Monitoring Online Dangerous Speech in Kenya

Insights from the Umati Project

Nanjira Sambuli / Kagonya Awori

Introduction

The Umati[71] project emerged out of concern that mobile and digital technologies may have played a catalysing role in the Kenyan 2007/08 post-election violence, and that the online dissemination of potentially harmful speech was inadequately monitored. In the build up to the 2007 Kenyan elections, avenues of propagating dangerous speech were generally limited to broadcast media transmissions, print media, SMS and email. Anecdotal evidence suggested that online spaces such as forums and blogs were also used to plan and incite violence on the ground. However, at that time, no system existed to track such data. Incendiary remarks by politicians and 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 have been in place through undertakings by Kenyan civil society as well as police authorities. However, the migration of inflammatory speech to online media remained neither monitored nor analysed.

Since the submarine fibre optic cables landed on Kenyan shores in 2009, Internet penetration has been on a steep increase.[72] Greater access to affordable Internet, especially through the use of smart and feature phones,[73] has seen increased use of social media in the country. Such platforms offer new spaces for people to express their opinions, especially during times of heightened anxiety such as election periods. With over 2 million active[74] Kenyan Facebook users as of April 2013[75] (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,[76] it can be deduced that social media is heavily used by Kenyans, and will continue to grow in popularity.

New media have diversified the audiences that engage in online communication. As these online spaces are a new medium for disseminating inflammatory speech, their influence on the actions of the audience warrants assessment. A possible result is the creation of a vicious cycle as audiences convene around hateful content, converse in self-selected groups and form new ideas or support their original biases with the hateful beliefs of others (see Ayala, pp. 17-21). However, there is a prospect of virtuous cycle creation, as new media spaces can also act as alternative information sources that neutralise the negative impacts of online and offline inflammatory speech. An example of this is noted in the findings section below.

Initially, the Umati project sought to better understand the use of dangerous speech in the Kenyan online space by monitoring particular blogs, forums, online newspapers, Facebook and Twitter. Online content monitored includes tweets, public status updates and comments, posts and blog entries. Umati was launched in October 2012, six months before the Kenya general elections (March 4, 2013) and exists in two distinct phases.

Phase I (September 2012 to May 2013) established the following initial goals:

- To monitor and understand the type of online speech most harmful to Kenyan society.
- 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.[77]

71 Umati is Kiswahili for 'crowd'.
72 Communications Commission of Kenya. (2013). *Quarterly Sector Statistics Report: First Quarter of The Financial Year 2013/14 (Jul-Sept 2013).* Available from: http://ca.go.ke/images/downloads/STATISTICS/Sector%20Statistics%20Report%20Q1%202013-14.pdf. (Accessed 2 Sept. 2014).
73 *Ibid.*
74 Number of people active on Facebook over a 30-day period.
75 Social Bakers Statistics. (2013). Available from: http://www.socialbakers.com/. (Accessed 12 April 2013).
76 Portland Communications. (2012). New Research Reveals How Africa Tweets. 1 Feb. Available from: http://www.portland-communications.com/2012/02/new-research-reveals-how-africa-tweets. (Accessed 2 Sept. 2014).
77 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.

- To define a process for online dangerous speech tracking that could be replicated in other countries.
- To further civic education on dangerous speech, and sensitise the Kenyan public in order that they are more responsible in their communication and interactions with people from different backgrounds.

Phase II (July 2013 to January 2016) further aims:

- To refine the Umati methodologies developed in Phase I and where applicable, increase scalability of the project through automation.
- To test the Umati methodology in other countries in order to improve and increase its global/contextual applicability.
- To explore non-punitive, citizen-centred approaches for reducing dangerous speech online.

Umati methodology for identifying dangerous speech

Umati uses Susan Benesch's definition of dangerous speech, that is, speech that has the potential to catalyse collective violence.[78] Benesch's 'Dangerous Speech Framework' offers the following key variables for identifying dangerous speech:[79] the speaker and his/her influence over a given audience – a political, cultural or religious leader or another individual with a significant following tends to have more influence over a crowd; a vulnerable audience subject to incitement by the influential speaker; the content of the speech that may be taken as inflammatory to the audience and be understood as a call to violence; the social and historical context of the speech – for instance, previous clashes or competition between two groups can make them more prone to incitement; and the medium of disseminating the speech, including the language in which it was expressed.

Umati built on the Benesch framework to form a practical identification method. Specifically, the project found that the following three components of the framework were the most relevant for the identification of online dangerous speech in Kenya:[80]

1. It targets a group of people. It is important to note that a hateful comment about an individual is not necessarily dangerous speech unless it targets the individual as part of a group. In our research, it was observed that dangerous speech towards a group can occur across various lines, including religion, tribe/ethnicity, gender, sexuality, political affiliation and race.
2. It may contain one hallmark of dangerous speech. Three hallmarks that are common in dangerous speech comments, as identified by Susan Benesch,[81] include:

   a. Comparing a group of people with animals, insects or vermin;

   b. 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'); or

   c. Suggesting that some people from another group are spoiling the purity or integrity of the speakers' group.
3. It contains a call to action. Dangerous speech more often than not 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.

Umati Phase I relied on a manual process of collecting and categorising online dangerous speech. Human input proved necessary for contextually analysing and categorising speech statements, which in turn facilitated the creation of an inflammatory speech[82] database. Between October 2012 and November 2013, up to eleven monitors scanned a collection of online sites in seven languages: English and Kiswahili (Kenya's official and national languages respectively); Kikuyu, Luhya, Kalenjin and Luo (vernacular languages from the four largest ethnic groups); Sheng (a pidgin language incorporating Kiswahili, local languages and English); and Somali (spoken by the largest immigrant community).[83] In Phase II, the Umati team has begun work on incorporating more automation in the data collection process where applicable. This is being explored through Machine Learning and Natural Language Processing techniques and tools, which if successful, will significantly increase the scalability and transferability of the Umati project going forward.

78 Benesch, S. (2013). Dangerous Speech: A Proposal to Prevent Group Violence. 23 Feb. Available from: http://voicesthatpoison.org/guidelines. [Accessed 2 Sept. 2014].
79 Not all variables must be present for speech to qualify as dangerous speech. Variables are also not ranked and may carry more or less weight depending on the circumstances. Each instance of speech must be evaluated in terms of the information available.
80 For further analysis see Awori, K. (2013). *Umati Final Report: September 2012-May 2013*. p. 27. Available from: http://www.research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf. [Accessed 2 Sept. 2014].
81 Benesch, S. (2008). Vile Crime or Inalienable Right: Defining Incitement to Genocide. *Virginia Journal of International Law*. 48 (3). Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1121926. [Accessed 2 Sept. 2014].
82 Inflammatory speech is used to refer to all three categories along the continuum: offensive speech, moderately dangerous speech and extremely dangerous speech.
83 The sources list currently covers 80+ blogs and forums, 350+ Facebook users, groups and pages, 400+ Twitter users, all major online Kenyan newspapers and YouTube channels for the five main Kenyan media houses.

Categories of inflammatory speech and their likelihood to catalyse violence

As monitors manually scanned online platforms for incidents of dangerous speech, they recorded the speech acts they perceived to be hateful in an online database. In this process, all dangerous speech statements were translated into English and sorted into three categories (in ascending order of severity):[84]

1. **Category one – offensive speech:** 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 catalysing violence.
2. **Category two – moderately dangerous speech:** comments are moderately inflammatory and made by speakers with little to moderate influence over their audience. Audiences may react differently; to some, these comments may be highly inflammatory, while to others, they may be considered barely inflammatory.
3. **Category three – extremely dangerous speech:** statements are made by speakers with a moderate to high influence over their audience. These statements are seen to have the highest potential to inspire violence, as they tend to constitute an action plan that can be understood and acted upon by the targeted audience. These statements are often stated as truths or orders. Umati categorised all statements with a clear or perceived call to beat, to kill, and/or to forcefully evict a particular group, or an individual because of their belonging to a particular group, as extremely dangerous speech statements.

It is important to note that a causal link is almost impossible to draw between dangerous speech and on-the-ground violence, due to the many factors that contribute to bringing about a physical violent act (see Grayman and Anderson, pp. 25-26). 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. 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 stakeholders. These included donor agencies, Umati partners and Uchaguzi key decision makers who were better equipped with mitigating the threats of violence that Umati collected. This process was triggered five times from January to April 2013 and on-the-ground teams mobilised based on the information passed to them.

Findings

- Over 90% of the inflammatory speech statements that Umati collected in 2013 were from Facebook. This has been attributed to the fact that Facebook is the most popular social media site in Kenya. Umati found however, that other factors come into play that accommodate dangerous speech on Facebook as opposed to the second most popular social media site, Twitter.[85] Most interestingly, a behaviour one of the authors named 'KoT cuffing' was observed on Twitter where '[offensive] tweets not acceptable to the status quo are shunned, and the author of the tweets, is publicly ridiculed. The end result is that the "offender" is forced to retract statements due to the crowd's feedback, and can even close his/her Twitter account altogether.'[86] KoT cuffing, a self-policing behaviour by Kenyans On Twitter (KoT), demonstrates that netizens themselves are capable of employing non-judicial means to counter online dangerous speech.
- Not surprisingly, it was possible to identify those that engaged in dangerous speech online, either via their real names, e.g. by use of their Facebook and Twitter accounts, pseudonyms which can be mapped to their email addresses, or through a traceable history of online activity using tracking software. Umati, however, did not attempt to uncover the true identities of online speakers, and instead focused on observing behavioural patterns of repeat dangerous speech offenders over short periods of time.
- Umati data reflected that in Kenya, ethnicity is a primary lens through which political, economic and social issues are viewed and reacted to by the public. Umati data showed that online discriminatory speech is mostly along ethnic lines. However, as different events transpired through 2013, most notably the Nairobi Westgate Mall attack,[87] Umati data shows that Kenyan online discriminatory speech has escalated along ethno-religious lines. What is crucial to note here is not that discrimination is mostly ethnic or religious,

84 Awori, K. (2013). *Umati Final Report: September 2012-May 2013*. p. 27. Available from: http://www.research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf. [Accessed 2 Sept. 2014]. The full categorisation formula, including the data entry form, can be viewed in the final report.
85 Further discussed in Awori, K. (2013). *Umati Final Report: September 2012-May 2013*. pp. 24-25. Available from: http://www.research.ihub.co.ke/uploads/2013/june/1372415606__936.pdf. [Accessed 2 Sept. 2014].
86 *Ibid.*
87 Daily Nation. (2013). Security forces move to end Westgate mall siege as death toll rises to 62. 23 Sept. Available from: http://www.nation.co.ke/news/Westgate-Mall-attack-alshabaab-terrorism/-/1056/2004630/-/kr74w0/-/index.html. [Accessed 1 Sept. 2014].

but that such discrimination often stems from political, economic and social tensions along various divides. Thus, analysis of dangerous speech should be put into context of other speech online, as rarely do such speech incidents happen in isolation. Moreover, efforts to tackle dangerous speech should focus on addressing the deeper-seated issues that drive people to engage in, disseminate and even act on such speech's provocations.

- While the languages used to disseminate dangerous speech are those that are widely understood in the country, Umati collected some instances of coded language that had been used in past election periods. Additional research is required to investigate this linguistic 'code-switching', which 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.
- Umati Phase II has taken a keen focus on counter-speech, based on emerging phenomena on how 'netizens' are dealing with inflammatory speech online. Umati is monitoring how public conversations take place online over time, how some of these conversations may move towards dangerous speech, and the resultant counter-speech efforts if any. This broader approach will help us better understand self-regulation mechanisms employed by online communities (see Ayala, pp. 17-21). Preliminary self-regulation mechanisms observed online include ridiculing a speaker or a narrative that attempts to inflame hate/misinform/disinform, e.g. the aforementioned KoT cuffing; flooding online spaces with positive counter messages that diffuse tensions arising from hateful messages; and the use of humour and satire to 'hijack' inflammatory narratives.

Conclusions

Observations of dangerous speech should be framed within the context of other conversations online, as inflammatory speech statements rarely happen in isolation. Online dangerous speech is a symptom of the much more complex offline socialisations and perceptions that precede online interaction. We are yet to find concrete instances of online dangerous speech catalysing events offline (see Grayman and Anderson, pp. 22-26). Nonetheless, as 'netizens' congregate and converse online, forming networks around issues of interest, the possibility of organising offline reactions to online conversations is likely.

As part of our third objective in Phase II, we will explore efforts to reduce online dangerous speech through online and offline civic engagement. Umati intends to engage with relevant stakeholders on matters pertaining to freedoms of speech and expression towards better understanding how these are understood and exercised by the Kenyan public. While we are primarily looking at online methods, we will build on experience from *NipeUkweli*[88] (Kiswahili for 'Give me truth'), which is an outreach campaign fashioned to explore proactive ways of mitigating dangerous speech both online and offline.

Going forward, we offer that findings from Umati can provide insight into how humanitarian NGOs can galvanise their crisis prevention efforts and help manage security risks, before and during highly polarised events such as general elections (see Grayman and Anderson, pp. 22-26). One possible avenue could be to promote fissures and spaces where citizens in conflict-prone areas can air out any misconceptions or grievances that would otherwise inform hate/inflammatory/dangerous speech, and even violence.[89] Efforts to tackle dangerous speech (and its consequences) should focus on addressing the deep-seated issues that drive people to engage in, disseminate and act on the provocations of such speech.

88 Njeru, J. N. (2013). *NipeUkweli: Outreach to Sensitize Communities on Dangerous Speech: Summary Report*. iHub Research. 20 March. Available from: http://www.ihub.co.ke/ihubresearch/jb_NipeUkweliSummaryReportMarchpdf2013-11-18-16-07-39.pdf. [Accessed 2 Sept. 2014].
89 A creative example of this is the 'Alternatives to Violence Program', in countries like Kenya and Rwanda: http://www.avpkenya.org. [Accessed 2 Sept. 2014].

From Kenya to Myanmar

Though Umati's methodology was designed to monitor online dangerous speech in Kenya, the project's methodology was adopted in early 2014 for a pilot study of online dangerous speech in Ethiopia. Various elements of the coding form were edited to suit the Ethiopian context.[1] Overall, the methodology was applicable and the same categorisation of dangerous speech into three spectra was employed.

Umati is currently piloting the project in Nigeria, ahead of the 2015 elections. We are working with local Nigerian civil society organisations, offering technical support, as the teams adopt the methodology for their context. The Umati team was also recently in Myanmar, sharing insights on setting up the project with civil society organisations such as MIDO[2] who are keen on monitoring and countering dangerous speech online. As the collection and analysis process continues to be improved in Kenya, the aim is that the methodology will remain explicit enough to be understood and redesigned for other country contexts. Findings drawn from Umati's experience in Kenya can guide organisations in managing risks in contexts where online media is a possible vehicle for catalysing dangerous speech and violence.

For further information on the Umati project, see http://www.ihub.co.ke/umati

1 *Gagliardone, I., Patel, A. and Pohjonnen, M. (2014). Mapping and Analyzing Hate Speech Online: Opportunities and Challenges for Ethiopia. Programme in Comparative Media Law and Policy, University of Oxford. Available from: http://pcmlp.socleg.ox.ac.uk/sites/pcmlp.socleg.ox.ac.uk/files/Ethiopia%20hate%20speech.pdf. [Accessed 2 Sept. 2014].*
2 *http://myanmarido.org/en. [Accessed 2 Sept. 2014].*

European Interagency Security Forum (EISF)

EISF is an independent network of Security Focal Points who currently represent 66 Europe-based humanitarian NGOs operating internationally. EISF is committed to improving the security of relief operations and staff. It aims to increase safe access by humanitarian agencies to people affected by emergencies. Key to its work is the development of research and tools which promote awareness, preparedness and good practice.

EISF was created to establish a more prominent role for security risk management in international humanitarian operations. It facilitates exchange between member organisations and other bodies such as the UN, institutional donors, academic and research institutions, the private sector, and a broad range of international NGOs. EISF's vision is to become a global reference point for applied practice and collective knowledge, and key to its work is the development of practical research for security risk management in the humanitarian sector.

EISF is an independent entity currently funded by the US Office of Foreign Disaster Assistance (OFDA), the Swiss Agency for Development and Cooperation (SDC), the Department for International Development (DFID) and member contributions.

www.eisf.eu

Editors

Raquel Vazquez Llorente and Imogen Wall.

The editors welcome comments and further submissions for future publications or the web-based project. If you are interested in contributing, please email eisf-research@eisf.eu. Imogen Wall can be contacted at imogenwall@hotmail.com.

Acknowledgments

The editors would like to thank Lisa Reilly, EISF Coordinator, for her input and advice, and especially for her comments on the initial drafts. We would also like to extend our gratitude to Tess Dury, for her research support at the initial stages of the project, Brian Shorten for sharing his expertise with us, and Crofton Black for his early guidance and, as always, his continuous support.

Suggested citation

Vazquez Llorente R. and Wall, I. (eds.) (2014) *Communications technology and humanitarian delivery: challenges and opportunities for security risk management*. European Interagency Security Forum (EISF).

Cover photo: Mary Kiperus, community health worker, uses a mobile phone for reporting to the local nurse. Leparua village, Isiolo County, Kenya. February, 2014. © Christian Aid/Elizabeth Dalziel.

Disclaimer

Other EISF Publications

If you are interested in contributing to upcoming research projects or want to suggest topics for future research please contact **eisf-research@eisf.eu**.

Briefing Papers

Security Risk Management and Religion: Faith and Secularism in Humanitarian Assistance
August 2014
Hodgson, L. et al. Edited by Vazquez, R.

Security Management and Capacity Development: International Agencies Working with Local Partners
December 2012
Singh, I. and EISF Secretariat

Gender and Security: Guidelines for Mainstreaming Gender in Security Risk Management
September 2012 – *Sp. and Fr. versions available*
Persaud, C. Edited by Zumkehr, H. J. – EISF Secretariat

Engaging Private Security Providers: A Guideline for Non-Governmental Organisations
December 2011 *Fr. version available*
Glaser, M. Supported by the EISF Secretariat (eds.)

Abduction Management
May 2010
Buth, P. Supported by the EISF Secretariat (eds.)

Crisis Management of Critical Incidents
April 2010
Buth, P. Supported by the EISF Secretariat (eds.)

The Information Management Challenge
March 2010
Ayre, R. Supported by the EISF Secretariat (eds.)

Reports

The Future of Humanitarian Security in Fragile Contexts
March 2014
Armstrong, J. Supported by the EISF Secretariat

The Cost of Security Risk Management for NGOs
February 2013
Finucane, C. Edited by Zumkehr, H. J. – EISF Secretariat

Risk Thresholds in Humanitarian Assistance
October 2010
Kingston, M. and Behn O.

Joint NGO Safety and Security Training
January 2010
Kingston, M. Supported by the EISF Training Working Group

Humanitarian Risk Initiatives: 2009 Index Report
December 2009
Finucane, C. Edited by Kingston, M.

Articles

Incident Statistics in Aid Worker Safety and Security Management: Using and Producing them
March 2012
Van Brabant, K.

Managing Aid Agency Security in an Evolving World: The Larger Challenge
December 2010
Van Brabant, K.

Whose risk is it anyway? Linking Operational Risk Thresholds and Organisational Risk Management
June 2010, (in Humanitarian Exchange 47)
Behn, O. and Kingston, M.

Risk Transfer through Hardening Mentalities?
November 2009
Behn, O. and Kingston, M.

Guides

Security Audits
September 2013 – *Sp. and Fr. versions available*
Finucane C. Edited by French, E. and Vazquez, R. (Sp. and Fr.) – EISF Secretariat

Managing The Message: Communication and Media Management in a Crisis
September 2013
Davidson, S., and French, E., EISF Secretariat (eds.)

Family First: Liaison and Support During a Crisis
February 2013 *Fr. version available*
Davidson, S. Edited by French, E. – EISF Secretariat

Office Closure
February 2013
Safer Edge. Edited by French, E. and Reilly, L. – EISF Secretariat

Forthcoming publications

Office Opening Guide