…failure to do so risks raising expectations unreasonably… possibly to a dangerous level and diminishes the credibility of your service." While GSMA was talking about SMS services, based on our small interaction with the community around this NipeUkweli project, this definitely rings true. Luckily, through frank discussions with the community we made sure that they realized that their asks were beyond our abilities as a research initiative. We did promise to share their stories and desires with relevant other organizations and we have been doing so with human rights organizations. We hope that perhaps this might eventually help these communities to be heard and assisted.

Through this experience, we were reminded that it is important for technology and even research companies to remember to manage expectations and not over-promise when interacting with users and research participants. Otherwise, your reputation and the credibility of your services/products will be diminished. Make sure you have frank and open communications with the communities you engage so that everyone is on the same page about the scope of interaction. If you over-promise and don't deliver, you will be making it harder not only for yourselves, but also for organizations who will want to work with these communities in the future.


umati