

Gender-Based Violence in 140 Characters or Fewer: A #BigData Case Study of Twitter

Hemant Purohit^{1,2}, Tanvi Banerjee^{1,2}, Andrew Hampton^{1,3}, Valerie L. Shalin^{1,3},
Nayanesh Bhandutia⁴ & Amit P. Sheth^{1,2}

¹Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Dayton, OH, USA

²Department of Computer Science and Engineering, Wright State University, Dayton, OH, USA

³Department of Psychology, Wright State University, Dayton, OH, USA

⁴United Nations Population Fund Headquarters, NYC, NY, USA

Corresponding Authors:

Hemant Purohit, Amit Sheth, Valerie Shalin: {hemant, amit, valerie}@knoesis.org

ABSTRACT

Humanitarian and public institutions are increasingly relying on data from social media sites to measure public attitude, and provide timely public engagement. Such engagement supports the exploration of public views on important social issues such as gender-based violence (GBV). In this study, we examine Big (Social) Data consisting of nearly fourteen million tweets collected from the Twitter platform over a period of ten months to analyze public opinion regarding GBV, highlighting the nature of tweeting practices by geographical location and gender. The exploitation of Big Data requires the techniques of Computational Social Science to mine insight from the corpus while accounting for the influence of both transient events and sociocultural factors. We reveal public awareness regarding GBV tolerance and suggest opportunities for intervention and the measurement of intervention effectiveness assisting both governmental and non-governmental organizations in policy development.

Keywords: computational social science, gender-based violence, social media, citizen sensing, public awareness, public attitude, policy, intervention campaign

KEY FINDINGS & IMPLICATIONS (see Table 1 for examples)

Social Media Content:

1. Substantial GBV related content exists in social media.
2. Spikes in GBV content reflect the influence of transient events, particularly involving celebrities.
3. Gender, language, technology penetration, and education influence participation with implications for the interpretation of quantitative measures.
4. GBV content includes humor and metaphor (e.g. in Sports) that reflect both attitude and behavior.
5. Content highlights the role of government, law enforcement and business in the tolerance of GBV.

Relevance to GBV Policy:

6. Social Media provides an alternative for measuring GBV attitude and behavior that is cheaper, faster, and broader than conventional survey-based methods.
7. Regional socio-cultural context influences both the measurement and the interpretation of data.
8. Computational methods for context-sensitive monitoring, modeling and interpreting GBV social media content are highly feasible.
9. The location and network of social media participants supports targeted regional anti-GBV campaigns.
10. Policy makers require tools to make social media content accessible in near-real time to monitor the effectiveness of anti-GBV campaigns.

[1.] INTRODUCTION

Gender-based violence (GBV), primarily against women, is a pervasive, global phenomenon affecting both developed and developing countries. Over 35% of the world's female population has experienced gender-based violence at some point in their lives (World Health Organization, 2013). According to the United Nations Population Fund (UNFPA)¹, "GBV is a serious public health concern that also impedes the crucial role of women and girls in development." However, anti-GBV sentiment is not universal, apparent in sexist chants during a professional sport² and the public response to convicted offenders³. While a United Nations campaign acknowledges GBV as a societal problem (United Nations, 2009), prevalence is very difficult to assess. The European Union's council report on GBV highlighted a persistent lack of comparable data across regions and over time (European Union, 2010), hampering both assessment and mitigation. Both the UNFPA⁴ and the European Union Agency for Fundamental Rights⁵ seek better data sourcing and policy design (European Union Agency for Fundamental Rights, 2012).

Mining large-scale online data from mobile technology and social media promises to complement traditional methods and provide greater insight with finer detail. Computational Social Science (Lazer et al., 2009) has leveraged such data to inform programs in a variety of domains, for example disaster response coordination (Purohit et al., 2014; Purohit et al., 2013; Vieweg et al., 2010), health (Culotta, 2014; De Choudhury et al., 2013), and drug abuse (Cameron et al., 2013). Here we examine the potential of Computational Social Science to address the problem of monitoring and mitigating GBV on a global scale (see Table 1), confronting the typical Big Data challenges of large-scale *volume*, *velocity* of content generation, *sparsity* of data behaviors, *variety* in language complexity and *heterogeneity* in participant demographics. Below, we provide several analyses including the volume, location, source gender, and a variety of content analyses to help monitor GBV, and both inform the design and assess the effectiveness of anti-GBV campaigns.

Traditional GBV monitoring methods face many challenges. Gathering statistics on GBV episodes is time consuming, collected under non-standardized protocols, and published in highly aggregated form. For example, the non-partner sexual violence prevalence data published by the World Health Organization⁶ is dated 2010 and reported by 21 aggregated geographic regions such as West Africa and South Asia. While the United Nations Office on Drugs and Crime data⁷ are less aggregated, and somewhat more recent with some data available as recently as 2012 and separated by country, the data are sparse and incomplete. For instance, sexual violence data are available for the Philippines and Nigeria for 2012, but the most recent data for India are from

¹ UNFPA agency's description about GBV: <http://www.unfpa.org/gender/violence.htm>

² <http://www.bbc.com/news/blogs-trending-31628729>

³ <http://www.telegraph.co.uk/news/worldnews/asia/india/11443462/Delhi-bus-rapist-blames-his-victim-in-prison-interview.html>

⁴ UNFPA agency: <http://www.unfpa.org/public/>

⁵ EU FRA agency: <http://fra.europa.eu/en/project/2012/fra-survey-gender-based-violence-against-women>

⁶ Data available here: <http://apps.who.int/gho/data/view.main.NPSVGBDREGION>

⁷ Data available here: <http://www.unodc.org/unodc/en/data-and-analysis/statistics/crime.html>

2010 and no data are presented for South Africa. The lag itself prevents monitoring change, either to detect unexpected increases in GBV, change in attitude, possibly due to recent events and awareness campaigns, or decreases resulting from mitigation efforts. Furthermore, these statistics reflect legal definitions, making direct comparisons among countries impossible due to differences in the definition and recording of offenses.

Table 1. Tweet samples from our analysis (see Section 3) with implications to inform and design targeted GBV intervention campaigns. M1-M3 illustrates existing automatic analysis capability, while M4-M8 result from partially automated analyses presenting the case for expanding computational methods.

Message	Implications
M1: RT @USER1: 1 in 3 women are raped/abused in their lifetime. RT if you rise to stop the violence. #1billionrising http://t.co/IXEEmQoLbO	<i>Volumetric analysis</i> of social media can help measure population engagement and effective penetration of designed campaigns in the community
M2: #StopRape Rape Crisis says many survivors of sexual abuse and assault still don't feel confident in the criminal justice system. CS	<i>Location analysis</i> to identify message origin (e.g., South Africa in M2) and route to appropriate agencies
M3: @USER2 Takes a MOMENT 2 Sign & ask Others 2,so DV & Rape Laws become Equal. TOGETHER we can Change History: http://t.co/yIWt5PgCI	<i>Gender detection analysis</i> of message authors (e.g., female in M3) to adjust apparent GBV attitudes for gender and suggest the content of anti-GBV policy and campaigns
M4: RT @USER3: Rape prevention nail polish sounds like a great idea but I'm not sure how you're going to get men to wear it	<i>Content analysis</i> sensitive to sociocultural considerations, such as the role of humor overall, and for acknowledging despair (e.g., in M4 from the Philippines)
M5: RT @USER4: 15 yard penalty for "unnecessary rape" http://t.co/yhzxtYzGP0	<i>Metaphor analysis</i> indicates the acceptance of GBV, echoed for example in sports, and suggesting opportunities for specific anti-GBV campaigns
M6: Valentine's day is really helping me sell these date rape drugs	<i>Entity recognition</i> based on knowledge-bases of GBV entities can help identify precipitating events (e.g., Valentine day in M6), apriori to design preventive campaigns
M7: @USER5 WB Govt must have ordered Police to protect the family of Rape-Victim. It is shameful for Mamata Banerjee O GOD GIVE WISDOM TO ALL	<i>Organization detection</i> , including government, law enforcement and commercial entities in relation to GBV can inform policy, e.g., M7 informs potential lack of police protection for a victim's family in West Bengal (WB)
M8: RT @USER6: It is not my job to coddle and "educate" young Black men when it comes to violence against women. Y'all wanna "teach"?	<i>Modeling of stereotypical association</i> can inform design of targeted campaigns, e.g., M8 author is stereotyping GBV violence with black men

Note. We anonymized user mentions as per the IRB guidelines.

Apart from the basic problems in the logistics of gathering GBV data, the data themselves are limited. Reliance on formal reports risks under-reporting by victims and witnesses, who may believe that domestic violence is a private matter (Nilan et al., 2014). In fact, such evaluative attitudes (Allport, 1935), left implicit in conventional data collection

74 methods, may provide a better metric of GBV risk and tolerance. Aggregation/generalization
75 across localities with different socio-economic properties masks important trends (Yoshihama,
76 Blazeovski, & Bybee, 2014) because sociocultural context including politics, history, religion, and
77 economy strongly influence attitudes (Nayak et al., 2003). Researchers suggest the need for
78 different prediction and mitigation models for different sociocultural contexts. For example, an
79 attitude of liberal sexual conduct combines with egalitarian values and economic development to
80 predict tolerance for GBV (Martinez & Khalil, 2012). Similarly, Corroon et al. (2014) found that
81 encouraging the Nigerian population to accept responsibility for their own health affects health
82 outcomes only in specific cities and urban regions.

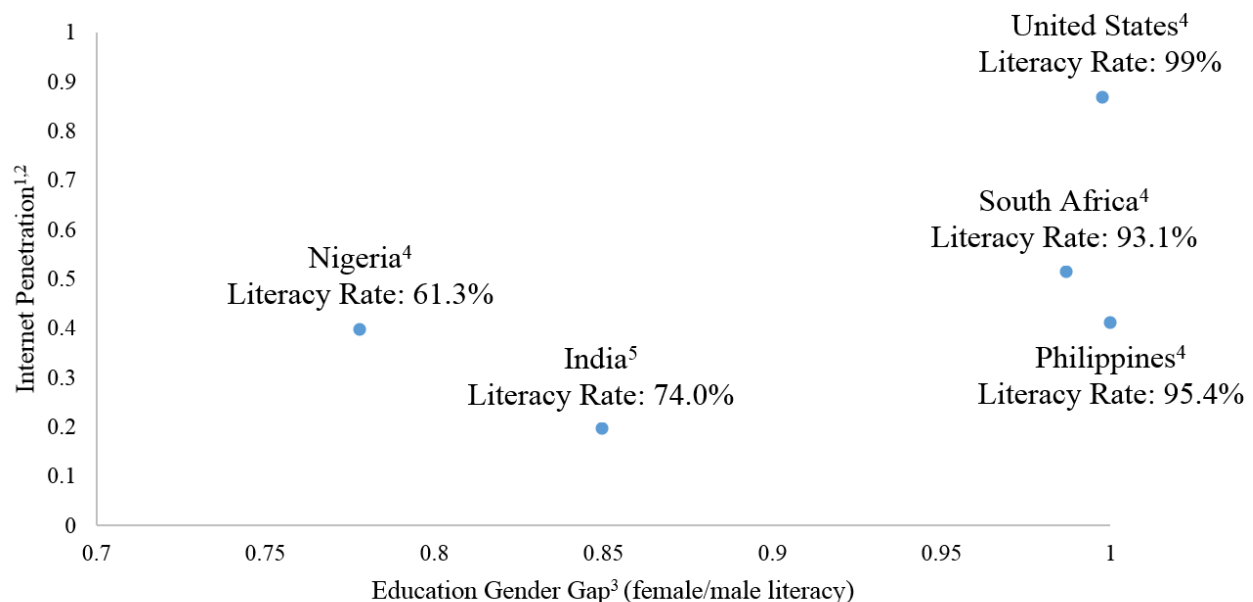
83 Social science survey methods complement the data gathering efforts of governmental
84 and non-governmental agencies. However, survey data also has several limitations. Surveys
85 reflect sampling biases and various confounds (Pierotti, 2013). For example, the rates of sexual
86 violence from a survey of patients being treated for sexually transmitted disease in South Africa
87 (Kalichman et al., 2005) bear uncertain relationship to the general population. Survey candidates
88 may refuse to participate, provide socially acceptable but inaccurate responses, or respond to
89 superficial yes/no questions using local mores and standards with high violence thresholds (Nilan
90 et al., 2014). For example, perpetrators may not consider beating to be an act of violence if the
91 victim failed to comply with behavioral norms. We highlight five additional limitations of survey
92 data of particular relevance to Computational Social Science methods. First, survey items tend to
93 address attitudes but not behavior, and therefore bear unclear relationship to the rate of GBV
94 episodes (Bhanot & Senn, 2007). In contrast, inasmuch as verbal abuse constitutes a form of
95 violence, social media posts can provide actual instances of behavior. Second, the methods fail to
96 account for transient global events, such as political or celebrity activity that can influence views
97 and responses, hindering comparison over time. Third, the items themselves presume an
98 established theory and standard measures of GBV, limiting the opportunity to discover latent
99 patterns that reflect attitude and behavior. For example, metaphor is a powerful reflection of
100 public opinion (Lakoff, 2008), but to our knowledge has not been explored in survey measures.
101 Fourth, survey methods constitute a highly labor intensive data collection method, necessitating
102 small samples while imposing cost and lag in data availability in a dynamic world. For example,
103 Pierotti's survey examined change over an average of five year intervals, ranging from three to
104 seven years (Pierotti, 2013). Finally, survey methods are by design static, discouraging
105 questionnaire modification to incorporate newly detected trends once the survey data collection
106 begins.

107 Social media provide a faster, cheaper and face-valid means to engage the public,
108 providing unprecedented large-scale access to public views and behavior (see Table 1). It
109 provides an ability to monitor attitudes in near real-time, to support timely mitigation efforts.
110 While use of social media give advantages with regards to speed (velocity), in some case
111 participation, broad sourcing and lower cost, studies cannot be tightly controlled with specific
112 statistical sampling, availability of demographic data may be limited or language use can skew
113 coverage—we discuss limitations later in section 4. Ultimately all three resources (formal reports,
114 surveys, and social media) require integration in order to assist policy design, prioritize attention

for interventions, and design region-specific programs to curb GBV. A logical first step is to understand what social media offer for GBV monitoring and the design of mitigation and policy.

Study Design Overview

Based on the suggestions of our UNFPA collaborators to identify a GBV related corpus, we selected three major themes that encompass gender violence concerns: physical violence, sexual violence, and harmful socio-cultural practices. Corresponding to these three themes, we created a seed set of keywords for data, crawling the Twitter social media microblogging platform. We also selected four countries with suspected elevations in GBV suggested by UNFPA experts: India, Nigeria, the Philippines, and South Africa, in addition to the U.S. given its Internet penetration. In this paper we assess the role of social media (data from Twitter) to identify public views related to GBV in these chosen parts of world, and tweeting practices by geography, time, gender, and events to inform concerned parties and assist GBV policy design.



¹<http://www.internetworldstats.com/stats>

²<http://www.digitalstrategyconsulting.com/india/>

³http://www3.weforum.org/docs/GGGR14/GGGR_CompleteReport_2014.pdf

⁴The World Factbook

⁵Registrar General and Census Commissioner of India

Figure 1. Summary of the education gap between genders, the penetration of Internet, and overall literacy rates in the diverse set of chosen countries.

The five countries present different contexts both for understanding social media data pertaining to GBV and for mitigation efforts. Figure 1 summarizes their variability across some key contextual dimensions: the education gap between genders, the penetration of Internet, and overall literacy rates. The figure illustrates the clustering of Nigeria and India for lower literacy rates, a greater education gap and lower Internet penetration. South Africa and the Philippines cluster with the U.S. regarding overall literacy and the reduced education gap, but reflect a

diverging range of Internet penetration. The graphs suggest the risk of sampling bias that affects data interpretation: an illiterate female citizen with no access to the Internet (likely in rural areas with unique GBV issues) may not be providing social media data, biasing the aggregated measures of attitude. Other differences may also be relevant. For example, India is among the top 20 of over 140 countries regarding female political empowerment while Nigeria is below average on this dimension. Additional influences on the use of social media not depicted in the figure include cultural influences on free speech. Sizable percentage of Nigerians for example may avoid public conversation about the Boko Haram atrocities⁸ due to fear of revenge.

In the next section, we employ quantitative and qualitative analyses to examine Twitter content related to GBV. Twitter supports the distribution of short messages called tweets that are a maximum 140 characters in length. The character limit influences message style and constrains communication practices. Therefore, tweets often contain URL links to web pages or blogs, sometimes relying on shortened URL versions from external services (e.g., <http://bit.ly/1C4HnMN>). A hashtag convention (e.g., #RapeJoke, #ChildMarriage) supports the identification of searchable user-defined topics. Other Twitter engagement features include retweet (or 'RT', a forwarding of someone's tweet). The device used to post a tweet may provide accessible, precise location indicators in some cases. Alternatively, accessible user profiles provide more general indicators of location and sometimes gender indicators such as author name.

A corpus of Big (Social) Data collected over a period of ten months included nearly fourteen million tweets. In this corpus, we examine volume, location, trends over time, gender participation, and content such as metaphor and humor. Our analyses present both challenge and opportunity to study phenomenon of gender-based violence. Challenges concern the need for computational methods to discern public perception and attitude from complex contextualized behavior. Opportunity lies in gaining fine-grained, region-specific insights concerning the prevailing GBV attitudes, and related policies along with potential approaches to mitigation.

[2.] METHOD

We first describe our data collection for the study, followed by a description of the analysis approach.

Data Collection

Based on domain expert guidance we collected data from the Twitter Streaming API (Twitter Developers, 2014), using its 'filter/track' method for the given set of keywords. We leveraged the keyword-set crawlers (data collectors) of our Twitris platform for mining collective social intelligence as described in (Nagarajan et al. 2009, Sheth et al., 2014). Twitris also provides real-time and scalable analyses of social data streams for specific topics of interest by end users (discussed in future work). UNFPA experts assisted in the definitions corresponding to the three themes of interest for GBV study and associated English terminology. Using the

⁸ For an overview, see the following: http://en.wikipedia.org/wiki/Boko_Haram

themes of physical violence, sexual violence, and harmful practices, we created a set of seed keywords and key multi-word phrases for data crawling (see Table 2).

Table 2. Seed set of key phrases for GBV related data crawling.

<i>Physical violence</i>	woman dragged, women dragged, girl dragged, female dragged, woman kicked, women kicked, girl kicked, female kicked, woman beat up, women beat up, girl beat up, female beat up, woman beaten, women beaten, girl beaten, female beaten, woman burn, women burn, girl burn, female burn, woman acid attack, women acid attack, girl acid attack, female acid attack, woman violence, women violence, girl violence, female violence
<i>Sexual violence</i>	sexual assault, sexual violence, rape, woman harass, women harass, girl harass, female harass, woman attacked, women attacked, girl attacked, female attacked, boyfriend assault, boy-friend assault, stalking woman, stalking women, stalking girl, stalking female, groping woman, groping women, groping girl, groping female
<i>Harmful Practices</i>	child marriage, children marriage, underage marriage, forced marriage, sex trafficking, woman trafficking, women trafficking, girl trafficking, female trafficking, child trafficking, children trafficking

For each single keyword K, the Twitter service provided messages containing any form of the keyword #K, k, K. For the multi-word phrase K, it provided the messages containing all the terms of K. Each message is associated with metadata, containing various tweet related and author profile characteristics, such as tweet origin location (latitude and longitude), time of posting, author profile description and location, number of author followers (users that subscribed to the author's updates), and followees (users to whose updates the author subscribed). Location metadata was crucial to geographical analysis. We first checked if tweet origin latitude-longitude were available from the device used to send the tweet, else we resolved the author profile location, if available, using Google Maps API. We used a bounding box of latitude-longitude for a country of interest to identify a country-specific tweet dataset. Using Genderize API (<http://genderize.io>), we collected genders of the authors of tweets. We first fetched the real names of the twitter authors using metadata of their Twitter handles. We extracted first names to detect author genders via calls to the Genderize API with first names as parameters. We do note limitations with the gender detection due to informal social media language use, and therefore discuss scope for improvement in our future work section.

Analysis Approach

With UNFPA guidance, we analyzed the data corpus of 13.8 million tweets from the ten months in non-uniform time slices, starting with a smaller pilot phase and adding two additional extended data collection periods due to supporting results from the pilot phase data.

- SLICE 1: Jan 1 2014 - Feb 15 2014: 1.5 months; Phase-1 (pilot phase)
- SLICE 2: Feb 15 2014 - June 31 2014: 4.5 months; Phase-2
- SLICE 3: July 1 2014 - Oct 31 2014: 4 months; Phase-3

To study the diverse set of data from the three phases, we employed mixed methods (Creswell, 2013) to reveal both patterns across the corpus as well as the content of specific

contributions. The focus of *quantitative* analysis is to provide data-driven insights of activity patterns in the social media community, by examining large-scale distributions of GBV content by geography, time, and gender. The recovery of meaning of a pattern requires more fine-grained analysis than mere statistical distribution. Therefore, the focus of our *qualitative* analysis is to reveal attitudes and behaviors across different countries and between genders. In both cases, our interpretation relies on context, regarding current events and socio-cultural considerations, ultimately supporting the need for context-sensitive computation for monitoring GBV content in social media.

[3.] RESULTS

Quantitative Analysis

We discuss four types of quantitative analyses in this section: volume, theme, sharing, and gender-based.

Volume analysis. To begin our study, we sampled an initial slice of data for 1.5 months from all over the world, which contained 2.3 million tweets worldwide related to GBV. We skip the detailed descriptive statistics report for brevity regarding this sample and instead summarize our observations. Regarding our interest in location-specific GBV data, we note relatively few tweets with device-based location information (1.91%). However, author profiles provide location related information as well, including which over half of the data (54.74%) for 1.3 million users had location information. These results motivated our more extensive data collection effort, spanning an additional 8.5 months.

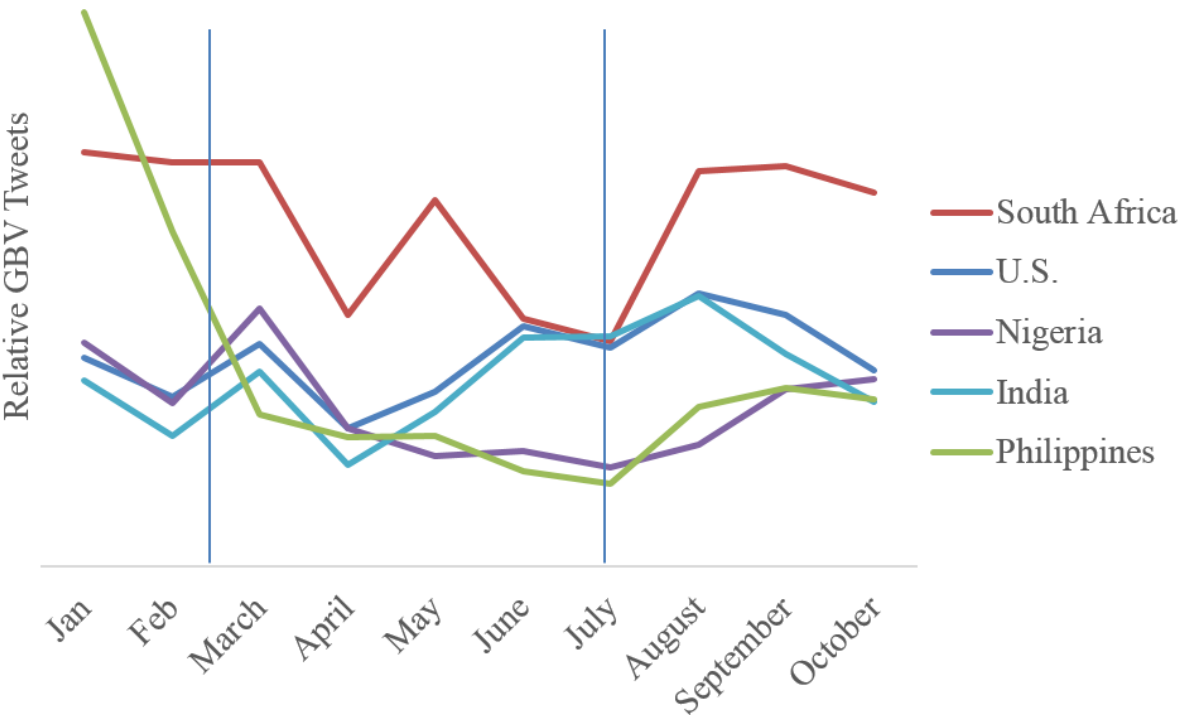
Table 3 provides the composition of the full data set by country. More than 10% of the social media traffic belongs to the five countries examined here. We note more than five times the traffic in the U.S. and India relative to the other countries. Moreover, the observed frequency ranking differs from the population demographic⁹ information for these countries (in descending order: India, the U.S., Nigeria, the Philippines, South Africa). We suspect that Internet penetration is influencing data collection. Worse, a frequency graph for all of the raw data over time would mask variability in the Philippines, Nigeria, and South Africa with smaller populations and variable Internet penetrability (We plan to address urban versus rural comparison in our future work.) Figure 2 therefore scales the raw data by two factors: population and Internet penetration. This allows us to consider the relative prevalence of GBV topics between countries and over time. Below we discuss some of the emergent patterns by country. Although we highlight the need to interpret these patterns with respect to a broad and complex knowledge base that includes current events, these are precisely the sorts of analyses that are best accomplished with computational tools.

⁹ http://en.wikipedia.org/wiki/List_of_countries_and_dependencies_by_population

246 Table 3. Volume of GBV related tweets by country.

Country of Origin	Number of Tweets
United States	698077
India	549033
Philippines	105550
Nigeria	134671
South Africa	104430

247 Note. This table represents 1,591,761 (11%) of the total 13,942,592 GBV tweets collected globally.



248 Figure 2. Tweet volume across countries from 1 January to 31 October 2014 related to GBV, and
249 controlling for the population and Internet penetration of each country. Vertical lines represent time
250 slices sampled. [Population expressed as a percentage of the smallest candidate (Nigeria). Nigeria = 1,
251 India ~ 25. The adjusted metric then is expressed as original tweet numbers divided by adjusted
252 population divided by Internet penetration.]

254 The trend in Fig. 2 shows that when controlling for population and internet penetration,
255 South Africa generates the most GBV related tweets overall, with the U.S. and India lagging
256 behind roughly the same amount, and Nigeria and the Philippines toward the bottom.

257 Several current events may explain the apparent peaks over time, revealed during a
258 parallel Google Trends search by country. In the Philippines, the ongoing saga of Vhong Navarro

dominated much of the discussion¹⁰. Navarro, a television personality, was assaulted in his home in the January of 2014 apparently in retaliation for attempted rape. The incident stirred concerns over rape culture, as many lashed out at the female accuser. However, the American movie star Jennifer Lawrence also factored highly in the most searched items in the Philippines regarding the hacking of her private email accounts.

Between March and July a vehemently anti-LGBT figure named Myles Munroe ranked among the most-searched topics in Nigeria¹¹. The largest Nigerian story from the liberal perspective (the kidnap of the Chibok schoolgirls in mid-April) factored only ninth in Nigeria. “Kidnapping” searches initially increased by about four fold in the month and a half following this incident, but then dropped to previous levels. Given the responsible group’s influence and tactics, we suspect citizens may have been reticent to discuss any wrongdoing openly via social media. We make this point to illustrate the need to consider socio-cultural influences on social media traffic. This suggests that we will require separate models of GBV in social media, by country.

South Africa’s Google searches reflected considerable violence¹². Among the top events driving traffic was the murder of a prominent soccer player in October during a burglary at his home. Oscar Pistorius also influenced the search patterns. The trial for the murder of his girlfriend began on March 3 and temporarily adjourned on May 20. The trial resumed on June 30 through August 8, with a judgment on September 12 and sentencing on October 21. At least two of the peaks in GBV related Twitter traffic are coincident with these events.

These observations suggest that any quantitative analysis of social media traffic must control for current events, in order to separate fundamental trends from local variability. Such control is well within the capability of computational analysis using topic-modeling approaches.

We note apparent correlations between certain countries: India and the U.S., as well as the Philippines and Nigeria at least for the bulk of the data. We observed a 48.1% overlap in the popular topics of the U.S. within the set for India for the month of August. Topics to determine overlap were based on the top 500 key-phrases (set of words in a message text) extracted using modified *tf-idf* based key-phrase extraction algorithm in Twitris (Nagarajan et al. 2009). Such content overlaps suggest the presence of mediating influences. To provide a source of hypotheses, we turn to demographic properties of the countries investigated in this study. We note a large Indian diaspora living in the U.S., potentially reacting to events originating in India. The story of a 15 year old Indian female acid attack victim received considerable attention following an Al Jazeera report in January (Dixit, 2014). The shared August peak coincides with an Independence Day speech by the Prime Minister of India (Narendra Modi) in which he urged suspension of the practice of questioning the families of girls regarding their social habits but never boys. Manual inspection of the U.S. corpus supports an explanation of the correlation between the U.S. and India trends based on the Indian diaspora. Tweets with apparent origins in

¹⁰ <http://www.google.com/trends/topcharts#geo=PH&date=2014>

¹¹ <http://www.google.com/trends/topcharts#geo=NG&date=2014>

¹² <http://www.google.com/trends/topcharts#geo=ZA&date=2014>

the U.S. may actually constitute amplification of tweets with origins in India, as the following example illustrates.

India: *What world calls shining #India is the worst place for women in terms of #Rape*
<http://t.co/F2oyKps60R> #BanBollywood #MediaMafia

US: *What world calls shining #India is the worst place for women in terms of #Rape*
<http://t.co/zZoB17p1L1> #BanBollywood #PakMediaHijacked

The U.S. and Indian patterns illustrate an important challenge in the interpretation of social media data—the trends do not necessarily reflect local events. Computational tools for monitoring GBV will need to conduct location analysis from the text to distinguish commentary about other countries. Model building must accommodate this distinction. However, given the tweet pedigree, it is not clear to whom we should attribute the attitude.

Table 4. Percentages of sexual violence messages discovered by time slice and by country.

Country	SLICE-1	SLICE-2	SLICE-3	ALL
United States	65.10%	64.78%	63.90%	64.44%
India	66.05%	65.41%	65.30%	65.45%
Philippines	77.68%	69.08%	69.09%	72.00%
Nigeria	55.80%	47.13%	49.08%	49.59%
South Africa	71.15%	66.96%	63.69%	66.24%
Total	65.67%	62.52%	61.11%	62.43%

Theme analysis. Our original data collection categorized the tweets into three thematic groups: physical violence, sexual violence, and harmful practices. Sexual violence dominates the sample of GBV tweets (72% overall in global corpus), so we provide their distribution over time in Table 4. We note a general decrease over time with some variability between countries. We suspect early spikes related to incidents like Vhong Navarro rape accusation in Philippines and the initial publicity surrounding the Boko Haram atrocities in Nigeria.

Sharing behavior analysis. Social media provides the opportunity to distribute information, potentially reflecting both the senders' judgment of information importance, and reliance on the voice of others. Sharing functions to amplify these voices, often the influential celebrities. We analyze two types of sharing behavior in the social media community surrounding GBV events: direct content resharing as *retweet (RT)*, and indirect sharing via references to the external resources using *URL*, such as news, blogs, articles, and multimedia.

324 *Table 5. Proportional RT and URL tweets by country.*

Country	RT	URL
United States	45.47%	42.25%
India	46.68%	42.63%
Philippines	40.74%	18.79%
Nigeria	26.28%	62.13%
South Africa	44.25%	31.49%

325
326 We calculated the percentage of GBV retweets relative to the total count of tweets for
327 each data sample as shown in Table 5. More than 40% of the GBV corpus is a retweet in the US,
328 India, the Philippines, and South Africa, amplifying information that senders consider to be
329 important. For comparison, Liu, Kliman-Silver, and Mislove (2014) found that retweets
330 generally constitute just over 25% of the total volume of tweets. Although we note variability in
331 retweet behavior between countries, the low retweeting frequency in Nigeria is particularly
332 remarkable (see Table 5). One might hypothesize that a low literacy country such as Nigeria, in
333 which senders are less able to compose messages, would have the highest retweet ratio. The
334 adjacent analysis of URLs suggests a different socio-cultural phenomenon at work, concerning
335 the identifiability of the responsible party. For GBV tweets containing URLs, Nigeria has the
336 highest percentage of tweets with URLs in comparison to other countries. Numerous
337 explanations should be tested, including literacy, credibility of the public press, and the
338 possibility that reliance on external resources somehow reduces the threat of being identified as
339 the responsible party.

340 **Author gender analysis.** We obtained gender identification for 37% of the overall users
341 (see Table 6). The reduced percentage is due to names missing in the Genderize API lexicon, as
342 well as unconstrained natural language features of social media content, such as use of special
343 characters in names, for instance, '@@shish' instead of 'Aashish', which is a male Indian name.

344 *Table 6. Gender-wise distribution of data for the overall global corpus.*

Statistic	Users	% of Total Users	Generated Tweets	% of Total Tweets
Total	3,036,576	100%	13,942,592	100%
- Gender Filtered*	1,148,329	37.82%	2,771,686	19.88%
-- Female author	563,016	18.54%	1,324,292	9.50%
-- Male author	585,313	19.28%	1,447,394	10.38%

345 *Filtered = where an author gender could be determined

346
347 Keeping in mind that statistically significant differences are certain with a large sample
348 dataset, the distribution of gender appears approximately equal in Table 6. We also note a tweet

frequency average for a female author as 2.352 tweets per author, while 2.472 tweets per author for a male. However, Figure 3 separates the gender distribution for the examined countries, and depicts the impression of gender inequality.

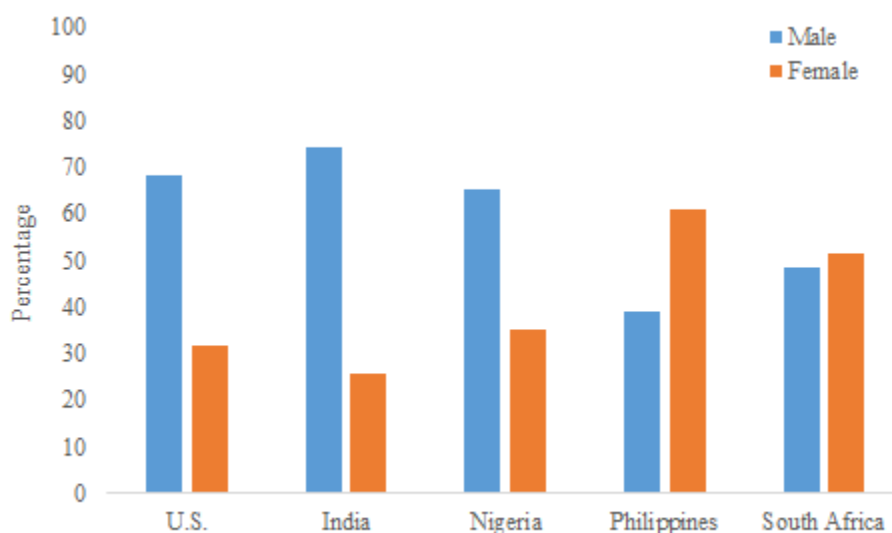


Figure 3. Relative distribution of tweets whose author's gender we could determine.

The name classification procedure that we employed suggests a gender bias in the samples across countries. Literacy serves as a partial explanation for the observed ratios, except for the U.S. Apart from the explanation, opinions collected in the U.S., India, and Nigeria reflect a male bias, while opinions collected from the Philippines and South Africa are more balanced or even reflecting slight female bias. It has corresponding implications for the assessment of GBV attitudes, the general reach of anti-GBV campaigns using social media and the ability to target potential perpetrators and activists for engagement separately.

Qualitative Content Analysis

Thus far we have described the corpus with respect to dimensions that we anticipated at data collection: country, theme of GBV event, sharing methods, and gender. We have made suggestions about the surrounding context, including transient events that might explain the observed patterns. In the remaining analyses, we look more closely at message content for indications of GBV attitudes and to clarify the requirements for future computational analysis capabilities.

Language indicators. Using Linguistic Inquiry Word Count (LIWC) software¹³ (Tausczik & Pennebaker, 2010), we analyzed language of the content of all tweets generated by both genders. We used the predefined LIWC dictionaries that tally word frequencies in categories such as anger, sexuality, sadness, health, etc. Content corresponding to these

¹³ Information available here: <http://www.liwc.net/>

particular categories, and in fact content across the majority of the LIWC categories appeared more frequently in tweets of male origin relative to tweets of female origin. However, we did note some trends in tweet content from females. Consistent with the research on gender issues in communication (Tannen, 1996), female authors here are more collective and socially oriented. Their tweets call for action and are more likely to express agreement.

Female (South Africa): *@USER7 Absolutely. If we follow each other I can DM you my email address. I applaud your speaking out on the rape epidemic in SA.*

Female (Nigeria): *I am worried abt our approach to d fight against rape. Permit me to vent here. @USER8 @USER9 @USER10 @USER11 #CurbingRape*

Female authors are also more likely to provide opinion on causality, as in the following example tweet from India:

Female (India): *The major factor behind #Rape in #India is the #Bollywood which incites feeling 2 cross the moral limits #BanBollywood #GreaterPakistan*

Gender specific analysis can be leveraged to design and promote anti-GBV campaigns. For example, tweets of female origin in India, although not guaranteed to be benevolent, could be amplified to extend their reach.

Following a manual review of a random sample of the corpus, we used computational analyses to describe the prevalence of attitude indicators across the whole corpus. Two of the attitude indicators examined here are the presence of humor indicators and GBV metaphors in sports. We also provide specific examples of tweet content from manual analysis of random samples.

Humor indicators. Humor might indicate a trivialization of an issue or an expression of underlying helplessness (Romero, 2010; Ancheta, 2007). We assessed the prevalence of humor references with related permutations of “haha” and “hehe”. The number of humor-flagged tweets by country appears in Table 7. The Philippines sample provides a greater proportion of humor indicators by far. This is consistent with Filipino culture¹⁴ in general. Moreover, the observed female bias in the Filipino sample reinforces the interpretation of humor as an expression of helplessness. Of course, laughter annotations do not provide a complete account of humorous references but rather a correlated indicator for jokes such as the following from the Filipino corpus:

Female (Philippines): *RT @USER12: Rape prevention nail polish sounds like a great idea but I'm not sure how you're going to get men to wear it*

Due to cultural differences in the role of humor, we cannot advocate comparison between countries. However, we do suggest that the prevalence of GBV humor may provide a useful metric for changes in attitude over time within a region.

¹⁴ A brief overview of traditional Filipino values can be found here: http://en.wikipilipinas.org/index.php?title=Philippine_Core_Values

415 Table 7. Analysis of humor-flagged tweets indicated by permutations of humor indicators.

Country	Flagged tweets	Total tweets	Percentage
Philippines	10418	105550	9.870%
South Africa	1801	104430	1.725%
USA	8067	698077	1.156%
Nigeria	1419	134671	1.054%
India	5195	549033	0.946%

416 Note. Lexicon set defined as {haha, ha ha, hehe, he he, he he he, hahaha, lol, lmao}.

417
418 **GBV and sports metaphors.** Sports involve competition, and violent metaphor is a
419 common device in sport discussion (Lewandowski, 2010, 2011, 2012). GBV metaphors may
420 appear in a sports corpus as an indication of dominance, such as the following example from the
421 South African tweet corpus:

422 South Africa: *The German Team is on a Steroid-induced anal-rape rage*
423 *against Brazil right now.*
424

425 Comparing the prevalence of GBV metaphors in sports across countries is difficult
426 because a given sport does not capture national attention equally across countries. We obtain
427 some insight regarding the GBV metaphor in tweets related to sports from exchanges concerning
428 the FIFA World Cup contest, held from 12 June through 13 July 2014. Soccer is the most
429 popular sport in Nigeria and South Africa, one of the top three favorite sports in India, and a top
430 ten favorite in the Philippines and the U.S.¹⁵. While all of the countries we examined have
431 eligible teams, only the U.S. (ranked 13) and Nigeria (ranked 44) participated in the final
432 tournament¹⁶.

433 We should not expect any FIFA related content in a corpus designed to capture GBV
434 issues. Yet, Table 8 illustrates the existence of tweets flagged for containing references to:
435 *football, futbol, soccer, worldcup, world cup, fifa, fifacup, soccercup*, as well as *pairs of team*
436 *names* participating in the tournament. Although the percentages are small, every country
437 examined provided such instances, in a ranking consistent with soccer popularity, with the
438 exception of the U.S. The finding requires a nuanced interpretation. On the one hand, rape as a
439 metaphor suggests a trivialization of the primary definition. On the other hand, at least one
440 dictionary¹⁷ indicates an archaic definition meaning plunder or violation, for example regarding
441 the environment. On the one hand violent metaphor is a common device in sport discussion. On
442 the other hand, it is just one of many common metaphors (Lewandowski, 2012). In fact, in
443 Lewandowski's (2010-2011) extensive analysis of violent, conflict oriented metaphor in football
444 in sports journalism, none of the 551 violent metaphors invoked rape.

¹⁵ <http://mostpopularsports.net/football-soccer-popularity>

¹⁶ http://en.wikipedia.org/wiki/2014_FIFA_World_Cup

¹⁷ <http://dictionary.reference.com/browse/rape>

A difference in the popularity of sport between countries is only one challenge to direct comparison of GBV metaphor between countries. The development of a knowledge base to support analysis poses further challenge, as different regions follow different teams with different players. As in the interpretation of humor, we believe the best use of sport metaphor data is to provide indicators of attitude trends over time by region, while adjusting for seasonal variations in sporting events or averaged annually.

Table 8. Analysis of sports-related tweets indicated by sports lexicon.

Country	Flagged tweets	Total tweets	Percentage
U.S.	1053	698077	0.0015%
India	657	549033	0.0011%
Nigeria	320	134671	0.0023%
Philippines	47	105550	0.0004%
South Africa	214	104430	0.0020%

Note: FIFA World Cup sample for each country in the period of June 12 to July 13 2014.

Lexicon: { *football, futbol, soccer, worldcup, world cup, fifa, fifacup, soccercup, [country-names pairs from http://en.wikipedia.org/wiki/2014_FIFA_World_Cup]* }

Policy-making and intervention insights via manual analysis. Our goal is to generate attitude metrics and inform mitigation campaigns automatically. All of the above examples were informed by computational analysis. For the following examples, we manually examined random subsets drawn from each of the three slices of our data corpus, as 200, 500, and 500 tweets. We end our presentation of GBV data by indicating the kind of content available in the corpus compatible with somewhat more specialized computational analysis. Developing the necessary content specific filtering methods (feasible using domain knowledge of politics, business, and behavioral theories) is particularly critical for appropriately routing content to specific policy making recommendation agencies. Some examples are following:

a.) Behavior pertaining to government/public officials/leaders

- @USER13 NCP leader doesn't know rape happens due to pervert mindset . what abt unreported rapes happen with minors & toddlers.

b.) Commercial considerations

- RT @USER14: .@USER15 Please don't allow violent hatemongers use your app to harass and exploit marginalized women. \"<http://t.co/DoRkbXFuN...>
- Spring Breakers isn't just a terrible movie, it reinforces rape culture <http://t.co/pxTAZftM8v>
- RJ Police already said that it's not a Rape case!But Media neglected it 4 creating SPICY news!#KnowTheTruth & Rise4Justice

c.) Persuasive/encouraging message content

- RT @USER16: Pregnancy, periods, breast cancer, being walked on, rape, harassment, abuse; females go through a lot. WOMEN ARE STRONG.
- @USER17 @USER18 Is it acceptable to use gang rape in an example?

d.) Stereotype association

- *RT @USER6: It is not my job to coddle and "educate" young Black men when it comes to violence against women. Y'all wanna "teach"?*

Automated filtering of GBV content according to such dimensions is challenging but feasible given the appropriate knowledge bases.

[4.] DISCUSSION AND FUTURE WORK

Here we revisit the limitations of conventional, survey-based methods for monitoring GBV attitudes to discuss the progress we have made, the limitations we face, and the promise of further advances in Computational Social Science.

Progress

We noted above the limitations of conventional methods in gathering GBV data for specific regions. Data are either highly aggregated, at the country level or missing. We have provided substantial amounts of GBV data for all five of our target regions. User profiles and GPS tags provide the capability of identifying content specific to much smaller units of analysis, e.g., type of socio-economic region or even city (Leetaru et al., 2014). This capability supports targeted GBV campaigns, where the prevalence and types of violations may vary. Below we note progress with respect to a number of additional concerns regarding conventional data gathering methods.

Reducing sampling bias. We noted the presence of bias in survey methods due to artificial sample selection. We have overcome some of the bias concern simply by the scope of data collection that is feasible with social media. We have the opportunity to observe male and female attitudes by specific regions, over time. However, bias does remain in our data collection methods. Some of the persisting bias is amenable to adjustment. For example, bias exists in the form of literacy assumptions by country and by gender within country and Internet penetrability. Awareness of these sampling issues allows us to amplify the weight of content from underrepresented participants, e.g., the contributions from females in lower penetration areas. This is possible because we can assess both gender and location in the available data. Of course, we cannot amplify content that does not exist. This limitation is fundamental, but it also identifies the regions where this limitation occurs and where we should focus other methods of data collection.

Reducing content bias. We suggested that survey participation itself constitutes a form of bias. Apart from the sampling issues noted above, the posted content on social media itself reflects a form of bias. Participants are still providing what they consider to be socially acceptable content. This is particularly apparent in the absence of commentary regarding Boko Haram and GBV in Nigeria. However, this observation alone can inform the course of action for policy makers, e.g., publicizing sensitive topics. Nevertheless, for the social media data that are available, participants are not providing guarded responses to external queries about their attitudes. Instead they express their attitudes directly, making them susceptible to analysis.

Analysis speed. We have over 1.5 million social media postings collected in 2014 from just five countries to support the claim that GBV issues generate commentary. Moreover, we have completely eliminated the need for labor-intensive data collection, and as a result have overcome the cost and lag limitations in data collection. We produced analysis within a year in contrast to survey methods, despite the absence of complete automated capability. This rapid turnaround enables the development of dynamic metrics to assess the results of campaigns designed to curb GBV. A rapid measurement capability can play a role in the promotion of effective efforts and the abandonment of those that are less effective, with real consequence to the alleviation of human suffering.

GBV attitude metrics. Survey items tend to address attitudes but not behavior. Social media data provide both attitudes and behavior, inasmuch as jokes and metaphor are both behavior and attitude. This provides us with potential measures of tolerance for GBV. Thus, the editing of socially acceptable content that constitutes a form of bias in data collection is the very same behavior that tells us what is considered acceptable. This provides a means to measure the effectiveness of anti-GBV campaigns, both those directly targeted to the potentially offensive jokes and metaphor, as well as those targeted to specific but apparently unrelated concerns such as the effect of holidays or weather.

Survey methods are by design static. Standard measures purport to provide evidence that is comparable across time and regions. However, standardization ignores the effect of context. Social media trends over time tell us that context cannot be ignored. For example the publicity surrounding a celebrity involvement in a GBV related event spikes social media commentary. The availability of such events (Tversky & Kahneman, 1973) may very well influence survey responses. But standalone surveys have no way to account for this influence. Our computational social science methods allow us to complement the interpretation of GBV commentary with adjustments for the influence of events at a short-time scale in order to discern long-term trends.

Finally, and completely outside the typical survey content, the analysis of social media also promises to assist GBV campaigns, both by targeting the views of specific groups and by providing content recommendations regarding law enforcement, politics, health services and commerce.

Limitations & Challenges

All data collection methods suffer from limitations, and ours is no exception. Here we note several concerns, some of which are amenable to computational solution and some of which are more fundamental.

Unconstrained natural language text. Our keyword-based crawling limits the completeness of the resulting corpus. Given the natural language of social media messages, we cannot guarantee collection of every single relevant message. Moreover, keyword selection matters. Countries vary the terminology they employ for different purposes. For example the word “rape” in tweets that originate in India generally refer to events in India, but “sexual assault” in the Indian corpus returns mostly American incidents. We are further constrained at present by restriction to the English language, missing the interpretation of GBV language

560 embedded in Tagalog for example (spoken widely across the Philippines). Furthermore, our
561 dependence on keywords glosses over the different definitions of rape across cultures.

562 **Global event sensitivity.** Both a feature and a limitation, we demonstrated the influence
563 of events on social media content, both within a region and between regions. World events
564 provoke the articulation of public opinion and provide an unprecedented large-scale opportunity
565 to gather opinion and attitude. At the same time world events create variations in magnitude that
566 require adjustments to frequency counts, in order to reflect enduring trends over time.

567 **Inter-regional comparison.** We commented earlier that differences in the legal
568 definition of GBV hindered comparison of GBV rates between countries. While we believe our
569 measurements within a region can be informative about attitude change over time, the ability to
570 compare GBV issues between regions still poses substantial challenge. Our two proposed
571 measures, sports metaphors and jokes illustrate the challenge. A given sport is not equally
572 important across countries (or even regions), so that the prevalence of more frequent violent
573 sport metaphor in one country relative to another may simply reflect population interest in the
574 sport. Allowing type of sport to vary between regions confounds sport with region, so that we
575 might be learning more about issues with the sport than the region. The meaning of jokes
576 notwithstanding (trivialization versus helplessness), we cannot compare the prevalence of GBV
577 jokes between countries without factoring in the prevalence of jokes between countries in
578 general. Thus, we have not escaped the local, sociocultural influences on measures that prevent
579 between region comparisons. Our contribution to this issue is to establish that it is not a
580 limitation of specific measures such as police reports and surveys. All measures reflect these
581 sociocultural influences.

582 **Location identification.** Location references of the messages are not always consistent
583 with the location of the author profile or GPS coordinates of source device. U.S. events appear in
584 the opinions of people in other countries, and vice versa. The technical aspect of this problem
585 resolves with techniques that discern event location from the message text or ancillary URL
586 content. This is not necessarily without remaining uncertainty, as people often assume shared
587 context location identification, e.g., with abbreviated names for familiar landmarks resulting in
588 referential ambiguity. However a far greater concern lies in the attribution of attitude when the
589 content corresponds to a remote event.

590 **Correlational logic.** Our argument has a correlational component. For example, we can
591 identify indicators of humor such as “hehehe” and “hahaha” using computational methods.
592 However, automatically detecting every instance of a joke embedded in social media content is
593 not computationally feasible. We assume that the frequency of humor keywords in a GBV
594 context correlates with the frequency of GBV jokes without such keywords. Although this
595 assumption is worth confirming using a manual classification of jokes, we note that the presence
596 of laughter indicators alone in this context is potentially offensive.

597 **Future Work**

598 We have demonstrated the potential of social media to inform policy makers regarding
599 attitudes, measuring the effect of campaigns, and even providing campaign concepts.
600

Computational Social Science holds promise for a range of economic development issues beyond GBV, covering a different set of themes, countries, and languages. However, much technical, theoretical, and practical work remains to realize the full potential of the medium for the GBV application.

Technical advances are required in order to distinguish message origins from message content and support more comprehensive gender detection. Substantial work remains in the development of specific knowledge bases to guide the detection and interpretation of jokes, as well as metaphor and commentary directed to particular entities such as government and business. The knowledge bases must be dynamic to capture the transient events that mask fundamental attitude. Theoretical guidance from social science is required to attribute humor and metaphor to either despair or tolerance or the lack thereof. The interpretation of retweets across regions raises the problem of attribution in more than a practical sense. The issue is whether and how to weigh the endorsement: as a property of the original sender, or the endorser who is amplifying the message. Does the endorsement reflect opinion of the endorser's location independent of birth origins? The resulting approaches require validation to address these issues.

Comparison of GBV threats between countries, though raising measurement issues, is not an exclusively methodological problem. Instead, it is a sociocultural issue, concerning the boundaries that violate globally established norms and determine the deployment of global resources. We lack the policy expertise to weigh in on such matters. However, we do endorse the development of adaptive regional, and even local models for GBV behaviors and attitudes. Although a large model encompassing all countries holds a certain appeal, it is fallacious to assume that all regions have the same underlying issues and beliefs. The number of countries constrains the number of contextual variables that one can test statistically. Local socio-economic conditions, literacy, Internet penetrability, gender, crime rates, religious beliefs, liberty and many other unexamined, but constant factors influenced the data and their interpretation.

Data accessibility for policy-makers. Near-real-time information benefits an urgent need to reduce GBV and the associated suffering. Conventional methods of disseminating public opinion, are reflected in written reports covering multiple years, and issued with considerable delay following data collection. Governments, NGOs, International Organizations, Aid Workers, etc. require far higher bandwidth access to the dynamic data and analysis in order to measure current public opinion and the effect of anti-GBV campaigns.

The Twitris collective social intelligence platform (Sheth et al. 2014) provides a foundation for delivering the necessary high bandwidth access. Twitris supports the presentation of thematic data along spatial and temporal dimensions (Nagarajan et al. 2009), network relationships among people related to the distribution of content (Purohit et al. 2012), and sentiment (Chen et al. 2012) as well as real-time trends¹⁸ shown in Figure 4.

¹⁸ A currently monitored GBV campaign is available at: <http://twitris2.knoesis.org/app/#campaigns/Gender%20Based%20Violence/dashboard> Register for full dashboard view.

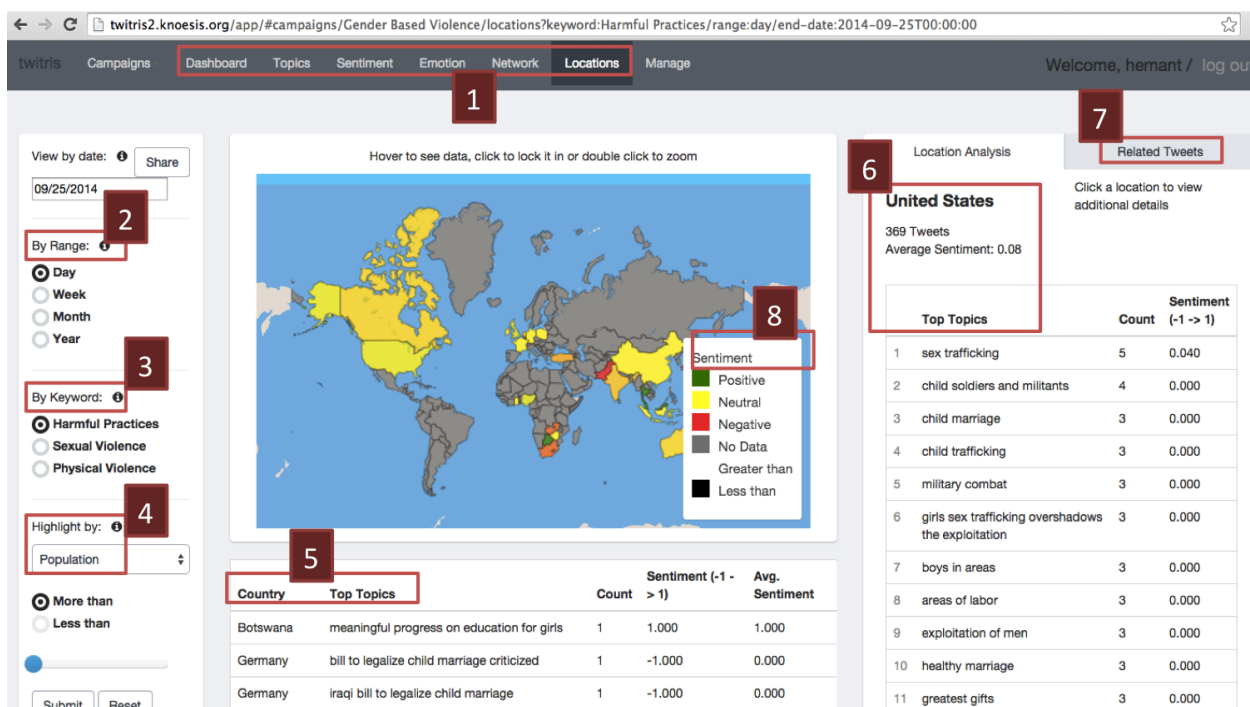


Figure 4. A snapshot of the Twitris social analytics platform showing on-going research to monitor and analyze the GBV data by a variety of dimensions- location (1, 8), time (2), types of content theme such as sexual violence (3), demographics based map visualization updates (4), top topics by country or region of interest (5), country specific demographic and top topics and related tweets (6, 7), sentiment based heat map (8), as well as fine-grained analysis of emotion expressed, who-talks-to-whom network for influential users about GBV topics, and real-time trends dashboard (1).

Twitris analyzes a topic such as GBV, and provides real-time scalable analyses of social data streams for greater insights and actionable information to improve intervention. For example, Twitris can monitor the real-time public sentiment and emotional reaction to criminal reports and the justice system response by region, permitting side-by-side comparison with other events. Under user guidance, Twitris automatically identifies the key topics meriting further attention. Twitris network analysis assists in the measurement of anti-GBV campaign diffusion to measure campaign effectiveness, as well as the identification of users who will spread targeted campaigns.

CONCLUSION

Big (Social) Data complement more controlled but slower survey-based data collection and analysis methods, whose conclusions may become obsolete in a dynamic world that continuously generates noisy data responding to transient events. The lag in surveys and noisy nature of data limits the use of conventional methods for measuring the effects of changing GBV attitudes and anti-GBV campaigns. Computational Social Science supports the collection and analysis of a large GBV corpus. We analyzed nearly fourteen million Twitter messages collected over a ten-month period. The large sample reduced bias relative to conventional data collection. We examined data by region and gender, identifying content such as humor and metaphor that

have implications for both the measurement of GBV attitudes as well as specific targets for anti-GBV campaigns. Our methods constitute an inexpensive way to engage with citizens at unprecedented scale, including the collection of public views regarding the behavior of government and business to revolutionize the conduct and measurement of anti-GBV campaigns.

ACKNOWLEDGEMENT

We acknowledge the partial support of the U.S. National Science Foundation Social-Computational Systems (SoCS) program for grant IIS-1111182 ‘Social Media Enhanced Organizational Sensemaking in Emergency Response’ for our study. We are thankful to our colleagues at the United Nations Population Fund NYC—especially Upala Devi, Judy Ilag, and Maria Dolores Martin Villalba and Kno.e.sis Center—especially Lu Chen, and current research interns Kushal Shah and Garvit Bansal from LNMIIT India, for their invaluable continued support in discussion and review for our GBV research.

REFERENCES

- Ancheta MRG. 2007. "Pig's nest" in an even bigger pen: Pugad Baboy as a case of subversion and renegotiation in Philippine comedy. *Humanities Diliman* 1:2.
- Aumiller S. 2011. Why is Twitter promoting violence against women? *Give Hope A Voice*. WordPress.com. Available at <http://givehopeavoice.wordpress.com/2011/08/01/why-is-twitter-promoting-violence-against-women/>
- Bhanot S, Senn CY. 2007. Attitudes towards violence against women in men of South Asian ancestry: Are acculturation and gender role attitudes important factors? *Journal of Family* 22(1):25-31.
- Bowes N, McMurran M. 2013. Cognitions supportive of violence and violent behavior. *Aggression and Violent Behavior* 18(6):660-665.
- Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, Gaurish A, Carlson R, Watkins KZ, Falck R. 2013. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. *Journal of Biomedical Informatics* 46(6):985-997.
- Corroon M, Speizer IS, Fotso JC, Akiode A, Saad A, Calhoun L, Irani L. 2014. The Role of Gender Empowerment on Reproductive Health Outcomes in Urban Nigeria. *Maternal and Child Health Journal* 18(1):307-315.
- Creswell JW. 2013. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage.
- Culotta A. 2014. Estimating county health statistics with Twitter. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (CHI'14):1335-1344.

De Choudhury M, Gamon M, Counts S, Horvitz E. 2013. Predicting depression via social media. In *ICWSM*.

Dixit A. 24 January 2014. How love got the better of India acid attack. *Al Jazeera India*.

European Union. 2012. Council conclusions on the eradication of violence against women in the European Union. Available at http://www.consilium.europa.eu/uedocs/cms_Data/docs/pressdata/en/lsa/113226.pdf

European Union Agency for Fundamental Rights. 2012. European council effort. Available at <http://fra.europa.eu/en/project/2012/fra-survey-gender-based-violence-against-women>

Factbook, C.I.A. 2013. *The world factbook*. Central Intelligence Agency, Washington D.C.

Google Trends. 2015. *Top charts*. Available at <http://www.google.com/trends/topchartsdate=2014>

HarassMap tool. Available at <http://harassmap.org/en/>

Heise L, Ellsberg M, Gottmoeller M. 2002. A global overview of gender-based violence. *International Journal of Gynecology & Obstetrics* 78(1):S5-S14, ISSN 0020-7292, [http://dx.doi.org/10.1016/S0020-7292\(02\)00038-3](http://dx.doi.org/10.1016/S0020-7292(02)00038-3)

ITU statistics 2014 on Mobile Tech adoption. Available at http://www.itu.int/net/pressoffice/press_releases/2014/23.aspx#.VCSELStdU7s

Kalichman S, Simbayi LC, Kaufman M, Cain D, Cherry C, Jooste S, Mathiti V. 2005. Gender attitudes, sexual violence and HIV/AIDS risks among men and women in Cape Town, South Africa. *Journal of Sex Research* 42(4):299-305.

Lakoff G. 2008. *Don't think of an elephant!: Know your values and frame the debate*. Chelsea Green Publishing.

Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M. 2009. Life in the network: The coming age of computational social science. *Science* 323(5915):721.

Leetaru K, Wang S, Cao G, Padmanabhan A, Shook E. 2013. Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday* 18(5).

- Lewandowski M. 2010-2011. The rhetoric of violence in Polish and English soccer. In I. Koutny and P. Nowak (eds.): *Language, communication, information* 87-89.
- Lewandowski M. 2012. Football is not only war. Non-violent conceptual metaphors in English and Polish soccer language. In Taborek, Tworek and Zieliński (eds.): *Sprache und Fußball im Blickpunkt linguistischer Forschung*, Verlag Dr. Kovač, Hamburg 79-95.
- Liu Y, Kliman-Silver C, Mislove A. 2014, May. The tweets they are a-changin': Evolution of twitter users and behavior. In *International AAAI Conference on Weblogs and Social Media (ICWSM)* 13:55.
- Martinez PR, Khalil H. 2012. Battery and development: Exploring the link between intimate partner violence and modernization. *Cross-Cultural Research* 47(3):231-267.
- Morrison A, Ellsberg M, Bott S. 2007. Addressing gender-based violence: A critical review of interventions. *The World Bank Research Observer* 22(1):25-51.
- Nayak MB, Byrne CA, Martin MK, Abraham AG. 2003. Attitudes toward violence against women: A cross-nation study. *Sex Roles* 49(7-8):333-342.
- Nilan P, Demartoto A, Broom A, Germov J. 2014. Indonesian men's perception of violence against women. *Violence Against Women* 20(7):869-888.
- Pierotti RS. 2013. Increasing rejection of intimate partner violence evidence of global cultural diffusion. *American Sociological Review* 78(2):240-265.
- Purohit H, Castillo C, Diaz F, Sheth AP, Meier P. 2013. Emergency-relief coordination on social media: Automatically matching resource requests and offers. *First Monday* 19(1).
- Purohit H, Hampton A, Bhatt S, Shalin VL, Sheth AP, Flach JM. 2014. Identifying seekers and suppliers in social media communities to support crisis coordination. *Computer Supported Cooperative Work (CSCW)* 23(4-6):513-545.
- Romero G. 2010. Why the Filipino laughs. *Research Lines*. Available at <http://www.ovcrd.upd.edu.ph/researchlines/2010/10/22/why-the-filipino-laughs/>
- SafeCity initiative. Available at <http://safecity.in/>
- Sheth A, Jadhav A, Kapanipathi P, Chen L, Purohit H, Smith GA, Wang W. 2014. Twitris- A system for collective social intelligence. *Encyclopedia of Social Network Analysis and Mining (ESNAM)*, Springer, Alhajj, Reda, Rokne, Jon (Eds.) ISBN: 978-1-4614-6169-2.

- Available at <http://www.springer.com/computer/communication+networks/book/978-1-4614-6169-2>
- Tannen D. 1996. *Gender and Discourse*. Oxford University Press, ISBN 0-19-508975-8; ISBN 0-19-510124-3.
- Tausczik YR, Pennebaker JW. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology* 29(1):24-54.
- Twitter Developers. 2014. Streaming API. Available at <https://dev.twitter.com/streaming/overview> Accessed Jan 11 2014.
- United Nations. 2009. *UN Secretariat campaign for violence against women*. Available at https://www.un.org/en/events/endviolenceday/pdf/UNiTE_TheSituation_EN.pdf
- United Nations Population Fund. 2013. *Addressing gender-based violence*. Available at <http://www.unfpa.org/webdav/site/global/shared/documents/publications/2013/final%20sexual%20violence%20CSW%20piece.pdf>
- United Nations Population Fund. *Gender-based violence*. Available at <http://www.unfpa.org/gender/violence.htm>
- United States Agency for International Development. 2008. *Indicators for gender violence*. Available at <http://www.cpc.unc.edu/measure/publications/ms-08-30>
- Vieweg S, Hughes AL, Starbird K, Palen L. 2010. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. *CHI '10: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1,079–1,088. doi:<http://dx.doi.org/10.1145/1753326.1753486>, accessed 26 December 2013.
- World Health Organization. 2013b. *Global and regional estimate of violence against women*. Available at http://apps.who.int/iris/bitstream/10665/85239/1/9789241564625_eng.pdf
- World Health Organization. (2013a). *WHO Fact sheet*. Available at <http://www.who.int/mediacentre/factsheets/fs239/en/>
- Yoshihama M, Blazeovski J, Bybee D. 2014. Enculturation and attitudes toward intimate partner violence and gender roles in an Asian Indian population: Implications for community based prevention. *American Journal of Community Psychology* 53:249-260.