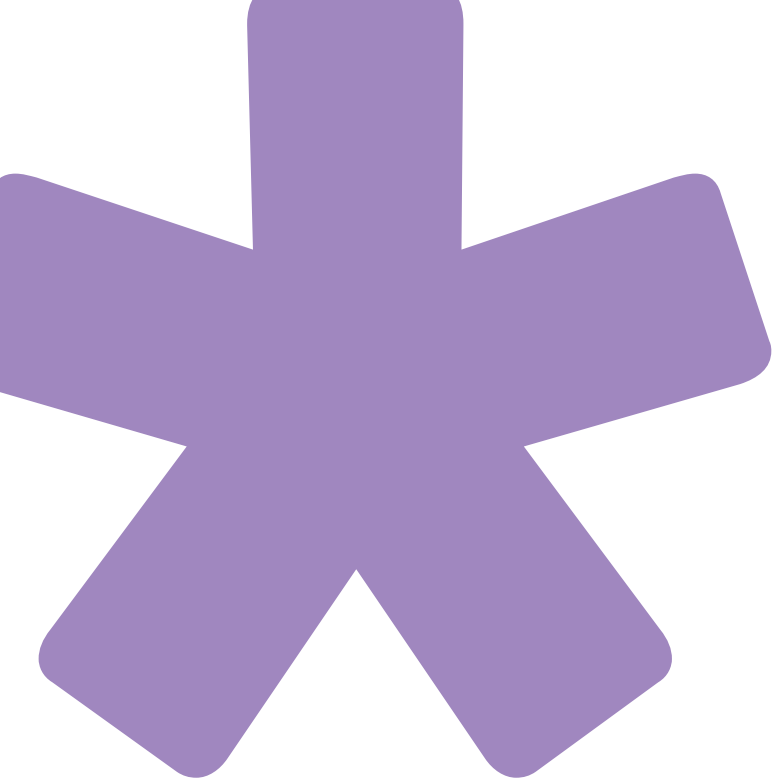


# Viability, Verification, Validity: 3Vs of Crowdsourcing

**TESTED IN ELECTION-BASED CROWDSOURCING**



*This research was funded by Canada's International Development Research Centre.*



**Written and compiled** by Nanjira Sambuli,  
Angela Crandall, Patrick Costello, & Chris Orwa

**Edited** by Eleanor Marchant, ASC, UPenn

**Advisory:** Anahi Ayala & Patrick Meier

**iHub Research © 2013**



# CONTENTS

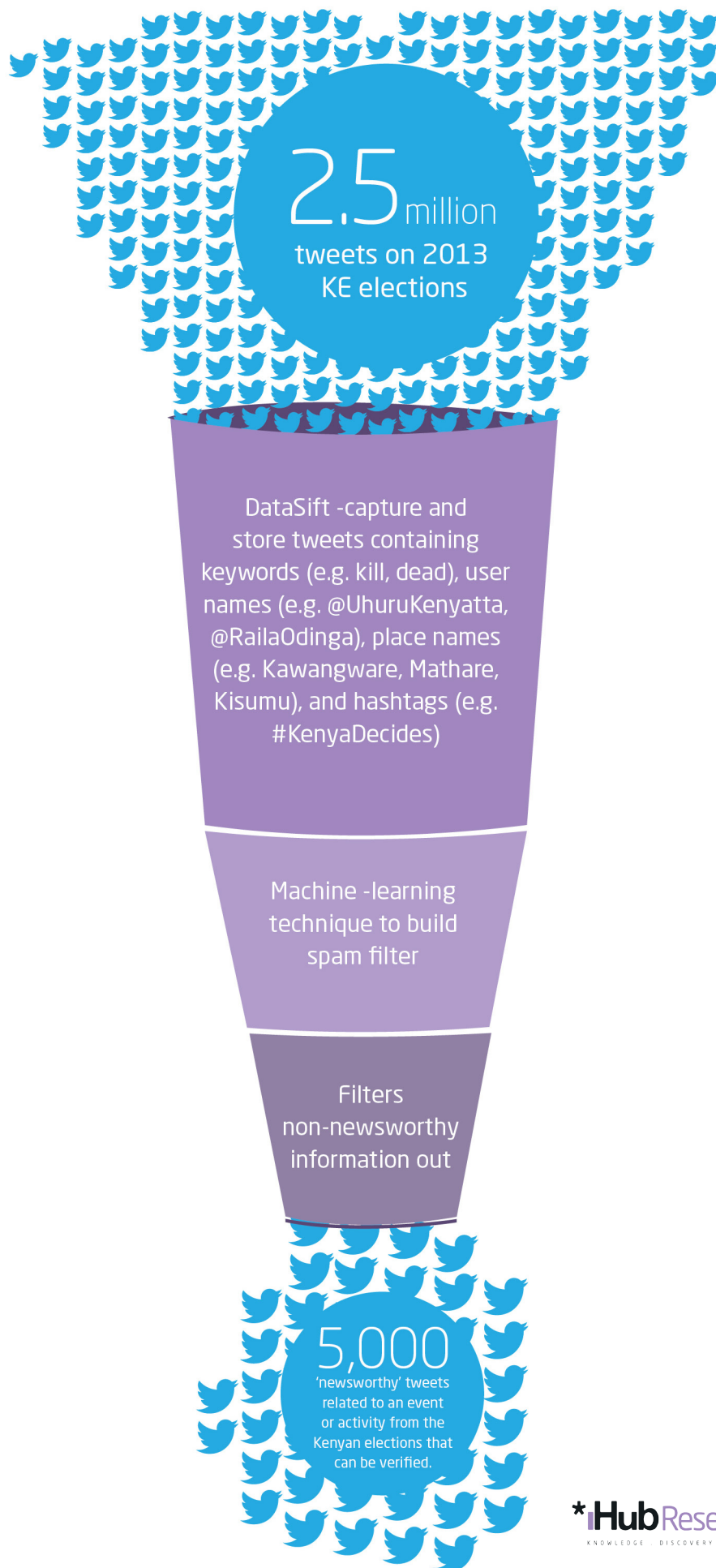
ACKNOWLEDGEMENTS	5
EXECUTIVE SUMMARY	7
GLOSSARY	11
INTRODUCTION	13
BACKGROUND	17
METHODOLOGY	23
FINDINGS	35
Passive Crowdsourcing is Viable During the Elections in Kenya	35
Twitter Breaks News During Elections	41
Mining of Twitter Data Without Machine Learning Is Not Feasible	44
CONCLUSION	47
REFERENCES	49
APPENDIX	51



# Acknowledgements

Thank you to our funder, IDRC, for supporting this research. We also appreciate the input and mentoring from Dr. Patrick Meier, Anahi Ayala, Dr. Chato, and Aditi. We thank Lillian Aluanga (Standard Media Group), Ruthnebokhi Nesoba and Matthew Eltringham (BBC), Kate Starbird (Tweak the Tweet), Markham Nolan (Storyful), Cynara Vetch (Al Jazeera), Cedric Gitura, Ken Macharia and Muba (Capital Group Kenya), Churchill Otieno (Nation Media Group), Marek Mracka (European Union Electoral Observer Mission-Kenya), and Beryl Aidi (Kenya Human Rights Commission) for their insights into their mainstream media/election observation digital operations.

*iHub Research August 2013*



# Executive Summary

Data and information can be a great leveler and democratizer, especially among developing economies where, until recently, many communities had limited exchange of information outside their immediate surroundings. If implemented well, crowdsourcing can be an incredibly important tool for improving access to information, with wide-reaching economic development impacts. The crowdsourcing approach to data collection enables very localized and relevant information to be collected and disseminated. From being able to find funding for a good idea (Kickstarter), to being alerted when the local water supply has been turned on (NextDrop), crowdsourcing can increase access to relevant information and has the potential to subsequently improve living standards as well as political and social engagement.

Our research project looks at a commonly held assumption that crowdsourced information (collected from citizens through online platforms such as Twitter, Facebook, and text messaging) captures more information about the on-the-ground reality than traditional media outlets like television and newspapers. We use Kenya's General Elections on March 4, 2013 as a case study event to compare information collected from the crowd with results collected by traditional media and other sources. The major aims of this study were:

- To assess if crowdsourcing information is viable in the Kenyan context.
- To understand what information, if any, Twitter provided beyond traditional media sources, and other crowdsourcing platforms, such as Uchaguzi.
- To understand if mining of social media data might be useful for traditional media.

## Our key findings include:

### 1. 'Passive Crowdsourcing' is viable during the elections in Kenya

We gathered approximately 2.57m tweets over the period March 3 - April 9, 2013, from which we filtered 12,000 'newsworthy' tweets, which represented 96 separate incidents. The analysis was performed using standard computer hardware and open source software. Twitter reported all incidents related to the election that were reported on traditional media, and in addition provided a large amount of data on the election that was not published in the traditional media.

### 2. Twitter breaks news during elections

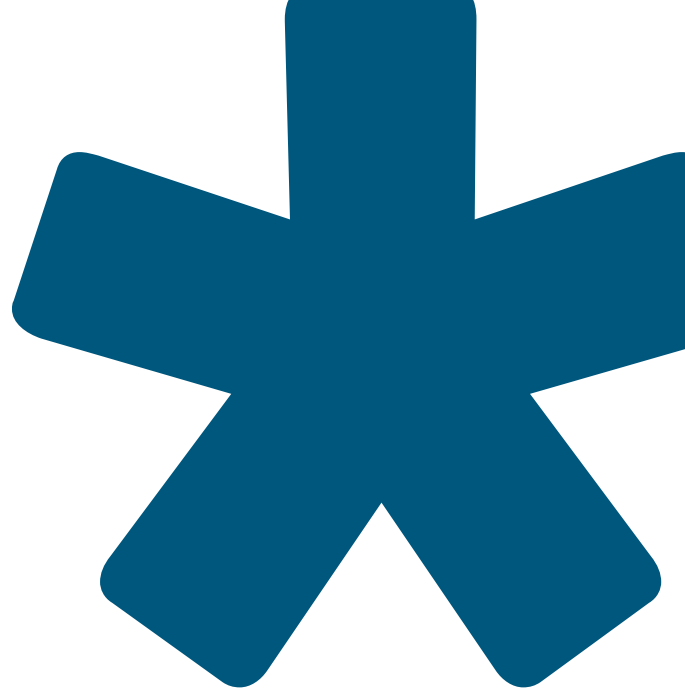
In our comparison of different information sources, we examined their time profiles and compared them to one another. By looking at the lead and lag relationship between when news breaks in the traditional media and when it breaks on Twitter, we found that Twitter either reports on the same day, or leads the traditional media reporting. When Twitter leads, it is by a margin of usually one day. Given that so many incidents happen within an election period, lead-time of one day can be quite important. This finding highlights that Twitter's value stems not only from increased incident coverage, but also from its ability to offer information in real-time.

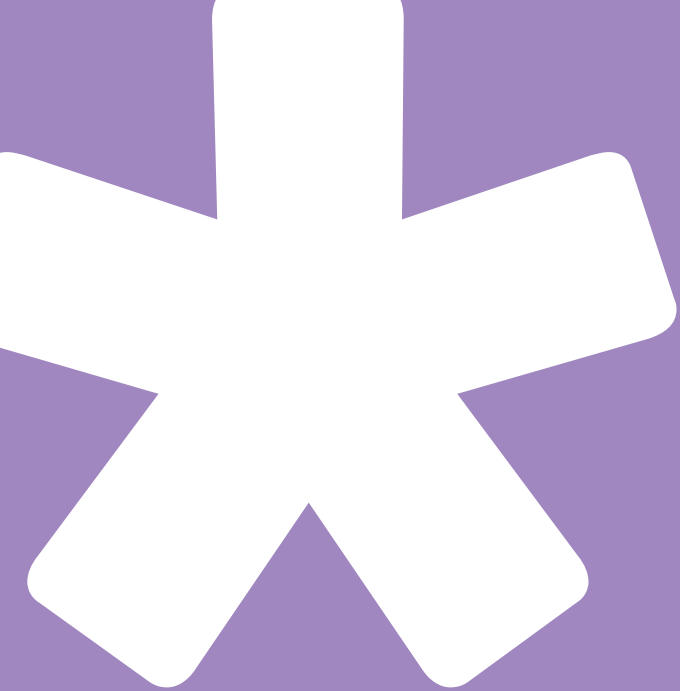
### 3. Mining of Twitter data without machine learning is not feasible

We have found that Twitter provides access to useful, first-hand, eyewitness information, not available in other media in near real-time. However, we also found that extracting this information needed data-mining techniques from the field of Machine Learning, requiring technical expertise. Simple searching of the data was not feasible.

We believe that this study is an important and timely one given the prevailing, often euphoric, rhetoric about the potential of crowdsourcing to provide access to information that might otherwise be overlooked. As a result of the work conducted for this project, we have created a draft Framework - the '3Vs of Crowdsourcing (During Elections?)' (Viability, Validity, Verification) - with the hope that it will help new and existing crowdsourcing deployments, media organizations, open data platforms, and other similar operations during election to better be able to assess if crowdsourcing during a particular election in a particular country is indeed a viable way to gather verifiable information. We look forward to testing this framework in other contexts in the future and welcome feedback and insights.







# Glossary

**Passive Crowdsourcing:** This method of crowdsourcing gathers information produced as a result of the existing behavior of a crowd of users on existing platforms, such as Twitter. Passive crowdsourcing ‘listens’ to what is being reported in an online public forum.

**Active Crowdsourcing:** This method of crowdsourcing issues an open call to specific individuals or to the general public, to participate in a crowdsourcing activity, sharing information into a designated platform or on social networks, which are either familiar or easy-to-use.

**Targeted Crowdsourcing:** This method of crowdsourcing issues an invitation to specific groups of people to share information on a crowdsourcing platform.

**Dedicated Platforms:** This method of crowdsourcing operates using a platform that has been purpose-built for crowdsourcing, or a particular crowdsourcing activity or task, e.g. a dedicated SMS or dial-in number. Examples include Ushahidi and Jangbeeshi.

**Non-targeted Crowdsourcing:** This method of crowdsourcing allows anyone who becomes aware of the crowdsourcing activity to participate.

**Non-dedicated platforms:** Platforms widely adopted used for communication or social networking, and through which crowdsourcing can be conducted, e.g. Twitter, Facebook, Reddit.

**Traditional media:** These include radio, television, newspapers, wire services, other print publications, and traditional news outlets online.

***NB:** For our research purposes, we look only at online publications by these media outlets and do not look at those in “real-time”, e.g. radio.*

**Mainstream media:** Media disseminated via the largest distribution channels, which therefore represent what the majority of media consumers are likely to encounter.

**Newsworthy:** Defined by the project as that which provides situational awareness of poll-related incidents, and is actionable. This definition could change based on the user, project goals and objectives.

**Incidents:** Defined by the project as occurrences related to violence, e.g. threats, demonstrations, protests, or attacks.

**Spam:** Defined by the project as data that is not a newsworthy ‘incident.’ This could include information that is about the Kenyan 2013 General Election, but that is not related to specific incidents (see above definition of an incident).

**Noise:** Data completely unrelated to the subject of the Kenyan General Elections, e.g. Spanish tweets captured using hashtags and place streams (words and phrases that have multiple meanings).

**Newsworthy to Noise Ratio:** This ratio is a measure used to compare the level of a desired signal (newsworthy) to the level of noise. Also known as a Signal-to-Noise Ratio (SNR).

$$\text{SNR} = \frac{P_{\text{signal}}}{P_{\text{noise}}},$$



# INTRODUCTION

Crowdsourcing is an increasingly popular concept as Internet access and mobile device ubiquity continue to increase across the globe. A term coined by Jeff Howe in 2006, crowdsourcing can be defined as the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.<sup>1</sup> As the concept has continued to evolve in its practice, a more integrated definition has been offered that highlights some key crowdsourcing features. Most notably, it is a type of participative online activity in which tasks are undertaken voluntarily as a response to an open call made to share information.<sup>2</sup> This definition of crowdsourcing entails mutual benefit, with the crowdsourcer obtaining information desired from a vast number of people in real-time especially when conducted online, through social media platforms. In turn, sharing information (via reports, publications, live maps or newswire articles) informs participants themselves. The value system that this creates has the potential to build a cycle that allows for future crowdsourcing deployments.

Information and Communication Technology (ICT)-based crowdsourcing has been used in disaster relief and crisis mapping situations. Examples include relief efforts during the 2010 earthquakes in Haiti<sup>3</sup> and Chile,<sup>4</sup> where the Ushahidi<sup>5</sup> platform was leveraged, with SMS messages being sent into the system, and a network of volunteers verifying the information and plotting it onto the publicly available crowdmap.<sup>6</sup> Twitter was also used to provide situational awareness and mobilize people to help<sup>7</sup> during the 2010-2011 Australian floods when Queensland police took to Twitter and Facebook to provide the public with regular updates and to deal with the spread of misinformation on Twitter.<sup>8</sup>

Crowdsourcing is also increasingly used as an election-monitoring tool. The Ushahidi platform was used in this way<sup>9</sup> in 2009 in India, Mexico, Afghanistan and Lebanon and in 2010 in Sudan and Togo.

In each country, the platform was used to collect reports from the general public about the election. During the 2012 Ghana Elections, Blogging Ghana,<sup>10</sup> a membership-based platform for Ghanaian bloggers and social media enthusiasts created an instance map powered by Ushahidi, known as “Ghana Votes 2012”,<sup>11</sup> designed to be a one-stop shop for election information collating reports from the elections body, civil society groups and citizens. In tandem, the African Election Project<sup>12</sup> developed a similar platform for mobile, called Jangbeeshi,<sup>13</sup> designed to collect election-related data directly from the polling stations. Data collection where trained election monitors used the Ushahidi platform has also been carried out during Namibia’s 2009 elections, and in Burundi’s 2010 elections.

Crowdsourcing clearly is a concept applied to a wide variety of tasks, but the commonality between them all is a broad reach of people in inexpensive ways. Using crowdsourcing, large amounts of data can be obtained quickly, and often in at least near real time. It can offer situational awareness for events, especially when information obtained is added to a map,<sup>14</sup> as is usually the case in crisis mapping. When applied to election monitoring, crowdsourcing has the potential to foster citizen engagement with the information—to dispute, confirm or acknowledge its existence. As citizens make their voices heard, the public sphere is widened through the strengthening of civil society.<sup>15</sup>

There are a number of factors that determine the success of a crowdsourcing initiative, as described in Ankit Sharma’s model of critical crowdsourcing success factors (2010). In Sharma’s model, motivational alignment of the crowd is central, whereas vision and strategy, linkages and trust, external environment, infrastructure, and human capital are peripheral factors. Ankit holds that for crowdsourcing projects to succeed, the crowd should be viewed as a key partner in the initiative.

Building upon Sharma’s model, our project investi-

gates and identifies factors that influence the viability, verification, and validity of crowdsourcing during elections. This has led us to construct a draft framework designed to help prospective crowdsourcers assess the viability of a crowdsourcing project during an election in a given context.

Our project specifically aimed to assess three aspects of crowdsourcing – what we’re calling the ‘3Vs of Crowdsourcing’ – with the following objectives:

1. Viability: In what situations, or during which events, is crowdsourcing a viable venture likely to offer worthwhile results and outcomes? We aim to understand and outline the features that make a particular election in a particular country viable for crowdsourcing.
2. Validity: Can crowdsourced information offer a true reflection of the reality on the ground? We aim to identify that conditions that might make real-time data validation possible.
3. Verification: Is there a way in which we can verify that the information provided through crowdsourcing is indeed valid? If so, can the verification process be automated? If so, we aim to devise a tool for doing so.

These objectives led us to construct the following key research questions:

1. What, if any, particular conditions should be in place for crowdsourcing of information to be viable during an election period?
2. Can crowdsourced information be validated during an election period? If so, what is the practical implementation of doing so?
3. How do different crowdsourcing methods contribute to the quality of information collected?

Our research used the 2013 Kenyan General Elections as a case study. Political incidents in Kenya have been noted to spark many online conversations, especially with the continued uptake of social media. Opinions, facts, rumors, and incidents are shared and reported online with increased frequency. We tracked such social media activity together with other media sources to better understand what information was generated in the build up to, during and after the March 2013 Kenyan election period.

1. Jeff Howe, "Crowdsourcing: A Definition," [www.crowdsourcing.typepad.com](http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html) (blog), June 2, 2006, [http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing\\_a.html](http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html).
2. Enrique Estellés-Arolas and Fernando González-Ladrón-de-Guevara, "Towards an integrated crowdsourcing definition," *Journal of Information Science* (2012): 189-200.
3. Patrick Meier, "Haiti and the Power of Crowdsourcing," [www.blog.ushahidi.com](http://blog.ushahidi.com/2012/01/12/haiti-and-the-power-of-crowdsourcing/) (blog), January 12, 2012, <http://blog.ushahidi.com/2012/01/12/haiti-and-the-power-of-crowdsourcing/>.
4. Chile Deployment of Ushahidi, <http://chile.ushahidi.com/>.
5. Ushahidi.com, <http://www.ushahidi.com/>.
6. Kate Starbird, "Digital Volunteerism During Disaster: Crowdsourcing Information Processing," *CHI 2011 Proceedings*, May 7 - 12, 2011, Vancouver, BC, Canada.
7. Yates and Paquette, "Emergency knowledge management and social media technologies: a case study of the 2010 Haitian earthquake," *International Journal of Information Management* 31, no. 1 (2011): 6 -13.
8. France Cheong and Christopher Cheong, "Social Media Data Mining: A Social Network Analysis of Tweets During The 2010-2011 Australian Floods," *PACIS 2011 Proceedings* (2011), Paper 46.
9. Ushahidi, "Election Monitoring Using The Ushahidi Platform," [http://community.ushahidi.com/uploads/documents/c\\_Ushahidi-Election\\_Monitoring\\_Guide.pdf](http://community.ushahidi.com/uploads/documents/c_Ushahidi-Election_Monitoring_Guide.pdf).
10. Ghana Decides, <http://africanelections.blogspot.com/2012/10/ghana-elections-2012-jangbeeshi-mobile.html>.
11. Ghana Votes Ushahidi Deployment, <http://gh-votes2012.com/>.
12. African Elections Project, <http://www.africanelections.org/>.
13. African Elections Project, "Jangbeeshi Mobile App," October 17, 2012, <http://africanelections.blogspot.com/2012/10/ghana-elections-2012-jangbeeshi-mobile.html>.
14. Patrick Meier, "Verifying Crowdsourced Social Media Reports for Live Crisis Mapping: An Introduction to Information Forensics," 2011, <http://irevolution.files.wordpress.com/2011/11/meier-verifying-crowdsourced-data-case-studies.pdf>.
15. Patrick Meier, "Traditional vs. Crowdsourced Election Monitoring: Which Has More Impact?," 2012, <http://irevolution.net/2012/08/07/traditional-vs-crowdsourced-election-monitoring/>.





# BACKGROUND

## Crowdsourcing & Social Media

With social media and crowdsourcing has come the era of the 'citizen journalist'. The concept of citizen journalism is based on social media tools allowing users to become creators, by making the publishing and distribution of multimedia content both free and easy, for amateurs and professionals alike. Due to the availability of technology, citizens can often report breaking news more quickly than traditional media reporters.<sup>17</sup>

By default or design, we are quickly getting to a point where anyone with access to a mobile phone or computer, Internet or social media platform can be an information disseminator about what is happening around them. This has disrupted how news is communicated, and the traditional custodians of news have to adapt to new ways of retaining the attention of the public.

Twitter and Facebook are good examples of social media platforms used to share incidents as they unfold. Although tweets are often filled with social conversation and chatter, they may also be used to share politically relevant information and report news.<sup>18</sup> Twitter is becoming a valuable tool in disaster and emergency situations with increasing evidence that it can function not just as a social network, but also a news service.<sup>19</sup> In emergency situations, tweets can provide either first-person observations or bring relevant knowledge from external sources.<sup>20</sup> Information from official and reputable sources is regarded as valuable and hence is actively sought and propagated.<sup>21</sup>

## Crowdsourcing & News Gathering

Traditionally, news gathering has been a labor-intensive process, comprising of establishing networks of trustworthy sources, deploying newsroom teams to various places or having correspondents spread out within various geographical areas. The advent of social media has disrupted how news is created, discovered, shared and distributed.

Traditional media can no longer ignore the power of social media and the information generated at any given time. Some traditional media are now players in the social media sphere in a number of ways, such as social newsgathering, conversation aggregators and establishing story leads. As traditional media establish their presence on online media platforms, they also introduce themselves to a world where ordinary citizens can share information they perceive to be of interest to the media outlet, as well as comments, opinions and reactions about particular news broadcasts or articles. This can offer such a media outlet a new network of sources, whose credibility and trustworthiness could be established over time using a number of mechanisms.

Some international media houses have realized the value of social media, with a number even dedicating resources to tapping into the wealth of citizen journalism. In 2005, BBC set up its User-Generated Content (UGC) Hub to sift through images, footage and eyewitness accounts emailed to them. However, in the past few years, people have become more prone to distributing material themselves via Twitter, Facebook and YouTube.<sup>22</sup> With the subsequent decline in number of contributions proffered to them, the UGC's task has therefore moved to newsgathering; staffers use search terms, see what's trending on Twitter and look at the images and footage their online trusted contacts are dis-

cussing on the platform.<sup>23</sup> This has seen the UGC Hub become a core of social newsgathering, as well as a crucial source of eyewitness information. Through the networks of contacts they have built over time, the UGC Hub gained unprecedented access to information.

Al Jazeera English also has a dedicated daily show about the social media community. 'The Stream' <sup>24</sup> taps into the potential of social media to disseminate news and acts as an aggregator of online sources and discussion, seeking out unheard voices, new perspectives from people on the ground and untold angles related to the most compelling stories of the day.<sup>25</sup> CNN's iReport is a similar compilation of news items submitted by citizen journalism.<sup>26</sup> Finally, Storyful, founded in 2010, is the first news agency of the social media age.<sup>27</sup> It helps journalists, broadcasters and publishers filter breaking news, trending stories and local sources from the noise of social media,<sup>28</sup> and provides tools to monitor social conversations and trends.

In Kenya, most local traditional media outlets have been vamping up their online presence, with websites, which are some of the top visited sites in the country,<sup>29</sup> and an increasing social media presence, notably on Facebook, YouTube, and Twitter.<sup>30</sup> These local media outlets use such platforms to re-share information they have disseminated on their television or radio stations or in their newspapers. However, rarely do these outlets explicitly use their social network presence as a source of news, except for the occasional sampling of tweets or Facebook posts or comments on live broadcasts. Traditional media outlets in Kenya are increasingly asking their followers to share their opinions on particular issues on their Facebook or Twitter pages, but they do not yet mine social media data for information other than the aggregation of sentiments.

Individual journalists tend to also be popular online; they engage with their followers and participate in commentary. Most carry the disclaimer that their views do not reflect those of their employer. In the case of one outlet (Capital Media Group), all employees are required to have an online presence (especially on Twitter), and are expected to be part of the community and its conversations. In this way, Capital Media Group has positioned itself as a digital media outlet with no print publication at all, but instead a focus on radio stations whose programs are supplemented with digital media presence and conversations.<sup>31</sup>

## Twitter

The micro-blogging service launched in 2006 has approximately 550 million active users worldwide,<sup>32</sup> with about 200 million monthly active users.<sup>33</sup> An average of 400 million tweets are sent everyday globally<sup>34</sup> and 60% of the monthly active users log on using a mobile device at least once every month.<sup>35</sup> Users tweet on a wide range of topics, and can tweet at anyone with a public profile, from prominent personalities to corporate brands. The public timeline conveying tweets from users worldwide is an extensive real-time information stream. Breaking news and coverage of real-time events are all shared under the 140-character limit. Tweet structures vary from plain text, to use of hashtags that mark key topics or keywords, to use of links to other news sources, photos and videos.

Extensive research studies have been conducted to further analyze characteristics of information sharing on Twitter. Kwak et al. (2010) set out to investigate whether Twitter is a social network or a news media. They found that people share information differently on the site than they do on other social networking sites, making it more of a news media than a social network. The authors also found that Twitter users search for up-to-the-second information and updates on unfolding events.

Politics and political events such as elections have increasingly become popular subject matter on Twitter. Politicians, for instance, are using Twitter to interact and engage directly with their constituents, bypassing traditional media as an information intermediary. Political commentary is no longer a reserve for political analysts; Twitter is home to political insights, opinions and exchange among 'ordinary' citizens across the globe. It has also become a medium through which reactions and sentiments are relayed. Tumasjan et al. (2010) set out to investigate whether Twitter is used as a forum for political deliberation and whether online messages on Twitter mirror offline political sentiment. Using the 2010 German federal election as context, they found that the number of messages mentioning a party reflected the election results. An analysis of political sentiment in tweets demonstrated close correspondence to the parties' and politicians' political positions, indicating that the content of Twitter messages plausibly reflects the offline political landscape.<sup>36</sup> A separate study by Lamos (2010) explored what could be discerned about voting in-

tentions from Twitter content, by extracting positive and negative sentiment from select keywords (such as party and politicians' names), then mapping this sentiment to voting intention percentages (from voting intention polls conducted), using the UK 2010 General Election as a case study. Lampos found that social media contains content relating to political opinion, and that when the content of tweets is enriched by attaching synonymous words, a significant improvement on signal is made and inference becomes easier.

## #KOT: Kenyans On Twitter and the 2013 Kenyan General Elections

Kenya has been ranked the second top tweeter in Africa by volume of tweets sent, after South Africa. Research by iHub Research's online hate speech monitoring group, Umati, found that Kenya's online conversations appear to be a window into offline opinions and sentiments and likely reflect conversations that might take place in a bar or family setting.<sup>38</sup>

The 2013 General Election was the first to be conducted under Kenya's new constitution. A Presidential Debate organized by a media alliance preceded the election by a month. Kenyans took to online media to share their thoughts on candidates' responses to questions, their agendas and their visions for the country. Kenya's 'middle class', who are widely believed to constitute the bulk of the online population, have often come under criticism for the high levels of hate speech that were observed online during the months leading up to and following the elections.<sup>39</sup>

The online activity on and around March 4th, the election day, largely entailed Kenyan tweeters sharing their 'journey to the ballot box': waking up early, bearing the long queues, patience despite technical hitches in the voting process. This kind of testimony ostensibly served to prove to critics and fellow 'tweeps' that they were dutiful to their country. The strong support for the importance of upholding peace that preceded the election manifested in the online community through their condemnation of

foreign correspondents who anticipated a repeat of the 2007/2008 post-election violence. Any oversensationalized reporting on Election Day and in the build-up to Election Day was met with criticism and subjected to ridicule and satire. Hashtags such as #tweetlikeaforeignjournalist<sup>40</sup> and #SomeoneTellCNN<sup>41</sup> were widely employed to quell false or misleading reporting during the period, even those originating from Kenyans.

Due to technical issues, the results of the elections took longer to be released than expected. During this tense time, several rumors circulated online such as the false accusation that the CEO of the electoral body, the Independent Electoral and Boundaries Commission (IEBC), was being held in the dungeons of a government building (Nyayo House). The CEO later appeared on TV to prove to all that, although tired, he was alive and well. As Kenyans awaited the final election results, some took to social media to express their anxiety, impatience and frustrations. That such an avenue for expression existed may have contributed to the ultimately peaceful acceptance of the final election results as well as the Supreme Court's dismissal of a petition, led by Presidential Candidate Raila Odinga, disputing the election. With the exception of a few incidents of violence, the election period was therefore largely peaceful.

This plethora of online activity during the election period enabled crowdsourcing to be a viable option for gaining insight into citizen sentiments, voting activity, and on-the-ground incidents in a number of areas in the country before, during and after the election process. Calls to 'crowdsource' were mostly conducted by traditional media outlets, which invited Kenyans to tweet or share information online using various hashtags or on their various pages (Facebook, websites), as well as by Ushahidi's election deployment, Uchaguzi.

16. Gaurav Mishra, "Digital Activism & the 4Cs Social Media Framework," *GlobalVoices Advocacy* (blog), May 11, 2009, <http://advocacy.globalvoicesonline.org/2009/05/11/digital-activism-the-4cs-social-media-framework/>.
17. Citizen Journalism, Mashable, <http://mashable.com/category/citizen-journalism/>.
18. Carlos Castillo, Marcelo Mendoza, and Barbara Poblete, "Information Credibility on Twitter," *WWW 2011 Proceedings* (2011): 675-684.
19. Yates and Paquette, 2011.
20. Sarah Vieweg, Amanda Hughes, Kate Startbird, Leysia Palen, "Microblogging during two natural hazards events: what twitter may contribute to situational awareness," *CHI 2010 Proceedings* (2010): 1079-1088.
21. Cheong and Cheong, 2011.
22. Neiman Report - Harvard University, "Truth In The Age of Social Media," 2012.
23. *ibid.*
24. Aljazeera, "The Stream," <http://stream.aljazeera.com/>.
25. Aljazeera, "The Stream Biography," <http://www.aljazeera.com/profile/the-stream.html>.
26. About CNN iReport, <http://ireport.cnn.com/about.jspa>.
27. Storyful, <http://storyful.com/>.
28. *ibid.*
29. "Top Sites in Kenya," *Alexa.com*, <http://www.alexa.com/topsites/countries/KE>.
30. *ibid.*
31. Capital Digital Media, <http://www.capitaldigital-media.co.ke/site/who.php>.
32. Statistic Brain, "Twitter Statistics," May 7, 2013, <http://www.statisticbrain.com/Twitter-statistics/>.
33. Stan Schroeder, "Happy 7th Birthday, Twitter," March 21, 2013, <http://mashable.com/2013/03/21/happy-7th-birthday-Twitter/>.
34. *Ibid.*
35. Twitter Advertising, "New Compete study: Primary mobile users on Twitter," February 11, 2013, <https://blog.Twitter.com/2013/new-compete-study-primary-mobile-users-Twitter>.
36. Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welp, "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment," *ICWSM 2010 proceedings*, May 16, 2010, <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewPaper/1441>.
37. Portland, "New Research Reveals How Africa Tweets," February 1, 2012, <http://www.portland-communications.com/2012/02/new-research-reveals-how-africa-tweets/>.
38. Kagonya Awori, "Umati Final Report," *iHub Research*, July 2013.
39. *Ibid.*
40. Xiao Mina, "Hashtag memes: Breakings the single story through humour," March 27, 2013, <http://m.aljazeera.com/story/2013326132026281740>.
41. Nanjala Nyabola, "Kenya tweets back: #Some-onetellCNN," *Aljazeera.com*, March 8, 2013, <http://www.aljazeera.com/indepth/opinion/2013/03/20133684021106816.html>.





# METHODOLOGY

## Overview - What we did

This research study combined qualitative interviews with a quantitative analysis of crowd-sourced data. For the qualitative component, we administered semi-structured questionnaires to 12 relevant individuals at online and traditional media organizations. These interviews focused on the organizational methods for validating data across different scenarios (Election Day, regular daily reporting, international crises, etc.) as well as associated costs. On Election Day (March 4, 2013), the research team was also present on-site (BBC situation room and Uchaguzi situation room) to observe the Election Day data collection process conducted by the different media outlets first-hand. We also used ethnographic methods to gather data through observation and immersion within the crowd mapping and traditional media communities. By observing the process, we gained greater empathy and insight on the data collection process in order to build the framework with the user in mind. The study also included in-depth interviews with 85 Kenyan citizens in 3 locations around the country, approximately 5 weeks after the new President was sworn in.

For the quantitative component, we collected and ran a cross-comparative analysis on data sourced from Twitter, Uchaguzi, online traditional media, and our fieldwork. We looked at the different information captured by the different sources and looked for ‘newsworthy incidents,’ defined by the project as that which provides situational awareness of poll-related incidents, and is actionable. We decided to focus the project on this definition of ‘newsworthy information’ because we wanted useful, verifiable information that could be used (and responded to) by agencies such as the International Committee of the Red Cross (ICRC) or media houses. We were cognizant of the difficulty in verifying sentiment information. Although such information could still be relevant for a news story, we

decided to limit our project to look only at potentially actionable information, and look specifically for ‘newsworthy incidents.’

The culmination of our research was a comparison across different sources of election information. The various data sets we compared are listed below in more detail:

## Data Sources Assessed During the Kenyan General Election 2013

### Data Mining from Twitter (passive, non-targeted crowdsourcing)

We used a third-party Twitter application called Datasift to capture and store tweets using Kenyan election-related keywords, user names, place names and hashtags from March 3 (the day before the elections) until April 9 (the day of the Presidential Inauguration). Each of these streams was selected upon observation of conversation trends prior to and during the election period. Keywords were determined based on activities and events we wanted to capture that would offer situational awareness (such as violence, riots, killings, tension). NB: We were not interested in opinion-related information or sentiment. We monitored usernames of traditional media outlets as well as those of influential/prominent Twitter users that were likely to amplify events as were shared to them through their various networks/followings. We also used place names, of towns, cities, estates, constituencies and counties to ensure we captured information from as many parts of the country as possible. We monitored hashtags that were determined by both Twitter users and mainstream media outlets to aggregate conversation. We adapted the streams to capture as much relevant information as



possible.

This data was then cleaned and mined for newsworthy information in the weeks after the election. Verification was not a prerequisite condition for obtaining the information. We were interested in aggregate data, employing only the above-mentioned filters to capture tweets generated largely within the country, and specific to the Kenyan election context. We determined if the tweets were generated within Kenya based on the user profile's set location and also based on any geo-tagged content.

## Online traditional media

We looked into the online publications of traditional media outlets as they are leading sources of information online. Standard Media, *The Daily Nation* and Capital FM are among the top 25 most visited sites in Kenya.<sup>42</sup> Having conducted in-depth interviews with these media outlets, we found their online sites were the most viable channels to retrieve information for our research objectives, since local traditional media outlets do not collect or aggregate 'raw' data towards reporting, and it was more efficient than collecting printed copies of their publications. From the interviews, we also learned that each of these local media outlets verify all published event reports through on-the-ground networks of correspondents and authorities.

We first conducted a manual search through the four leading traditional media online publications in Kenya (Standard Media, *The Daily Nation*, Capital FM and *The Star Kenya*). Using sets of keywords that mirrored those we used to mine the Twitter data, as well as searching through publications per date (from March 3rd to April 9th, 2013), we retrieved 26 unique 'newsworthy' incidents relevant to the research project.

We then set up a Google Custom Search to crawl the web for more information sources, casting a wider net by incorporating international media outlets that were likely to and had been carrying news on the Kenyan election. These included *The New York Times*, Al Jazeera English, BBC, CNN and *The Guardian* to name a few. From these, we identified 5 additional 'newsworthy' incidents.

A third search on Google News was also conducted as a final sweep. We found that a manual search on Google returned better results, and we were able to

find 9 additional 'newsworthy' incidents that had initially been found only in the mined Twitter data.

## Data from crowdsourcing platforms that made an open call to the public to share (active, targeted crowdsourcing)

Raw data was obtained from the Uchaguzi platform,<sup>43</sup> which saw collaboration for election monitoring between citizens, election observers, humanitarian response agencies, civil society, community-based organizations, law enforcement agencies and digital humanitarians. A key partner of Uchaguzi, CRECO, employed on-the-ground election monitors to verify Uchaguzi incidents. Uchaguzi also involved a digital and local team of volunteers who reviewed the citizen reports and categorized them as per structured workflows.<sup>44</sup> After the Uchaguzi deployment ended, the iHub Research team conducted a post-deployment analysis of the Uchaguzi data using the same keywords and process as was used for the Twitter analysis (detailed below).

## Data from fieldwork

We identified three locations where incidents were identified and reported by all of the above three sources. These included Mathare, an urban slum in Nairobi; Kisumu, a growing town along Lake Victoria; and the coast region. Members of our research team traveled to these locations at the end of May/early July 2013 and conducted in-depth interviews with 85 citizen respondents to gain insights into the activities that occurred on the ground before, during and after the election period. Through these investigations, we also sought to find out if respondents used social media or other mediums (SMSs, phone calls) to share, alert or report incidents that they may have witnessed. A limitation of the fieldwork was the fact that the data mining had not been completed by the time the field locations were selected and therefore we were unable to go to the location where we found incidents that were only reported on Twitter and not on traditional media or Uchaguzi.



# Twitter Analysis

## Overview

Twitter and Facebook have proven extremely valuable sources of information in a range of other social contexts, such as disaster response. As part of our research, we aimed to understand how information from non-targeted crowdsourcing platforms, such as Twitter, contributed to the pool of information available in the context of the Kenyan Elections.

We chose to analyze Twitter over other non-targeted platforms due to the platform's 'openness' (the ability to mine the data on Twitter is much easier as compared to Facebook), usage in Kenya, and the availability of earlier academic work on viability and verification of Twitter in other contexts such as disaster situations, that we could draw on.

There are advantages and difficulties to using passive crowdsourcing platforms such as Twitter, due to their nature as a general social media platform and news medium,<sup>45</sup> rather than being dedicated to an event or topic. Advantages include among others 1) that there is a well-established user base already using the platform making it easy for users to report incidents spontaneously as they occur, and 2) that reports can be gathered from users unaware that information is being used for crowdsourcing. Disadvantages include 1) that data feeds from the platform need to be filtered for relevant information as the full stream is huge, consisting of information postings on a multitude of topics from all users in the globe, and 2) that verification of information is tricky as users are free to publish content, including false or misleading content, as they wish.

A central caveat about the usefulness of crowdsourced information is the verifiability of the provided reports. In our present research study, we were not able to find direct and reliable machine learning methods to verify incidents. Nonetheless, other factors such as reliability of a Twitter user based on established networks of information sources might be used to give the information higher trust, but these might be time-intensive. In the next stage of our research, we plan to delve deeper into this issue of the reliability of the Tweeter as we believe developing tools to help with this will be central to the usefulness of Twitter as a crowdsourcing tool.

The following sections describe the method we used for analyzing Twitter data. The findings section that follows discusses results from the analysis of the data we obtained. We plan to produce additional reports in the future that will go into more depth on some of the findings covered here.

## Capture & Storage of Twitter Data

To capture and store data from Twitter, we considered the two available options: developing an in-house collector based on the Twitter API, or choosing from among the different commercial data vendors. We chose to use the commercial data vendor and analysis company, *DataSift*,<sup>46</sup> based on:

- The requirement of a reliable collection device for high velocity and volume of Twitter data, and the very tight project setup time (under 2 weeks). Given the time and resource constraints, and advice received from other scholars in this field, we decided purchasing data collection services was our best option.
- *DataSift* had a very transparent user interface and data export facilities, and straightforward pay-as-you-go payment option.

Since the full Twitter stream contains all tweets from the globe, a large portion would not contain information relating to the Kenyan election. Hence we set up data collection streams on Twitter so that we would collect a selection of tweets more likely to contain relevant information. Simple trials using the *DataSift* platform based on 'and/or' search filtering criteria either produced very broad streams that collected too much data, or very narrow streams that removed too much data. We opted for broad collection streams to capture as much relevant information as we could, at the cost of also capturing redundant *noise* data. The next section describes how we filtered the noise out of the data streams.

We designed the *DataSift* collection streams to capture tweets that contained words we thought likely to be contained in 'newsworthy' tweets. We set up four different collection streams, as we wanted to test which sets of tweets captured the most useful information. Table 1 describes the collection streams we used, and the amount of data collected from each.

*Table 1: Collection streams used to capture tweets. The total number of unique keywords/phrases used in each stream is displayed in parentheses under collection streams.*

Collection Steams	Description of filter	Example keywords	Number of Tweets
Hashtags (40)	Hashtags specific to the election	#kenyadecides, #KEElections13, #Decision2013	372,449
Places (289)	Names of all towns, constituencies and counties in Kenya	Kisumu, Mathare, Mombasa	861,270
Keywords (98)	Words we thought likely to indicate newsworthy incidents	Kill, riot, violence, ballot	1,297,092
Username (37)	Twitter users we thought likely to comment on newsworthy information, including official Twitter profiles of traditional media outlets, journalists and prominent Twitter personalities.	@UKenyatta, @NiNanjira	39,793

*For full list of keywords, see table at the end of the document, Appendix 1.*

## Data Filtering

Before proceeding to analyze the election-related information in the Twitter data we collected, we first had to filter out non-relevant information as a large proportion of the data was unrelated to the election, due to the broad nature of the data collection streams we used.

For example, tweets such as:

*@xyzshow killed it. #Dida guy too hilarious. Thumbs up.*

were captured in our data streams due to the presence of the word 'killed', which was contained in the keywords stream, but the tweet is not about a newsworthy election-related incident.

Manual searching of the Twitter data is not a feasible way of sifting through this noise for real-time data analysis, or even post-event data analysis in general. We therefore used data-mining technology to build smarter filters that could 'mine' our data given nuanced search criteria. The design of our approach was heavily influenced by other data mining research, especially work by Carlos Castillo (ChaTo), Aditi Gupta, and Patrick Meier.

The criteria we used for labeling an incident as 'newsworthy' was:

1. It contained information that reflected incidents that we deemed an election watcher would consider actionable, e.g. reports of violence with a location information, reports of vote rigging.
2. We did not include tweets with opinions, or general non-actionable negative sentiment, e.g. opinion on vote counts, complaints on voting queue lengths.

To build a smart newsworthy filter, we used Supervised Learning techniques from the field of Machine Learning (ML), to train an algorithm to look for newsworthy tweets.

### Supervised Learning

Supervised Learning is a technique where a human trains a computer to label from a set of given categories. A quick summary of the technique is:

1. *Features* of the data are constructed. These are usually different numerical properties of data members that can be used to differentiate data. In the case of Twitter data, these typically fall into *user-based features* such as usernames and user ids of who posted the tweet, time of tweet posting, and *message-based features* such as the number of words in a given tweet, counts of a specific word in a tweet, number of positive or negative sentiment words.
2. A selection of data from a full set of data is labeled or annotated by humans with labels from a set of different categories. This labeled data is called the *training set*.
3. The labeled data and features are fed to various *learning* algorithms that *train* the computer to label the data in a (hopefully) analogous way to how the human annotators did.
4. The *trained* algorithm is then applied to the rest of the dataset to predict how the data should be labeled.
5. The computer-labeled data can then be viewed to see how effective the algorithm was.
6. If needed steps 1-5 can be iterated to improve performance.

We did this by first constructing several types of features for each tweet:

- *meta data-based*: user id , klout score, time of tweet;
- *tweet text-based*: 'Bag of words' uses word counts of all words contained in 'corpora' of tweet.

Using the criteria above, we labeled 91,000 tweets as 'newsworthy' or not in order to create a training set of data. We found we could label at an average of 1,200 tweets/hour. We tried different methods for labeling including *linear search*, *keyword search* and *machine learning*. Machine learning was the only labeling method we tried that we concluded would be feasible for both real-time and post-data analysis of the Twitter data we collected.

### Linear search

For linear search, we took samples from the collected data and manually labeled them either newsworthy = true or newsworthy = false. Using this method, we found a very low 'Newsworthy-to-Noise Ratio' of 110 newsworthy tweets: 87,000 noisy tweets. In other words, around 0.1% of the tweets in our sample were newsworthy. This process took approximately 70 hours to label at 1,200 tweets/hour. This demonstrated that manual linear searching is unfeasible for any kind of data analysis. Indeed, labeling the full dataset using this manual linear search method would take an estimated 270 days assuming labeling for 8 hours/day.

### Keyword search

We then performed a keyword search in an attempt to improve the 'Newsworthy-to-Noise Ratio', and to speed up the labeling of newsworthy tweets. This process involved running a search of the tweets across several keywords such as 'kill', 'riot' and place names as described above. From the resulting set, we randomly selected 4,000 tweets and labeled these. This gave a ratio of 400 newsworthy tweets: 3,600 noisy tweets. In other words, around 10% of tweets in our sample were newsworthy - a good improvement on the linear search. However, 10% of the 2.57 million tweets is still 257,000 tweets, which would take 27 days to label manually. As with linear search this is again unfeasible for real-time analysis, and at the edge of what is feasible for post-data analysis using multiple label-

ers. It would cause particular problems if we wanted to change our search question as we would then likely need another 27 days for re-labeling, which makes this method unfeasible for post-data analysis in general.

### Trained Algorithm

Finally, using the 91,000 labeled tweets, we then trained Machine Learning algorithms to filter our data for 'newsworthy' tweets. We tried different algorithms, and found that using a linear Support Vector Machine (SVM) implemented in the Python module, scikit-learn,<sup>48</sup> with message-based 'bag of words' feature was effective. Applying this trained algorithm to the dataset resulted in a predicted 17,000 newsworthy tweets. Sampling 1,400 of these predicted tweets randomly and annotating gave a ratio of 950 newsworthy tweets: 450 noisy tweets. In other words, about 68% of our sample tweets were found to be newsworthy. Computational time for feature construction, training and predicting on full 2.57 million tweets took less than 3 minutes. Labeling of output predictions took approximately 1.2 hours.

Adding the 1,400 newly labeled tweets to the 91,000 already labeled tweets, retraining the algorithm and applying to the full set resulted in 12,208 predicted newsworthy tweets. We then took a sample of 2,000 from these and labeled them, resulting in a ratio of 1,696 newsworthy tweets: 304 noisy tweets. In other words, about 85% of the predicted tweets using the algorithm were found to be newsworthy. It would have been possible to perform another iteration, including this set of labels to improve the Newsworthy-to-Noise Ratio even further; however, we found that 85% accuracy was enough to proceed with further analysis, and we did not want to 'over-train' our classifier and filter out useful information.<sup>49</sup>

We further found that the computational time it takes to retrain and predict on the full dataset is less than 3 seconds, and that the time it takes to label to test accuracy is 1.7 hours. Crucially, this demonstrated that Supervised Learning is a feasible technique for real-time analysis. We were able to train a classifier that processes 2.57 million tweets in less than 1 second with 85% precision on the 'newsworthy' classification. This also demonstrates that Supervised Learning is also feasible for post-data analysis. The total labeling time to construct the classifier was  $90 + 4 + 1.5$  hours = 95.5 hours, and computational time less than 6 minutes. Note that the labeling time could be greatly reduce since the linear labeling of 87,700 tweets was more than what was needed to get the accuracy required in our classifier.<sup>50</sup> The results of this entire training procedure are summarized in the following tables.

The results of the training procedure are summarized below:

Table 2: Newsworthy detection by different methods

Labeling Method	Newsworthy	Noise	Total	Newsworthy/ Total Collected
Linear	110	87,000	87,110	0.13%
Keyword	400	3,600	4,000	10%
Trained 1st iteration (sample)	950	450	1,400	68%
Trained 2nd Iteration	1,696	304	2,000	85%

Table 3: Search method times and feasibility

Labeling Method	Size of set to label (tweets)	Approximate Time to Process and Label	Feasible for Real-Time Analysis	Feasible for Post-Data Analysis
Linear	2.57 million	270 days	No	No
Keyword	257 thousand	27 days	No	Not really. Maybe possible for single query of data
Machine Learning	2.57 million	Less than 4 minutes of computational, 1.5 hours of labeling	Yes, can predict all tweets less than 1 sec with 85% precision	Yes, took 75 hours of labeling time, which can be re- duced greatly

We also looked at how well each of the original collection streams that we set up on Datasift performed. Looking at Table 4, we see that:

- The 'Places' stream collected by far the most newsworthy tweets, with the highest percentage of 'news-worthy' among all the collected tweets.
- The 'Hashtags' and 'Keywords' streams collected comparable numbers of tweets, but 'Hashtags' had almost twice as many newsworthy tweets among those it collected as compared to 'Keywords.'
- The 'Usernames' stream collected the fewest number of tweets, likely because it had the lowest number of broad filters set up (37). But a relatively high percentage of the tweets captured were newsworthy.

From this we can conclude that 'Places' was the most effective stream for collecting newsworthy tweets, followed by 'Usernames,' 'Hashtags,' and 'Keywords.' This is a useful result for those interested in designing general collection streams for identifying 'newsworthy' actionable data in the future. This is laid out in Table 4 and Figure 1 below.

Table 4: Newsworthy Tweets per Stream

Collection Stream	Total Number of Tweets	Number of Newsworthy Tweets	Percent of Tweets that were Newsworthy
Hashtags	372,449	1,397	0.38%
Keywords	861,270	1,865	0.22%
Places	1,297,092	8,748	0.67%
Username	39,793	198	0.50%
<b>Total</b>	<b>2,570,604</b>	<b>12,208</b>	<b>0.47%</b>

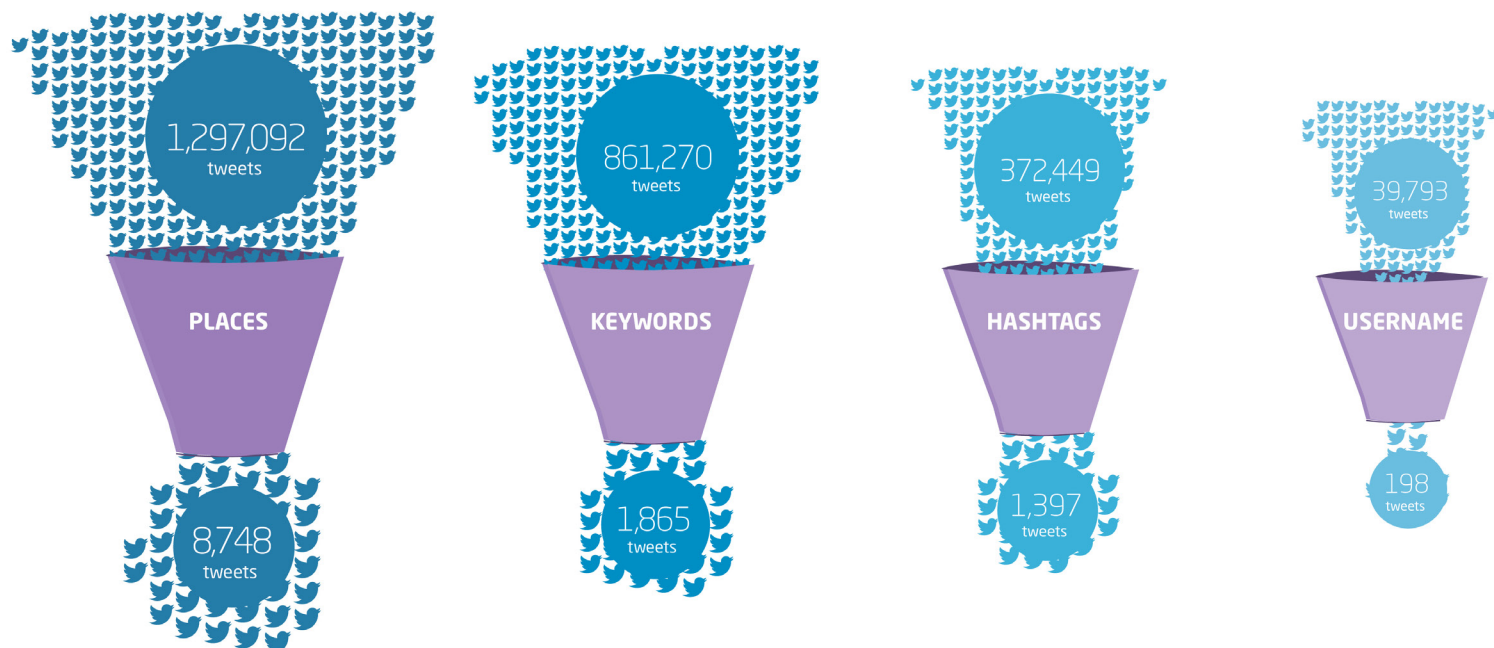


Figure 1. Visualization of the four original collection streams used on the Datasift platform illustrating the total tweets captured by each stream as input compared to the newsworthy incident they found as output.

## Conclusions about the filtering process

Through our assessment of the filtering process, we can draw the following conclusions:

- 1) that the Machine Learning Supervised Learning approach was successful in finding and filtering useful information from Twitter data;
- 2) that the 'Places' keyword stream collected the most useful information on Datasift; and
- 3) that we were successfully able to create a filter with high precision (our filter catches approximately 0.6% of newsworthy tweets while a linear search only catches approximately 0.13%).

Ultimately, one limitation is that we may have classified our filter too strictly. That is, although we have filtered accurately so a high percentage of the tweets we collected were newsworthy, we may have over-trained and out filtered, missing some tweets that we should have collected. In Machine Learning jargon, we have good precision, but do not know the level of recall. We were also able to conclude that there is a great deal of useful information on Twitter that goes beyond that found on other information sources, like in traditional media. The next section on Event Clustering and Analysis will explore this final point more closely.

## Event Clustering

Once we isolated the newsworthy tweets, we divided these into different incidents, using automated clustering techniques. This enabled event analysis across different dimensions, including the time profile of when incidents were tweeted,

allowing us to compare the speed of coverage of Twitter with other news platforms, like Uchaguzi and traditional media.

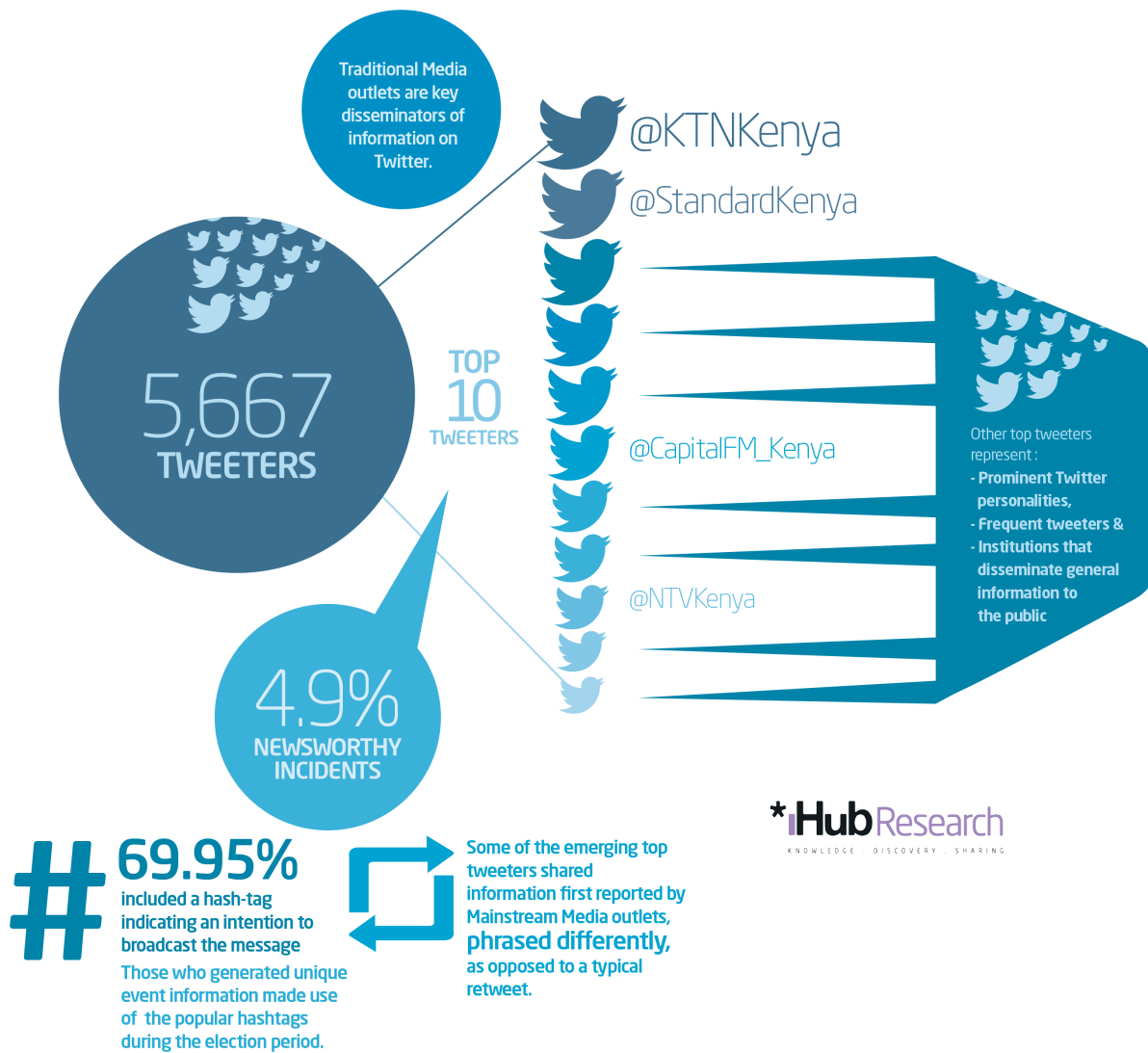
We considered two different methods for finding clusters among the newsworthy tweets.

- The first was Google Refine, which would allow us to group tweets together. Using the key collision method in Google Refine with fingerprint, n-gram fingerprint, metaphone 3 and cologne-phonetics options, 96 clusters were generated which represented unique incidents during the election period. Key collision relies on phonetic algorithms to define cluster centroids. Applied recursively, the algorithms produce all-inclusive clusters. Inspecting these categories showed a good differentiation of tweet clusters.
- The second was using K-means algorithm on word counts of tweets. This algorithm clusters tweets into a user-specified number of clusters. This was also an effective method, but required the user to tune the number of clusters. If the chosen number of clusters was too low (e.g. 5 clusters), then the clustering was too coarse, and if the chosen number of clusters was too high (e.g. 500 clusters), then the clustering was too high.

We opted for the Google Refine method as it automatically chose the number of clusters, and gave a good breakdown of the tweets. With the 96 clusters chosen, we then grouped these into locations, which resulted in a 2-level hierarchy of clustering labels, location -> event, which could be compared with other media source. We discuss the findings of this comparison in the next section.



42. Alexa, <http://www.alexa.com/topsites/countries/KE>.
43. Ushahidi was one of three key partners that participated in the Uchaguzi Kenya 2013 deployment. More information about Uchaguzi can be found at [Uchaguzi.co.ke](http://Uchaguzi.co.ke). Ushahidi and iHub are separate organizations. Ushahidi is a funder for the iHub space.
44. Rhoda Omenya, "iHub Research Uchaguzi Kenya 2013 Monitoring and Evaluation Report," *iHub Research*, July 2013, [http://www.ihub.co.ke/ihubresearch/jb\\_UchaguziMEFinalReport-pdf2013-7-5-14-24-09.pdf](http://www.ihub.co.ke/ihubresearch/jb_UchaguziMEFinalReport-pdf2013-7-5-14-24-09.pdf).
45. Kwak, Lee, Park, Moon, "What is Twitter, a social network or a news media?", *19th Conference on WWW Proceedings*, <http://doi.acm.org/10.1145/1772690.1772751>.
46. Datasift, <http://datasift.com/>.
47. Including Decision Trees, Random Forest, Naive Bayes implemented in Weka and Python modules.
48. Scikit Learn, "Machine Learning in Python," <http://scikit-learn.org>.
49. In ML jargon, we were worried that if we increased the classifier's precision, we would reduce its recall.
50. We are currently looking into quantifying the training set size vs accuracy tradeoff.



# FINDINGS

## 1. Passive Crowdsourcing is Viable During the Elections in Kenya

Our study found that in the case of the Kenyan 2013 election, Twitter provide more information than other sources, although the level of severity of the incidents was different compared to traditional media; Twitter data contained smaller, real-time interest stories. Regardless of whether they were also picked up by traditional media or captured by Uchaguzi, such stories from Twitter would be worth verifying and following up on as they might be of interest to citizens, particularly if they lived in areas where incidents were taking place. It is important to keep in mind, however, that the data collected through Twitter was not innately verifiably through the data itself. If verified information were needed, on-the-ground monitors, as employed by Uchaguzi, would likely be necessary.

### 1.1 The main tweeters of the news are news outlets

There were 5,667 tweeters that reported newsworthy election-related incidents. Of these, the top ten included prominent Twitter personalities, frequent tweeters and institutions that disseminated general information to the public; almost half of the top ten tweeters were traditional news outlets posting on their official Twitter channels. These included Standard Kenya (1st), KTN Kenya (2nd), Capital FM Kenya (6th) and NTV Kenya (8th). This shows that traditional media is still a key disseminator of newsworthy information, even on Twitter, highlighting their ongoing important role on not only traditional platforms, but also social media platforms.

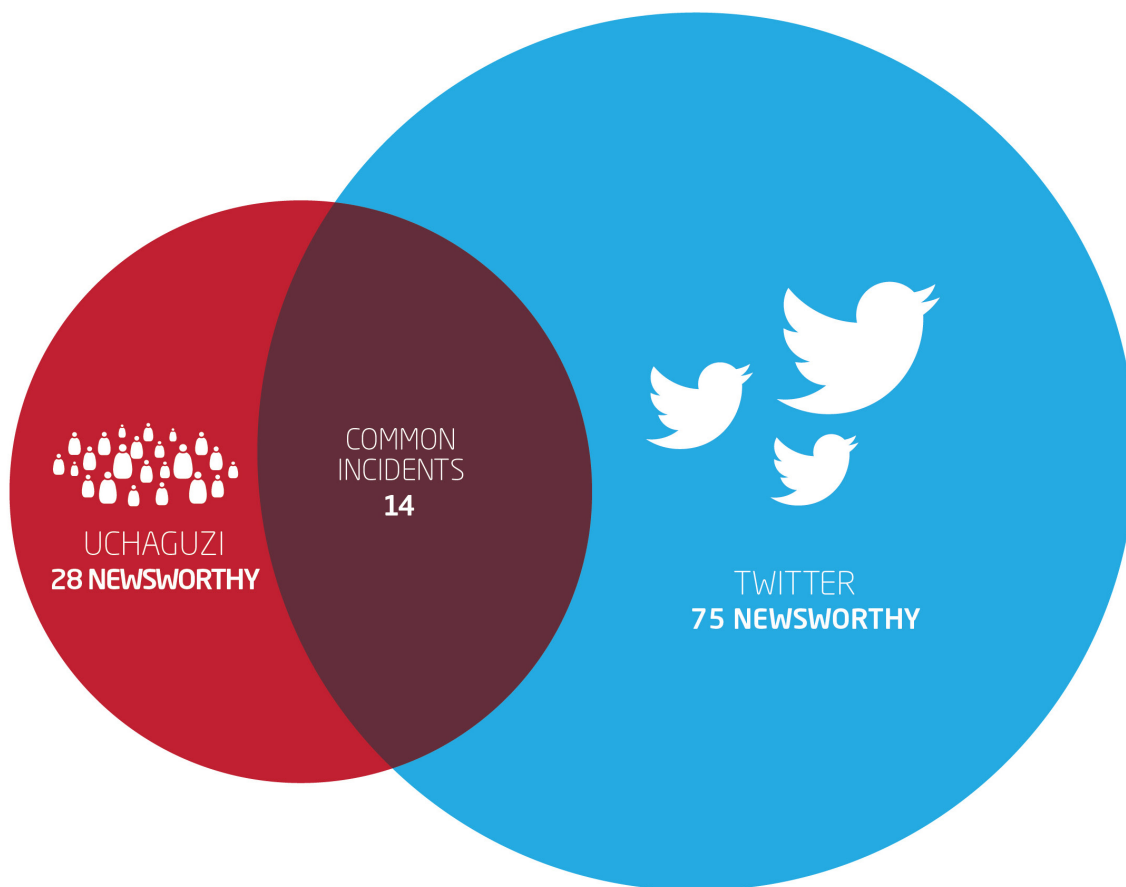
These top ten tweeters generated 5% of the newsworthy incidents captured. Out of the incidents they reported, 70% included a hashtag, indicating an intention to broadcast the message. Some of the emerging top tweeters shared information first reported by traditional media outlets, rephrased instead of simply retweeted. Others who generated unique event information made use of popular hashtags during the election period.

One particular tweeter (@OnOurRadar) seemingly had a network of informers whose event information they amplified, based on the structure of their tweets, which indicated the event, its location and the name of the person who shared the information with them. All ten tweeters were from Kenya (based on their provided user location), although none of their tweets were geo-enabled.

Table 5: Summary Comparison Of Incidents Across News Sources

	Twitter	Uchaguzi	Traditional Media	Fieldwork
Twitter	75	14	22	6
Uchaguzi		28	5	4
Traditional Media			24	4
Fieldwork				13

Entries in the table above indicate the number of incidents that the news source in the row and the column have in common. So diagonal entries indicate the number of incidents for a particular news source. Twitter, for example, captured 75 different incidents. Comparing Twitter and traditional media, we see that they had 22 incidents in common, which is almost all of the 24 incidents found on traditional media. This table only shows a two-way intersection of incidents. It does not show information on incidents that were reported by more than 2 sources.



*Total Unique Incidents from all sources = 96*

*Figure 2. A comparison of the number of newsworthy incidents found on Twitter versus Uchaguzi.*

## TWITTER OUTSTRIPS UCHAGUZI ON ACTIONABLE INFORMATION BUT HAS MORE NOISE

As illustrated in Figure 2, data collected straight from Twitter retrieved more newsworthy incidents (75). Nonetheless, Uchaguzi captured 28 ‘newsworthy’ incidents that were not found on Twitter; these were incidents that occurred in more remote areas of the country such as Samburu and Garissa. Use of an active crowdsourcing platform such as Uchaguzi therefore appears to be especially relevant when one is interested in obtaining information from areas where residents may not be already using social media platforms such as Twitter.

As Table 6 depicts, Uchaguzi also had a better signal-to-noise ratio, due to the differing characteristics of passive crowdsourcing and active crowdsourcing. Since Uchaguzi engaged in active crowdsourcing, directly asking the crowd for information, it is not surprising that they received a higher proportion of relevant reports. In comparison, passive crowdsourcing, which we engaged in for our data mining from Twitter, entails sifting through a broader flood of information to try to find relevant information.

*Table 6. Uchaguzi and mined Twitter reports, incidents, and noise*

	Twitter	Uchaguzi
Number of Total Reports	2,570,000	2,087
Number of ‘Newsworthy’ Reports	12,000	37
Number of Noisy Reports	2,558,000	2,050
Newsworthy-to-Noise Ratio	0.004691	0.018049

## TWITTER NEWS IS MORE LOCALIZED THAN TRADITIONAL MEDIA NEWS

Interviews we conducted with members of traditional media revealed that during the Kenyan election process they carefully consider their audience when deciding whether to publish a report about an incident or not. Therefore, traditional media houses appeared to have lower propensity to publish localized incidents, regardless of whether they might be considered ‘newsworthy’ by a local audience. Examples of Twitter posts about incidents that were not captured by traditional media include:

- “Gunshots at Kawangware primary school,”
- “Looting in Kawangware 56,”
- “5 AP [Administration Police] drunk and harassing the public at Kibera Kichinjio,”
- “2 perceived CORD supporters one G4S and another KK security beaten seriously injured in Kibera,”
- “Attack of the two security guards occurred around Kibera’s Laini Saba.”

While these incidents are highly localized and would not be of interest to international news, they are important for election monitors to note as they may be relevant for humanitarian organizations and authorities to potentially follow-up on.

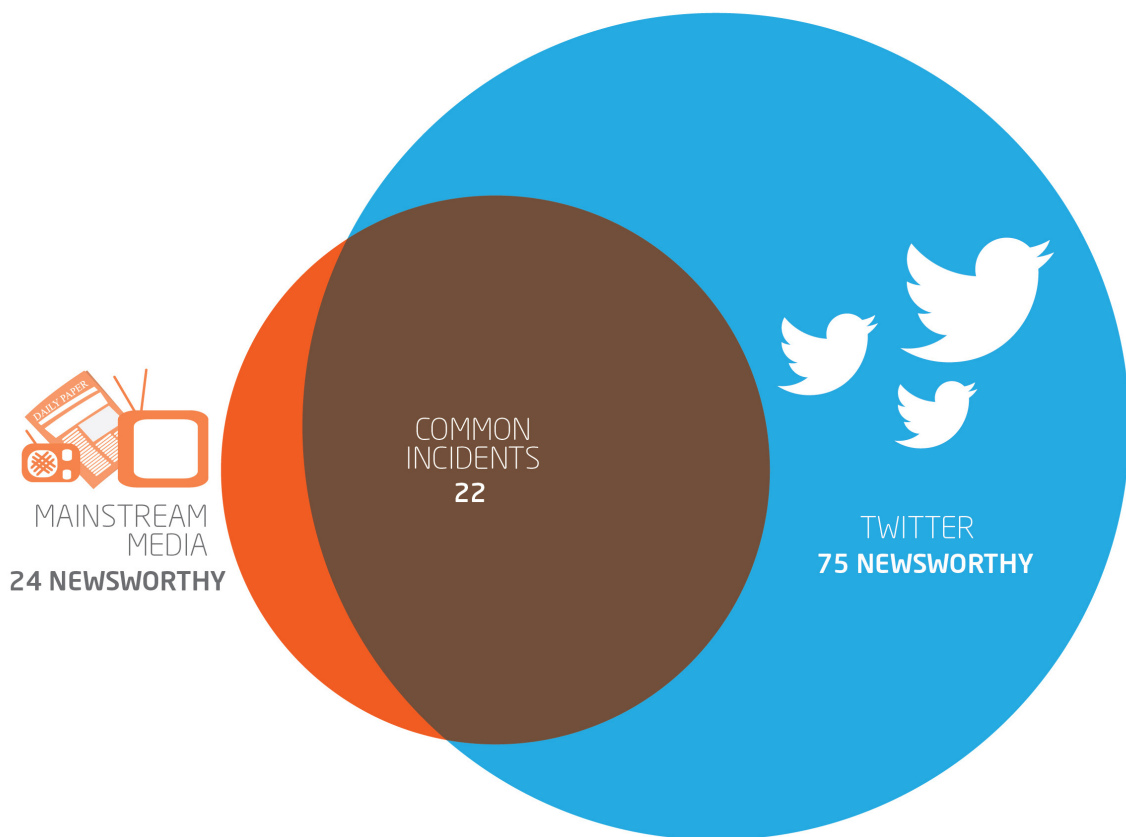
These individual encounters of local violence may be particularly useful when aggregated, forming a larger story worthy of traditional media coverage or providing insights into the state of an area for purposes of situational awareness. One example is the spontaneous violence that our Twitter analysis indicated occurred in Kawangware, an area in Nairobi, during the election period (See side box), but no traditional news

outlet covered the story.

Another example of the “localized” and “personalized” nature of Twitter data is the additional information provided by Kenyans on Twitter about the state of violence in Nairobi’s Mathare slum area. While the major news outlets provided the dominating headlines on the overall state of affairs, Twitter carried more personal accounts including:

- *“#KOT Just heard unconfirmed reports that 1 person was killed this morning @3am by #Mathare 3C thugs..”*
- *“Crew and I just got chased out of #Mathare by ODM supporters who are angry at #IEBC and media. But they are still calling for peace.”*
- *“Tension high in some parts of Mathare as several gunshots goes off!”*
- *“David Kariuki in Mathare slum reports a stabbing + Pangani police’s response, we are awaiting further information + confirmation #KenyaDecides”*
- *“@KenyaPolice Plans are underway for youth at Mathare Area 4 to cause problems after the supreme court ruling 2morrow. @MajorEChirchir”*
- *“Some pple have just blocked the rd here in Mathare leading police throwing tear gas @ntvkenya @KTNKenya @citizentvkenya”*

These and other tweets narrate a story and give an indication of what might need to be confirmed or verified on the ground.



Total Unique Incidents from all sources = 96

Figure 3. A comparison of the number of newsworthy incidents found on Twitter versus Traditional Media.

## Kawangware:

### **Bullet Shots Only Reported Through Twitter**

The first tweet on Kawangware was reported at 6:31 AM on March 4, 2013, “there r bullet shots at Kawangware pri. School, what is really going on...is it intimidation by the police?”. The tweet was followed up nine hours later by another tweet reading, “Mike we were at Kawangware primary school. Police officers fired into air on two occasions a minute or so apart.” Five days into the election period, other reports of violence surfaced on Twitter: “Kindly send police at Riruta Junction near Kawangware...nt doing so well. Rioting youth from both side. scared @CitizenTVNews @ntvkenya #IEBC.” On the same day, another Twitter user reported a group of youth walking from Kawangware towards the Central Business District, “Apparently it’s a procession from Dago/Kawangware heading towards town along Ngong Rd.”

These incidents in their individual capacity might not have amounted to an incident worth covering in the traditional media, but strung into a sequence they depict a worsening situation. At 8 PM on the same day, a different Twitter user alerts the Twitter account of a media outlet of the killing of a watchman in Kawangware: “@KTNKenya a watchman has been killed at 46 stage Kawangware.” Then suddenly, a barrage of tweets giving situation at Kawangware stream in, “@kenyapolice #Kawangware near the terminus fujo...smthin has to be done...”. Days later on March 27th, more tweets indicate violence in the area: “@kenyapolice @redcrosskenya there is a woman being mob justiced at delivarance Kawangware”, “So much tension in Kawangware, demonstrators causing havoc”, and “A couple was attacked outside their home in Kawangware by a gang who stole their vehicle.”

The day of the Supreme Court ruling about Candidate Odinga’s petition also saw numerous violent incidents in the same area as captured on earlier Twitter during the elction. More Twitter users shared their experiences, “Bagas rioting at Kawangware 56 stage after petition ruling. Police have intervened”, “Violence in Kawangware 56. Houses being torched”, “People’s houses going up in flames in Kawangware right now... SMFH”, and “@ItsMainaKageni A man was killed by muggers near Kawangware primary this morning., another had serious cuts near elshadai @KenyaPolice.”

In this case, Twitter served as an early warning system about the level of violence in the area through the testimony from members of the public. Piecing together the individual experiences made a story worthy of a headline. This also highlights another potent feature of Twitter: its ability to provide specific details of a story as seen or perceived by an eyewitness. Incorporating sentiment analysis techniques applied to such tweets in the future may enable the prediction of incidents, potentially providing traditional media and relevant authorities with information that they might miss through other information sources.

## TRADITIONAL MEDIA CAN COVER REMOTE AREAS WHERE TWITTER IS INACCESSIBLE

Our research showed that traditional media, like the data from Uchaguzi, sometimes contains information that might not otherwise be available on a social media platform like Twitter. In order to be a Twitter user an individual must have a phone or laptop, 3G Internet connection and being a user of social media, prohibiting sections of the population from participating. Passive crowdsourcing of social media for election monitoring therefore can only work for places where the crowd is already generating data. If a particular crowd does not have access to the platform, then they cannot generate the information, and other ways may need to be devised for sourcing information from that location.

In such scenarios, traditional media can leverage their news gathering networks to illuminate incidents in more rural, far-flung areas of the country where the proportion of social media users is lower. For example, one such incident was the killings in the coastal town of Malindi on Election Day. No information related to these incidents was found on Twitter, showing the limitations of passive crowdsourcing for capturing information from regions of little or no connectivity, or where passive crowdsourcing platforms, like social media, are not widely used. Determining if passive crowdsourcing is viable based on your needs is therefore very important to do prior to choosing a crowdsourcing method. Our framework provides a starting point to determining the types of crowdsourcing that might be most suitable for your needs and context.

## INTERNATIONAL AND LOCAL MEDIA DIFFER IN GRANULARITY

It is important to remember that the traditional media is not a monolithic whole; there are distinct differences in terms of what information different outlets choose to disseminate. Our research showed that international media outlets tended to summarize widely and did not report on very granular incidents. This is because of the global nature of the audience that they serve. For example, on Election Day, in the international media *The Guardian* reported, ‘Kenyan elections marred by Mombasa violence,’ *The Telegraph* reported, ‘Kenyans defy machete violence to vote in historic

election,' and *Al Jazeera* reported, 'Kenya deploys forces to contain violence.' In contrast, in the local media Capital FM reported, 'Security Bolstered after MRC Attack Leaves 14 Dead,' and headlines in *The Daily Nation* included 'Long queues in Embakasi Constituencies' and 'Two seized for alleged voter bribery.' The wider audience of the international media limited what they considered to be a newsworthy event to only major incidents, while local media encompassed both the major incidents and the granular details of the election.

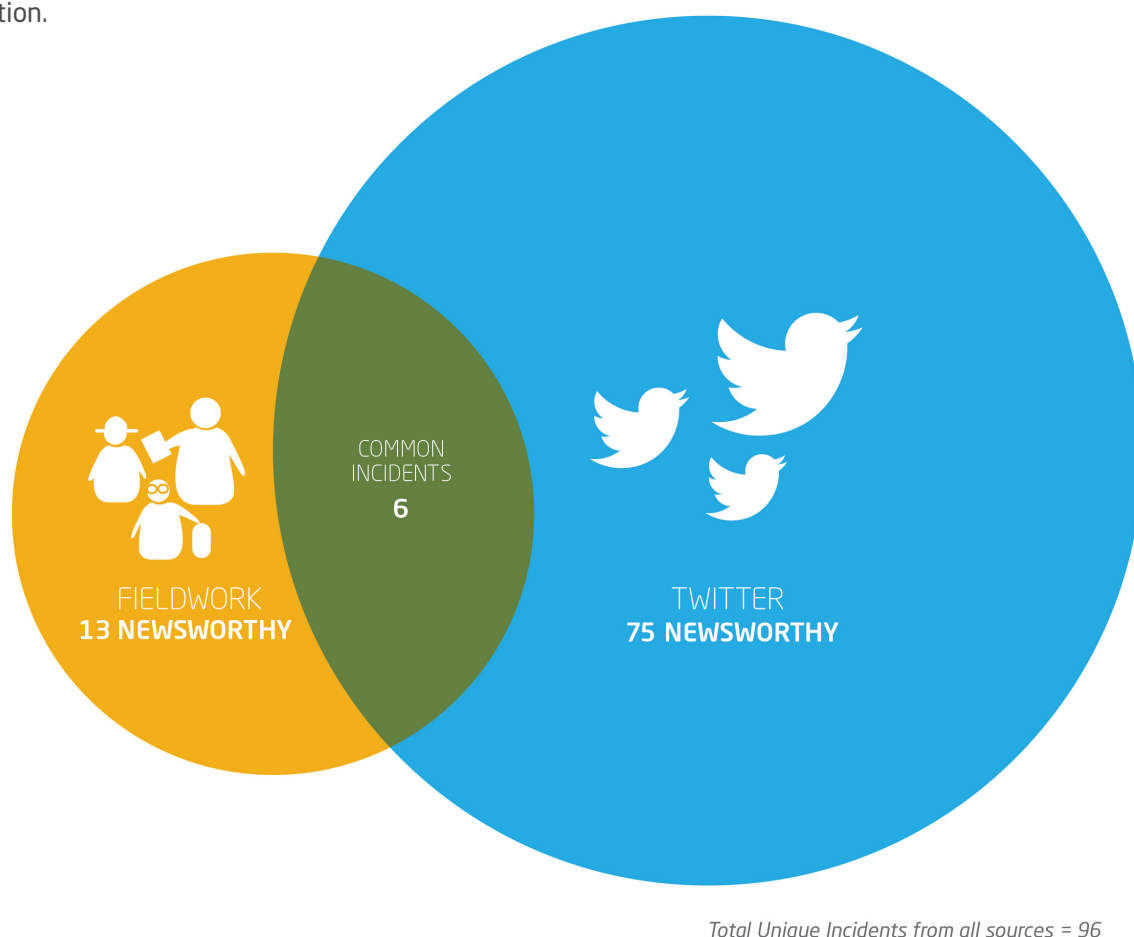


Figure 4. A comparison of the number of newsworthy incidents found on Twitter versus Fieldwork.

## IT IS EASIER TO FIND EYEWITNESS ACCOUNTS ON TWITTER THAN IN PERSON

During the fieldwork we conducted in Kisumu, our research team actually found it very difficult to obtain first-hand or even second-hand accounts of violence. Although many of our 85 interviewees acknowledged witnessing some forms of violence, often their accounts lacked the detail and even the personal quality of posts on Twitter. This might be due to the time lag between when the incident happened and when our research team was in the field (over 5 weeks later), possibly resulting in people forgetting details. It could also be the case that individuals feared reporting about a violent incident in person. As a result, we concluded that verifying incidences by in-person interviews with victims and residents of areas where violence had been reported on Twitter was not necessarily the most effective method. It was also difficult to find the appropriate individual who might have witnessed an event, and the cost and time it took to travel to the field were both high. However, we did find it easier to obtain a higher level of detail in Mathare and the coast region. This could potentially be due to the more well-established social networks our researchers had in these regions than in Kisumu, or to the more pervasive nature of the violence in those locations.



## 2. Twitter Breaks News During Elections

In addition to comparing the quantity and quality of the information gathered from these different data sources, we also looked at how Twitter and traditional media differed across time. By looking at the relationship between the lead and lag times of when news broke in the traditional media and on Twitter, we found that Twitter either reported on the same day as traditional media outlets, or led in reporting the story. When Twitter led, it was by a margin of approximately one day. In the age of the 24-hour news cycle for television broadcast news, lead-time of one day during an election period can make quite an important difference. Our finding therefore highlights that Twitter's value stems not only from increased incident coverage, but also from its ability to offer information in real-time. This is visualized in Figure 5 below, which shows the timing by day of incident reports on Twitter versus online traditional media for the Kenyan 2013 election.

Due to capacity, we were unable to conduct this comparison at finer intervals than one day. In future research, we will consider conducting time comparisons at finer time intervals to see who breaks a story first by the hour, particularly in places where both mediums reported incidents on the same day. Also, due to the primary focus of this project on content produced through social media, we did not run this analysis on the other information sources (e.g. Uchaguzi versus Traditional Media). In our future work, we plan to run this analysis on the other data sets.

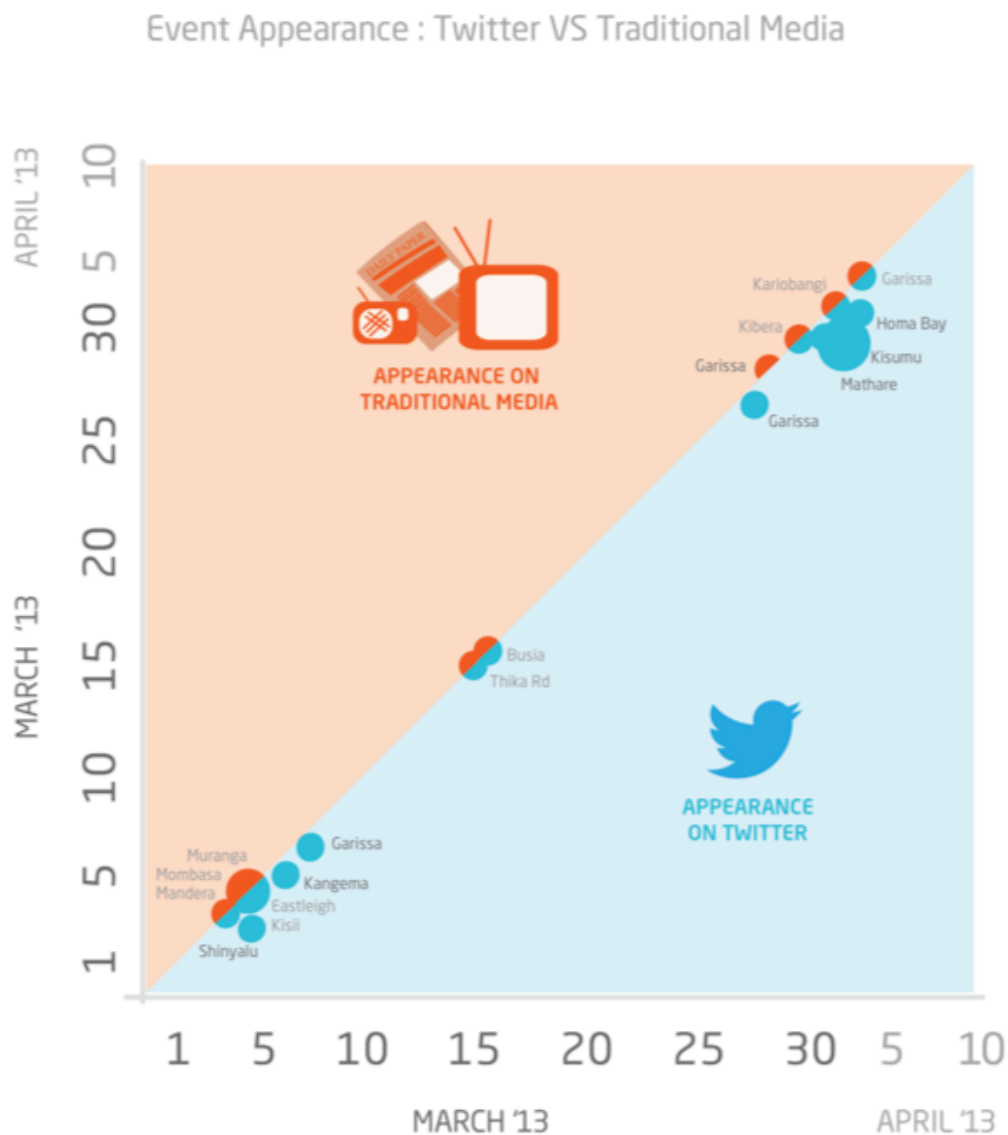


Figure 5. Timeline of Newsworthy Incident Appearances in Traditional Media versus Twitter

Points on the diagonal line in the figure show incidents being reported by both traditional media and Twitter on the same day. The blue region indicates where Twitter first reported a story; the orange region shows the same for the traditional media. We can see that most dots are either on the diagonal or slightly inside the blue area, which demonstrates that Twitter either reported on the same day or slightly ahead of traditional media. We expect that this time lag may have been due to the more thorough verification and editing process that traditional media outlets undertook before publishing a story in print or online. One important limitation of our finding is that we were unable to include breaking news on television and radio broadcast channels in our assessment of traditional media. This was because, at the time of the study, we did not have the capacity to constantly monitor these outlets during the election period. If we had, it is possible that this timeline would have been significantly different.

A second time analysis we performed looked at how individual incidents were reported on Twitter over time. Figure 6 shows one result of this analysis with the volume of tweets collected about the protests in Kisumu on March 30 over time.

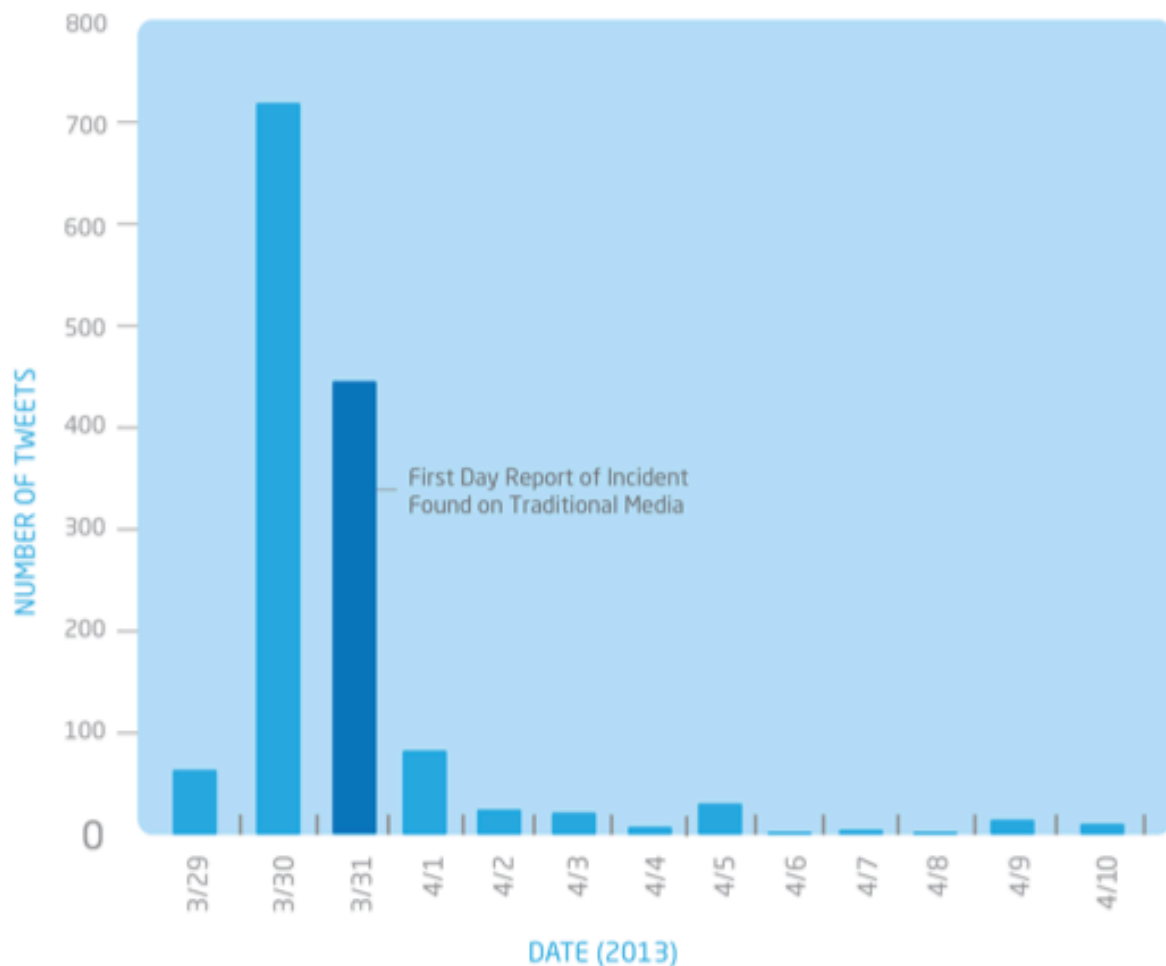


Figure 6. Timeline of Kisumu Incident in March and April 2013

We see that there is a high level of tweeting activity on March 30th, the day of the protests, but that this drops off sharply over the next two days. We found that this pattern of a quick spike in Twitter activity during an event, with quick drop off after the event was common across other incidents during the election period. In the case of the protest in Kisumu case, the highest level of reporting on Twitter occurred a day before the incident was reported by the traditional media on March 31st.

Another interesting finding from our analysis of this event is that Twitter activity actually preceded the event itself; most of these tweets posted on March 29th were reports of tension in Kisumu, potentially serving as a useful predictor for the event. Indeed it seems that our Twitter data captured the moment when tension turned to violence on the day of the incident as seen in the following three tweets posted over a time interval of around 30mins on March 30th, 2013.

**At 17:30:57, the first tweet reported, 'Tension at carwash Kisumu,' showing only tension and no actual incident.**

**At 17:41:19, the second tweet reported, 'Happening now.....demonstrations in kisumu city near car wash,' illustrating the occurrence of an event.**

**At 17:47:12, the third tweet reported, 'Kisumu in its element!! Pple running n screaming, Shots fired....' showed the final stage of the escalation of the incident into violence. A large volume of tweets about the incident and the violence followed after the third tweet.**

This result illustrates how Twitter may be able to provide leading indicators of incidents in real time, faster than by simply monitoring reports in traditional media. However, although our filter picked up these reports of leading tension, the search criteria we used in this study were tuned to only look for actionable information, rather than sentiment or expression of feeling such as tension. So picking up such leading indications of incidents was in fact atypical for our study. The filter would need to be altered to pick up more such sentiment in order to more reliably serve as an incident-predictor mechanism. However, we believe a filter incorporating

this sort of sentiment analysis would be feasible and is an area of research we intend to pursue in future studies.

In summary, Twitter broke news as compared to traditional media during the elections, in some cases by up to a day. It appears that Twitter's advantage to traditional media is its ability to break news in real time. Even where Twitter and traditional media both covered an incident on the same day, Twitter was likely to have started covering it earlier in the day, and if trained properly, Twitter has the potential to predict an incident occurrence.

### 3. Mining of Twitter Data Without Machine Learning Is Not Feasible

As outlined above, our analysis found that in the case of the Kenyan 2013 election Twitter provided access to useful, first-hand eyewitness information, often not available in other media outlets in near real-time. However, we also found that to extract this information, we needed technical expertise in data-mining techniques from the field of Machine Learning. Simple searching of the data was not feasible as discussed earlier in the methodology section and as demonstrated below in Table 7.

Table 7: Processing time for Machine Learning

Search method	Time taken	Number of Newsworthy Tweets	Search time for whole data set	Viable for real time analysis	Viable for post-data analysis
Linear search	90 hours	100	270 days	No	No
Keyword search	4.5 hours	400	27 days	No	In a very limited way
ML, supervised learning	Less than 6 minutes computational time for both iterations, 1.5 hours labeling	12,208	Less than 1 second	Yes	Yes

Based on the above table, you can see that labeling the data manually from start to finish would take approximately 270 days. A 90-hour search only found 100 newsworthy tweets. This method is not feasible for real-time, or post-data analysis. Similarly, searching the data using combinations of keywords would take a minimum of 15 days, for each similar question criteria to newsworthy. Searching for 4 hours produced 400 newsworthy tweets. Again, this is not feasible for real-time analysis, and barely feasible for post-data analysis in a limited fashion. Combining search results with Machine Learning techniques allowed us to train search algorithm in less than 3 minutes, searched 2.57 million tweets in less than 1 second, predicting 17,780 newsworthy tweets.





# CONCLUSION

The three main goals of this study were to 1) test the viability of passive crowdsourcing in the Kenyan context, 2) to determine which information-gathering mechanism (passive crowdsourcing on Twitter, active crowdsourcing on Uchaguzi, or online publications of traditional media) produced the best real-time picture of the on-the-ground reality, and 3) to develop a framework to help aspiring crowdsourcers to determine whether crowdsourcing is viable in their context and if so, which techniques will offer verifiable and valid information.

By conducting a quantitative analysis of the data collected from Twitter during the Kenyan election period, we found that passive crowdsourcing is indeed viable in the Kenyan election context, but only using machine learning techniques. Mining Kenyan Twitter data during an election scenario looks to be a very valuable and worthwhile technique when looking for timely, local information. However, mining is only possible if the user has knowledge of machine learning techniques since, without such techniques, the manual process can take up to 270 working days.

In contrast, although the platform collected less data in absolute terms, active crowdsourcing through Uchaguzi collected a high proportion of relevant actionable data, since mechanisms were in place to approve and verify the information obtained. We therefore conclude that both crowdsourcing techniques can be useful mechanisms to collect information in Kenya during an election, but one may be more appropriate than the other depending on whether verified information is of paramount importance or not. There is also a potential difference in the resources needed to conduct each of these techniques that should be considered.

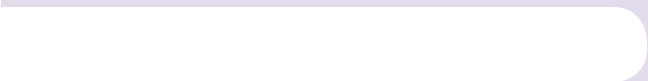
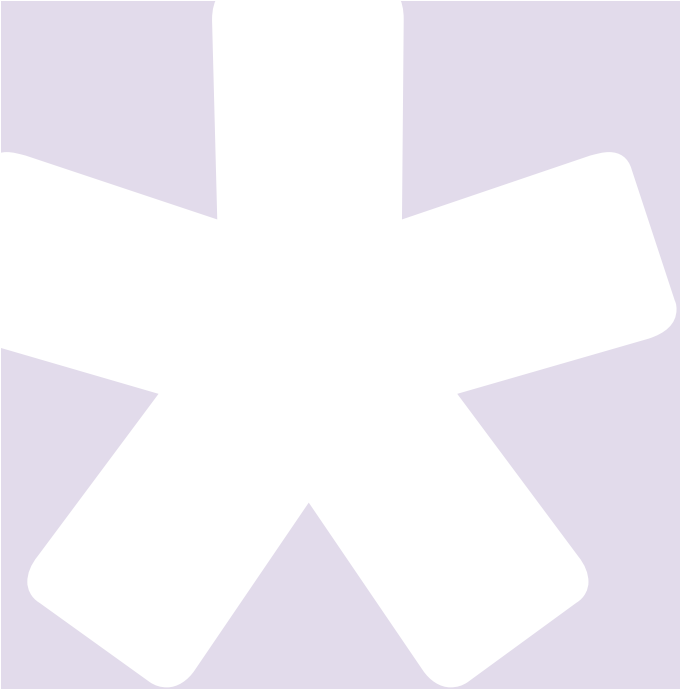
The second objective of the study was to understand what information, if any, Twitter provided beyond traditional media sources, and other crowdsourcing platforms, such as Uchaguzi. We found that Twitter reported incidents as fast or faster than traditional media (as measured in days), though these reports had the disadvantage of not being previously verified like traditional media or Uchaguzi. Twitter contained sets of information/localized information useful to particular interest groups that may not be broadcast by traditional media. Aggregation of such content could string together newsworthy information on a grander scale.

Our third objective of this study was to determine whether there are particular conditions that need

to be in place in order for crowdsourcing using online and mobile technologies to be a viable way to gather information during an election. By looking at the particular case of the 2013 Kenyan election, we found that indeed there are factors and considerations that are useful in assessing whether there will be an adequate online 'crowd' to source information from. These include, among others: 1) the availability of, and access to, Internet, 2) the adoption and penetration of mobile phone telephony, and 3) the extent and culture of social media networks usage. We further found that it is crucial to consider what type of data is required by the aspiring crowdsourcers before deciding how to gather data. For our project for instance, we desired data from multiple sources for comparative analysis, so we used both passive crowdsourcing to compare to existing active crowdsourcing project, Uchaguzi.

Based on these findings as well as existing knowledge we identified through our literature review, we created a '3Vs Crowdsourcing Framework for Elections' made for practitioners such as journalists or crowd-mappers. The aim of the framework is to provide guidelines for any crowdsourcers, new or experienced, who are interested in seeing if crowdsourcing is a viable option in a particular location and if so, what type of crowdsourcing is appropriate. This framework helps to carry out an effective crowdsourcing activity by prompting the potential crowdsourcer to investigate the factors that facilitate the sharing of information by 'ordinary citizens,' who generate the bulk of crowdsourced information.

The framework we have created outlines the elements aspiring crowdsourcers should look at before beginning a crowdsourcing project in a given country including: the data desired by a crowdsourcer, as well as the level of freedom and risk, and the technology, infrastructure, demographic and behavioral/cultural factors of the country where crowdsourcing is being proposed. We hope that this 3Vs framework will be tested and revised through practical application in different country contexts. As a first test case, we have retroactively applied the draft framework to the Kenyan 2013 General Election context. In future scenarios, we hope potential deployers will test the potential viability of their crowdsourcing initiative using the framework prior to implementation. Based on the work achieved thus far, we look forward to engaging the wider crowdsourcing community to testing the 3Vs Framework in other countries approaching elections.





# REFERENCES

Castillo, C., Mendoza, M., & Poblete, B. "Information Credibility on Twitter." *Paper presented at the WWW 2011, Hyderabad, India. (2011).*

Cheong, C.; Cheong, F. "Social Media Data Mining: A Social Network Analysis Of Tweets During The 2010-2011 Australian Floods." *PACIS 2011 Proceedings*. Paper 46. (2011) <http://aisel.aisnet.org/pacis2011/46>.

Estéllés-Arolas, E.; González-Ladrón-De-Guevara, F. "Towards an integrated crowdsourcing definition." *Journal of Information Science* 38 , no. 2 (2012): 189-200.

Gao, H., Barbier, G., & Goolsby, R. "Harnessing the Crowdsourcing Power of Social Media for Disaster Relief." *IEEE Intelligent Systems*, 10 (2011).

Kwak, H.; Lee, C.; Park, H.; Moon, S. "What is Twitter, a social network or a news media?." In *Proceedings of the 19th international conference on World wide web (WWW '10)*. ACM, (2010): 591-600. <http://doi.acm.org/10.1145/1772690.1772751>.

Lamos V. "On voting intentions inference from Twitter content: a case study on UK 2010 General Election." *CoRR*, April 2012.

Meier P. "Verifying Crowdsourced Social Media Reports for Live Crisis Mapping : An Introduction to Information Forensics." *irevolution.com* (blog). 2011. <http://irevolution.files.wordpress.com/2011/11/meier-verifying-crowdsourced-data-case-studies.pdf>.

Meier P. "Traditional vs. Crowdsourced Election Monitoring: Which Has More Impact?" *irevolution.com* (blog). 2012. <http://irevolution.net/2012/08/07/traditional-vs-crowdsourced-election-monitoring/>.

Petrovic, S., Osborne, M., McCreadie, R., Macdonald, C., Ounis, I., & Shrimpton, L. "Can Twitter replace News-wire for breaking news?." In *International AAAI Conference on Weblogs and Social Media*. 2013. Retrieved from <https://aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6066>.

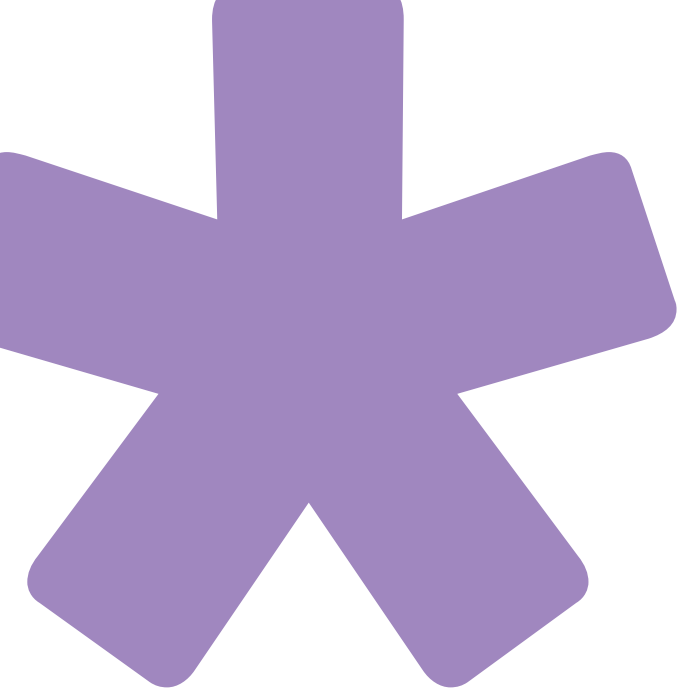
Sharma A. "Crowdsourcing Critical Success Factor Model Strategies to harness the collective intelligence of the crowd." *irevolution.com* (blog). 2010. <http://irevolution.files.wordpress.com/2010/05/working-paper1.pdf>.

Starbird K. "Digital Volunteerism During Disaster: Crowdsourcing Information Processing." *Workshop Position Paper for CHI '11 Workshop on Crowdsourcing and Human Computation at the 2011 Conference on Human Factors in Computing Systems (CHI 2011)*. 2011. Vancouver, Canada.

Tumasjan, A.; Sprenger, T.; Sandner, P.; Welpel I. "Predicting Elections with Twitter : What 140 Characters Reveal about Political Sentiment." *International AAAI Conference on Weblogs and Social Media*. May 2010. <https://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1441>.

Vieweg, S. "Microblogged contributions to the emergency arena: Discovery, interpretation and implications." 2010. *Paper presented at the Computer Supported Collaborative Work*.

Yates, D., & Paquette, S. "Emergency knowledge management and social media technologies: A case study of the 2010 Haitian earthquake." *International Journal of Information Management*, 31, no. 1 (2011): 6-13.



# Appendix


Complete List of all keywords, usernames, places and hashtags used are all listed at <http://goo.gl/rYiwrn>.

Table: Keywords and Names of Places that generated significant noise in our data capture

Word	Kenyan Context	Other Context
Kura	Means 'vote' in Swahili, Kenya's national language	A resort in Java, Indonesia.
Nandi	A highland area, and a county in Kenya's Rift Valley	A village in northern Tanzania.
Meru	A dialect of the Kalenjin tribe	A town located in Selangor, Malaysia  A small town in Jharkhand, India.  Verb conjugation meaning 'am' in Spanish.
Soy	A Kenyan county and town.	The fruit of the soybean plant.
Dida	A language of the Ameru people of Kenya	
Suba	An electoral constituency in Kenya (Uasin Gishu County)	A place in Malaysia.
Teso	Name of a presidential candidate in the 2013 Kenya General Elections  A constituency, a language, and a people of Kenya  An ethnic community in Kenya.	A textile company in the UK (that had called for a strike at around the KE General Election period).

<i>Location Cluster</i>	<i>Traditional Media</i>	<i>Twitter</i>	<i>Uchaguzi</i>	<i>Fieldwork</i>
BabaDogo	0	0	0	1
Buruburu	0	1	0	0
Busia	1	1	0	0
Dandora	0	0	1	0
Nairobi	0	1	0	0
Eastleigh	1	1	1	0
Eldoret	0	1	0	1
Garissa	1	1	0	0
Githurai	0	1	0	0
HomaBay	1	1	0	0
Kamba	0	0	1	0
Kangema	1	1	0	0
Kariobangi	1	1	0	0
Kariokor	0	1	0	0
Kawangware	0	1	0	0
Kiambu	0	0	1	0
Kibera	1	1	1	1
Kilifi	1	1	1	1
Kisii	1	1	0	0
Kisumu	1	1	1	1
Kitengela	0	1	0	0
Kondele	0	1	1	0
Makongeni	0	0	1	0
Malindi	1	0	1	0
Mandera	1	1	0	0
Marsabit	0	1	0	0
Mathare	1	1	1	1
Migori	1	1	0	0
Mombasa	1	1	1	1
Moyale	0	1	0	0
MtElgon	0	0	1	0
Muranga	1	1	0	0
Ngongrd	0	1	1	0
Parklands	0	1	0	0
Pumwani	0	1	0	0
Riruta	0	1	0	0
Rongai	0	1	0	0
Shinyalu	1	1	0	0
ThikaRd	1	1	1	0



For more information on this project, contact **\*iHubResearch** - Nairobi, Kenya  
[www.research.ihub.co.ke](http://www.research.ihub.co.ke) |  @iHubResearch

*This research was funded by Canada's International Development Research Centre.*



© iHub Research - 2013