



CAN BIG DATA BE USED FOR EVALUATION?

A UN Women feasibility study



ACKNOWLEDGEMENTS

We thank:

- Shravanti Reddy (Lead) and Alexandra Capello of the UN Women Independent Evaluation Service for guidance and support.
- The UN Global Pulse team for advice related to selection of data sources and facilitating access to Twitter data.
- Kristen Sample, substantive expert for global corporate evaluation of UN Women's Political Participation and Leadership (WPP) program for her insights about the WPP evaluation in Mexico.
- Andrea Azevedo, gender specialist and monitoring and evaluation professional for her insights about the campaigns in Mexico and Pakistan.
- Shazia Abbasi of Shazia Abbasi Consulting and Qamar Iqbal of Association for Gender Awareness & Human Empowerment (AGAHE) for their insights on the WPP campaigns in Pakistan.

Our advisory group:

- Michael Bamberger, Independent Evaluation Expert
- Rebecca First-Nichols, Deputy Director, Data2X
- Navin Haram, Programme Specialist, Strategic Planning, Programming and Effectiveness Unit, UN Women
- Robert Kirkpatrick, Director, UN Global Pulse
- Karin Mattsson and Roxana Flores, Mexico Country Office, UN Women
- Andrew Means, Head of beyond.uptake, Uptake Foundation
- Veronica Olazabal, Director of Measurement, Evaluation and Organizational Performance, The Rockefeller Foundation
- Papa Seck, Policy Specialist, Research and Data, UN Women
- Sangeeta Thapa, Deputy Representative, Pakistan Country Office, UN Women
- Peter York, Principal and Lead of BCT Partners' Data Analytics

Authors: Dr Claudia Abreu Lopes

Dr Savita Bailur

Giles Barton-Owen

Produced by the Independent Evaluation Service

IISBN: 978-1-63214-129-3

Lead Manager: Shravanti Reddy

Editor: David Marion

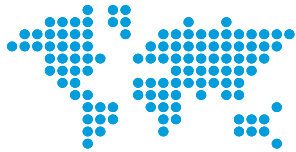
Design and layout: Yamrote Alemu H.

© 2018 UN Women. All rights reserved.

The views expressed in this publication are those of the author(s) and do not necessarily represent the views of UN Women, the United Nations or any of its affiliated organizations.

FEASIBILITY STUDY

CAN BIG DATA BE USED FOR EVALUATION?



INDEPENDENT EVALUATION SERVICE

UN WOMEN

New York, April 2018

TABLE OF CONTENTS

FOREWORD	5
ACRONYMS	6
GLOSSARY OF TERMS	7
EXECUTIVE SUMMARY	8
1. INTRODUCTION	12
1.1. Background : why big data?	13
1.2. Aims and objectives of this study	14
1.3. Pilot cases	14
1.4. Sources of big data	15
1.5. Challenges and risks of relying on big data as a source	18
1.6. Project team	19
2. METHOD	20
3. CASE STUDY DETAILS	25
3.1 Data sources	26
3.2 Pilot study in Mexico	27
3.3 Pilot study in Pakistan	27
4. DATA PROCESSING, ANALYSIS AND RESULTS	29
4.1 Twitter analysis and results for Mexico	30
4.2 Facebook analysis for Pakistan	39
4.3 Analysis of radio data	42
5. SUMMARY OF FINDINGS	43
6. RECOMMENDATIONS AND FUTURE WORK	44
APPENDIX: ALGORITHM TO DISCOVER CO-OCCURRENCES INVOLVING WORDS	46
REFERENCES	47

FOREWORD



The UN Women Independent Evaluation Service strives to continuously improve its evaluation practice and contribute to the wider evaluation community in support of gender equality and human rights responsive evaluation.

Most recently, improving our evaluation practice has meant that we need to explore how new technologies can be harnessed to improve our data collection sources and analytical capacities to deliver more credible and useful evaluations.

This study was commissioned with the above in mind to explore the feasibility of incorporating big data sources as part of the evidence base for UN Women evaluations. It represents a crucial first step toward understanding how to use big data from social media and other sources that can potentially be used for analyzing contributions to the achievement of gender equality and women's empowerment by looking at two cases – Mexico and Pakistan – related to a recent evaluation of women's political participation and leadership.

The study is an important contribution to the evaluation community's better understanding of how to make use of big data sources by providing an analysis of the pros and cons of some potential data sources, initial step-by-step protocols for their use, and recommendations based on lessons learned about using big data sources in a meaningful way for evaluation. This is an important first step on a promising method but requires further study, discussion and consideration before it can be mainstreamed as part of standard evaluation processes.

I am pleased to share this study with you and hope that you will benefit from its findings to move this area forward. I would like to take this opportunity to thank the authors and lead of the study for their hard work and contribution to improving our understanding of the possibilities for future evaluation practice.

Sincerely,

A handwritten signature in black ink, reading 'Verasak Liengsriwat'.

Verasak Liengsriwat

Director a.i., Independent Evaluation and Audit Services

ACRONYMS

AGAHE	Association for Gender Awareness & Human Empowerment
API	Application Programming Interface
DRC	Democratic Republic of Congo
FGD	Focus Group Discussions
GUID	Globally Unique Identifier
IES	Independent Evaluation Service
KII	Key Informant Interview
NLP	Natural Language Processing
ONU	Organizacion de Naciones Unidas
POS	Point-of-Sale
RCT	Randomized Control Trial
SDG	Sustainable Development Goal
SMS	Short Messaging Service
UN	United Nations
UNDG	United Nations Development Group
WPP	Women's Political Participation and Leadership
WVS	World Values Survey

GLOSSARY OF TERMS

API	Application Programming Interface (API), consisting of a set of functions, routines and protocols that define how software components communicate with each other, allowing software applications to be built from the initial code.
Construct validity	The degree to which an inference can be made from an operationalisation to the theoretical construct that it is supposed to measure (Trochim, 2006).
Interrupted time series	Research design that implies repeated measurements obtained before and after a certain period of time that is marked by one or several interventions (Ferron and Rendina-Gobioff, 2015).
Low resource languages	Less-often studied languages for which resources for natural language processing (NLP), such as machine-readable corpora, annotated data, grammar, dictionaries, treebanks and POS tagging, are scarce.
Randomized block design	Research design in which cases are sub-grouped by common attributes (forming blocks) and randomly assigned to different conditions (e.g., intervention vs control) within each block, allowing evaluating the effect of the intervention for particular blocks (Saville and Wood, 1991).
Propensity score matching	Statistical technique for observational studies that relies on a large number of covariates to predict the effects of an intervention, without the need to use a true experiment (Austin, 2011).
Nomological networks	Representation of constructs used in a study, their measurements and the relationships among them (Cronbach and Meehl, 1955).
Sentiment analysis	Technique based on natural language processing (NLP) to classify text in terms of its polarity or tonality by identifying expressions people use to evaluate or appraise persons, entities or events (Pang and Lee, 2018).
Social network analysis	Technique to represent and analyse the networks between individuals and groups (Wasserman, and Faust, 1994).
Support vector machine	Supervised learning model, with associated algorithms that analyse data used for classification tasks.
Tokenization	Process of breaking down text into its constituent elements.
Topic models	Technique based on statistical models to discover hidden topics embedded in text documents (Melville et al., 2013).

EXECUTIVE SUMMARY

The objective of this study was to investigate the feasibility of leveraging big data sources – particularly Twitter, Facebook and radio data – to improve the evaluation of gender equality and women’s empowerment initiatives. In particular, this study seeks to understand the role of big data to evaluate the contribution of UN Women to women’s political participation and leadership (WPP). Taking Mexico and Pakistan as two case studies, which present different challenges to access of big data sources and distinct barriers to WPP, we documented the process of accessing, analysing and triangulating big data sources with traditional data as a feasible means to provide more credible and robust insights about the effectiveness of UN Women interventions.

The community for international development evaluation has spent decades developing and refining tools for collecting and using data to determine whether social interventions work. Evaluation data often are limited to the evidence available through traditional evaluation methods constrained by rigid timeframes and scarce resources. In this respect, big data offers an additional evidence base to triangulate with and complement traditional methods. In 2017, UN Women’s Independent Evaluation Service (IES) commissioned a study to determine to what extent big data could help strengthen traditional UN Women evaluations, with three key aims:

- Determine if it is possible to improve the evaluation of UN Women’s work using additional evidence streams from big data, mainly focusing on social networking/media and news platforms.
- Apply the new United Nations Development Group (UNDG) “Principles for Big Data and the Sustainable Development Goals”¹ (SDGs) and the “Risks, Harms and Benefits Assessment Tool”² to provide feedback for their use and

refinement in regard to gender equality and women’s empowerment issues.

- Support understanding of how UN Women and its partners might effectively use big data to support future evaluation efforts on WPP and in other thematic areas.

The feasibility study was commissioned with three main uses in mind:

1. Support UN Women IES’s understanding of how and when it can incorporate big data within corporate and decentralized evaluation processes.
2. Share with like-minded organizations and partners to build learning and knowledge on the use of big data methods for evaluation.
3. Consider how to incorporate relevant data analysis and findings, specifically for evaluating WPP.

METHOD

This feasibility study focused on combining big data sources with traditional data sources to validate big data indicators. The limitations of including social media data as compared to other big data sources include selection of meaningful indicators; exclusions or under-representation of certain groups, due to access, restricted use or online harassment; population and usage drifting that may compromise analysis of trends; and impossibility of using control clusters to compare. A social media analysis was proposed for the study in four stages, with two pilot case countries and data sources selected to test the feasibility – Twitter data for Mexico and Facebook data for Pakistan:³

¹ UNDG Principles for Big Data and the SDGs: <https://undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/>.

² Risks, Harms and Benefits Assessment: <https://www.unglobalpulse.org/privacy/tools>.

³ The use of radio data was also explored in Pakistan.



1. Develop and test a measurement model to select the best big data indicators (e.g., keywords and/or hashtags) that correlate with well-established indicators from traditional data, meaning they are measuring the same construct.
2. Describe the universe of tweets relevant for UN Women campaigns in Mexico and Facebook posts in Pakistan, in terms of demographics and language and identify biases.
3. Analyse results (engagement, topics, sentiment) across geographies and and/or over time to derive insights about the contribution of UN Women interventions, disaggregating by gender.
4. Triangulate and complement insights from big data sources with findings from the UN Women corporate evaluation, based on traditional methods.

SUMMARY OF FINDINGS



Findings on Twitter

- Twitter appears more appropriate for evaluating UN Women's interventions aimed at fostering political participation and attitudes towards gender equality.
- Social network analysis can help to reveal the online network of users and their degree of influence within their network. This type of analysis may be able to answer questions related to the reach and spread of information through Twitter.
- Given the short life of hashtags, longitudinal analysis based on the same hashtags is not meaningful.
- Analysis and interpretation of conversations within a cultural context can be enhanced by focus groups with Twitter users and/or validated by media and communication experts from the country.



Findings on Facebook

- Private or semi-private discussions may pose ethical issues because they can reveal sensitive personal details that could place users at risk.
- Many pages from organizations do not contain much discussion; pages associated with political or social issues have biased samples, as people self-select strongly based on their views (e.g., political and social issues).
- Other sources hold more promise, such as radio data, responses to SMS campaigns and responses to newspaper articles online.



Findings on radio data

- Radio can be a significant social venue.
- Historical streaming of radio data is not always present.
- Radio programmes can be designed to gather useful information for evaluation through voice or SMS.
- Requires careful recording and coordination to ensure large volume of data is available for analysis, but can be highly relevant and rich (e.g., documenting community conversations).



RECOMMENDATIONS AND FUTURE WORK

Recommendation 1

Understand the bigger picture of big data in a country before considering it as a source for evaluation.

- Twitter and Facebook have digital architectures encouraging certain styles and degrees of engagement that need to be understood with all their cultural specificity before embarking on an evaluation using these platforms. Different big data platforms could prove culturally insightful for understanding the context variables to attend to when evaluating using big data.
- Understanding representativeness is not a binary question about those who are on social media or not. There is also the question of varying degrees of social media use that will influence the representation of social media users.
- Another issue is assuming that majority languages are representative of countries (e.g., Urdu or Punjabi in Pakistan), which will exclude those who use other languages on social media.
- A challenge of using big data for evaluation in low-income countries is that representative responses of many will be those offline that can be gathered through street plays and offline campaigns.
- Findings need to be interpreted through the lens of the cultural, language and media context where individuals belong. If possible, discuss the results through key informant interviews (KII) or focus group discussions (FGD) with people who provided the data.

Recommendation 2

Big data should be incorporated in the design of the evaluation from the outset.

- Identifying data sources early in the design stage allows for planning and collecting traditional data to compensate for coverage problems.
- Natural experiments with big data require effective control and intervention groups that should be closely monitored during the life cycle of an intervention.
- Access, scraping and preparation of big data sources that respect best ethical practices take most of the time allocated for analysis. Be realistic about the time and cost involved in gaining access, scraping and analysing big data.
- Consider whether open source tools are available or new models or tools need to be built. For example, consider strongly if a language is well resourced and if not, whether it is worth using big data or not, as it will be very time-consuming to build any analytical model.

Recommendation 3:

Big data should precede traditional data when sequencing and evaluating.

- Start with a study of the demographics of the social media platform and topic to understand the nature and extent of the exclusions; consider using traditional data sources to obtain information from groups poorly represented.
- Big data analysis will help with the scoping stage of the evaluation, disclosing general trends and surprising case studies.
- Triangulation of data will be more effective if different methods are aligned.



Recommendation 4:

Big data can be shaped in ways that enhance its value.

- For example, launching social media campaigns, using hashtags and posts that trigger meaningful reactions and engagement from diverse groups.
- Set up interactive radio shows that invite participation from audiences through SMS and social media, while gathering individual demographics.
- Big data platforms can be both the intervention tool (e.g., hashtag campaigns) and the source of evaluation data. This does not present any problem, but studies using big data analysis need to make this distinction explicit, and adjust methods and conclusions according to the objective of the evaluation.

Box 1: Recommendations

1. **Understand the bigger picture of big data in a country** before considering it as a source for evaluation.
2. Big data should be **incorporated in the design of the evaluation** from the outset.
3. Big data should **precede traditional data** when sequencing and evaluating.
4. Big data can be **shaped in ways that enhance its value.**



1

INTRODUCTION

1.1 BACKGROUND - WHY BIG DATA?

The community for international development evaluation has spent decades developing and refining tools for collecting and analysing data to determine whether social interventions work. For example, UN Women's Independent Evaluation Service (IES) uses methods such as surveys, interviews and focus groups. Evaluation data often are limited to the evidence available through these methods constrained by rigid timeframes and scarce resources. In this respect, big data offers an additional evidence base to triangulate with and complement traditional methods.

Big data emerged in developed societies as a private sector tool for corporate analytics, for example, to improve decision making based on customer digital feedback and activities. But in the last ten years, big data has been evolving in parallel as a source for understanding the social world by disclosing large-scale patterns and generating models to make predictions about uncertain events.

Box 2: Definition of big data

DEFINITION OF BIG DATA

Big data is often seen as 'data exhaust,' in other words, 'the digitally trackable or storable actions, choices, and preferences that people generate as they go about their daily lives.'⁴ As opposed to traditional data (e.g., surveys or census), which are collected with a specific intention and follow a structured format with valid and reliable measurements, big data generally:

- Is produced as a by-product of people's digital behavior.
- Is not gathered in a way guided by a research question so requires interpretation after the event.
- Imperfectly matches the entire universe of cases. Representativeness is generally not a factor in big data collection, as there is no sampling strategy involved, and non-coverage is often a concern when assessing data quality.
- Is often accessible in real time (at the time the data are produced). However, data analytics may require some days or weeks.
- Can be analysed by combining different data sources; multiple datasets can illuminate new insights and/or validate indicators, or help with triangulation.
- Can be harnessed to improve decision making. Appropriate guidance and frameworks can help translate insights from big data into value for organizations, governments or business.

(Adapted from Hilbert, 2016)

⁴ UN Global Pulse (May 2012) Big Data for Development: Challenges and Opportunities: www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseMay2012.pdf

1.2 AIMS AND OBJECTIVES OF THIS STUDY

In 2017, UN Women's IES undertook a corporate thematic evaluation of its global efforts to advance women's political participation and leadership (WPP), one of six core thematic impact areas that UN Women aimed to contribute to at the global, regional and national levels. The aims of the corporate evaluation were to assess UN Women's cumulative contribution towards women's ability to lead and participate in decision making at all levels and to provide evidence from past practices to inform future strategic planning and implementation on WPP.

This study was commissioned in June 2017 to determine to what extent big data could help strengthen traditional UN Women evaluations. In particular, we sought to test the feasibility of leveraging big data and data analytical techniques to:

- Determine if it is possible to improve the evaluation of UN Women's work using additional evidence streams from big data, mainly focusing on social networking/media and news platforms.

- Apply the new United Nations Development Group (UNDG) "Principles for Big Data and the Sustainable Development Goals" ⁵ (SDGs) and the "Risks, Harms and Benefits Assessment Tool" ⁶ to provide feedback for their use and refinement in regard to gender equality and women's empowerment issues.
- Support understanding of how UN Women and its partners might effectively use big data to support future evaluation efforts on WPP and in other thematic areas.

The feasibility study will be used to:

- Support UN Women IES's understanding of how and when it can incorporate big data within corporate and decentralized evaluation processes.
- Share with like-minded organizations and partners to build learning and knowledge on the use of big data methods for evaluation.
- Incorporate relevant data analysis and findings into the overall data analysis and reporting phases of the WPP evaluation if possible. ⁷

1.3 PILOT CASES

As the objective of the corporate evaluation was to assess the contribution of UN Women to WPP programmes globally, six countries were selected for site visits (Democratic Republic of Congo, Egypt, Malawi, Mexico, Pakistan and Zimbabwe). The two countries selected in this feasibility study, Mexico and Pakistan, were shortlisted from the group of these six countries considered in the corporate evaluation.

The selection of these two countries resulted first, from an extensive desk review of the materials collected during the UN Women IES's visits to Mexico and Pakistan in 2017. This included strategy documents, interview transcripts of UN Women Country Office staff, government officials, other UN agencies, international organizations and civil society organizations.

⁵ UNDG Principles for Big Data and the SDGs: <https://undg.org/document/data-privacy-ethics-and-protection-guidance-note-on-big-data-for-achievement-of-the-2030-agenda/>.

⁶ Risks, Harms and Benefits Assessment: <https://www.unglobalpulse.org/privacy/tools>

⁷ Given the timing of the study and its findings, it was not possible to incorporate big data analysis into the evaluation.









Second, in terms of sources of big data, Mexico was chosen to be an archetype of countries with a high internet penetration (65.3 per cent of the population)⁸ and a well-resourced language with accessible libraries and textual analysis tools (Spanish). Outside the US, Mexico is one of the top three countries of Twitter users (25.7 million⁹ in 2017, representing approximately 20 per cent of the population), alongside Brazil and Japan. Spanish, the de facto national language in Mexico for which comprehensive language resources exist, presented no challenges for analysis of content.

In turn, Pakistan was chosen to represent countries with lower internet penetration (22 per cent of the population in December 2017)¹⁰ and languages with limited libraries and tools (e.g., Urdu). Twitter uptake is low, with 3.1 million users in July 2016 (approximately 1.6 per cent of the population), but

Facebook has gained popularity in Pakistan for the past five years, with over 31 million users by the end of 2017¹¹ (approximately 16 per cent of the population). Although internet access through the third generation of wireless mobile telecommunications technology (3G) that include mobile phones, computers and other portable devices, has been improving (24.5 per cent of the population in 2018),¹² it is still confined mostly to urban centers. The majority of the population resides in rural areas (67.5 per cent), according to the last available population census (1998).¹³

To validate our choice of countries, we scoured the UN Women's Mexico and Pakistan social media presence (Twitter and Facebook) to seek hashtags and pages that may have been associated with specific UN Women campaigns.

Box 3: Social media profile of Mexico and Pakistan

 MEXICO	 PAKISTAN
 25.7 M Twitter users in 2017	 3.1 M Twitter users in 2016
 One of the top three countries of Twitter users	 More than 31 M Facebook users by end of 2017
 65.3% of the population has internet access	 22% of the population has internet access

1.4 SOURCES OF BIG DATA

This feasibility study used Twitter data for Mexico and Facebook data for Pakistan. There are fundamental differences between these two platforms. Twitter is better for sharing information quickly and publicly, with short opinions, links and news headings. Facebook is more suited for longer lasting interactions, in-depth discussion and personal sharing within a context of friendship networks.

Twitter data is more appropriate for evaluating UN Women's interventions based on political participation and attitudes towards gender equality. It has been shown, for example, that Twitter political sentiment reflects political preferences of users (Tumasjan et al., 2010), but some authors urge caution about using social media to predict election results (Jungheer et al., 2016). In certain contexts (e.g., characterized by high freedom of expression),

⁸ See for example: <https://www.internetworldstats.com/stats2.htm>.

⁹ See for example: <https://www.emarketer.com/Article/Twitter-User-Base-Latin-America-Continues-Grow/1013924>.

¹⁰ See for example: <https://www.internetworldstats.com/stats3.htm>.

¹¹ See for example: <https://www.geo.tv/latest/131187-Over-44-million-social-media-accounts-in-Pakistan>.



¹² See: <https://pta.gov.pk/en/telecom-indicators>.

¹³ One problem in Pakistan is the registration of mobile subscriber identification modules, commonly called SIM cards. Due to non-availability of national identity cards and cultural practices, often SIM cards used by females are registered in the name of male family members. These practices make it impossible to estimate penetration of 3G by gender.

Twitter may offer a quick solution to gauge the offline political landscape or socio-political attitudes (Schober et al., 2016). In turn, Facebook's strength lies in stimulating political participation (Busseta et al., 2017). Notwithstanding, social media data presented the clear advantage of being easily accessible compared with, for example, radio data.¹⁴

Selection of data sources was also influenced by the nature of UN Women's campaigns in the two countries and by the interest in understanding challenges of different types of social data for evaluation. To decide on the most feasible data sources, we went through an analytical process, identifying and weighing the pros and cons of each potential data source, as documented in Table 1.




Table 1: Comparison of some big data sources

NOTES	PROS	CONS
Twitter 		
<ul style="list-style-type: none"> Twitter users are typically younger, wealthier, more educated and more likely to live in urban areas than Facebook users.¹⁵ Reportedly, 92% of all activity and engagement happen within the first hour of posting a tweet - so discussion tends to be more 'real-time' than in Facebook.¹⁶ Activity tends to be measured in terms of number of engagements (activities related to the tweet) and impressions (how many times a tweet was seen). 	<ul style="list-style-type: none"> Given the near real-time nature, Twitter is a good barometer of topical issues. It is possible to conduct social network analyses to identify influencers within networks and to analyse how information spreads within it. 	<ul style="list-style-type: none"> Twitter represents only a small proportion of populations, particularly in low-income countries. Tweets are limited to 140 or 280 characters depending on the language, making it difficult to elaborate opinions. Intensely personal issues are less likely to be discussed, depending on the user's openness and privacy concerns. Twitter's public application programming interface (API) poses limitations in accessing historical data older than two weeks. Demographic data is very limited (e.g., only 1% of tweets accessed from public API contain geographical data). It is impractical to obtain informed consent from users.
Facebook 		
<ul style="list-style-type: none"> Discussions tend to be within friends and/or community groups and over a period of time that can stretch to days and weeks. Activity tends to be measured in terms of posts, likes, shares and comments. 	<ul style="list-style-type: none"> Allows analysis of conversations that contain more or less elaborated opinions. Possible to access historical data in public pages. Data can be collected through Facebook apps that require informed consent from users. Allows analysis of social networks if one has access to users' Facebook friends' data. 	<ul style="list-style-type: none"> Discussions tend to be siloed in posts and within a small group of users. Facebook's public API only allows data to be scraped from public pages. Demographics are available for the majority of users (gender, age, location).

¹⁴ Note that our major restriction was the timeframe – particularly in terms of procuring access to certain datasets – as the project duration was three months.

¹⁵ See for example: <https://www.bloomberg.com/gadfly/articles/2016-02-12/social-studies-comparing-twitter-with-facebook-in-charts>.

¹⁶ See for example: <http://www.visualscope.com/twitfb.html>.

NOTES	PROS	CONS
Whatsapp 		
<ul style="list-style-type: none"> Currently a closed system without external access through an API. It is only possible to access and analyse data from groups to which one belongs. 	<ul style="list-style-type: none"> It provides data about topical discussions among a group of users from an unlimited historical period of time. Data from groups one belongs to can be downloaded directly from the application. 	<ul style="list-style-type: none"> A username is not a reliable unique identifier, as people can change their username within and across groups. Network data is not available. It is possible to obtain informed consent from users.
Radio data 		
<ul style="list-style-type: none"> Audio data from radio broadcasts, as well as audience interaction during shows (call-ins, SMS), allow measuring levels of engagement with topics and a range of opinions. 	<ul style="list-style-type: none"> In many ways, radio data is most reflective of broader cross sections of communities and marginalized voices due to the low levels of technology and literacy required, compared to 'elite capture' of social media. It is possible to reach a large number of people who form the audience of the radio shows and engage them through other communication platforms (e.g., SMS or social media) to collect demographics or other self-reported data. 	<ul style="list-style-type: none"> No established practices (or tools) for recording, storing or analysing radio shows, particularly in low income countries. Gaining access to radio data requires planning to select and record radio shows, which can be time-consuming. Automatic analysis of audio data requires speech-to-text models that are not yet available for most languages.
News data 		
<ul style="list-style-type: none"> Comments to news articles online offer a possibility to analyse reactions related to the topics of the news. 	<ul style="list-style-type: none"> Provides more focused data around a topic than social media. Offers opportunity to gauge the pulse of public opinion on particular topics. 	<ul style="list-style-type: none"> Data sources themselves may be biased due to self-selection of users to certain groups (as with other platforms). Comments may be filtered and moderated, rendering them unrepresentative of the spectrum of opinions. Many comments are anonymous, making it difficult to perform socio-demographic analysis.

Given the specificities of each data source, it is important to highlight the need for a framework for understanding how countries and populations adopt and use each platform and for what purposes. Issues of representativeness would come to light as well as changes in the ways that the platforms are

used overtime (platform drifting). To use social data for evaluation, it is imperative to map what each platform is capturing and for whom, from a cultural lens.

1.5 CHALLENGES AND RISKS OF RELYING ON BIG DATA AS A SOURCE

With increased access to technology (e.g., mobile phones and social media) individuals in low-income countries are increasingly likely to leave digital traces. The widespread use of free analytical tools allows digital data to be accessed and analysed by individuals, governments and organisations across the world. Researchers and evaluators need to understand how insights drawn from the data might impact people's lives and to be responsible in how they use and present these conclusions to commissioning organisations and the general public.

The skillset necessary to access and derive meaningful insights from this data, however, is not evenly distributed over regions or organizations. Analysis of big data requires specialised data skills and experience working with unstructured and large datasets. Besides proficiency in certain software or programming languages, analysts need to be well-versed in methods of differentiating between 'signal' (meaningful data) and 'noise' (irrelevant data) and to be able to translate findings into actionable insights. The process of data analysis needs to be presented in ways that allow relevant findings to emerge and be interpreted by evaluators or policy makers with no data science background. Such skills often can be expensive and/or inaccessible.

A number of risks and challenges of relying on digital data for evaluation can be identified. First, there is a major risk of 'black holes' of data where entire demographics can be missed because of restricted access and use. As an example, globally, Twitter tends to be used slightly more by males and by young age groups (predominantly by 18 to 29 year olds).¹⁷ Evaluating a programme solely based on social media data risks disregarding entire groups, contributing to 'elite capture' and skewed representation.

Second, there are a number of factors that constrain women's ability to participate in social media or other interactive platforms. Access to radio or mobile phones can be restricted for women due to issues of ownership and control (e.g., men deciding when and what to listen to on the radio). In contexts of domestic violence against women, use of communication platforms are often monitored and possibly barred.

Third, with increasing numbers of people online, we see an equivalent increase in misogyny on social media. Cyber bullying, trolling (making a deliberately offensive or provocative online post with the aim of upsetting someone), flaming (making personal attacks on someone online) and threats against women, particularly celebrities, are becoming all too common. Fox et al. (2017) state that anonymity is a major contributing factor; this is especially true for Twitter, which unlike Facebook, does not enforce real-name policy. Fox et al. borrow from Glomb et al.'s (1997) the concept of 'ambient sexism', i.e., the notion that attitudes and behaviours prevalent in a negative environment can affect all those in the environment even if not directed towards one particular individual. Fear of being bullied means fewer women may be willing to go online or use public social media.

Elite capture, restricted use and ambient sexism lead to biases and exclusions on conclusions or recommendations, which may perpetuate existing social inequalities, including gender inequalities. Before we decide to embark on an evaluation based on big data, we need to ask ourselves: Who is not heard? Whose realities are not reflected in the data? What is the impact of these exclusions when using data to inform programmes, advocacy or policy change?

¹⁷ See for example: <https://www.omnicoreagency.com/twitter-statistics/>.



Fourth, there are risks related to population drift (change in who is using social media platforms), usage drift (change in how people are using them), and system drift (change in the system itself – number of characters from Twitter), that make it hard to use big data sources for studying long-term trends (Salganik, 2018).

Fifth, there are ethical issues related to consent and legitimacy (cf. Williams et al. 2017 for an ethical framework for publishing Twitter data). Legitimate data use is one of the key pillars of responsible data handling, while consent is one of the main bases for legitimate data use (UN Women, 2018). People who provide data might not be in a position to make informed choices or provide consent. They might not be aware of the implications of their data being collected or used, have little or no awareness of their digital rights, and have even less power to influence the process of data gathering (Antin et al., 2014). The focus on the broader societal benefits of working on ‘good causes’ sometimes contributes to avoiding a more critical discussion of how data are managed or used, and whether accessing or collecting data will benefit the people who provide them. Approaches to ensure responsible data use in such situations include: (i) transparency and education about the possible risks and benefits of data use and non-use; (ii) conducting risk assessments of the use of data in a given context; and (iii) well-informed consent (UN Women, 2018).

Finally, further ethical questions for the analysis of social data involve data aggregation, right of privacy, data security and data capacity. Data aggregation is one of the most common ways to both protect individual privacy and present conclusions. Data teams need to be mindful of the risks of data use, while ensuring that the level of aggregation does not

diminish the quality of the data. (UN Global Pulse, 2015). Data teams should employ the principles of data minimization, necessity and proportionality when aggregating data to ensure that only a minimum necessary level of detail in data is used to achieve the intended positive outcome of the data use (UN Women, 2018).

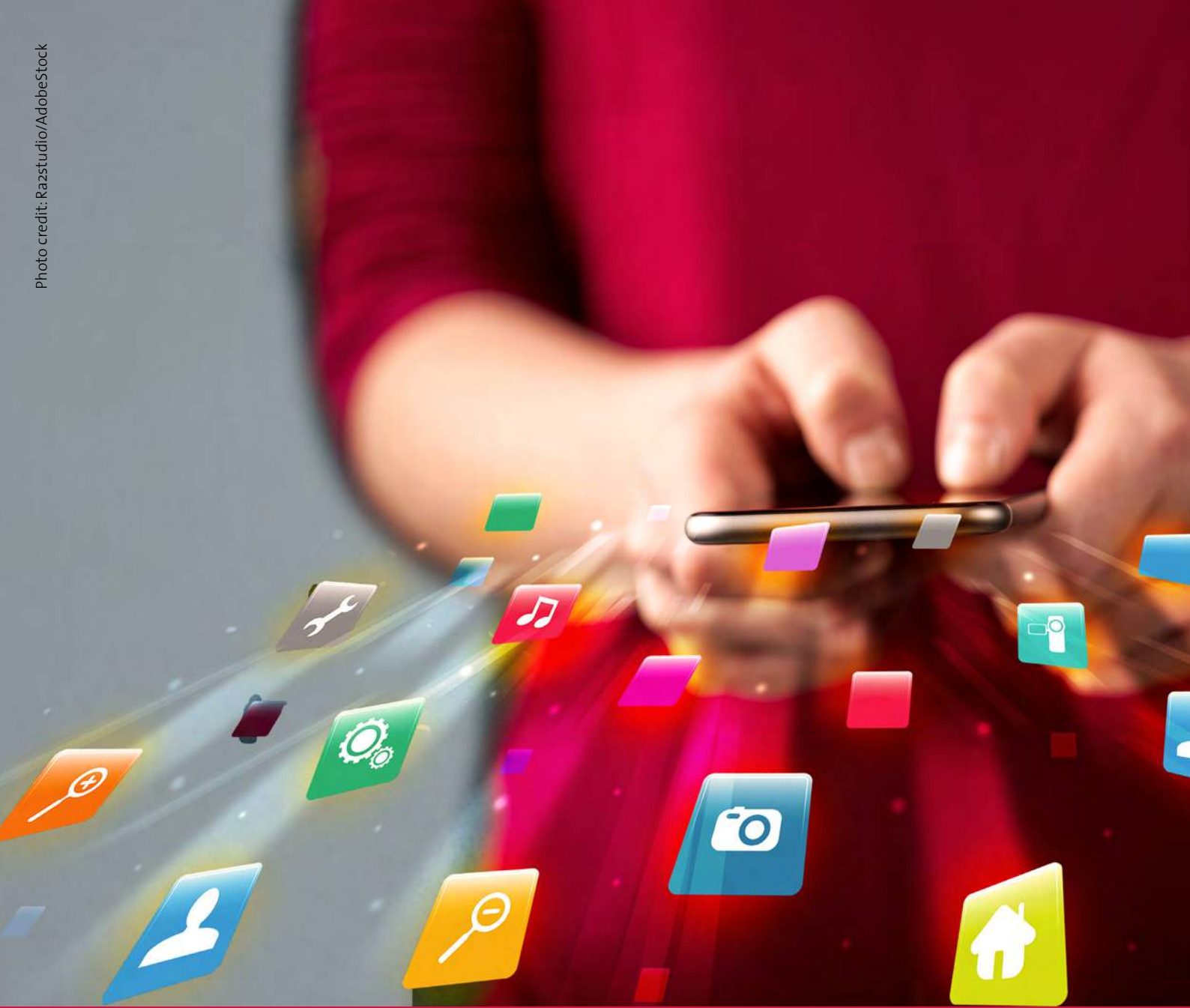
To prevent data breaches, data teams need to perform ongoing vulnerability assessments of their systems and undergo regular trainings on data privacy and security. Global Pulse recommends the following five rules when handling Twitter data (Global Pulse, n.d) that were strictly followed in this study:

1. Never copy any type of non-aggregated data (e.g., individual tweets), although it is ok to copy aggregated results.
2. Always use encrypted volumes to store the non-aggregated data.
3. Always keep two identical copies of all non-aggregated data downloaded.
4. Always treat data as ‘personally identifiable information.’ For example, although tweets are public and are shared by Twitter users with consent, they may contain highly-sensitive personal details that in the wrong context could place users at risk.
5. Never single out any one person in the results – i.e., do not publish analysis that contains individual user account names or that tracks the movements of individuals over time.

1.6 PROJECT TEAM

UN Women IES commissioned and managed the overall study (Shravanti Reddy/Lead Manager and Alexandra Capello) and guided the research team, consisting of a Project Lead, Senior Researcher and Data Scientist. We also received technical support

from UN Global Pulse and were aided by a group of advisors with expertise in data science, evaluation, international development and humanitarian affairs.



2 METHOD

This feasibility study recommends combining big data sources with traditional data sources to validate big data indicators (cf. Callegaro and Yang 2018; Salganik, 2018; Schober et al., 2016). When combining surveys and big data sources, Salganik (2018) distinguishes between (i) ‘amplified asking,’ when a predictive model allows combining survey data from a few people with big data from many people and (ii) ‘enriched asking,’ when survey data is built around a big data source that contains some important measurements but lacks others. Both strategies imply data linkages where the same unique identifier is used across survey data and big data (e.g., linking a survey respondent with a particular Twitter handle). This may not always

be possible in an evaluation context, particularly when using secondary data sources, such as well-established cross-national surveys by other organizations. However, if UN Women is conducting its own surveys, these strategies can enhance the value of big data sources. When it is not possible to link traditional and big data sources per individual, it may still be possible to link them through some sort of common identifiers, such as segments of users or geographical units. There are some caveats with this method that will be discussed later in this section.

We recommend a social media analysis to be done in four stages:

Box 4: Four stages for social media analysis

1

MEASUREMENT MODEL TO SELECT THE BEST BIG DATA INDICATORS

Proposed indicator to measure SDG 5.5



Percentage of seats held by women in elected offices

Traditional data sources to measure political participation...

Valid and trusted international survey results e.g. World Values Survey



... are used to validate indicators from social media

Explore concepts from Theory of Change and map them onto possible social media indicators (e.g., hashtags and keywords)



Hybrid approach traditional + big data

Validate the big data indicators with traditional data i.e. measuring the same construct



2

DESCRIBING THE UNIVERSE OF RELEVANT TWEETS (& FACEBOOK POSTS)



Analysis of tweets

All tweets/posts with relevant hashtags and keywords



Engagement : likes, shares and comments

Demographics of users

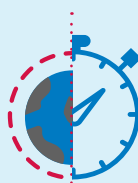
Sentiment of tweets/posts/comments

3

ANALYSIS OF RESULTS ACROSS REGIONS & OVER TIME

Regional comparison of engagement

For example, comparing engagement in Mexican districts where UN Women invested more resources with other districts



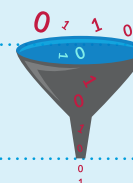
Longitudinal analysis of engagement, demographics and sentiment

Mapping a timeline of UN Women events in the target regions onto the patterns of social media results over time

4

TRIANGULATE & COMPLEMENT EVALUATION FINDINGS WITH BIG DATA

The results of the big data analysis can be triangulated with the results of the corporate evaluation and interpreted alongside traditional data, such as interviews that provide a deeper understanding of the contribution of UN Women to observed changes.



1 Measurement model to select the best big data indicators (e.g., topics or hashtags) that correlate with traditional data, meaning they are measuring the same construct.

The proposed indicator to measure SDG 5.5 – “Ensure women’s full and effective participation and equal opportunities for leadership at all levels of decision-making in political, economic, and public life” – is the “Percentage of seats held by women and minorities in national parliament and/or sub-national elected office according to their respective share of the population.” This data is compiled by the Inter-Parliamentary Union on the national level and UN Women on the sub-national level.¹⁸ Mexico occupied sixth place in the ranking of 193 countries, and Pakistan occupied 96th place in 2017.

As examples of traditional data for this feasibility study, Latinobarometro¹⁹ and World Values Survey (WVS)²⁰ provide reliable and valid measures of

political participation that can be used to validate indicators from social media. WVS gives national figures of political participation for Mexico and Pakistan. Latinobarometro provides indicators only for Mexico, but disaggregated for different states.

For example, Table 2 presents the comparison of the female vote in national elections for Mexico and Pakistan, using the last wave of WVS (6th wave, 2010-2014). Political participation is higher in Mexico with 66.8 per cent of women reporting in 2012 that they only voted in elections at the national level (with a margin of error +/- 3 per cent). One of the limitations of international surveys is their periodicity; they do not overlap with the time frame of the last election (2015 for Mexico and 2013 for Pakistan).

Table 2 : Comparison of results related to the female vote in Mexico and Pakistan, using WVS wave 6

	TOTAL	Mexico	Pakistan
Vote in Elections: National level			
Always	52.3%	66.8%	27.3%
Usually	24.6%	19.8%	32.8%
Never	22.0%	13.2%	37.1%
No answer	0.6%	-	1.7%
Don't know	0.5%	0.2%	1.0%
Sample size female	1,583	1,001	582
<i>Selected samples: Mexico 2012, Pakistan 2012</i>			

The hybrid approach of combining big data with traditional data sources allows to test the construct validity of big data indicators, ensuring that an inference can be made from an operationalization – the volume of tweets with hashtags/keywords

related to political participation – to the theoretical construct that the indicator is supposed to measure (e.g., political participation). We suggest selecting the constructs of the study by mapping nomological networks, which represent concepts relevant to the

¹⁸ <http://archive.ipu.org/wmn-e/arc/classifo10118.htm>.

¹⁹ Latinobarometro is an annual public opinion survey that polls citizens in 18 Latin American countries (<http://www.latinobarometro.org/lat.jsp>). Example of Latinobarometro question related to political participation: “In the last presidential election what did you do?” (I voted in the last election/I decided not to vote in last election/I was stopped from voting in last election/I didn’t have time to vote/I didn’t vote for other reasons) (<http://www.latinobarometro.org/latContents.jsp>).

²⁰ The World Values Survey consists of nationally representative surveys conducted in almost 100 countries, which contain almost 90 per cent of the world’s population, using a common questionnaire (<http://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=Findings>).

evaluation and their relationships, e.g., interest in politics, willingness to vote and access to key decision-making positions within formal political processes.

The constructs need to relate to the UN Women Theory of Change (ToC) for WPP. The ToC suggests three necessary conditions for building electoral frameworks:

- Electoral frameworks and arrangements that promote gender balance in elections.
- A cadre of interested, diverse and capable women political leaders is formed.
- Women and men are perceived as equally-legitimate political leaders as men in society.
- Women are promoted as leaders in gender-sensitive political institutions.

If these four conditions are met, the ToC projects that women will be politically empowered to realize their rights because women will have political agency and leadership in decision-making processes.

Along these lines, survey measures are used to benchmark indicators derived from social data to empirically test suitability of measuring political participation. This type of construct validity is called convergent validity as it refers to the degree to which two measures of constructs that theoretically should be related are, in fact, related. For example, counting tweets with hashtags related to gender equality with positive sentiment to infer positive attitudes towards gender equality is only possible if a positive and strong correlation is obtained between counts of tweets and valid attitudinal measures derived from survey questions across several groups. Since it is impossible to match the indicators from surveys with social media by individual cases, we can look at correlations within clusters of users with certain attributes, such as the regions where they live. Both the hashtags and the survey results can be aggregated by districts and the correlation between the two measures calculated. If a strong correlation is obtained for the group of districts, we would use big data indicators for other districts for which traditional data is not available with a certain degree of confidence. However, evaluation using aggregated data not tied to individuals will always increase the threats to the validity of results related to spurious correlations.

2 Describe the universe of tweets relevant for UN Women campaigns in Mexico (or relevant posts in Facebook in Pakistan).

We analysed all tweets contained in hashtags from UN Women campaigns and events disseminated in social media to see how they resonated on Twitter. The analysis considered level of engagement with the hashtags, demographics of Twitter users and sentiment of tweets for Spanish tweets.

Although not done in this study, topic modelling and social network analysis would provide further evidence, offering a different angle of analysis on the range of topics of the tweets and the spread through social networks of users. It, for example, could shed light on how people discuss topics/ideas and how these ideas spread through social networks, if relevant for the conclusions of the evaluation.

3 Analyse results across geographies or over time to derive insights about the contribution of UN Women programmes.

After the first step of measurement validation is established, it is possible to compare engagement or sentiment for particular regions and to analyse the evolution of certain indicators. One example is comparing Mexican districts where UN Women invested more resources with other districts, or analysing engagement on Twitter with hashtags related to the parity reform in Mexico over time.

To enable some degree of attribution to UN Women programmes, it is necessary to match cases (e.g., geographical areas) with similar political participation levels, demographic characteristics or media access to implement a matched block design. Comparing intervention and control regions within blocks would help to rule out other possible explanations for the observed results. A considerable challenge is to find matching cases: for instance, if the criterion for selecting regions to be targeted by UN Women is low participation levels among women, there would be no other regions with low participation levels as a comparison control.

However, when these block-matching regions are available, it may not be possible to ascertain positive change in regions targeted by UN Women to the impact of an intervention/campaign. Attribution or even contribution of UN Women interventions to specific outcomes would not be feasible because other development organizations and actors might also have been targeting the same regions.

What is specific for UN Women interventions is the particular sequence of events over time. Mapping a timeline of UN Women events in the target regions onto the patterns of social media results over time would make it possible to see changes associated to critical moments, following an interrupted time series design. Such a research design implies repeated measurements gathered before and after an intervention to evaluate its impact.

4 Analyse results across geographies or over time to derive insights about the contribution of UN Women programmes.

The results of the big data analysis can be triangulated with the results of the corporate evaluation and interpreted alongside traditional data, such as interviews that provide a deeper understanding of the contribution of UN Women for observed changes.

This triangulation analysis may complement initial findings, for example: big data results showing impact through large scale patterns that cannot be discerned using qualitative methods. Some divergences may also be explained, for example: big data failing to demonstrate impact because the population that benefited directly or indirectly from the programme does not use social media.



3

CASE STUDY DETAILS

3.1 DATA SOURCES

The datasets comprise traditional data and social media data and were selected according to their relevance for political participation among women in the two countries (Table 3). In the recommendations section, we suggest ways to analyse other sources of big data, such as radio stream, and to combine social media data with more traditional data (e.g., publicly accessible datasets by other organizations, electronic surveys and polls, and voters' registration data) to shed light on accuracy of Twitter indicators to predict certain political outcomes.

When selecting traditional data such as survey data it is important to consider datasets that provide geographical information for the respondents to be connected to the big data sources. Linking datasets by geographical unit is a more viable alternative to linking datasets through individual unique identifiers, as we rarely have different data sources available for the same subjects.

To link data by region, the data needs to be representative at the geographical level and not only to the country level to provide valid results. Some cross-national surveys (e.g., WVS) do not provide geographical data and others (e.g., Latinobarometro) are designed for the results to be representative of the country as a whole, but not for country regions. In cases where it is not possible to match social media and survey by geographic units, direct triangulation of big data with traditional data is not possible.

An alternative route yet to be explored is to link data sources by individuals. Respondents, for instance, can voluntarily give their Facebook name or Twitter handle when answering a survey. Although this is not the case in most of the international surveys, it may be possible if UN Women is implementing its own surveys.

Table 3 : Examples of sources of traditional and big data relevant for UN Women Evaluations in Mexico and Pakistan

	Mexico	Pakistan
Traditional data		
International surveys (e.g., WVS)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Electronic surveys and polls (UN Women Global Survey)	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Voters' registration data		<input checked="" type="checkbox"/>
Key informant interviews	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Desk reviews	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Big data		
Twitter data	<input checked="" type="checkbox"/>	
Facebook data	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Radio livestream		<input checked="" type="checkbox"/>
News data	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

3.2 PILOT STUDY IN MEXICO

The UN Women programme in Mexico focused on high-profile events, involving political leaders and social media campaigns. UN Women supported the parity reform²¹ in Mexico that consisted of an amendment to oblige political parties to observe and respect the principle of gender parity in the composition of candidates' lists for elective office.

This feasibility study was limited to the social media campaigns captured by a series of hashtags used by the Country Office in Mexico for different purposes over a long period of time. The object of study was not the UN Women programme or work, but just one aspect of the campaigns. The hashtags were from different initiatives not necessarily linked to the UN Women strategy on WPP. In some cases, they originated in other countries and by actors other than UN Women. We analysed only tweets with hashtags in Spanish that the UN Women Mexico Twitter account used. We selected the following hashtags:

#IgualdadeDeGenero	#ODS
#NiUnaMenos	#ODS5
#Planeta5050	#MujeresPoderosas
#México5050	#Agenda2030
#DemosElPaso	#ÚNETE
#NinasNoEsposas	#ATENEA

Not all hashtags are UN Women campaign-related. Some were initiated outside of UN Women (#NiUnaMenos). Some are “multi-initiative” (#ODS5). Some are not related to WPP for instance: #Niunamenos, #UNETE and #NinasNoEsposas are linked to ending violence against women and girls. #ATENEA is directly linked to UN Women’s work in WPP, as are #Planeta5050 and #México5050 (even though these two are used more broadly as linked to Agenda 2030).

3.3 PILOT STUDY IN PAKISTAN

In 2013, UN Women partnered with several civil society organizations, as well as with the local and national government in Pakistan, to encourage civic engagement in the general election and particularly from women, through the media campaign Daughters of Pakistan.

In our follow-on interviews for this study with two campaign organizations in Pakistan (Association for Gender Awareness & Human Empowerment – AGAHE) and Shazia Abbasi Consulting, we heard that social media analysis – understandably – would target only a segment of the population due to low general use (e.g., lack of access to mobile and smart phones; low digital literacy and low relevance). Both interviews with the organizations²² also indicated that while Twitter use remained generally low and remained within upper echelons of society in Pakistan, as recognised in the risks above, Facebook

use increased dramatically since the 2013 elections, particularly within youth demographics.

In the future, therefore, Twitter may be a likely big data source in terms of analysing comments and engagement of certain pages, such as those of popular celebrities or news. Bearing these limitations in mind, we conducted a pilot analysis of the UN Women Facebook page in Pakistan.²³ It is worth noting that the choice of the Facebook page should be guided by the principle of political neutrality. It would not be recommended to analyse Facebook pages of political parties or candidates, as it can be interpreted as supporting the featured party. Furthermore, these pages typically have extremely biased samples, because people self-select strongly based on their views. Consequently, the analysis of the discussion is rarely useful, as it represents a homogenous view.

²¹ On 10 February, 2014, Mexico, passed an amendment to Article 41 of the Federal Constitution stating that political parties should put in place “...rules to ensure gender parity in the nomination of candidates in federal and local congressional elections.”

²² Interviews with Shazia Abbasi, Shazia Abbasi Consulting, 19 September 2017 and with Qamar Iqbal, Association for Gender Awareness & Human Empowerment (AGAHE), 20 September 2017.

²³ <https://www.facebook.com/unwomenpakistan/>.



In addition, both interviewees noted that radio should be considered as a source of big data. In particular, responses to radio shows (e.g., phone-ins) represent a potentially valuable source when such civic engagement discussions did take place. While UN Women produced awareness-raising short videos and jingles (e.g., on the importance of voting) that were played on television and radio, they did not monitor the responses. More recently, the Country Office in Pakistan has been developing content for and disseminating messages on

inclusion, gender equality, health, education and WPP through community-based radio in Quetta. While such live discussions take place, particularly on popular FM radio (e.g., FM 94, 95 and 96 in Punjab), we could not find evidence of these being recorded and easily accessible for analysis. There is a language challenge, as such discussions are likely, in the main, to be in Urdu or Punjabi (or Pashto, Sindhi and Balochi).



4

DATA PROCESSING, ANALYSIS AND RESULTS

In this section, we present a breakdown of the steps and associated challenges for accessing and analysing social media data with the view of offering a protocol for data analysis available for other UN Women evaluations using big data. We include the data processing method for Twitter and Facebook. For Twitter analysis, we also present the outcomes for each step. For the Facebook data, we describe the

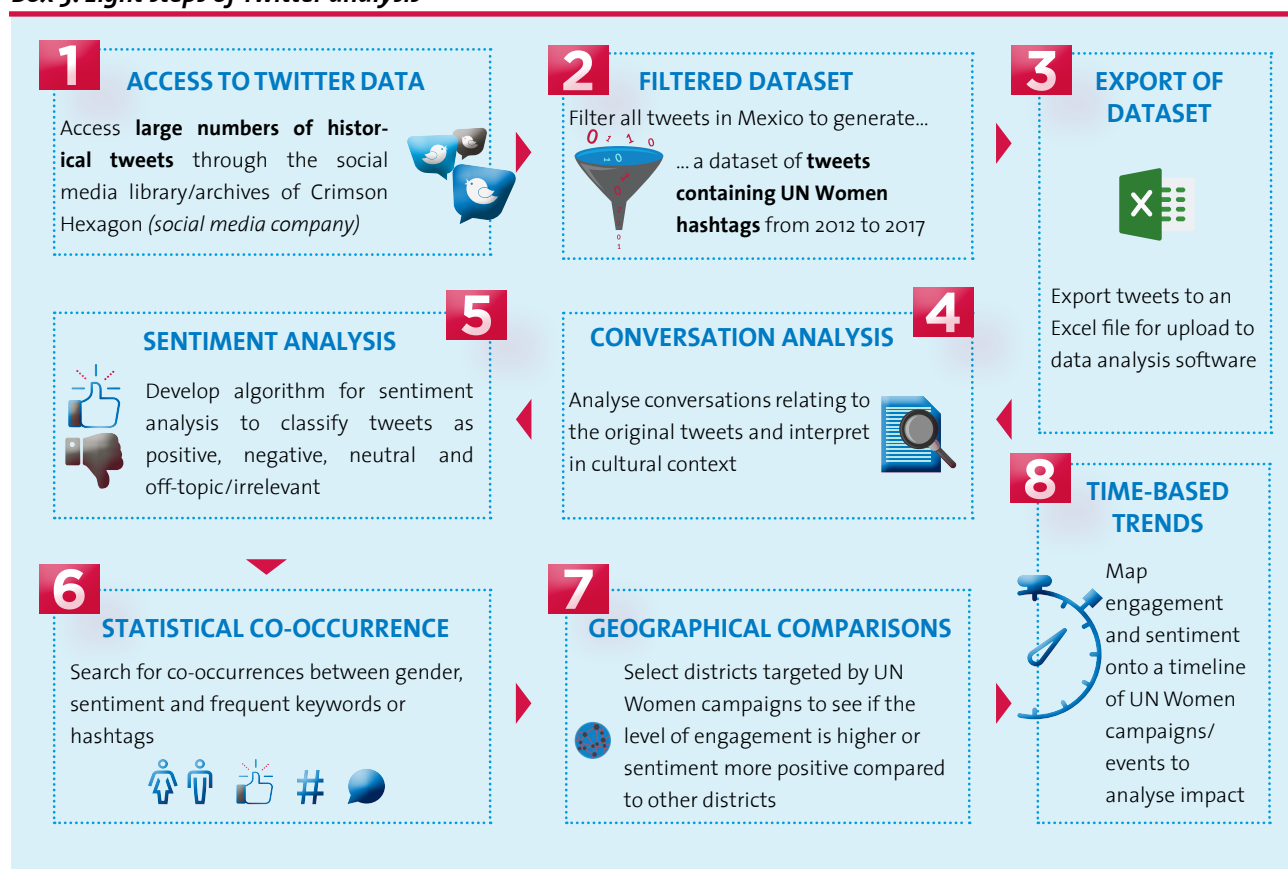
stages of the data analysis but not the outcomes, as it was not possible to develop language models for Urdu in the time frame of this study. Analysis of radio streaming is discussed in terms of the process and recommendations; the analysis was not implemented due to the impossibility of obtaining radio data within the time frame of this study.

4.1 TWITTER ANALYSIS AND RESULTS FOR MEXICO

The Twitter analysis consists of eight steps explained below. For each of the steps, we explain the procedure and the results that accompanied them. Results are discussed in the next section (5: Findings). The outcomes in this study comprise the number of tweets with hashtags used by UN Women Mexico and the sentiment of the tweets using the hashtags. The sentiment may evaluate

attitudes towards gender equality issues (broader than female political participation), but this needs to be established through triangulation with other data sources. Attitudes are not necessarily related to behaviour, since material barriers, social norms or other ways of social control (by church, political groups or governments) may influence behaviour.

Box 5: Eight steps of Twitter analysis





STEP 1: ACCESS TWITTER DATA

Twitter's website search is fairly limited in features, not allowing filtering by region or country, and it does not have any time-based tools. As a result, even exploring initial trends or datasets –including getting a sense of the level of conversational engagement – can be challenging.

Accessing large numbers of historical tweets poses a number of challenges. Twitter's public API, by default, does not allow searching tweets older than two weeks. If one wants to access large volumes of tweets through the public API, gathering streaming tweets is more convenient, but with the drawback of accessing tweets only from the current date to a date in the future. With adequate planning, one can start collecting Twitter data as soon as an intervention begins.

Several channels permit the gathering of tweets through the Twitter public API for free, such as libraries from Python (tweepy, python-twitter), packages from R (Twitter) or other dedicated software (e.g., COSMOS).²⁴ Further issues were found with the public Twitter API, as it does not provide the ability to recall whole conversations (responses to a tweet) but only original tweets, thereby requiring a scraping approach to obtain tweets with responses. The public API has only limited information related to location and gender of users, however, several tools enable the prediction of gender based on Twitter handles, some of which are implemented by Twitter data providers.

Given the impossibility of accessing historical Twitter data through the public API, access to Twitter data in this study was done through Crimson Hexagon, a social media company that provides a social media library²⁵ with an archive of tweets dating back from 2012. Crimson Hexagon provides parsed location data (e.g., district and state), as well as gender, for a large number of users (mainly based on prediction models). However, the quality of this 'guessed' data may not immediately be trusted, as the algorithms for predictions are not provided. We demand some caution when using demographics data of tweets.

The tweets are searchable using 'monitors' that are queries related to a set of variables, such as hashtags and location. Crimson Hexagon does not have an on-demand search tool. This considerably slows down the exploration of tweets based on keywords.

Access to the historical Twitter API may provide quicker extraction than Crimson Hexagon can offer, especially with larger volumes of data (datasets with 10,000 to 50,000 tweets are not an issue for either method). However, this data is not parsed, and demographic data is only available for a small group of users (about 10 to 20 per cent for gender and 1 per cent for location). In order to get unrestricted access, an application process through Global Pulse should be started early to avoid delays in accessing the data.

STEP 2: FILTER AND INSPECT THE INITIAL DATASET

The sampling frame for Twitter data consisted of the universe of all tweets in Mexico from which a subset of tweets was selected. We selected tweets from 2012 to 2017 that contained at least one of the hashtags that was used by the Twitter account of UN Women Mexico (cf. Table 4). Hashtags became a method of identifying the theme or subject of a tweet or other social media posts in 2007 and have

continued to evolve in usage over the last decade. In particular, use of hashtags vary inevitably as social or political issues evolve, as pertinent issues have a title that suits a tag (e.g., #feminicidio).

There are other situations when hashtags of keyword need to be discovered. We suggest combining manual and automatic techniques. For

²⁴ This software is developed by the Social Media Lab from University of Cardiff UK with funds from UK Economics and Social Research Council. It is free for academic institutions and non-profits; see <http://socialdatalab.net/software>.

²⁵ <https://www.crimsonhexagon.com/>

example, start with manually selected hashtags and keywords relevant for the topic (e.g., validated by domain experts) and through co-occurrence, discover other hashtags and keywords, adding them to the initial group. There are more sophisticated approaches to extracting tweets, based on discovery of 'surprise' words in domain-related tweets (selected manually) that deviate from their expected distribution or based on the overall corpus of tweets (e.g., Method 52 described in Bartlett et al., 2014).

Using Twitter's search, various hashtags and search terms were found, until a rough idea of a good search term (selection and combination of hashtags) was established. After this, Crimson Hexagon was

used to construct a monitor limited by hashtags, a timeframe (2012-2017) and the area restricted to Mexico. The dataset included followers engaged by users, new followers, and followers' followers in conversations and hashtag contributions.

After this data has been processed by Crimson Hexagon, one should have a long history of tweets using the hashtag(s), so that one can establish if the numbers are worth processing and correspond chronologically with the campaigns and events on the ground. The final dataset consisted of a total of 65,512 tweets, including tweets with identified hashtags and responses to these tweets.

Table 4: Hashtags,²⁶ number of tweets containing the hashtag and date of first and last tweet

Hashtag	Number of tweets	Date first tweet	Date last tweet	Lifecycle of tweets (months)
#IgualdadeDeGenero	24	2012-02-23	2017-07-03	64
#NiUnaMenos	36710	2012-11-25	2017-08-20	57
#Planeta5050	5971	2015-03-06	2017-08-20	30
#DemosElPaso	17148	2015-03-08	2017-08-20	29
#NinasNoEsposas	3966	2016-02-03	2017-08-20	19
#ODS	482	2016-02-28	2017-07-05	16
#ODS5	64	2016-03-08	2017-07-05	16
#Agenda2030	35	2016-03-08	2017-08-17	17
#México5050	2945	2016-04-09	2017-08-20	16
#ATENEA	13	2016-05-19	2017-05-21	12
#MujeresPoderosas	4	2016-06-21	2017-03-08	9
#Únete	12	2016-10-18	2017-08-08	10

²⁶ The hashtags count considered variations of the original (e.g., small caps and misspellings).

The resulting dataset showed that two hashtags were used from 2012 (#IgualdadeDeGenero and #NiUnaMenos), but most hashtags were used for the first time in 2015 and 2016. Before November 2015, the count of tweets is fewer than five per month (Figure 1).

The most popular hashtags are #NiUnaMenos (started in 2012 with 36710 tweets) and #DemosElPaso (started in 2015 with 17148 tweets).

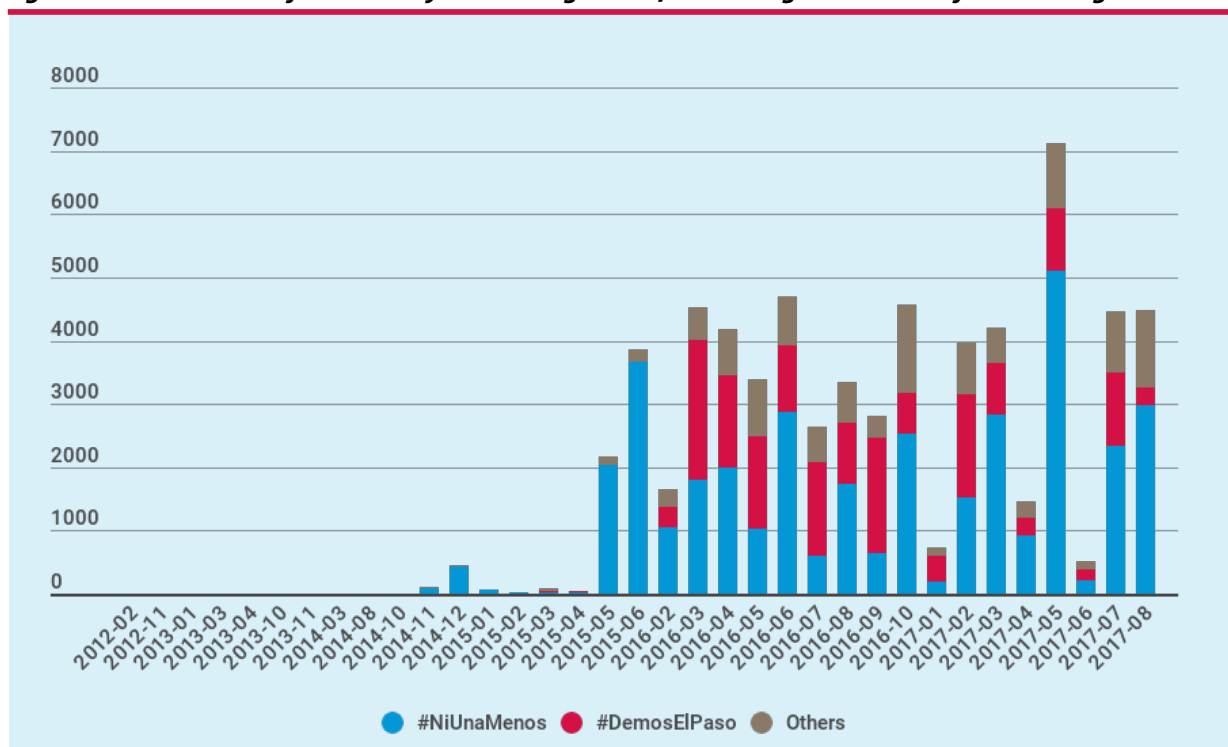
- **#NiUnaMenos** is a social media movement that campaigns against gender-based violence toward women. Originated in Argentina as a reaction by the public for a femicide case, the movement spread to other Latin American countries. This is a very generic hashtag used for numerous initiatives, activism and specific activities across the region.
- **#DemosElPaso** was launched on International Women's Day (8 March 2015) with the objective of showing steps the government is taking to

help women and girls to achieve their potential, e.g., creating programmes to eradicate violence against women; enhancing women's participation in decision making; creating or changing laws/policies; and launching campaigns that promote gender equality.²⁷

The most engagement with the set of hashtags was observed during three distinct periods:

- **November 2014 to June 2015** with a peak on June 2015, due mainly to #NiUnaMenos.
- **February to November 2016** with a peak on June 2016, due mainly to #NiUnaMenos and also #DemosElPaso.
- **February to August 2017** with a peak on March, May and August 2017, due mainly to #NiUnaMenos and #DemosElPaso.

Figure 1: Mexican tweets from January 2012 to August 2017 containing at least one of the hashtags



²⁷ - <http://www.unwomen.org/-/media/headquarters/attachments/initiatives/stepitup/stepitup-brochure-es.pdf?la=es&vs=2219>

Only 748 tweets (1.1 per cent of the dataset) are authored by UN Women Mexico (@ONUMujeresMX) but they generated 18,681 retweets (28.5 per cent of the dataset). The most used hashtag by UN Women Mexico was #DemosElPaso (537 tweets), used for the first time in February 2016, marking the start of the uptake of this hashtag by the public (Figure 2). However, the number of retweets of original tweets with #DemosElPaso from the UN Women Mexico account is 12,933 which corresponds to 75 per cent of all tweets with the hashtag, showing that the spread of this hashtag through Twitter can confidently be attributed to the UN Women Mexico campaign.

In turn, #NiUnaMenos was used in only 12 tweets by UN Women Mexico, appearing for the first time in April 2016. This result shows that #NiUnaMenos was used primarily by the public long before UN Women Mexico started using it (since November 2012). The number of retweets with the hashtag #NiUnaMenos from UN Women Mexico is 232, which corresponds to less than 1 per cent of tweets with this hashtag. This shows that the pulse of #NiUnaMenos campaign on social media was not (directly) related to any UN Women Mexico campaign.

Other hashtags frequently used by UN Women Mexico were #Planeta5050 (187 tweets), #NinasNoEsposas (104 tweets), #Mexico5050 (64 tweets) and #ODS (37 tweets). All these hashtags were taken up by the public either with re-tweets or by generating original tweets.

Additional hashtags not used to filter the Twitter data also appear alongside the identified hashtags. Some of these additional hashtags could have been added to the queries to filter the dataset in a subsequent step. But as they were not used by UN Women Mexico, we considered that they were not as helpful to understand the pulse of social media campaigns. Examples are #8M (831 tweets), #NosotrasParamos (773 tweets) and #24A (646 tweets) that refer to marches and protests against gender-based violence and salary/work inequalities. Other hashtags that appear in the dataset are: #VivasNosQueremos (3638 tweets), #Feminicidio (1755 tweets), #DiaInternacionalDeLaMujer (1631 tweets), #SiMeMatan (1266 tweets) and #DíaNaranja (83 tweets).

Box 6: Wordcloud of hashtags used to filter twitter data (2012-2017)



STEP 3: EXPORT, ORGANIZE AND CHARACTERIZE THE DATASET

Exporting the data from Crimson Hexagon is limited to 10,000 tweets at a time, but as one can import specific timeframes, getting a large quantity of data is straightforward. This procedure is time consuming because it requires repeating the queries multiple times. Once out of Crimson Hexagon, an Excel file is created that can be imported to any data analysis

software (Figure 2) with certain variables to inspect their characteristics and analyse associations (Table 5). These were made into Python objects for ease of processing by Pandas, which is a data analysis library, and into R dataframe for data visualisations (both Python and R are programming languages widely used among data scientists).

Figure 2: Dataset obtained from Crimson Hexagon (imported to R as a dataframe)

category	ch_gender	city	content	country	emotion	followers	following
Basic Neutral	F	Puebla	RT @ONUMujeresMX Vamos por un #Planeta5050 #D...	Mexico	NA	1059	219
Basic Neutral	NA	Mexico City	RT @RuidoEnLaRed #NiUnamenos Marcha en la UNAM...	Mexico	NA	338	1871
Basic Neutral	NA	Mexico City	RT @Tania_Tagle Por favor lean todo este hilo. Te no...	Mexico	NA	2024	1303
Basic Positive	NA	NA	RT @Alejolgao #NiUnaMenos esto tiene que parar de ...	Mexico	NA	150	168
Basic Negative	NA	Mexico City	RT @Patysd12 @CARLOS_CAVA Ahora resulta q las m...	Mexico	NA	9921	8634
Basic Neutral	F	Mexico City	RT @ONUMujeresMX Mismos derechos para todas y t...	Mexico	NA	1020	2951
Basic Neutral	NA	Mexico City	RT MASmx2016: En la marcha de la #UNAM #NIUNAM...	Mexico	NA	166	69
Basic Neutral	NA	Cuernavaca	RT @ParodiaSoto #NiUnaMenos parceiros es hora de r...	Mexico	NA	1251	2034
Basic Neutral	M	NA	RT @ONUMujeresMX Empoderar a las mujeres es emp...	Mexico	NA	56	317
NA	NA	Mexico City	#NosEstanMatando #NiUnaMenos #AsÃmispornos htt...	Mexico	NA	1496	946
Basic Neutral	F	Mexico City	RT @bombonetass @Cruzamaranta @CardeMonter @...	Mexico	NA	1649	537
Basic Neutral	F	Mexico City	RT @ONUMujeresMX Â¿SabÃas que el 21.5% de las m...	Mexico	NA	2297	1985
Basic Neutral	NA	NA	RT @radioamlo #Viernes5 Marcha Interna Contra la Vi...	Mexico	NA	3166	2051
Basic Negative	M	Mexico City	RT @MontseNarro Ser mujer en #MÃxico es saber q...	Mexico	NA	329	546

Table 5: Fields for Twitter query in Crimson Hexagon

Variables	Description	Summary
GUID	Globally Unique Identifier (from Twitter)	-
Date (GMT)	Date and time of the tweet	Between Feb 2012 and Aug 2017
URL	Web address	-
Content	Tweet text	-
Author	Twitter handle	65,512 different users
Name	Twitter name	65,512 different users
Country	Country of user	100% in Mexico
State/Region	State/Region	Mode is Distrito Federal with 59.7% tweets
City/Urban Area	City/Urban Area	Mode is Mexico City with 62.8% tweets
Emotion	Types of emotion (if any)	Mode is sadness with 159 tweets
Klout Score	Rate of online social influence (1-100)	Mean=39.3, standard deviation=12.49
Gender	Gender of user	55.1% users are female; 44.9% are male
Posts	Counts of posts by the user	Median=7735, max=1018618
Followers	Counts for followers	Median=554, max=6854524
Following	Counts for following accounts	Median=633, max=288914



The tweets in the dataset were authored by 65,512 different users, mainly women (55.1 per cent) and users living in the Mexico City area. The users' average rate of online social influence is 39.3 (out of 100). Depending on the objective of the evaluation,

social network analysis would help to reveal the online network of these users and their degree of influence within their network. These analyses could answer questions related to reach and spread of information through Twitter.

STEP 4: GATHER AND ANALYSE CONVERSATIONS

By scraping Twitter, the conversations relating to the original tweets gathered through Crimson Hexagon are recorded, including date/time, content and user handle. In this way, one obtains a lot more data than the raw hashtags, particularly spontaneous and human reactions. These can be analysed in similar ways to the original tweets and potentially separated for conversational or topic analysis. These conversations need to be interpreted in their

cultural context, including the social media culture of a particular country. As such, the analysis can be enhanced by discussing the conversations through focus groups with Twitter users and/or validated by media and domain experts from the country. In this dataset, re-tweets are very frequent (76 per cent of dataset). There are only a few tweets that are responses to others, but not enough to justify a conversation analysis.

STEP 5: SENTIMENT ANALYSIS

Using Stompol, a well-established Mexican political tweet dataset, a sentiment model is trained using machine learning²⁸ that predicts the sentiment of each message (e.g., for methods for sentiment classification, see Badhane, Dalal and Doshi, 2015). A Spanish tokenization, a process that breakdowns the text into stemmer (reduced words) and separator, is pre-applied to reduce the number of words. This potentially loses some subtlety in language, but it means that the training set is not too big. The resulting categories are positive, negative, neutral and off-topic/irrelevant.

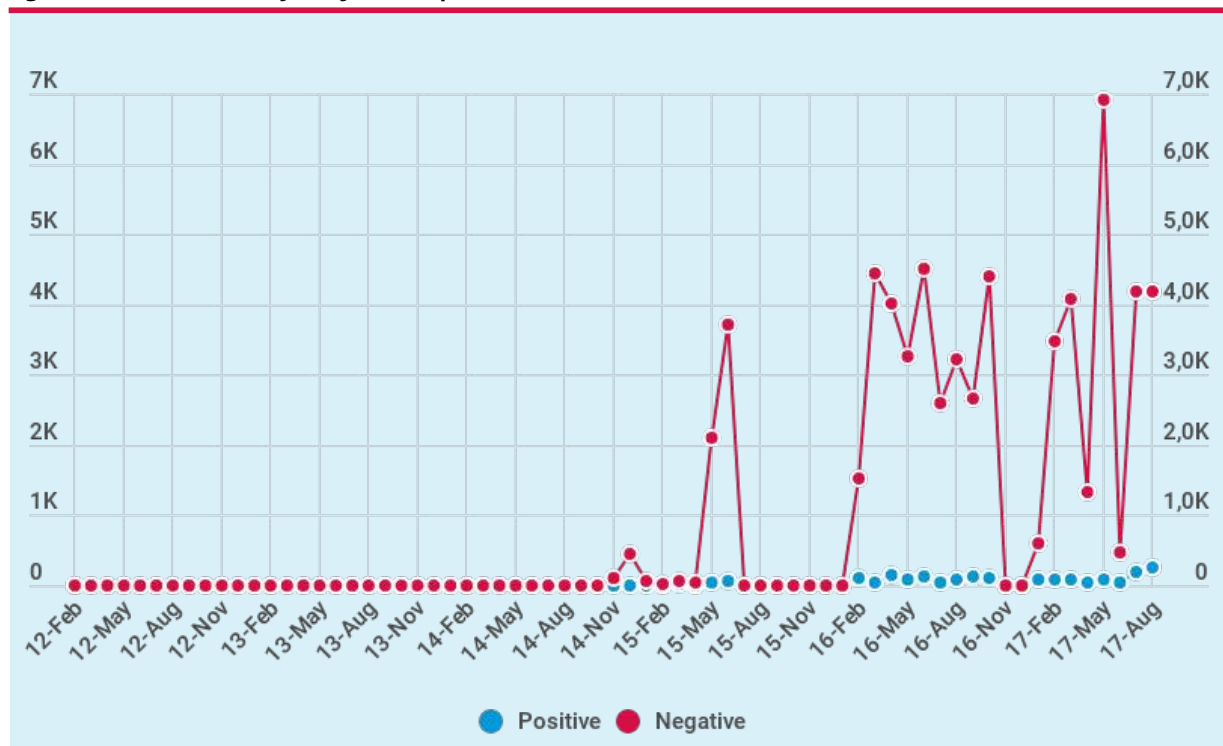
It is important to note that algorithms for sentiment and feelings are comparatively weak and unreliable when compared to human coders due to interpretation, sarcasm and other contextual linguistic features. Human coders are preferable if they have tools to code in a quick and cost-efficient way as part of a crowd-coding project where accuracy is enhanced and inter-coder reliability established (Haselmayer and Jenny, 2017).

The sentiment of the tweet reflects the nature of the topic. For example, topics related to fight, inequality or violence have a negative code due to the emotional connotation of these words (negative). As such, the sentiment does not reflect whether the user agrees or disagrees with the topic, but whether the tweet refers to a negative topic.

Figure 3 depicts the analysis of sentiment over time only for tweets with a clear, positive or negative tonality. When disaggregated by gender, the same pattern is obtained suggesting that there is no noteworthy difference on the sentiment of men and women regarding this topic. The volume of tweets increases from November 2014 as seen before (Figure 2), and this increase corresponds to tweets with negative topics. Given that the most frequent hashtag is #NiUnaMenos, which refers to violence against women, this result is not surprising.

²⁸ It consists of a support vector machine (SVM), which is a supervised learning model with associated learning algorithms that analyses data used for classification.

Figure 3: Sentiment analysis of tweets per month



STEP 6: STATISTICAL CO-OCCURRENCE OF WORDS, SENTIMENT AND DEMOGRAPHICS

Once one has a dataset labelled with sentiment plus gender plus content, it is easy to inspect co-occurrences between demographics (gender and district), sentiment, hashtags and words. The

procedure and algorithm to look at co-occurrences based on words is explained in the Appendix. One might look to see if overall replies are more negative or tend to use very different terminology.²⁹

STEP 7: GEOGRAPHICAL COMPARISONS

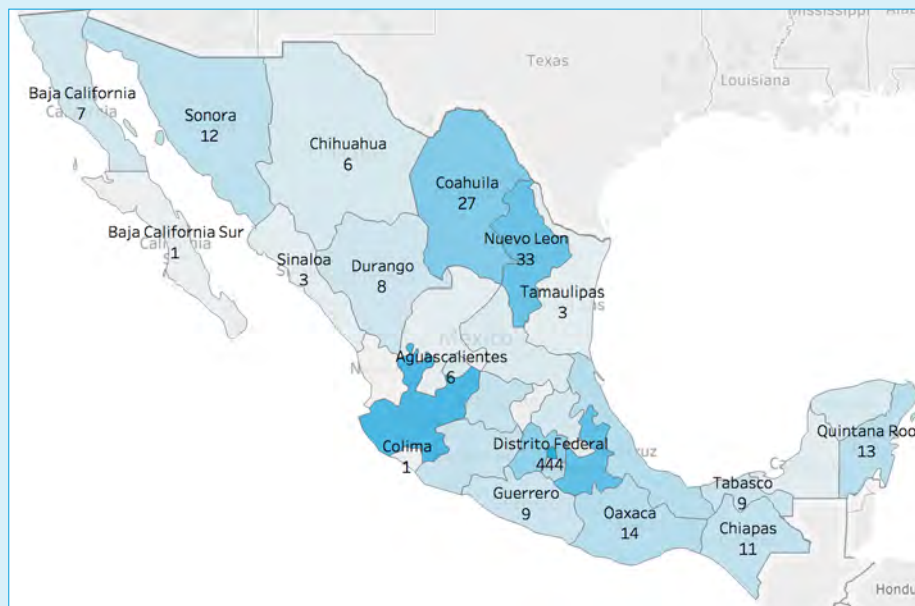
This step comprises implementing algorithmic models, such as comparison of block matched groups, to evaluate the feasibility of attributing particular trends observed on Twitter to UN Women interventions or campaigns. The objective is to determine ‘weak’ attribution based on natural experiments within larger datasets. When it is not possible to conduct a randomized control trial (RCT), the robustness of the evaluation depends on the design that comprises the datasets that will be used, the reasons for their selection, their characteristics and biases (representativeness) and how well they align with the ToC. In addition, the variables related to the intervention should be rich and varied, and there must be no interference with the real-world implementation.

Figure 4 depicts heatmaps for Mexican districts, according to the number of tweets with positive and negative sentiment in 2016 and 2017. One could compare different districts to understand if people in districts targeted by UN Women campaigns are showing more engagement with gender equality issues on Twitter (either positive or negative topics). It is important to consider baselines to compare increase rather than absolute engagement. For example, to evaluate the impact of a campaign/intervention implemented in 2007, one could match two districts with similar levels of positive and negative sentiment in 2016 and compare them in 2017 to see if the districts/area targeted presents higher levels of engagement compared to the ‘control’ districts.

²⁹ – Additionally, to understand if the data is reflecting particular interventions, one might establish a model where each event created pulse (perhaps a pulse based on an exponential decay function), and compare this predicted model with the observed tweets to establish correlation.

Figure 4 - Heatmaps for tweets with positive and negative sentiment in 2016 and in 2017

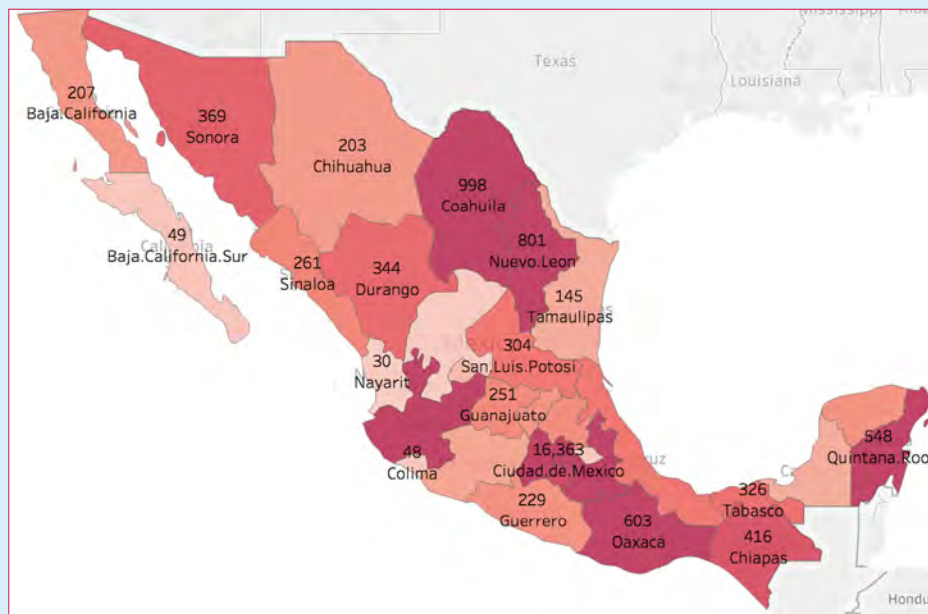
Positive sentiment 2016



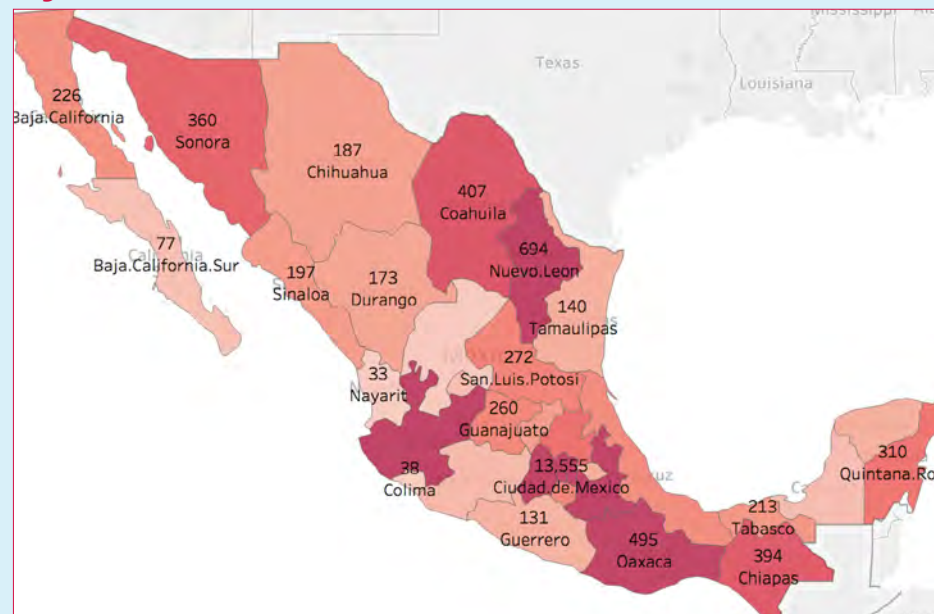
Positive sentiment 2017



Negative sentiment 2016



Negative sentiment 2017



One possible approach to matching blocks is to select different regions within the same country that correspond to contrasting cases for UN Women interventions. Characteristics of these areas (matching covariates), such as population size, female participation levels and voter registration, need to be controlled for to remove effects of these covariates in the comparison of groups. It will be challenging, however, to find regions that may be appropriate matches, as they can vary considerably in terms of population, cultural and economic characteristics. Yet, it is possible if a natural experiment is incorporated in the design of the intervention.

Machine learning techniques can be applied to big datasets very effectively to simulate experiments

addressing selection bias. Based on the idea of naturally occurring experiments, propensity score matching allows analysis of data from an observational study to find cases and regions that can be contrasted. For instance, different sub-groups comprised of people who share very similar Twitter engagement patterns and prior sentiments about the topic of interest, make them equally likely (i.e., matched propensity) to engage in a Twitter hashtag group or conversation. Some within these sub-groups of people may or may not engage in a specific UN Women campaign and/or present different levels of engagement. These naturally occurring variations on engagement within each matched sub-group will provide counterfactuals for cause-and-effect inferential analyses.

STEP 8: TIME-BASED TRENDS

Since all the data is time coded, it is trivial to plot various metrics, such as engagement by gender, or sentiment over the relevant timeframes and regions. In order to understand the impact of UN Women campaigns/events on social media, the Twitter data can be combined with a timeline of interventions. One should bear in mind that the impact of UN Women interventions can be

confounded with other events that occurred at the same time. Twitter outcomes (e.g., engagement, sentiment) and longitudinal trends for other data (e.g., voters' registrations) can be plotted to inspect whether the trends agree with those for Twitter and are linked to UN Women interventions, following an interrupted time-series design.

4.2 FACEBOOK ANALYSIS FOR PAKISTAN

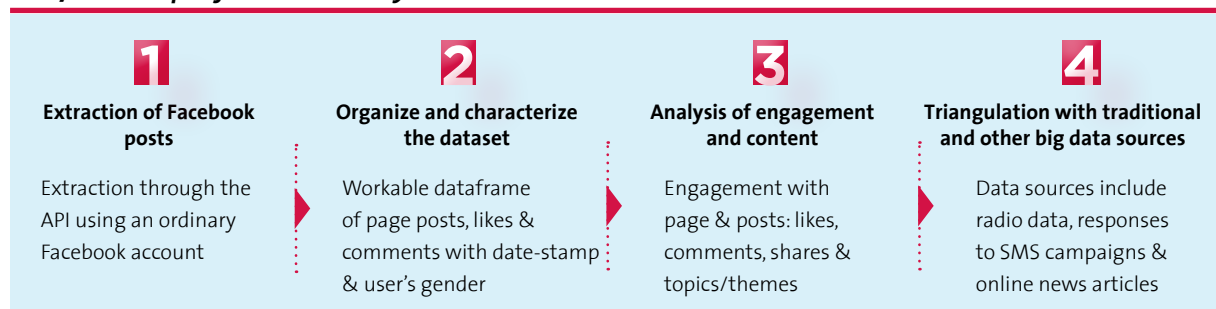
Facebook is the most popular social media platform in Pakistan, as elsewhere in the world.³⁰ Globally, it mostly consists of private or semi-private discussions which may pose ethical issues as it may reveal sensitive personal details that may place users at risk. Public Facebook pages associated with political parties or candidates are sensitive for inclusion in the evaluation (as it may question UN Women impartiality) and present homogeneous views. Facebook groups based on location or a topic might have a more diverse engagement, but they tend to be limited in numbers of users and reach. Large pages for institutions generally have poor engagement, even though they have large numbers of followers or likes.

For this feasibility study, we selected the UN Women Facebook Pakistan page. UN Women Pakistan has the second largest followership among UN Women Country Offices with over 50,000 followers (greater than India and the Asia-Pacific Regional Office). There have been successful advocacy campaigns with more than one million views. We analyse the engagement metrics from 2016 to 2017. It is important to note that as Facebook is expanding in Pakistan, patterns of engagement may also change, making it invalid to compare trends over time, particularly among socio-demographic groups.

The Facebook analysis consisted of four steps:

³⁰ Facebook use dwarfs Twitter - over 31 million Facebook users in Pakistan, and 3.1 million Twitter users, although there are concerns around duplicate and fake Facebook IDs. See: <https://www.geo.tv/latest/131187-Over-44-million-social-media-accounts-in-Pakistan>.

Box 7: Four steps of Facebook analysis



STEP 1: EXTRACTION OF FACEBOOK POSTS

Extraction from Facebook is done simply through the API tool which can be tapped into fully without the need for special access. An ordinary Facebook account is used and an app registered for the

extraction. The online temporary API key tool was used for simplicity to generate a temporary API key for the downloading of data.

STEP 2: ORGANIZE AND CHARACTERIZE THE DATASET

The API³¹ call to page posts and comments provides a paginated list of posts for a page and their comments and likes. The page identification is found by browsing the page and looking for the 15-digit number at the end of the URL. The conversion

of these responses into workable and useful conversations is a simple task. Facebook provides information such as the commenter's name, which allows basic gender guessing and time and date for the comments.

STEP 3: ANALYSIS

In contrast to the number of followers, we found a low engagement in the UN Women Pakistan Facebook page through likes, shares and comments (Figure 5). Most of the posts were from 2017 (654) and only 72 in 2016. The language of the post is mainly English (over 90 per cent of the posts) and, therefore, most of the comments are also in English although there are a considerable number of comments in Urdu. This poses some questions in terms of reaching and encouraging people who speak vernacular languages to engage with the page. Most of reactions to the published posts are likes (on average 58 in 2017) and shares (on average 44 in 2017) and a few comments (on average five comments per post in 2017).

Human interest stories tend to attract the most comments, such as a Facebook post on 11 October

2017, which reads: "Meet Fatima ... my relatives used to mock my father for raising six daughters. He never treated us any less than his sons." This post attracted 206 reactions, 26 shares, and 28 (all positive) comments. Posts on Facebook, such as those on gender equality citing SDG 5, gained 15 reactions but no comments.

One of the key factors in deciding the sophistication and ambition of an analysis plan is based around the level of resourcing and available tools for a particular language. Some languages (English mainly) are extremely well-resourced, with automated understanding, robust speech transcription and beyond all easily available with a moderate amount of work. Many more languages (widely-spoken languages, such as Spanish as Mandarin) have some easily-available resources: ready-to-use stemmers,

³¹ The page's posts with comments are extracted by an API call to /v2.10/{page_id}/posts?access_token={token}&fields=comments,message,created_time,likes.limit(100).

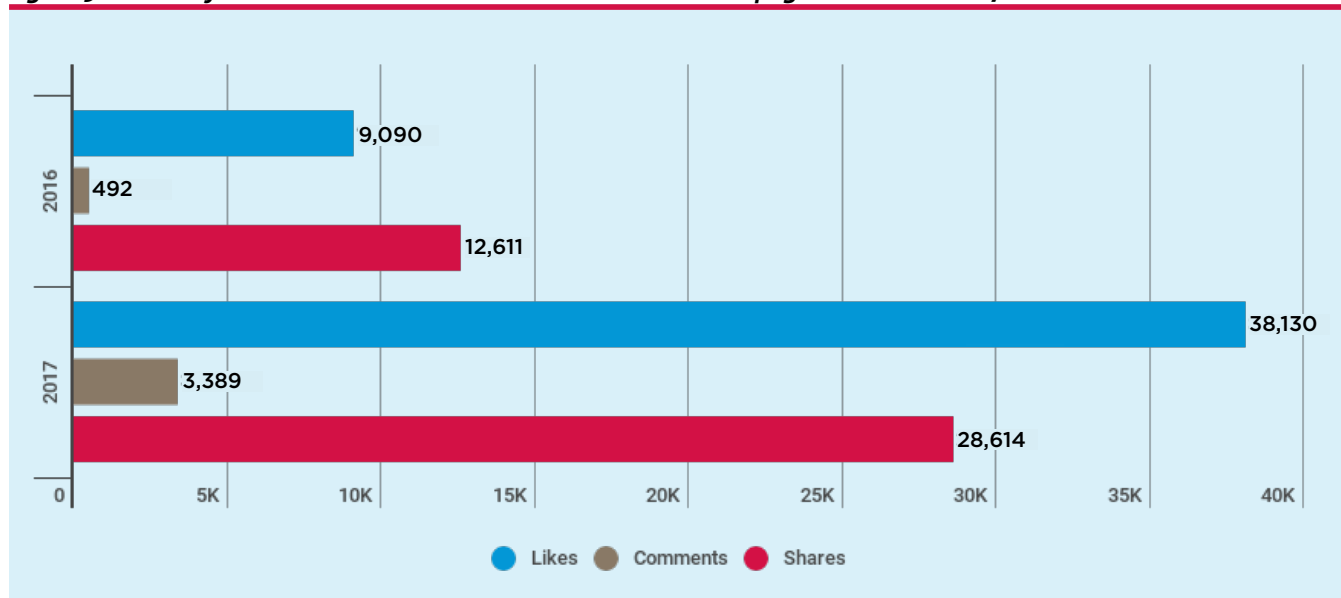
well-curated stop-word lists and pre-labelled datasets. Less- resourced languages make a quick analysis challenging, as one cannot build off easily-accessible previous work. Analysis is still possible, but may have lower accuracy or reveal less interesting insights, especially with lower volumes of data (e.g., around 10,000 tweets).

Although Urdu is a less well-resourced language than Spanish, many types of analysis and cleaning can be carried out on it before it is analysed. The main objective of these cleaning steps is to make the vocabulary of words smaller, so that analysis

efforts are quicker or more effective (machine learning works better on smaller vocabularies). This allows subsequent semi-automated steps to go quicker and more effectively.

Depending on the target of the analysis, the next steps are often to identify the most common words within the dataset as either stop words, topic words, or other, and possibly the polarity of various words (indicating negative or positive inclination). This allows a basic sentiment analysis and to track various topics over time.

Figure 5: Counts of reactions in the UN Women Pakistan Facebook page in 2016 and 2017



STEP 4: TRIANGULATION WITH OTHER DATA SOURCES

In conclusion, what our initial feasibility study found for Pakistan was that while social media – particularly Facebook – analysis would provide some useful insights, we should not discount other sources, including radio data, responses to SMS campaigns (though we could not access data for this) and responses to newspaper articles online.

However, an additional challenge is that the most representative responses may well be those offline, such as street plays and campaigns, of which we have no big data at all. This is the ever-present challenge of using big data to evaluate impact in low-income countries.

³¹ The page's posts with comments are extracted by an API call to `/v2.10/{page_id}/posts?access_token={token}&fields=comments,message,created_time,likes.limit(100).`



4.3 ANALYSIS OF RADIO DATA

Although radio is a less-commonly gathered social media source, in some communities the community radio is one of the big social venues, with well-known hosts and regular callers and with several segments for feedback and opinionated speech from audience members.

The initial idea was to collect radio data for Pakistan, as UN Women campaigns were widely disseminated through radio and TV. But, as there was no interactive component in the radio show campaigns and radio streaming data was not historically available (before and just after 2013 elections), we opted for using Facebook data for Pakistan.

In future evaluations, we suggest radio programmes be designed to gather information that will be useful for the evaluation, either from voice or SMS. Several open-access platforms (e.g., RapidPro) allow the integration of a shortcode, usually a free number that is managed by the organization carrying the evaluation while the radio stations have access to a platform to manage the volume and content of the calls and SMS. By inviting participation from audiences, for example through simple questions or contribution to debates, radio stations can be turned into social data hubs, gathering data in a social and relevant context at a large scale. This type of data offers a rich big data source that can

be linked to other data through amplified asking, for example an SMS survey that asks information about the demographics, knowledge or behaviour of participants. Some careful considerations should be taken to establish an archiving system for feedback and for the radio shows, as many stations do not record shows.

Analysis of the radio shows takes more hands-on work, but is tackled primarily using an automated transcription tool, such as Google Cloud's Speech API, which has facilities for many languages, including Pakistani Urdu. This API provides time-stamps for every word, allowing later analysis of the speaker (including an estimate of gender) and/or tonality. Using these time stamps, individual words or sentences can be cut out and analysed by tools, such as UN Global Pulse's audio gendering tools. Other tools exist for analysing tonality and other metrics, which may help with more in-depth analysis. Getting the volume of data needed to justify such involved analysis, however, would require careful recording and coordinating. This data could be highly relevant and rich, featuring live question-and-answer sessions and opinions of many people in the community.



SUMMARY OF FINDINGS & RECOMMENDATIONS

5. SUMMARY OF FINDINGS

The findings highlight that social media data should be seen as evaluative due to their temporal dimension and capture of people's opinions embedded in their social context. But due to population under-coverage, self-selection bias and focus on the digital sphere of action, social media data should not be analysed in isolation. Big data should be considered as another source of rich data, which is more connected to individual's realities and with wider topic coverage, compared with staged and controlled studies requiring complex data collection.

Findings on Twitter

- Twitter appears more appropriate for evaluating UN Women's interventions aimed at fostering political participation and attitudes towards gender equality.
- Social network analysis can help to reveal the online network of users and their degree of influence within their network. This type of analysis may be able to answer questions related to the reach and spread of information through Twitter.
- Given the short life of hashtags, longitudinal analysis based on the same hashtags is not meaningful.
- Analysis and interpretation of conversations within a cultural context can be enhanced by focus groups with Twitter users and/or validated by media and domain experts from the country.

Findings on Facebook

- Private or semi-private discussions may pose ethical issues because they can reveal sensitive personal details that could place users at risk.
- Many pages from organizations do not contain much discussion; pages associated with political or social issues have biased samples, as people self-select strongly based on their views on those issues.
- Other sources hold more promise, such as radio data, responses to SMS campaigns and responses to newspaper articles online.

Findings on radio data

- Radio can be a significant social venue.
- Historical streaming of radio data is not always present.
- Radio programmes can be designed to gather useful information for evaluation through voice or SMS.
- Requires careful recording and coordination to ensure large volume of data is available for analysis, but can be highly relevant and rich (e.g., documenting community conversations).

6. RECOMMENDATIONS AND FUTURE WORK

Recommendation 1

Understand the bigger picture of big data in a country before considering it as a source for evaluation.

- Twitter and Facebook have digital architectures encouraging certain styles and degrees of engagement that need to be understood with all their cultural specificity before embarking on an evaluation using these platforms. Different big data platforms could prove culturally insightful for understanding the context variables to attend to when evaluating using big data.
- Understanding representativeness is not a binary question about those who are on social media or not. There is also the question of varying degrees of social media use that will influence the representation of social media users.



- Another issue is assuming that majority languages are representative of countries (e.g., Urdu or Punjabi in Pakistan), which will exclude those who use other languages on social media.
- A challenge of using big data for evaluation in low-income countries is that representative responses of many will be those offline that can be gathered through street plays and offline campaigns.
- Findings need to be interpreted through the lens of the cultural, language and media context where individuals belong. If possible, discuss the results through key informant interviews (KII) or focus group discussions (FGD) with people who provided the data.

Recommendation 2

Big data should be incorporated in the design of the evaluation from the outset.

- Identifying data sources early in the design stage allows for planning and collecting traditional data to compensate for coverage problems.
- Natural experiments with big data require effective control and intervention groups that should be closely monitored during the life cycle of an intervention.
- Access, scraping and preparation of big data sources that respect best ethical practices take most of the time allocated for analysis. Be realistic about the time and cost involved in gaining access, scraping and analysing big data.
- Consider whether open source tools are available or new models or tools need to be built. For example, consider strongly if a language is well resourced and if not, whether it is worth using big data or not, as it will be very time-consuming to build any analytical model.

Recommendation 3:

Big data should precede traditional data when sequencing and evaluating.

Start with a study of the demographics of the social media platform and topic to understand the nature and extent of the exclusions; consider using traditional data sources to obtain information from groups poorly represented.

- Big data analysis will help with the scoping stage of the evaluation, disclosing general trends and surprising case studies.
- Triangulation of data will be more effective if different methods are aligned.

Recommendation 4:

Big data can be shaped in ways that enhance its value.

- For example, launching social media campaigns, using hashtags and posts that trigger meaningful reactions and engagement from diverse groups.
- Set up interactive radio shows that invite participation from audiences through SMS and social media, while gathering individual demographics.
- Big data platforms can be both the intervention tool (e.g., hashtag campaigns) and the source of evaluation data. This does not present any problem, but studies using big data analysis need to make this distinction explicit, and adjust methods and conclusions according to the objective of the evaluation.

APPENDIX: ALGORITHM TO DISCOVER CO-OCCURRENCES INVOLVING WORDS

The algorithm to discover which words correlate with various categorisations (e.g., gender, location, sentiment) is as follows:

1. The tweets are divided into populations for comparison and the overall set of tweets.
2. The tweets are 'wordified' (divided into words, filtered for stop-words, stemmed and cleaned for punctuation, with URLs removed).
3. For the corpus, the most frequent words are identified with their frequency. Those which appear in more than 1 per cent of messages are kept and called the 'vocabulary.'
4. For each tweet, a vector of words is created, and for each word that appears in the 'vocabulary,' its value is the occurrence of those words in the tweet.
5. The occurrence of the words in the vector are normalized (in proportions), so that the sum of the occurrences of words in each tweet is always one (or zero in the case a message contains no words in the vocabulary).
6. These vectors of word occurrences are then averaged, based on the overall distribution of words in the 'vocabulary' and for the segment of the population we are interested in, to establish an average word frequency.
7. Our null hypothesis is that the probability of occurrence of a word in our sub-group of the population is equal to the probability of occurrence in the overall population.
8. Our overall distribution of word occurrence is modelled as a binomial distribution.
9. For each word, we calculate the difference in the occurrence of the word between the sub-group of the population from the overall population and divide it by the standard deviation, obtaining a standardized score (z-score) for each word.
10. Those with a z-score over two standard deviations are considered significant at $p < .01$.
11. The words are ranked by $\log(z) * F_{pop}$ (F_{pop} is the frequency in the target population) to even out rarely-used words.

REFERENCES

- Antin K., Byrne R., Geber T., van Geffen S., Hoffmann J., Jayaram M., Khan M., Lee T., Weinberg F., Wilson C., Rühling B., Rahman Z. and Simeoni C. 2014. "Shooting Your Hard Drive into Space and Other Ways to Practise Responsible Development." <https://responsibledata.io/wp-content/uploads/2014/10/responsible-development-data-book.pdf>.
- Austin, P. C. 2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". *Multivariate Behavioral Research* 46(3), 399–424. <http://doi.org/10.1080/00273171.2011.568786>.
- Bartlett, J., Miller, C., Reffin, J., Weir, D. and Wibberly, S., 2014. "Social media is transforming how to change society." In *Vox Digitas*, pp. 1-133. London: Demos. http://www.demos.co.uk/files/Vox_Digitas_-_web.pdf?1408832211
- Bhadanea, C., Dalalb, H. and Doshi, H. 2015. "Sentiment Analysis: Measuring Opinions." *Procedia Computer Science* 45: 808-814.
- Bossetta, M, Segesten, A-M.; Trenz, H-J. 2017. "Engaging with European Politics through Twitter and Facebook: Participation Beyond the National?" M. Barisione, and A. Michailidou, *Social Media and European Politics*, pp. 53-76. London: Palgrave Macmillan.
- Callegaro, M. and Yang, Y. 2018. "The Role of Surveys in the Era of "Big Data." D.L. Vannette, J.A. Krosnick (eds.):175-192. *The Palgrave Handbook of Survey Research*.
- Couper, M. P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3): 145-156.
- Cronbach, L. J. and Meehl, P. E. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52(4): 281-302.
- Ferron, J. and Rendina-Gobioff, G. 2015. "Interrupted Time Series Design." *Wiley StatsRef: Statistics Reference Online*. 1–6.
- Fox, J., Cruz, C. and Lee, J. Y. 2015. "Perpetuating Online Sexism Offline: Anonymity, Interactivity, and the Effects of Sexist Hashtags on Social Media." *Computers in Human Behavior* 52: 436-442.
- Glomb, T. M., Richman, W. L., Hulin, C. L., Drasgow, F., Schneider, K. T. and Fitzgerald, L. F. 1997. "Ambient Sexual Harassment: An Integrated Model of Antecedents and Consequences." *Organizational Behavior and Human Decision Processes* 71(3): 309-328.
- Groves, R. M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly*, 75(5): 861-871.
- Haselmayer, M. and Jenny, M. 2017. "Sentiment Analysis of Political Communication: Combining a Dictionary Approach with Crowdcoding." *Quality and Quantity*, 51: 2623-2646.
- Hilbert, M. 2016. "Big Data for Development: A Review of Promises and Challenges." *Development Policy Review* 34: 135–174.
- Jungherr, A., Harald S., Oliver P. and Pascal J. 2016. "Digital Trace Data in the Study of Public Opinion an Indicator of Attention Toward Politics Rather Than Political Support." *Social Science Computer Review* doi:10.1177/0894439316631043.
- Melville, P., Chenthamaraksha, V., Lawrence, R. D., Powell, J., Mugisha, M., Sapra, S., Anandan, R. and Assefa, S. 2013. "Amplifying the Voice of Youth in Africa via Text Analytics." *Proceedings of the 19th International Conference on Knowledge Discovery and Data Mining*.
- Pang, B. and Lee, L. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2(1): 1–90.
- Quercia, D., Askham, H. and Crowcroft, J. 2012. "TweetLDA: Supervised Topic Classification and Link Prediction in Twitter." Contractor, N.S., Uzzi, B., Macy, M.W., Nejd, W. (Eds). *Proceedings of the ACM Web Science Conference*. New York: ACM. 2012: pp. 247–250.
- Saville D.J. and Wood G.R. 1991. *Randomized Block Design. Statistical Methods: The Geometric Approach*. Springer Texts in Statistics. New York: Springer
- Salganik, M. 2018. *Bit by Bit: Social Research in the Digital Age*. New Jersey: Princeton University Press.
- Schober, M. Pasek, Guggenheim, L., Lampe, C. and Conrad, F. 2016. "Social media analysis for social measurement." *Public Opinion Quarterly*, 80(1): 180-211.
- Trochim, William M. *The Research Methods Knowledge Base*, 2nd Edition. <http://www.socialresearchmethods.net/kb/> (version current as of 20 October 2006).
- Tumasjan, A., Sprenger, T. O., Sandner, P. G. and Welp, I. M. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*, Washington, DC, USA, 23-26 May.

- UN Global Pulse. 2012. Big Data for Development: Challenges & Opportunities.
- UN Global Pulse. 2016a. Data Innovation Risk Assessment Tool.
- UN Global Pulse. 2016b. Integrating Big Data into the Monitoring and Evaluation of Development Programmes.
- UN Global Pulse. n.d. Our Data Privacy & Data Protection Principles. <http://www.unglobalpulse.org/privacy-and-data-protection-principles>.
- UN Women. 2018. Gender Equality and Big Data: Making Gender Data Visible. <http://www.unwomen.org/en/digital-library/publications/2018/1/gender-equality-and-big-data>.
- Wasserman, S. and Faust, K. 1994. Social Network Analysis: Methods and Applications. Cambridge: Cambridge University Press.
- Williams, M., Burnap, P. and Sloan, L. 2017. "Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation." *Sociology*, 51(6): 1149–1168.

Claudia Abreu Lopes
Big Data Consultant (Lead)

Claudia is an Affiliated Lecturer in Statistics and Methods at the University of Cambridge (UK) and Senior Advisor for Research at Africa's Voices Foundation. She has coordinated several media and development projects that harness the widespread use of mobile phones and social media to consult citizens for governance and public health outcomes. Claudia holds a Ph.D. in Social Research Methods from the London School of Economics and Political Science.

Savita Bailur
Senior Researcher

Savita is currently Research Director at Caribou Digital, an ICTs and emerging markets consultancy. She is a researcher with around fifteen years of experience in ICTs and development. Her recent work has focused on technologies for transparency and accountability, freedom of information and open data, with previous work on community radio, telecentres and mobile use in development. Savita is also a Visiting Fellow and External Lecturer in the Department of Media and Communications. Savita holds an MSc and Ph.D. in Information Systems from London School of Economics and Political Science.

Giles Barton-Owens
Data Scientist

Giles is a chief technology officer at Psyomics Ltd. based in Cambridge (UK). He specialises in low-resource language analysis and social media data extraction and analysis. He has worked with multiple start-ups and university projects after graduating in Computer Science from Cambridge University, UK.

UN WOMEN IS THE UN ORGANIZATION DEDICATED TO GENDER EQUALITY AND THE EMPOWERMENT OF WOMEN. A GLOBAL CHAMPION FOR WOMEN AND GIRLS, UN WOMEN WAS ESTABLISHED TO ACCELERATE PROGRESS ON MEETING THEIR NEEDS WORLDWIDE.

UN Women supports UN Member States as they set global standards for achieving gender equality, and works with governments and civil society to design laws, policies, programmes and services needed to implement these standards. It stands behind women's equal participation in all aspects of life, focusing on five priority areas: increasing women's leadership and participation; ending violence against women; engaging women in all aspects of peace and security processes; enhancing women's economic empowerment; and making gender equality central to national development planning and budgeting. UN Women also coordinates and promotes the UN system's work in advancing gender equality.



**Planet 50-50 by 2030
Step It Up for Gender Equality**

220 East 42nd Street
New York, New York 10017, USA
Tel: 212-906-6400
Fax: 212-906-6705

www.unwomen.org
www.facebook.com/unwomen
www.twitter.com/un_women
www.youtube.com/unwomen
www.flickr.com/unwomen