DECEMBER 2022

DISCUSSION PAPER SERIES NO. 2022-55

Addressing Data Gaps with Innovative Data Sources

Jana Flor V. Vizmanos, Jose Ramon G. Albert, Mika S. Muñoz, Arlan Brucal, Riza Teresita Halili, Angelo Jose Lumba, and Gaile Anne Patanñe

CONTACT US:

RESEARCH INFORMATION DEPARTMENT
Philippine Institute for Development Studies

18th Floor, Three Cyberpod Centris - North Tower
EDSA corner Quezon Avenue, Quezon City, Philippines

publications@pids.gov.ph
(+632) 8877-4000

https://www.pids.gov.ph

Addressing Data Gaps with Innovative Data Sources

Jana Flor V. Vizmanos
Jose Ramon G. Albert
Mika S. Muñoz
Arlan Brucal
Riza Teresita Halili
Angelo Jose Lumba
Gaile Anne Patanñe

PHILIPPINE INSTITUTE FOR DEVELOPMENT STUDIES

December 2022

Abstract

With the advent of digital transformation, ICT innovations have also led to a "data revolution" wherein more data is being captured, produced, stored, accessed, analyzed, archived, and reanalyzed, and at an exponential pace. An examination of new data sources, including big data and crowd-sourced data, can complement traditional sources of statistics and can unlock insights that can ultimately lead to interventions for better outcomes by informing policies and actions toward attaining robust, sustainable, and inclusive development. This study will examine PIDS website download data and Twitter data to illustrate stories that can be obtained from new data sources and explore how access, analysis and use of new data sources can be promoted. Several quantitative tools are used on these new data sources, including (a) market basket analysis for website download data, (b) text mining (and sentiment analysis) for web scraped Twitter data, and (c) other big data analytics tools. Policy issues, including risk management for use of these new data sources, are also discussed.

Keywords: data revolution, big data, new data sources, social media data, market-basket analysis, web scraping, text mining, sentiment analysis

Table of Contents

List of Tables

List of Figures

Addressing Data Gaps with Innovative Data Sources

Jana Flor V. Vizmanos, Jose Ramon G. Albert, Mika S. Muñoz, Arlan Brucal, Riza Teresita Halili, Angelo Jose Lumba, and Gaile Anne Patanñe*

1. Introduction

With the advent of digital transformation, ICT innovations and the growing use of digital tools, including the internet, have also led to a "data revolution" wherein more data is being captured, produced, stored, accessed, analyzed, archived, and re- analyzed, and at an exponential pace (Independent Expert Advisory Group on a Data Revolution for Sustainable Development 2014). New data sources, including big data[1] and crowd-sourced data[2], can complement traditional sources of statistics (such as censuses, sample surveys and administrative data) to monitor and analyze the public sector's development targets such as national development plans and progress in attaining international commitments such as the Sustainable Development Goals (SDGs), also referred to as the Global Goals. New statistical methods and tools are also being developed side by side with the availability of the tsunami of data. New data requirements, including more disaggregated data and granular data,[3] are also being demanded by data users (ADB 2021a).

Meanwhile, institutions in both the public and private sectors, including the Philippines Institute for Development Studies (PIDS) are accumulating data at an exponential rate, sometimes as a by-product of an administrative function or some communication medium with its clients, and in other cases, through data collection systems designed not necessarily for generating statistics. Data is growing especially amid the growing use of technology, including the use of the internet and digital platforms (Albert 2021). Yet little data analytics are being performed on data holdings. At PIDS, the database of website download that identify download transaction IDs, time stamps and information on the PIDS papers being downloaded by stakeholders is growing.

While the Philippines has identified its development priorities in the Philippine Development Plans and its commitment to international declarations, such as the attainment of the SDGs by 2030, data gaps persist for monitoring the country's development plans and the Global Goals even amid the emerging data revolution that has led new data sources, such as big data (Albert 2014). Many governments have long recognized the need for data and statistics to inform policy and effect development outcomes, but data gaps persist. Without high-quality

* The first three authors are research associate, senior research fellow, and research analyst, respectively, of the Philippine Institute for Development Studies (PIDS). Meanwhile, the next four authors are country office economist, Pintig Lab project manager, data scientist, and intern, respectively, at the United Nations Development Programme (UNDP) - Philippines. The views expressed here are the authors' own and do not necessarily reflect the positions of the organizations that the writers are associated with.

[1] Although there is no standard definition of big data, it can be viewed as digital fingerprints that are unfiltered by-products or exhaust from the use of information and communication technology (ICT) tools. These digital tools include electronic devices (smart phones, tablets, laptops), social media, blogs, search engines, as well as (fixed and mobile) sensors and tracking devices (including climate sensors and global positioning system or GPS).

[2] Crowdsourced data collection is a participatory method of building a dataset with the information and opinions, of a large group of dispersed people usually sourced via the Internet. (https://dimewiki.worldbank.org/Crowd-sourced_Data)

[3] "Disaggregated data refer to data that can be used to generate statistics and indicators for population groups defined by (or disaggregated by or broken down further into) one or more dimensions or characteristics (commonly sex, geographic area, and/or age); a related term is "granular data," which represents the idea of data about smaller chunks or pieces of a larger population." (ADB 2021a, p.3)

data, especially disaggregate data (e.g., by sex, disability status, ethnicity, socio-economic class, and other relevant disaggregations) providing the right information on the right things at the right time, the assessment of needs, the design of strategies, policies, programs, activities, and projects (PAPs), as well as monitoring and evaluating policies and PAPS becomes close to impossible.

Faced with a growing demand to produce evidence-based policy research, PIDS has been part of the data ecosystem, but the Institute, together with other organizations in the public and private sectors, can take advantage of new, innovative data sources. PIDS, in particular, if it harnesses new data sources can provide decision-makers with near real time information for use in policy development.

Given many data gaps in national development plans, and in monitoring the SDGs, especially disaggregated data, at the macro level, as well as the need for PIDS to know its clients better, at the micro level, an examination of new data sources can unlock insights that can ultimately lead to interventions for better outcomes. Development data ultimately takes its meaning from use in the entire policy cycle. From the diagnosis or identification of problems in society, to the formulation of policy options, to the adoption of a specific policy from several policy alternatives to the design and implementation of policies and other program interventions, to the assessment of policies whether monitoring or evaluation, we need data to describe conditions, and assess performance. Policy may be viewed as what governments do or does not do, why they do it, and what difference does it make. The development of policies should include a rational analysis of what works and what does not. Policies that are based on systematic evidence are widely viewed to produce better outcomes than those that are not informed by data and statistics. Decisions at these various stages of policy development, including adjustments on policies, should be better informed by available development data and statistics(ADB 2021a, p.19.). Insights gained from new data sources, provided they undergo curation, can inform policies and actions toward attaining robust, sustainable and inclusive development.

2. Official statistics and big data

Data and statistics for development have traditionally been sourced from surveys (whether censuses or sample surveys)[4], administrative reporting systems[5], and other compilations of secondary data following established concepts, definitions, methods, and classification systems. The choice of data source in statistics production is often guided by considerations on cost and reducing the burden on respondents of surveys and censuses.

The increased use of ICT, the internet and other frontier technologies of the Fourth Industrial Revolution have led to a resulting hyperconnectivity that connects persons to persons, people to machines, and machines to machines. Based on the Digital 2022 July Global Statshot Report from DataReportal[6], as of July, there are 5.34 billion unique mobile phone users, 5.03 billion

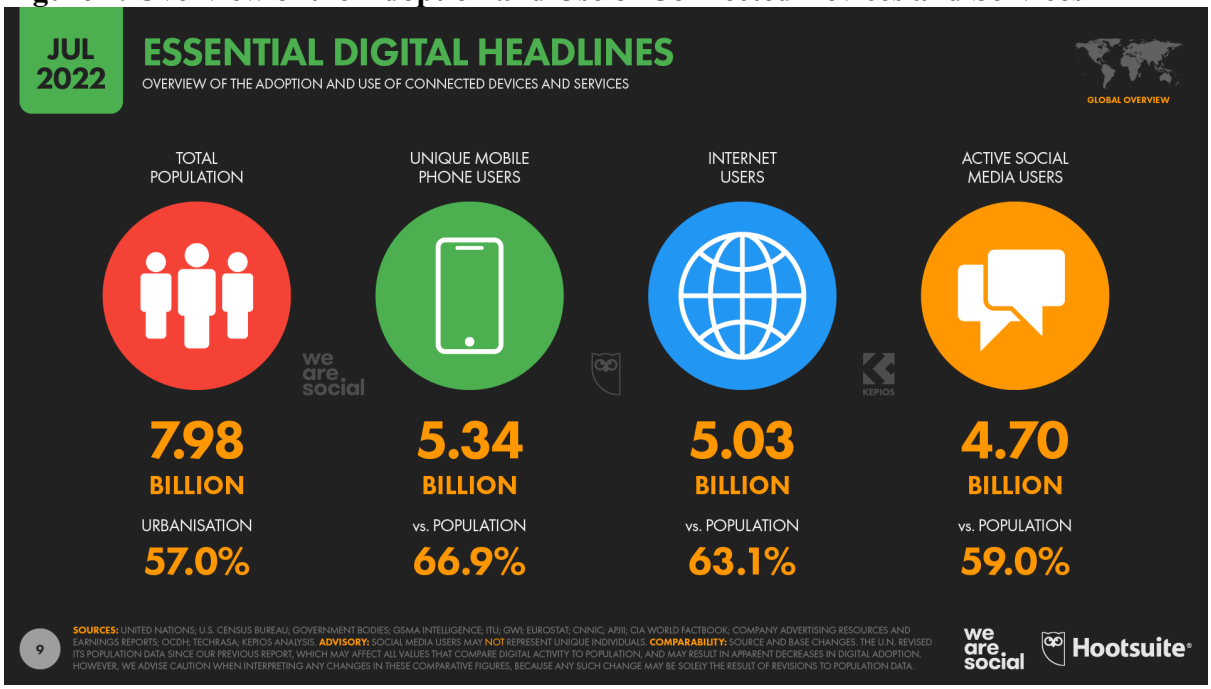
[4] A survey is a systematic method for gathering information from a target population of interest for purposes of producing quantitative descriptors of the attributes of the population; A census is a survey with data collection involving a complete enumeration of the population, meanwhile if data are collected in a survey only from a subset of the population, then the survey is called a sample survey.

[5] Administrative data are data holdings collected typically by national government agencies and local governments for the purposes of administering taxes, benefits or services.

[6] https://datareportal.com/

internet users, and 4.7 billion active social media users across the world (see **Figure 1**). The use of mobile phones, the internet and social media has been rising over the years, and this has led to a deluge of big data, characterized by what Hilbert (2013) called the three V's—volume (i.e., the amount of data), velocity (i.e., the speed of data), and variety (i.e. the range of data formats including personal documents, SMS messages, photos, videos, maps, financial or social transactions). A fourth V has also been added veracity (trustworthiness of the data) in the wake of the post-truth era. According to Albert *et al.* (2019), Big data can be categorized largely into three main sources:(i) human-sourced information (e.g., social networks); (ii) process-mediated data (e.g., search engines, commercial transactions), and (iii) machine-generated data (e.g., mobile phone location). All these voluminous, fast-paced, and complex data, however, are often by-products of transactions from hyperconnectivity (thus mere data exhaust), and as such, are often unstructured and do not necessarily relate to a target population, unlike traditional data sources of official statistics.

Figure 1. Overview of the Adoption and Use of Connected Devices and Services



Source: DataReportal (2022), "Digital 2022 July Global Statshot Report," retrieved from https://datareportal.com/reports/digital-2022-july-global-statshot

Further, the connectivity has allowed organizations across the world to also make use of crowdsourced data than through traditional data collection approaches given the relatively low cost of crowdsource data collection. There are trade-offs, however, between sample size and sampling issues (i.e. representativeness) in crowdsourced data. The reliability of crowdsourcing data is often questioned because of the lack of an underlying sampling frame. An example of crowdsourced data is data on citizens' feedback from smart cities such as Jakarta Smart City obtained through digital technology to improve local service delivery. Another example is data on violence against women obtained by the Safecity app (used in India) where women can post online about incidences of harassment and assault in public spaces across cities in India (including geo-locations of these events).

Today's world has experienced an upsurge in data capture, storage, access, analysis, archive, aggregation and reanalysis compared to the world of the mid to late 20th century. Every tweet, like, comment, photograph and video posted or shared on social media platforms, every movie

or tv show watched on YouTube or Netflix, every email or SMS sent, every banking or financial transaction made in the internet and/or through digital devices, contains data that people are (knowingly or unknowingly) sharing about themselves. While access to and storage of a large volume of data for analytics have existed in business for quite a while, the use of big data for development only gained traction recently with the recognition that this non-traditional data source can fill in the data gaps for monitoring the SDGs (Martinez et al. 2018; Albert and Martinez 2018; Albert et al. 2019). Although big data can be structured (i.e. in pre-defined formats), unstructured (i.e. not organized in a pre-defined manner until they are extracted for analysis), or semi-structured (i.e. a mix of both structured and unstructured data), often big data is unstructured (e.g. audios, videos, photos, or messy text heavy information). Analytics can thus be challenging since big data can be subject to significant amounts of noise, that is, big data may not represent the underlying population of interest (Silver 2012; Cox et al. 2018).

Cognizant of the wealth of digital information and Big Data sources that could be used by NSOs, governments, companies and individuals and the challenges of data curation and triage (i.e., finding the useful portion of data in big data), the UN Statistical Commission (UNSC) established the Global Working Group (GWG) on Big Data for Official Statistics, also called the United Nations Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD), in 2014, as an outcome of its 45th meeting. The UN GWG is mandated to formulate an ongoing global programme on Big Data for official statistics, and in so doing, it also promotes capability building, training and sharing of experiences on practical use of Big Data especially for policy applications. Under the governance of the UN-CEBD, the UN Global Platform was established via cloud enabling statisticians and developers across the globe to collaborate, share and develop new technologies on Big Data sources and methodologies. One of its four regional hubs is located in Asia and the Pacific, jointly established by the China's National Bureau of Statistics (NBS), the People's Government of Zhejiang Province and the UN Department of Economic and Social Affairs (UN DESA).

3. Methodology and Innovative Data Sources

The analytical framework for this study requires three tiers: (i) asking questions (at the macro level on development issues, and at the micro level on PIDS clients); (ii) gathering insights; and (iii) communicating results following a hierarchy of (data) needs. From gathering data and generating facts to report on questions, understanding relationships between facts will be examined, and finally giving value to insights by communicating results of data analytics derived.

The study focusing on the following data from alternative data sources that are readily available or can be derived from existing online platforms:

(a) PIDS web download data;
(b) twitter data and other web scraped data; and
(c) traffic congestion data.

The first set of data will feed into gaining insights on PIDS clients, while the second and third set of data will be used to address data gaps on various development issues. Several quantitative tools are used on these new data sources, including market basket analysis for website download data, text mining (and sentiment analysis) for twitter and web scraped data, and other big data analytics tools.

3.1. PIDS-web download data

An analysis of the combinations of research documents downloaded by PIDS clients can identify what studies, or groups of studies, tend to be associated with each other when clients download these research documents. A total of 64,207 website publications, including discussion papers, policy notes, research paper series, and articles from Philippine Journal of Development, were downloaded from the PIDS website from April 18, 2019 to August 9, 2022. In this report, these PIDS website download data is subjected to "association rule mining", which essentially involves examining the association between different "items" downloaded, to find frequent patterns in the PIDS website download transaction database.

According to online traffic data from Similarweb[7], PIDS website traffic registered over 55,100 visits for September 2022, with a 43.22% increase compared to August. This is likely on account of our many events during the Policy Development month. Among PIDS website visitors, there are more female than male audiences, and many are young: 44% aged 18-24 years old. Comparing this with actual PIDS web download data **(Table 1**), about a fifth of website visitors who shared some of their basic information are aged 19-35, female, with postgraduate degree, or employed full time. Thus, validating the earlier profile of visitors.

Table 1. PIDS Website Downloader Profile: April 18, 2019 to August 9, 2022

Downloader Profile		Frequency	Distribution (%)
Age			
	Below 18	2,037	3.2
	19-35	14,188	22.1
	36-50	4,961	7.7
	51-65	2,403	3.7
	66 and above	400	0.6
	Missing data	40,218	62.6
	Total	**64,207**	**100.0**
Sex			
	Female	12,328	19.2
	Male	9,906	15.4
	Prefer not say	1,067	1.7
	Prefer to self-describe	117	0.2
	Missing data	40,789	63.5
	Total	**64,207**	**100.0**
Education			
	No schooling	453	0.7
	Elementary	115	0.2
	High School	1,839	2.9
	Vocational	141	0.2
	College	10,341	16.1
	Postgraduate	11,108	17.3
	Missing data	40,210	62.6
	Total	**64,207**	**100.0**
Occupation			
	Employed (Full-time)	12,650	19.7
	Employed (Part-time)	891	1.4
	Homemaker	101	0.2

[7] https://www.similarweb.com/website/pids.gov.ph/#overview

Downloader Profile		Frequency	Distribution (%)
	Self-employed	1,260	2.0
	Student	7,673	12.0
	Retired	378	0.6
	Others	878	1.4
	Missing data	40,376	62.9
	Total	**64,207**	**100.0**

Note: Authors' tabulation

The PIDS web download dataset contains data per download instance. Each download is assigned an ID and includes information on type of publication downloaded, focus area, keywords, as well as the title, authors, and ID of the publication. Website visitors who wish to download PIDS publications are given the option to supply certain personal information such as name, reason for downloading the publication, e-mail address, age, gender, education, and occupation. To process the dataset for analysis, each website visitor is assigned a unique ID using his/her e-mail address. However, providing a unique identifier among website visitors is a challenge, as in the cases of using auto-generated e-mail addresses or providing different e-mail addresses from the ones they previously supplied during their first download.

In terms of broad development issues, Governance (11.9%) Health (11.8%), Labor and Education (10.2%) are top publication themes downloaded from the PIDS website, with roughly one-thirds of PIDS website visitors download publications under these topics (**Table 2**). However, considering that the number of downloads may be dependent on the number of available publications on a particular theme/sector, it is important to examine the frequency of web downloads relative to the total publications per theme. For instance, while there are only four (4) identified publications under International Relations and Foreign Policy, each of these publications was downloaded 126 times, on average. Meanwhile, PIDS publications on Governance have a frequency of 74 downloads relative to total publications, despite being the most downloaded theme. Thus, when download data are examined in relation to total publications, we find the top three themes to be instead : (i) International Relations and Foreign Policy; (ii) Fiscal Policy and Taxation; (iii) Technology and Innovation.

Table 2. PIDS Web-Downloaded Publications by Theme

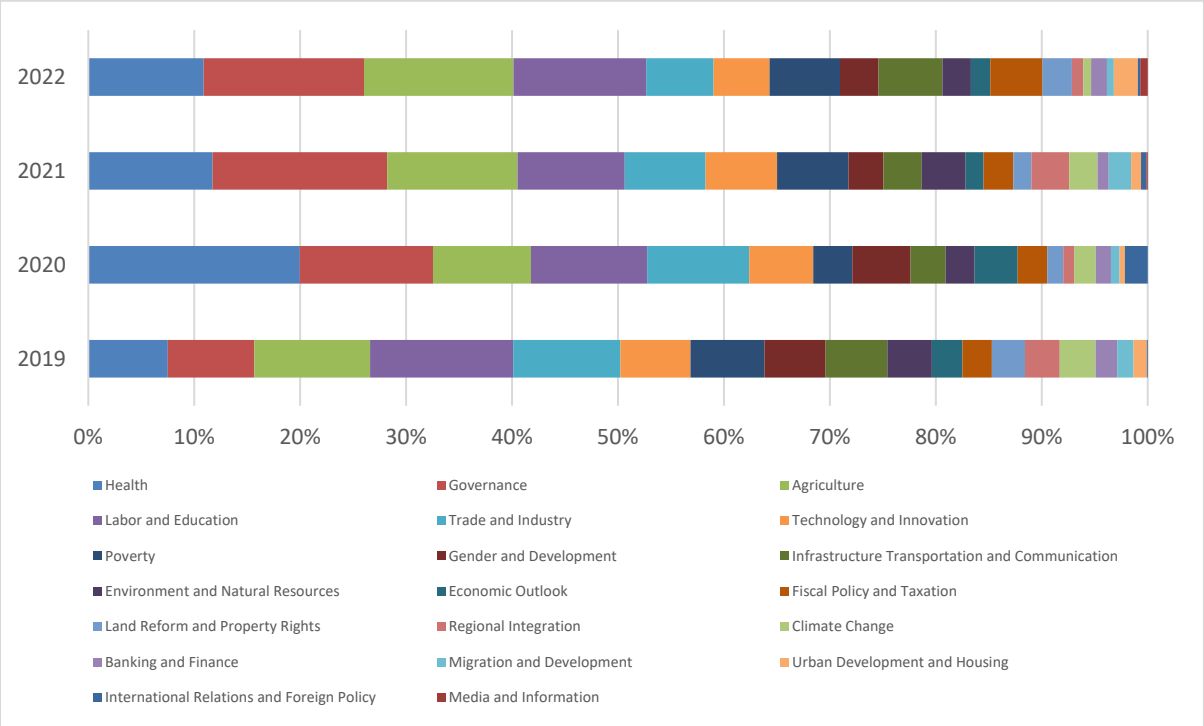
Publication Theme	Number of web downloads* (a)	Rank	Number of total publications (b)	Rank	Ratio of downloads to total publications (a:b)	Rank
Governance	7,636	1	103	4	74	8
Health	7,572	2	105	3	72	9
Labor and Education	6,518	3	118	2	55	13
Agriculture	6,485	4	101	5	64	10
Trade and Industry	4,824	5	100	6	48	15
Technology and Innovation	3,545	6	36	10	98	3
Poverty	3,305	7	58	8	57	12
Gender and Development	2,566	8	31	11	83	6
Infrastructure Transportation and Communication	2,489	9	122	1	20	19
Environment and Natural Resources	1,958	10	39	9	50	14
Fiscal Policy and Taxation	1,811	11	18	15	101	2
Economic Outlook	1,525	12	16	17	95	4
Regional Integration	1,285	13	21	14	61	11
Climate Change	1,276	14	28	13	46	16
Land Reform and Property Rights	1,203	15	29	12	41	17
Banking and Finance	823	16	60	7	14	20
Migration and Development	756	17	10	18	76	7
Urban Development and Housing	631	18	18	15	35	18
International Relations and Foreign Policy	503	19	4	19	126	1
Media and Information	95	20	1	20	95	5

Note: Authors' tabulation of PIDS website download data for *April 18, 2019-August 9, 2022.

Observing patterns by theme as well as by year (**Figure 2**), publications discussing Labor and Education are found to be the most downloaded content from the PIDS website in 2019 at 13.5%, while Health used to rank 5th at 7%. However, at the onset of the COVID-19 pandemic in 2020, Health took over the first spot, comprising 20% of website publication downloads

followed by Governance at 12.6%. The following year, more downloads were attributed to Governance at 16.5%, overtaking Health at 11.7%. The popularity of Health publications reflects PIDS clients' view of the urgency to formulate health policies amid the pandemic, which is still on-going, aside from other issues in the health sector including non-communicable diseases that already attracted some attention prior to COVID-19.

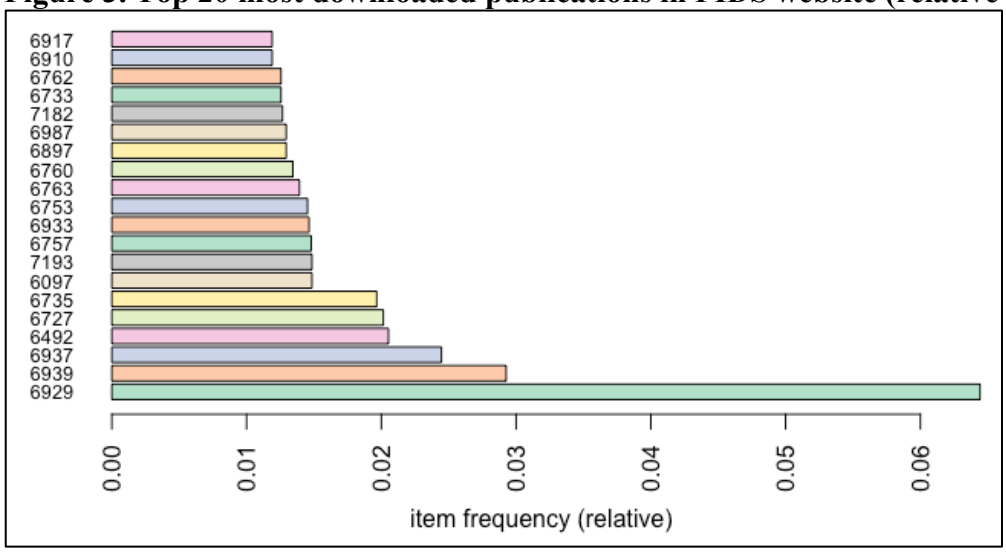
Figure 2. PIDS Website Publication Downloads by Theme by Year (%)



Note: Authors' visualization summary of PIDS website download data

Meanwhile the Item Frequency Histogram (**Figure 3**) displays how many times an item, i.e., PIDS research outputs downloaded, has occurred in our dataset as compared to other publications downloaded by visitors to the PIDS website. The relative frequency plot shows, in particular, that Publication ID 6929, a Discussion Paper on the Situation Analysis of Early Childhood Care Development-First 1000 Days (ECCD-F1KD) Initiatives in Selected UNICEF-KOICA Provinces constitutes around 6% of PIDS website downloads for the time period examined. This is followed by Publication ID 6939, a Policy Note on the Issues and Concerns in the Implementation of the Performance-Based Bonus at the Department of Department of Education (DepEd) and Publication ID 6937, a Discussion Paper on Expanding Health Insurance for the Elderly of the Philippines, at around 3% and 2.5%, respectively.

Figure 3. Top 20 most downloaded publications in PIDS website (relative value)



Note: Authors' visualization summary of PIDS website download data

3.2. Social media, twitter data, and web scraping

With 4.70 billion social media users, corresponding to nearly three-fifths (58 %) of the world's total population, spending two and a half hours daily on social media to keep up to date with current events, connect with family and friends, or search for entertainment, social media is a wealthy source of information for research, marketing, and development of applications. Public conversations on social media, particularly tweets on Twitter, can be extracted using application programming interface (API) tools to analyze data from social media. Unlike other social platforms, nearly every user's tweets are completely public; Twitter's API allows a programmer to do complex queries such pulling out every tweet about a certain topic within the last ten minutes, or pulling a certain Twitter user's non-retweeted tweets (Kudarvalli and Fiaidhi, 2020).

As of 2022, Twitter had a total of 238 million monetizable daily active users around the world based on the social media company's investor earnings report for the second quarter. Further, according to Hootsuite's 2022 Digital Trends Report, Twitter is the world's seventh favorite social media platform and Twitter.com is the ninth most visited website globally. Three-fifths (61.0%) of active Twitter users around the globe use the social media platform to keep up to date with news and current events while nearly two-fifths (37.4%) follow or research brands and products (DataReportal 2022a). According to a similar report (DataReportal, 2022b) focusing on the Philippines, there are 92.05 million social media users as of January 2022 (though these users need not pertain to unique users), and Twitter is the 8th most popular social media platform in the country with 10.5 million users. In terms of marketing, the potential audience that can be reached with ads on Twitter is estimated at 11.55 million, which constitutes roughly 10% of the country's population.

While social media platforms like Twitter are complex websites, there are available tools to web scrape tweets and other data on Twitter. Web scraping, the process of extracting data from a website, can be programmed to extract data objects from Twitter (e.g., tweets, author name, retweet count and like counts). Python, a commonly used programming language with its broad ecosystem of well-maintained libraries, is often used in web server development, data science, and automation. Analyzing Twitter data by scraping tweets using the Twitter API platform can

provide insights on global to local topics as well as events, gain information to better profile a target audience, and identify trends and important conversations on Twitter.

The Twitter API platform provides broad access to public Twitter data that users have chosen to share publicly. The Twitter API can be accessed by applying for a Twitter developer account and describing the intended use of Twitter data for a particular project. Approval of application is free but can take roughly a week or more, depending on additional information that may be requested by Twitter to evaluate the developer account and the level of API access request. There are three access levels for Twitter API (as shown in **Table 2**). Essential level requires only signing up using a Twitter developer account and allows for one app in a project to retrieve up to 500 thousand tweets per month. Meanwhile, elevated access offers more flexibility and permits three apps in a project and can retrieve up to 2 million tweets per month, provided the developer account specifically applied for additional access within the developer portal. Meanwhile, academic research access allows up to 10 million tweets to be retrieved per month, with permission to access full-archive search tweets and tweet counts, as well as access to advanced filter operators. While these features are not available to essential and elevated levels, the academic research level is limited to only one project and one app.

Table 3. Twitter API access levels and versions

Twitter API access levels and versions	Essential	Elevated	Academic Research
Getting access	Sign up	Apply for additional access within the developer portal	Apply for additional access
Price	Free	Free	Free
Access to Twitter API v2	✔	✔	✔
Access to standard v1.1	✔ (Limited access - only media endpoints)	✔	✔
Access to premium v1.1	✘	✔	✔
Access to enterprise	✘	✔	✔
Project limits	1 Project	1 Project	1 Project
App limits	1 App per Project	3 Apps per Project	1 App per Project
Tweet caps	Retrieve up to 500k Tweets per month	Retrieve up to 2 million Tweets per month	Retrieve up to 10 million Tweets per month
Filtered streamrule limit	5 rules	25 rules	1000 rules
Filtered stream rule length	512 characters	512 characters	1024 characters
Filtered stream POST rules rate limit	25 requests per 15 minutes	50 requests per 15 minutes	100 requests per 15 minutes
Search Tweets query length	512 characters	512 characters	1024 characters
Access to full-archive search Tweets	✘	✘	✔
Access to full-archive Tweet counts	✘	✘	✔
Access to advanced filter operators	✘	✘	✔

Option to manage a team in the developer portal	✗	✓ (Requires an organization type account)	✗
Access to the Ads API	✓ (Requires additional application)	✓ (Requires additional application)	✗
Authentication methods	OAuth 2.0 with PKCE App only	OAuth 2.0 with PKCE OAuth 1.0a App only	OAuth 2.0 with PKCE OAuth 1.0a App only

Source: Getting Started: About the Twitter API https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

Other web scraping tools are available without using a Twitter API. For instance, SNScrape, a social networking service scraper in Python, allows one to scrape basic information such as a user's profile, tweet content and can also scrape public posts from other prominent social media networks like Facebook, Instagram, and others. While there are no limits to the number of tweets you can retrieve and that scraping tweets beyond 7-day window of standard Twitter API can be done, some functionalities are not present that may be important for analyzing certain topics (e.g., no data on geolocation for spatial analysis). Further, Section 4 on Using the Services of Twitter's Terms of Service states that:

> *"You may not do any of the following while accessing or using the Services: ...(iii) access or search or attempt to access or search the Services by any means (automated or otherwise) other than through our currently available, published interfaces that are provided by Twitter (and only pursuant to the applicable terms and conditions), unless you have been specifically allowed to do so in a separate agreement with Twitter (NOTE: crawling the Services is permissible if done in accordance with the provisions of the robots.txt file, however, scraping the Services without the prior consent of Twitter is expressly prohibited)."*

Although there may be concerns about privacy issues on the use of Twitter data, but with the enactment of Republic Act 10173, also known as the "Data Privacy Act of 2012", the Philippines ensures lawful processing and protection of personal data. During the Countdown to Data Privacy Conference held last October 11, 2017 by the National Privacy Commission (NPC), Former Commissioner Ivy Patdu stressed caution in processing personal data shared in social media, emphasizing that "even if information is publicly available, one cannot assume that it is free to use social media data for any purpose" (*PhilStar*, par.7)[8]. Thus, while any data that is publicly available can be scraped, it is important to be familiar with existing laws as well as the terms and policies of the target website subject for web scraping to avoid data privacy and copyright issues.

3.3. Traffic congestion data

Metro Manila is one of the megacities in Asia with a long-standing issue of excessive traffic congestion. The region is integral to the Philippines as it is the center for many of the political, economic, and social activities in the country, accounting for 31.5% of the entire economic output in 2021, according to the Philippine Statistics Authority (PSA 2022). Moreover, Metro

[8] https://interaksyon.philstar.com/national/2017/10/11/102741/on-data-privacy-can-you-use-public-information-on-social-media-freely/

Manila is the most densely populated region with a density of 21,765 persons per square kilometer - 60 times greater than the national figure (PSA 2021). Aside from the population in Metro Manila, residents from nearby provinces like Bulacan, Rizal, Laguna, and Cavite also commute to the city daily, typically to work.

According to Ang (2022), Metro Manila is currently ranked eighth among cities in the world with the highest levels of traffic congestion. The same report suggests that an average Filipino spends around four (4) days in traffic jams annually. This estimate is lower than that of 2019, the year before the onset of the pandemic, when commuters in the metro lost as much as 257 hours or 10 days and 17 hours in traffic (Subingsubing, 2020). According to a study by the Japan International Cooperation Agency (JICA) (2020), the traffic in Metro Manila and adjacent areas of Bulacan, Laguna, Cavite, and Rizal, costs Php 3.5 billion per day due to foregone productivity. This amount is expected to triple by 2030 if no developments will be made to mitigate the problem.

Traffic congestion also intensifies air pollution due to increased vehicle emissions. This degradation in ambient air quality has been linked to excess morbidity and mortality for drivers, commuters, and individuals, especially for those living near major thoroughfares (Zhang and Batterman, 2013). Data from the World Health Organization (WHO)[9] reveals that as of 2014, an estimated 7 million people die every year due to air pollution.

Metro Manila also happens to have the highest mortality rate due to respiratory and cardiovascular diseases, with an estimate of 5000 annual premature deaths (Montano, 2016). Moreover, a 2012 study by Washington University in St. Louis noted that long commutes eat up exercise time, thereby establishing an association with weight gain, lower fitness levels, and elevated blood pressure – all of which are strongly associated with heart disease, diabetes, and some types of cancer. Lastly, traffic also threatens the commuters' overall well-being as it causes stress, fatigue, and hindrances to pursuing other wellness activities (e.g., meeting friends and family, pursuing hobbies, etc.).

Despite policies and interventions made by the government, the issue of traffic congestion has yet to be addressed fully as it continues to be a major problem in Metro Manila. Thus, this section of the study aims to:

- provide exploratory data analysis of traffic congestion;
- produce insights for improving the efficiency of the transportation system; and,
- generate inputs for predicting traffic congestion and measuring the associated impact on society, economy, health, and the environment

The data used for the analysis in this section is based on Waze[10] traffic congestion data, which is accessed by the UNDP Pintig Lab via the Development Data Partnership (DDP) - a concerted effort by multilateral organizations and tech companies to facilitate the use of third-party data for purposes of international development (Development Data Partnership, n.d.). It is composed of individual jam reports with the following key attributes:

[9] https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution

[10] Waze is a subsidiary company of Google that provides satellite navigation services to netizens on smartphones, tablets and other ICT devices that support the Global Positioning System.

Table 4. Attributes of Traffic Congestion Data

Attribute	Description
id	Unique ID of a report
ts	Date and time (UTC) of the report's occurrence
city	City in which report occurred
street	Street in which report occurred
startNode	Area in which a user starts their navigation
endNode	Destination chosen by a user
speedKMH	Speed in kilometers per hour of a report
length	Length of jam in meters
delay	Difference between jam and free flow speed in seconds
level	Traffic congestion level (categorical: 0 for free flow and 5 for blocked)
geo	Line string of latitude-longitude coordinates of a report

As of the latest extract, there are 35,729,550 reports from April 2019 to April 2022 covering 9,297 streets across 61 cities. Areas covered are primarily from the Greater Manila Area, Cebu City, and Davao City as traffic congestion is more pronounced in these locations

4. Empirical Findings

Discussions on this section will focus on the use of the following data analytic methods from the aforementioned non-traditional data sources:

1. Market-basket analysis of PIDS web-download data
2. Sentiment analysis and text mining:
    a. web scraped news related to violence against women in the Philippines
    b. tweets on Philippine tourism
3. Traffic congestion analysis using data from Waze

4.1. Market-basket analysis of PIDS web-download data

While there maybe suspicion that certain publications are frequently downloaded together, the question is, how do we identify these associations? We employ a method called association rules analysis, which three common ways to measure association:

(i) Support, which reflects how popular an itemset is, is measured by the proportion of transactions in which an itemset appears, where an itemset maybe one or more publications downloaded;
(ii) Confidence, which refers to how likely item Y is downloaded when item X is downloaded expressed as {X -> Y}; this is measured by the proportion of transactions with item X, in which item Y also appears; and

(iii) Lift, which measures how likely item Y is downloaded when item X is downloaded, while controlling for how popular item Y is.

In using Association Rule Mining on a given set of website download transactions, the goal will be to find all rules with:

(a) Support greater than or equal to minimum support threshold level
(b) Confidence greater than or equal to minimum confidence level

This analytical tool involves a two-step approach:

(I) Frequent Itemset Generation: Find all frequent item-sets with "support" >= pre-determined minimum support count
(II) Rule Generation: List all Association Rules from frequent item-sets. Calculate Support and Confidence for all rules. Prune rules that fail min_support and min_confidence thresholds. The association rules will need to generate item sets, i.e. combinations of PIDS products jointly downloaded with the use of an apriori algorithm that is computationally designed to find subsets which are common to at least a minimum number of item sets.

In practice, the amount of computation to generate rules depends on the minimum support specified. Note that Frequent Itemset Generation is the most computationally expensive step because it requires a full database scan. Aside from this standard "market basket analysis", it is important to also examine if transactions may be made by the same PIDS client (using the email address of the client as the link variable for download transaction data), and for this analysis, some transactions may be merged, with a re-analysis with the same methodology.

The Apriori algorithm for Association Rule Mining identifies item sets that occur with a support greater than a pre-defined value and calculates the confidence of all possible rules based on those item sets, in this case publication themes. Support is set at 0.01 and confidence at 0.1. For the association rule

{X -> Y}

we will use LHS to stand for the left-hand side, the antecedent X, while RHS stands for right-hand side, the consequent Y.

In **Table 5**, the rule suggests that if a website visitor downloads a Health-themed publication, the **support** of the rule, the probability that a website visitor will also download a Governance content given that he or she downloaded a Health-tagged publication, is estimated at 28%, referred to as the **confidence** of the rule. As regards **lift**, a value greater than 1 suggests that a Health Publication download increase the chances that Governance publications will also be downloaded in a given transaction.

Table 5. Summary Results of Association Rule Mining of Downloaded PIDS Publications by Theme

Focus Area: Health

LHS	RHS	support	confidence	coverage	lift	count
Health	Governance	0.07	0.28	0.23	1.27	1210
Health	Labor and Education	0.06	0.26	0.23	1.36	1132
Health	Agriculture	0.06	0.25	0.23	1.30	1082
Health	Trade and Industry	0.05	0.22	0.23	1.45	921
Health	(NULL)	0.05	0.20	0.23	0.96	850
Health	Technology and Innovation	0.04	0.17	0.23	1.41	729
Health	Poverty	0.04	0.17	0.23	1.60	722
Health	Infrastructure Transportation and Communication	0.03	0.14	0.23	1.64	605

Health	Gender and Development	0.03	0.14	0.23	1.44	581
Health	Environment and Natural Resources	0.03	0.11	0.23	1.66	489
Health	Fiscal Policy and Taxation	0.03	0.11	0.23	1.66	472
Health	Economic Outlook	0.03	0.11	0.23	1.80	463

Focus Area: Governance

LHS	RHS	support	confidence	coverage	lift	count
Governance	Health	0.07	0.30	0.22	1.27	1210
Governance	Labor and Education	0.07	0.30	0.22	1.51	1206
Governance	Agriculture	0.06	0.28	0.22	1.44	1145
Governance	Trade and Industry	0.06	0.27	0.22	1.81	1103
Governance	(NULL)	0.04	0.20	0.22	0.96	812
Governance	Technology and Innovation	0.04	0.20	0.22	1.64	809
Governance	Poverty	0.04	0.19	0.22	1.76	759
Governance	Gender and Development	0.03	0.15	0.22	1.63	628
Governance	Infrastructure Transportation and Communication	0.03	0.15	0.22	1.69	594
Governance	Fiscal Policy and Taxation	0.03	0.13	0.22	1.99	540
Governance	Environment and Natural Resources	0.03	0.13	0.22	1.89	534
Governance	Economic Outlook	0.03	0.12	0.22	1.98	486

Note: Authors' tabulation from the resulting Association Rule Mining

Findings at the individual publication-level involved setting lower thresholds (**Table 6**). The LHS represents a cart of publications. A comma represents "and" when there are two or more items in an "if" statement. There are 26 carts that had Publication IDs 6754 and 6759 with 6758. Support is the fraction of baskets that have all items referenced in a rule. Support for Table 6 was set to a minimum of 0.001. Of all the carts that had 6754 and 6759 in them, a confidence level of 93% means that these publications are also downloaded together with Publication 6758. The support set at 0.001 indicates too few associations can be accurately concluded. This is likely because there still is very little traction on PIDS websites as of the current analysis, but over time, PIDs should expect more website visits that will enable the Institute to gain more insights on associations of publications that are summarized based on the currently available download data.

Table 6. First Five Publications with Strongest Association Rules

LHS	RHS	Support	Confidence	Coverage	Lift	Count
6754 (Assessment of TRAIN's Coal and Petroleum Excise Taxes: Environmental Benefits and Impacts on Sectoral Employment and Household Welfare), **6759** (Economic Principles for	**6758** (Child Stunting Prevention: The Challenge of Mobilizing Local Governments for National Impact)	0.001420	0.928571	0.001529	147.89	26

LHS	RHS	Support	Confidence	Coverage	Lift	Count
Rightsizing Government)						
7171 (Lack of Innovation Cripples PH COVID Response), **7174** (Costs and Benefits of New Disciplines on Electronic Commerce)	**7172** (Land Tenure, Access to Credit, and Agricultural Performance of ARBs, Farmer Beneficiaries, and Other Rural Workers)	0.001092	0.909091	0.001201	252.27	20
6899 (Impacts of TRAIN Fuel Excise Taxes on Employment and Poverty), **6902** (Towards Inclusive Social Protection Program Coverage in the Philippines: Examining Gender Disparities)	**6903** (Improving Human Resource through Mutual Recognition in ASEAN)	0.001365	0.833333	0.001638	118.31	25
7154 (Online Work in the Philippines: Some Lessons in the Asian Context), 7155 (Digital Divide and the Platform Economy: Looking for the Connection from the Asian Experience)	**7156** (Impact of FTA on Philippine Industries: Analysis of Network Effects)	0.001037	0.904762	0.001147	212.45	19
7163 (Impacts of the Sustainable Livelihood Program's Microenterprise Development Assistance with Seed Capital Fund on Poor Households in the Philippines), **7172** (Land Tenure, Access to Credit, and Agricultural Performance of ARBs, Farmer Beneficiaries, and Other Rural Workers)	**7171** (Lack of Innovation Cripples PH COVID Response)	0.001147	0.875000	0.001310	254.38	21

Note: Authors' tabulation from the resulting Association Rule Mining

4.2. Sentiment analysis on violence against women-related news

According to the 2022 Global Gender Gap report conducted by the World Economic Forum (WEF) annually, the Philippines ranks 19th out of 146 countries assessed, garnering a score of 0.783. Since its inception in 2006, the Global Gender Gap Report has been measuring through a composite index the gaps between men and women based on four components, viz., (i) Economic Participation and Opportunity, (ii) Health and Survival, (iii) Educational Attainment, and (iv) Political Empowerment. The country fell two spots from its 2021 rankings, where it placed 17th out of 156 countries with the indicator on political empowerment remaining low. While the Philippines is the only Asian country in the top 20, this performance is still far from its highest recorded rank at 7th place out of 145 countries in the 2015 Global Gender Gap Index.

Figure 4. Economy Profile: Philippines, 2022 Global Gender Gap Report

Source: 2022 Global Gender Gap Report, WEF

One of the areas of concern on gender equality is the prevalence of Violence against Women (VAW). Two sources of data on VAW are VAW-related crime reports from the Philippine National Police (PNP) and the reported number of VAW cases served by the Department of Social Welfare and Development (DSWD), which both have displayed a downward trend in the last seven years (**Table 7**). A spokesperson for the Inter-Agency Council on Violence Against Women and their Children emphasized, however, that this observed sudden drop does not necessarily mean that VAW cases have dramatically fallen, nor does this signal low prevalence of gender-based crimes[11]. Examining administrative data must be exercised with caution, especially since the reporting process became possibly more challenging during the implementation of lockdown restrictions due to COVID-19 pandemic when women victims were forced to stay at home, closely monitored by their abusers.

[11] https://manilastandard.net/news/national/357417/fewer-violence-vs-women-cases-but-more-unreported.html

Table 7. VAW-related administrative data from PNP and DSWD

Year	Cases Served by DSWD	Cases Reported to PNP
2015	532,998	41,049
2016	355,133	40,684
2017	4,242	34,143
2018	5,883	18,947
2019	3,418	21,366
2020	1,035	15,828
2021	1,208	12,492

Source: Men and Women Factsheet, Philippine Statistics Authority

With growing use of internet and social media, news and media companies have also utilized online channels to deliver news. There are 561 VAW-related news contents web scraped from online news sources with strong presence such as ABS-CBN, The Philippine Daily Inquirer, Manila Bulletin, The Manila Times, and Rappler. These five news outlets registered approximately 80 million monthly website visits according to estimates from SimilarWeb, a free-version online platform that measures digital traffic (**Table 8**).

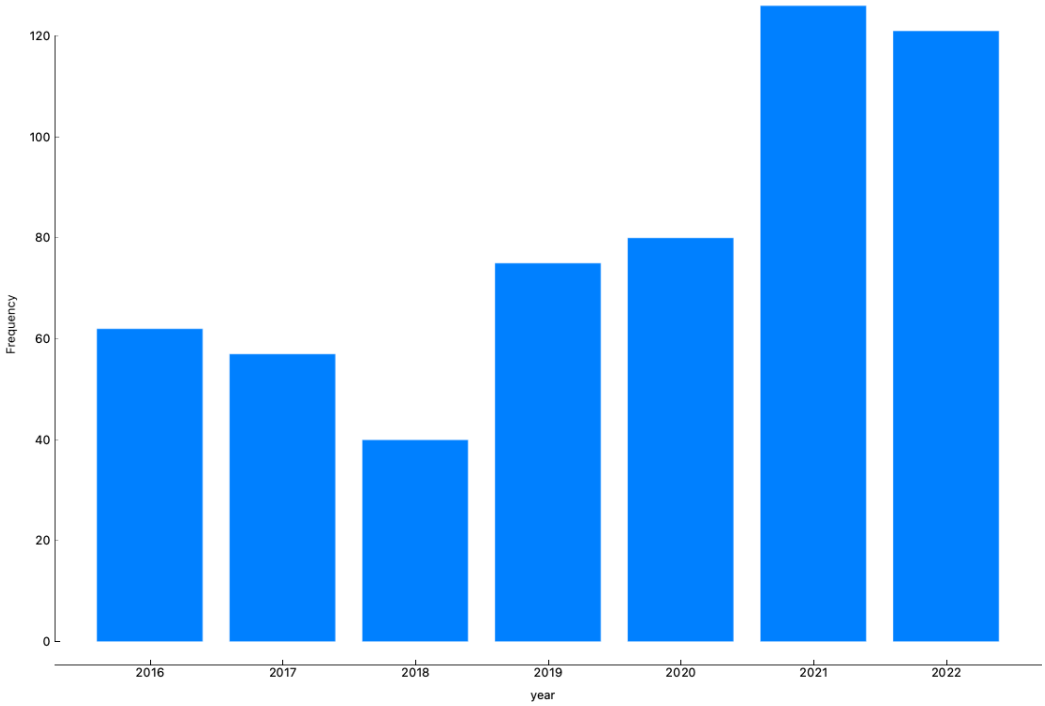
Table 8. Total Website Visits in Selected Online News Sources in the Philippines

News Source	Website	Website Visits (in millions)		
		August 2022	September 2022	October 2022
ABS-CBN News	news.abs-cbn.com	8.2	8.4	8.9
Manila Bulletin	mb.com.ph	4.0	4.8	5.0
Manila Times	manilatimes.net	5.2	6.1	6.5
Philippine Daily Inquirer	inquirer.net	61.8	53.2	47.0
Rappler	rappler.com	10.6	12.2	12.5
TOTAL		**89.8**	**84.7**	**79.9**

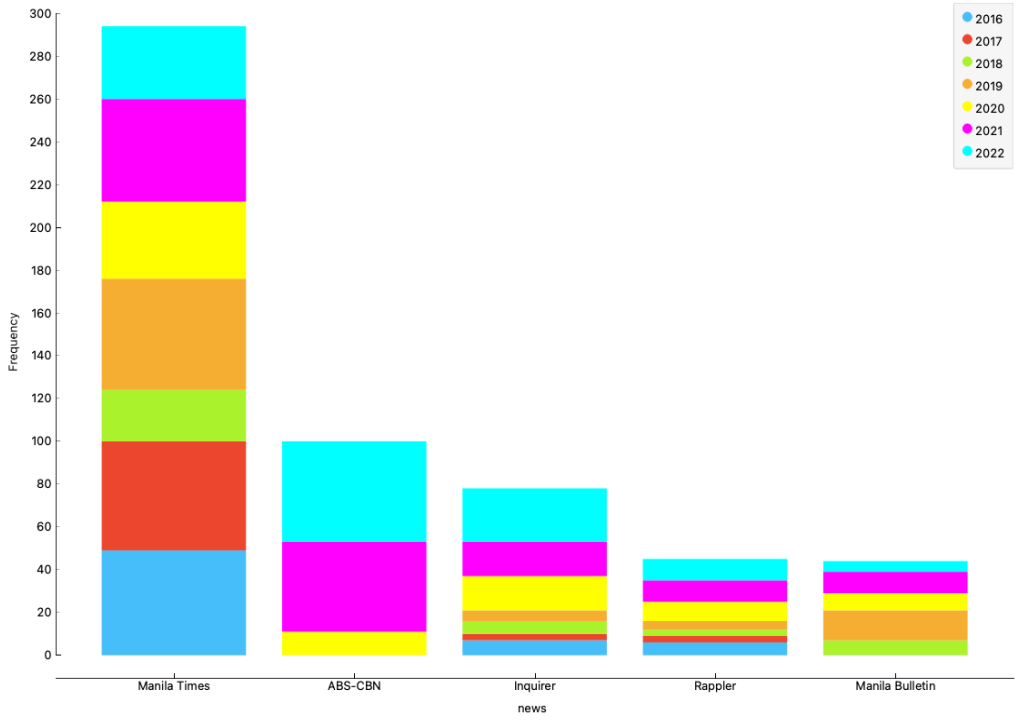
Source: Similarweb

This section attempts to analyze language, speech, and trends Rappler using sentiment analysis and topic modeling from online news articles related to VAW in the Philippines for the period 2016-2022. While reported cases from PNP and DSWD show a downward trend, coverage of VAW related-news from the five news outlets are highest in 2021 and 2022 compared to previous years (**Figure 5**). While the variance in the trends in these two sets of data may be merely some information bias, but it may also reflect the likely gap between reported VAW cases in government and actual VAW incidents.

Figure 5. VAW-tagged news articles from select online news sites, for period 2015-2022: (a) by year, (b) by source



(a)



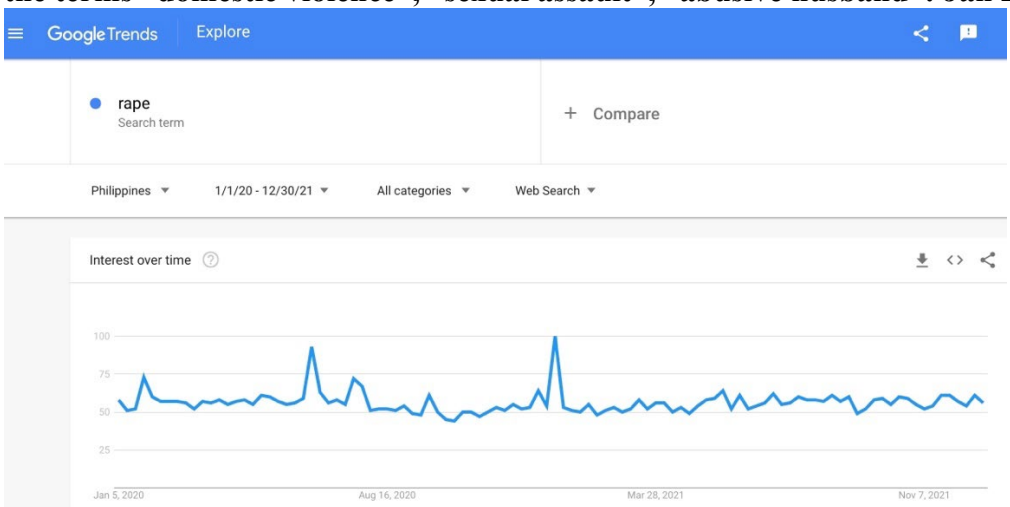
(b)

Sources: from ABS-CBN, The Philippine Daily Inquirer, Manila Bulletin, The Manila Times, and Rappler

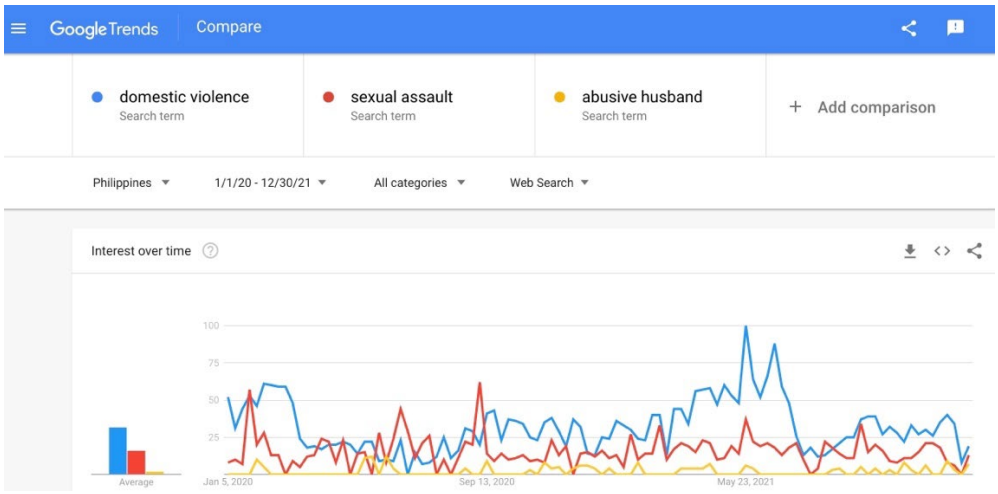
A rapid assessment conducted by UN Women (2021) in 13 select countries suggests a rising prevalence of VAW amid the pandemic, and a small proportion (1 in 10) of women reporting that they would seek help from the police if they experience domestic violence. Further, another study, this time involving the Philippines and 7 other countries, suggests an increase

of 10 percent in the average search volume for help-seeking keywords in 2020 in the Philippines compared to a year before (UN Women, UN Fund for Population Activities and Quilt.ai 2021). Further, the same report notes a 953 percent change in number of tweets with misogynistic language between October 2019 and October 2020 in the Philippines, These reports from UN Women provides evidence on the bias in trends in the government estimates of VAW incidents, and why trends in social media analytics presented here may give evidence of a shadow pandemic, i.e., conditions that suggest an increase in VAW incidents amid the pandemic. However, one should also note possible biases in social media data, with Google trends on the term "rape" not providing clear evidence of an increase during the pandemic, though there was slight rise, though erratic, in other related terms (**Figure 6**).

Figure 6. Google Trends in the Philippines on Searches for (a) the term "rape"', and (b) the terms "domestic violence", "sexual assault", "abusive husband": Jan 2020-Dec 2021



(a)



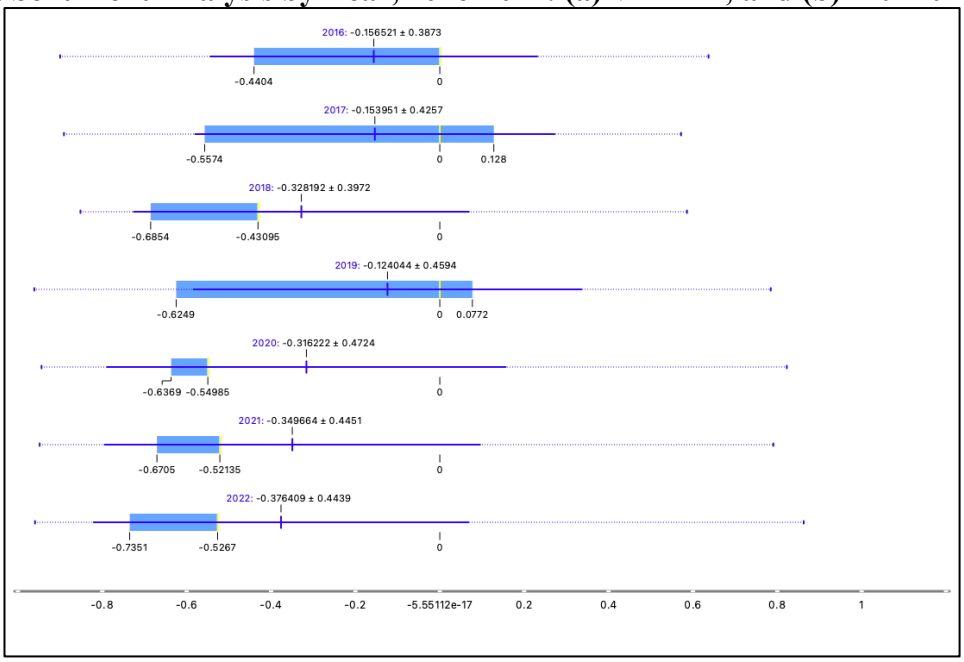
(b)

Sources: https://trends.google.com/trends/explore?date=2020-01-01%202021-12-30&geo=PH&q=rape; https://trends.google.com/trends/explore?date=2020-01-01%202021-12-30&geo=PH&q=domestic%20violence,sexual%20assault,abusive%20husband

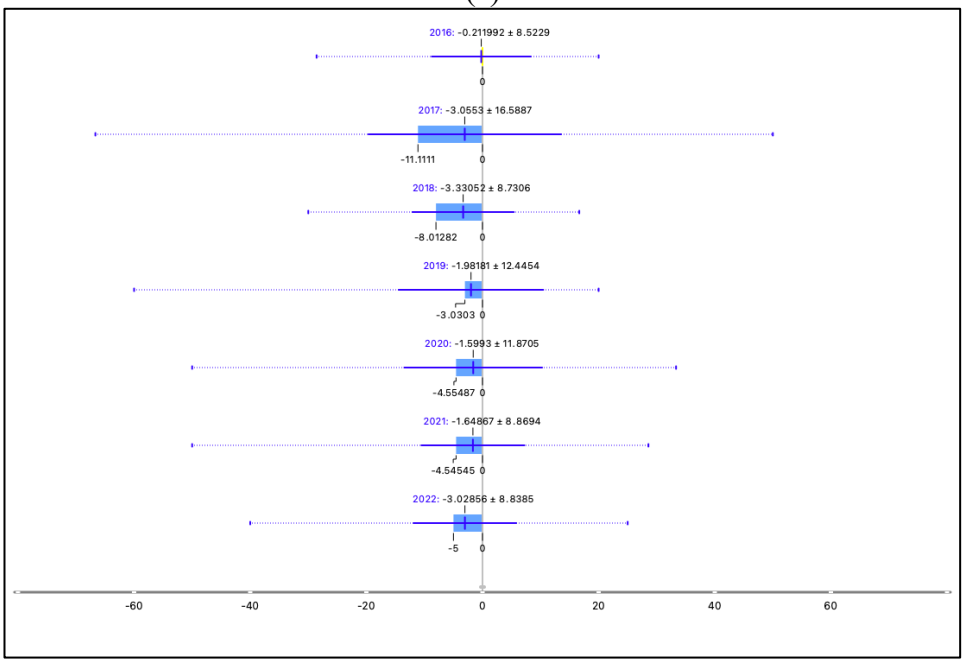
According to Boiy and Moens (2008), sentiment analysis, also called opinion mining, is a natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text. It can help monitor

citizens' perceptions on certain issues and implementation of public policies (Ceron and Negri, 2015). The Valence Aware Dictionary for sEntiment Reasoning (VADER) method uses a list of positive and negative words with scores depending on intensity and generates scores for each category (positive, negative, neutral) and appends a total sentiment score called a compound. Meanwhile, Liu Hu method focuses on polarity classification and computes a single normalized score of sentiment in the text (negative score for negative sentiment, positive for positive, 0 is neutral).

Figure 7. Sentiment Analysis by Year, 2016-2022: (a) VADER; and (b) Liu Hu



(a)



(b)

Observing the results by year, distribution of sentiment scores of VAW-related news contents are concentrated toward negative scores for both methods (**Figure 7**). It can be further observed that VAW-tagged news articles in the last three years reflect more negative compound scores

using the VADER method. However, positive and negative scores provide limited insights to address a policy issue.

On the other hand, an illustration of a word cloud from **Figure 8** also offers insights on how the issue on VAW is framed among news articles in the Philippines, with the size of the word reflecting its frequency or importance. The term "children" has been closely associated with VAW, which may be attributed to Republic Act 9262, a national legislation addressing violence against women and children (VAWC).

Figure 8. Word cloud of VAW-related news articles

Topic modelling was also explored to analyze these news contents and provide more information on most discussed topics related to VAW. Topic modelling is a statistical modelling technique that discovers abstract topics in clusters of similar words found in news articles. Using the Latent Dirichlet Allocation method, web scraped dataset of news leads is analyzed by illustrating the contents as random mixtures over latent topics, where each topic is characterized by a distribution over words. Topic 6 which generated the highest marginal topic probability at 22.2%, covered words such as "violence", "philippines", "child", "manila", "sexual", "anti", "cases", "year", "victims", "help" which could pertain to news contents related to the incidence of child sexual abuse cases in the Philippines (**Table 9**).

Table 9. Topic Modelling of VAW-tagged News Leads

Topics	Marginal Topic Probability	Keywords
1	7.8%	abuse, national, must, philippine life, report, accused, quezon, pnp, local
2	6.4%	state, ferdinand, jr, jalandoni, kit, two, thompson, abused, strengthen, victim
3	18.5%	women, children, city, act, protection, marcos, republic, cebu, opposing, recorded
4	6.3%	gender, right, ph, based, death, vulnerable, raised, opposed, advocates, responsive
5	6.8%	law, president, government, barangay, leni, violence, condemned, racism, case, programs
6	22.2%	violence, philippines, child, manila, sexual, anti, cases, year, victims, help
7	6.9%	police, 9262, fighting, npa, war, crimes, vote, projects, program, funding
8	7.8%	men, also, crime, new, saying, francis, physical, society, three, survey
9	6.8%	2022, marriage, trafficking, get, desiderio, complaint, ex, internet, inquirer, back
10	7.7%	feb, address, could, chief, social, russian, end, alexander, gesmundo, justice

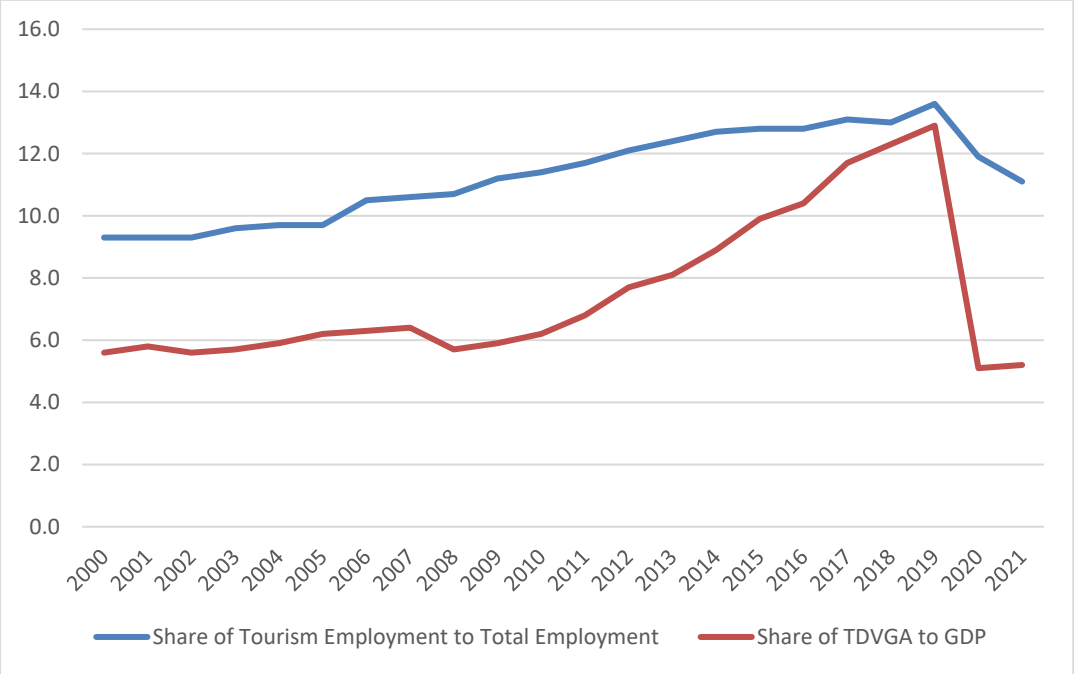
Note: Authors' tabulation

4.3. Text mining twitter data on Philippine tourism

Among the three major sectors of the economy, the services sector bore the brunt of the effects of the pandemic, given the mobility restrictions that were used by government to prevent the spread of infections. Within the services sector and likely across the entire economy, tourism has been one of most negatively-impacted sub-sectors. While other countries in the region promoted local tourism to mitigate the impact, this was hardly done in the Philippines. As restrictions for travel have started to ease and with leisure travel already allowed in most parts of the country and the entire world, examining the latest available tourism data may aid decision makers in formulating policies and programs to reinvigorate tourism and strengthen the competitiveness of the Philippine tourism industry.

Official sources of tourism statistics include tourism satellite accounts annually compiled by the Philippine Statistics Authority (PSA). **Figure 9** shows the trends in the share of tourism industries direct gross value added (TDVGA) to the gross domestic product (GDP), and the relative share of employment of the subsector to total employment in the country from 2000 to 2021. In 2020, the performance of the tourism sub-sector showed a sharp decline with data for 2021 hardly getting any better. While the PSA conducts other data collection activities on domestic travel, these activities are not frequent owing to cost issues. The Census of Population and Housing (CPH), for instance, collects information on long-term travel/migration but the CPH is a decennial data collection activity although it was last conducted in 2020. The Household Survey on Domestic Visitors (HSDV) was also last conducted in 2012.

Figure 9. Selected Derived Indicators from Philippine Tourism Satellite Accounts



Source: PSA

Tourist destination-specific data provide granular information on the tourism market. Tourism demand statistics available from the Department of Tourism (DOT), include monthly data on air visitor arrivals to the Philippines by country of residence but there is little to no information on specific tourist destinations in the Philippines. While administrative data at the local government level is a potential data source, reports on the number of tourist arrivals on certain tourist spots are also not be regularly updated.

Available data on tourist arrivals in Boracay reported on the website of the local government of Malay, Aklan show comparative statistics on tourist arrivals from two years ago (**Table 10**). Timeliness is a factor to consider in examining tourism data, since the peak season of tourist destinations may differ. There is limited tourist arrival from foreigners and overseas Filipino workers (OFWs) in 2020 amid mobility restrictions. More information beyond these statistics is needed to formulate responsive marketing and promotion strategies for Boracay as a top-of-mind tourist destination in the Philippines.

Table 10. Comparative Statistics of Boracay Tourist Arrival (CY 2019-2020)

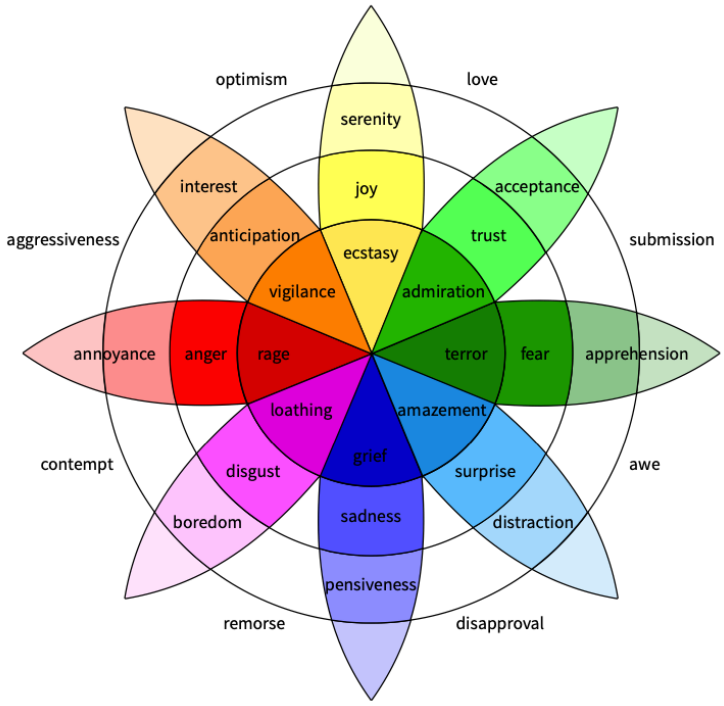
Month	2019				2020			
	Foreign	Local	OFW	Total	Foreign	Local	OFW	Total
January	101,613	57,534	8,467	167,614	98,301	60,213	8,556	167,070
February	115,143	52,872	4,680	172,695	40,436	57,146	6,252	103,834
March	92,835	74,226	5,146	172,207	6,921	27,783	1,730	36,434
April	91,981	123,007	7,342	222,330				
May	79,284	135,543	6,311	221,138				
June	83,292	100,144	6,008	189,444		81		81
July	93,297	72,495	6,553	172,345		668		668
August	95,703	62,435	4,845	162,983		1,631		1,631

Month	2019				2020			
	Foreign	Local	OFW	Total	Foreign	Local	OFW	Total
September	67,474	51,159	2,061	120,694		2,646		2,646
October	70,676	67,493	2,840	141,009		2,630		2,630
November	72,952	63,101	3,703	139,756		4,154		4,154
December	73,369	72,424	6,591	152,384		15,307		15,307
Total	**1,037,619**	**932,433**	**64,547**	**2,034,599**	**145,658**	**172,259**	**16,538**	**334,455**

Source: Malay, Aklan local government unit (https://malay.gov.ph/index.php/for-visitors/tourist-arrival)

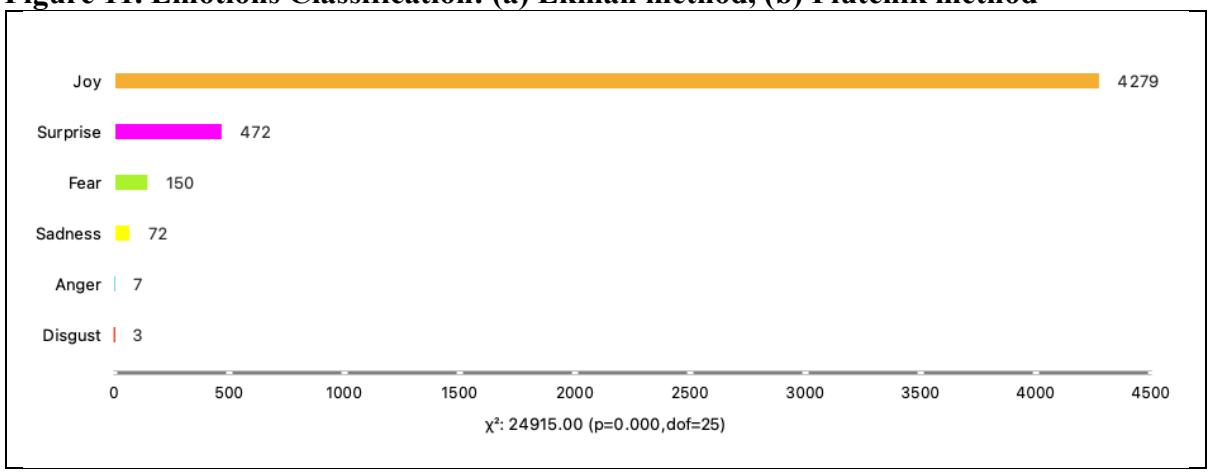
Despite issues on representativeness and limitations on scraping historical tweets, sentiments from 4,986 tweets related to Philippine tourism for the period July-September 2022 were collected using key words "travel" and "Philippines". Emotions classification is another method of sentiment analysis via emotion-recognition algorithms from volumes of trained datasets using neural networks, although there are still limitations in terms of identifying neutral and sarcastic tweets using the most commonly used methods in natural language processing: Ekman's basic set of emotions and Plutchik's wheel of emotions (Colneric, 2019). Paul Ekman, an American psychologist, is a pioneer on research of emotions from his study, "*An Argument of Basic Emotions*" in 1980. He identified six basic sets of emotions such as *anger, disgust, fear, joy, sadness, surprise*, that can be distinguished based on facial expressions, nervous system activity, and responses to certain contexts and tasks. Meanwhile, Robert Plutchik developed the Wheel of Emotions which defined eight basic, pairwise contrasting emotions, such as *anger, disgust, fear, joy, sadness, surprise, trust, anticipation*, that can be depicted in a circular representation with varying intensity levels (**Figure 10)**.

Figure 10. Plutchik's Wheel of Emotions

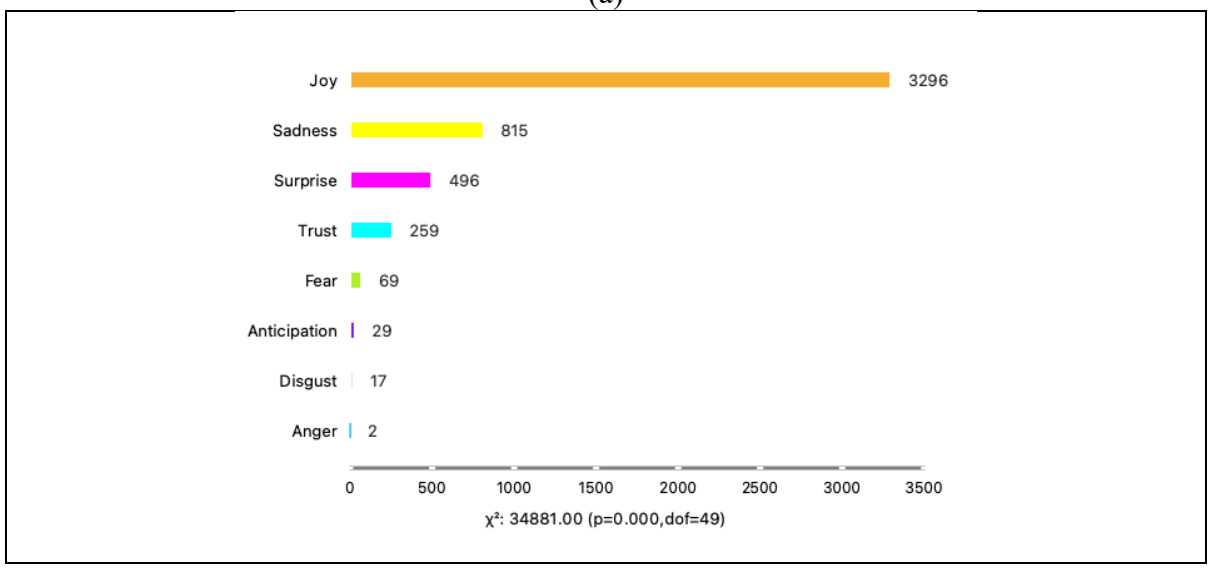


Emotionally labelling tweets can help assess satisfaction on Philippine tourism among twitter users, with both methods generating Joy as the most prominent emotion at 85.1% and 66.1% of the scraped tweets, for Ekman and Plutchik, respectively (**Figure 11**).

Figure 11. Emotions Classification: (a) Ekman method, (b) Plutchik method



(a)



(b)

Particular topics, hashtags, or words frequently associated with Philippine tourism are identified in **Table 11** in terms of the most frequently used tweets, and the entire list is illustrated in a word cloud in **Figure 12**. Tourist destination in the Philippines such as Manila, Boracay, and Cebu are among the top 20 words in terms of word count. "It's more fun in the Philippines", a tourism campaigned launched 10 years ago under the leadership of former DOT Secretary Ramon Jimenez, may still be an effective tourism tagline as the 60th most frequent word in the word count.

Table 11. Top 60 Most Frequent Words on Philippine Tourism-related Tweets

Rank	Word	Word Count	Rank	Word	Word Count
1	philippines	5325	31	get	172
2	travel	4834	32	thailand	170
3	island	509	33	country	170
4	manila	431	34	explore	168
5	boracay	413	35	see	167
6	beach	387	36	first	167
7	japan	316	37	palawan	166
8	singapore	271	38	philippinestravel	166
9	visit	245	39	people	164
10	asia	242	40	vacation	162
11	cebu	240	41	visa	161
12	tour	239	42	via	155
13	boracayisland	234	43	travelph	154
14	like	229	44	countries	152
15	world	223	45	tokyo	149
16	resort	215	46	trip	145
17	new	211	47	find	145
18	city	211	48	best	144
19	photography	211	49	year	141
20	time	210	50	back	139
21	us	209	51	korea	138
22	one	207	52	adventure	138
23	2022	205	53	nature	136
24	travels	203	54	instajapan	135
25	go	192	55	want	132
26	travelblogger	190	56	package	130
27	tourism	190	57	safe	128
28	read	185	58	next	125
29	south	183	59	covid	120
30	travelphotography	175	60	itsmorefunintheplippines	118

Figure 12. Word Cloud of Philippine Tourism-related Tweets

Boracay, the fifth most frequent word, is among the keywords on Topic 1 with a marginal topic probability of 33.8%--roughly a third of tweets cover discussions around or related to Boracay (**Table 12**).

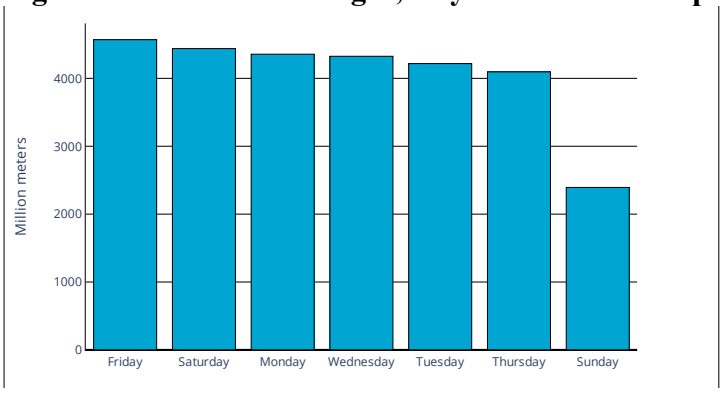
Table 12. Topic Modelling of Tweets on Philippine Tourism

Topics	Keywords	Marginal Topic Probability
1	philippines, travel, boracay, beach, asia, photography, cebu, boracayisland, 2022, destination	33.8%
2	singapore, japan, tokyo, travelblogger, instajapan, sunset, sea, days, long, trying	5.4%
3	people, tourism, dive, family, many, wrong, visited, half, industry, transportation	6.5%
4	island, world, resort, thailand, guide, malaysia, islandlife, want, next, itsmorefuninthephilippines	8.6%
5	travel, philippines, manila, like, read, new, get, philippinestravel, paradise, good	14.0%
6	us, super, pakistan, community, japanese, location, k, fully, libya, cannot	4.9%
7	explore, nature, canon, photooftheday, back, eos, bantayanisland, outdoors, lensculture, travelling	6.6%
8	visit, visa, year, beautiful, open, covid, place, apply, fun, taiwan	6.6%
9	countries, food, photo, even, watch, sept, india, meal, colors, asian	5.2%
10	tour, culture, first, would, trip, solo, bangkok, abroad, make, part	5.8%

4.4. Analyzing traffic congestion data

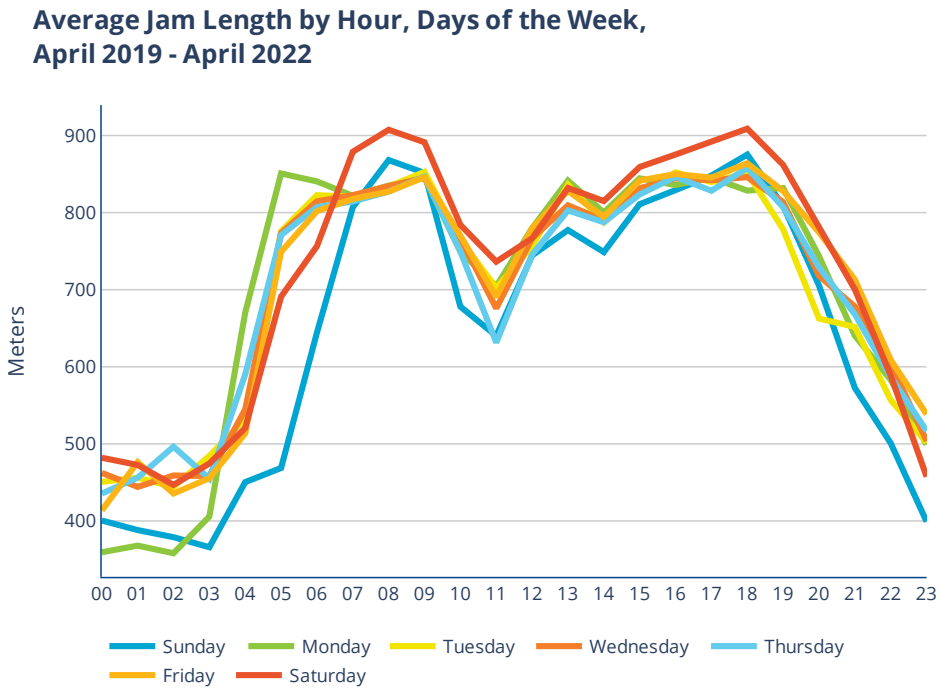
From April 2019 to April 2022, across days of the week, Friday is the most congested day in terms of total jam length, followed by Saturday and Monday (**Figure 13**). Friday and Monday mark the end and start of the workweek, respectively while Saturday is often spent by Filipinos addressing matters or activities beyond work or school (e.g., shopping, hanging out with friends, hobbies). On the other hand, Sunday, which is recognized in the country as a time for rest and family, is the least congested day.

Figure 13. Total Jam Length, Days of the Week: April 2019-April 2022



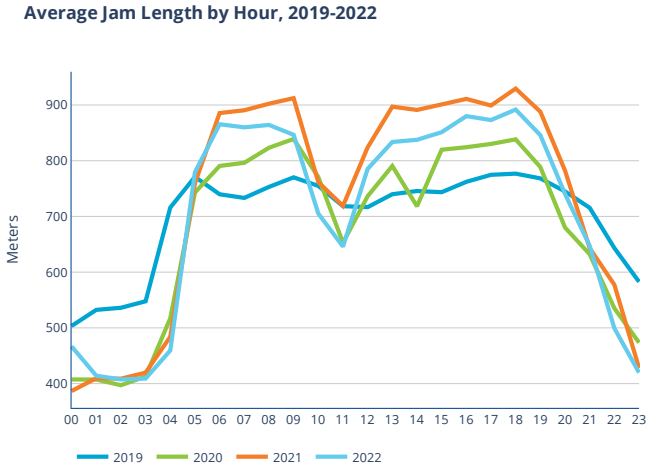
Looking at average hourly trends in jam length, there are two peak hours for all days of the week – one in the morning and one at night. In general, these peak hours are between 6:00 to 9:00 am as well as 6:00 pm. Saturday is recorded as having the highest average jam lengths and, quite surprisingly, the second highest is from Sunday (**Figure 14**).

Figure 14. Average Jam Length by Hour, Days of the Week: April 2019-April 2022

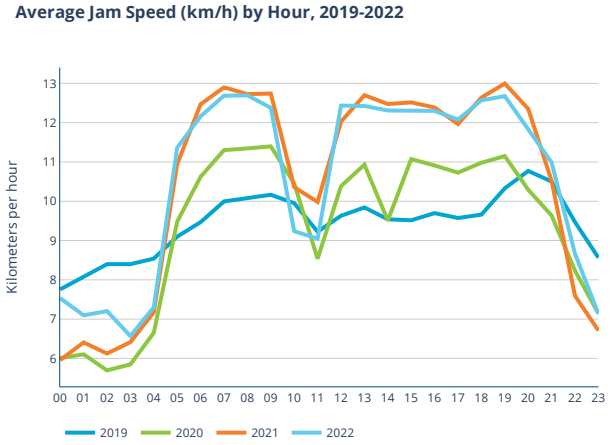


Trends in the hourly jam length, speed, and delay moved in a similar way for 2020, 2021, and 2022 (**Figure 15**). In general, average jam lengths and speed were less in 2020 compared to these two years, which is clearly due to strict mobility restrictions imposed at the onset of the pandemic. Average jam delays, on the other hand, were higher in 2020 compared to 2021 and 2022, which means that speeds in jams in the first year of the pandemic differed more compared to free flow speeds on the aggregate. Interestingly, the data reveals shorter average jam lengths in 2019 compared to the pandemic but movement in traffic was relatively slower this year as evidenced by less speeds and longer delays.

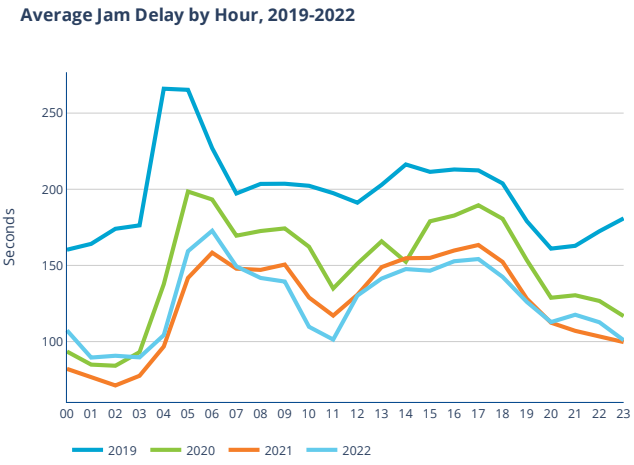
Figure 15. Traffic Jam Indicators by Hour, 2019-2022: (a) Average Jam Length (b) Average Jam Speed; (c) Average Jam Delay



(a)



(b)

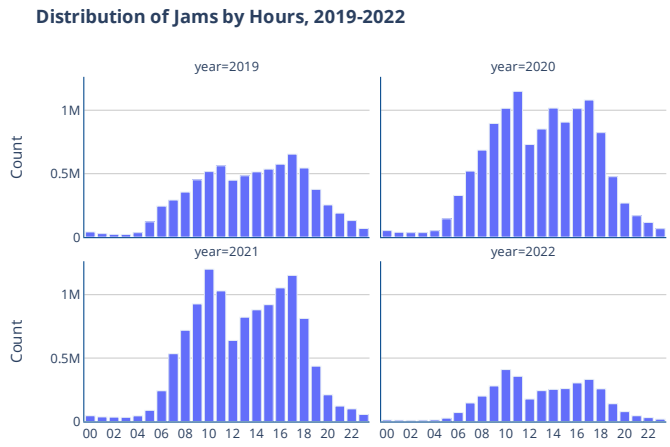


(c)

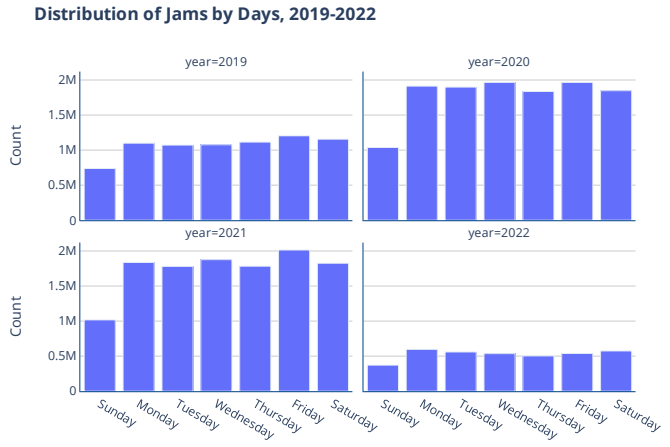
Although it is reasonable to expect longer jam lengths in 2019 compared to the pandemic period – particularly 2020 – since no stringent controls were imposed during this period, but data shows a contrasting result. This may be due to data-related matters or changes in conditions on the ground (e.g., new traffic management approaches, behavior of drivers, alternative routes, road closures, or more vehicles in recent years). Probing the former, the distribution across days, months, and hours of reports is checked for each year.

In terms of days and hours, no noticeable differences are seen in terms of composition (i.e., shares) but level differences are quite considerable (**Figure 16**). Regarding months, January, February, and March were not represented in 2019, with the first having the most shares for 2020 and 2021. Since January 2020 is technically not part of the pandemic period while the second most represented month in 2020 (December) marked a return to "normalcy" - with the reopening of establishments and resumption of certain socioeconomic activities – jam information in this period may be more similar to the pre-pandemic levels.

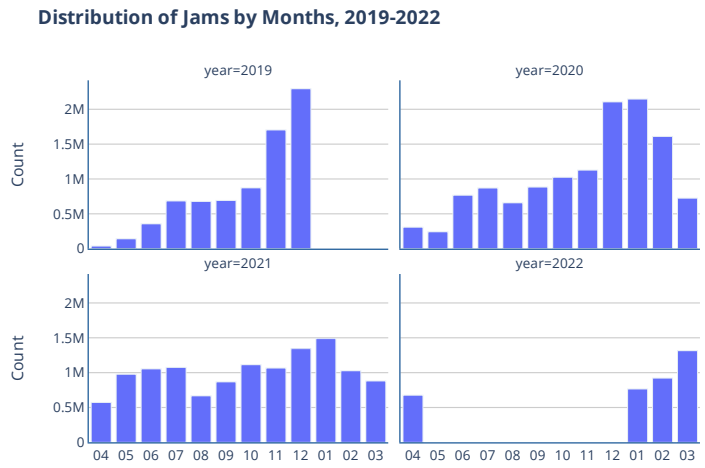
Figure 16. Distribution of Jams, 2019-2022: (a) by Hours; (b) by Days; (c) by Months



(a)



(b)



(c)

5. Policy Issues and Ways Forward

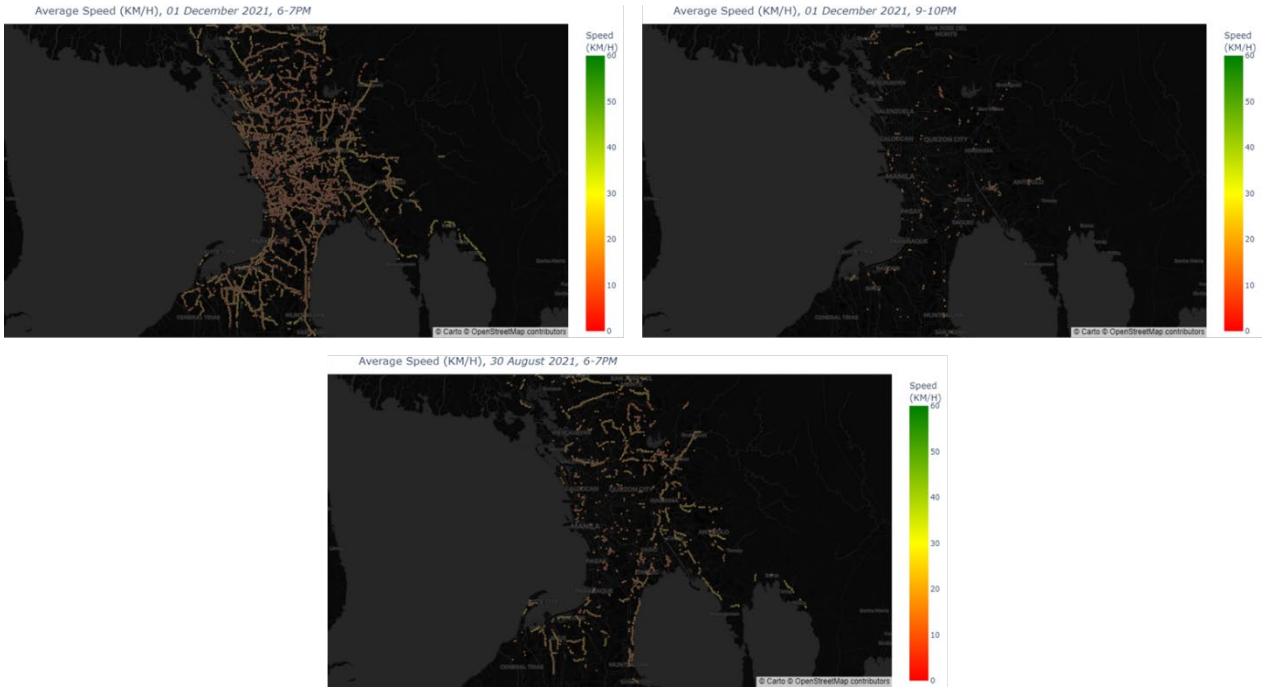
5.1. Continuous examination of data and capacity building for data analytics at PIDS

It will benefit the Institute to continue a regular examination of PIDS web download data, say on a semestral or annual basis by way of conducting a "market basket" analysis with association rules to train the algorithm to identify meaningful patterns of association beyond themes but also in individual publications. This information can help PIDS develop targeted campaigns for communicating publications, including the promotion of events. Capacity building programs for data analytics on new data sources can also help the Institute in harnessing innovative data sources to provide policy advises with near real time information, while uncovering limitations of these data sources.

5.2. Using traffic data for policy development

As the dataset examined in this report contains latitude-longitude coordinates of jam reports, geospatial analysis of traffic congestion can be done at a more refined resolution. For computational efficiency, location data can be bucketed into hexagonal grid cells, as was done in a 2018 study by the University of Pennsylvania. The use of grids will cover any road network of choice without overlaps and gaps while the choice of hexagons will allow analyses with connectivity and movement dimensions (Hackernoon, 2020).

Sample maps that can be generated using hexagonal binning are shown below:



To analyze traffic congestion using this approach, a primary road network of interest must first be chosen, which will then be represented in hexagonal grid cells of a given size. Jams information from the Waze dataset will then be assigned to the cells in which they occur before an aggregation operation (e.g., sum, mean, min, max) is applied. Transformation into this format will also allow the resulting dataset to be shared to PIDS and its institutional partners without compromising agreements with the DDP. End users who access these files can then use them as inputs for further analysis such as unsupervised or supervised machine learning methods, as well as an examination of the determinants of traffic jams.

From a policy perspective, traffic congestion in and of itself is not the main issue but it is the economic, health, and environmental costs associated with it that cause concerns for the government and society in general. Jams, however, do play a key – and sometimes fundamental – role in forming these wide-ranging development problems, which makes spatial and temporal analysis of it a valuable use case for understanding obstacles for socio-economic development.

For these types of analysis to be more practical and impactful, however, traffic congestion data needs to be merged with other data containing information on health, environment, weather and climate, economy, infrastructure, among others. This will lead to more nuanced analysis and predictive analytics, as well as allow measurement of associated impact across different sectors.

If found useful, regular updates from DDP can be requested to keep track of more current trends and emerging patterns on the subject matter. Once there is access to the data, effectively communicating the data and analysis to transport policymakers at the local and national level must also be prioritized. This may involve translating the data analysis into policy recommendations or mentorship for decision-makers in the public sector to fully utilize and harness the resulting analytics from these types of datasets.

5.3. Integrating new data sources with traditional data sources

It is important to recognize that new data sources complement but cannot replace traditional data sources (such as surveys and censuses) that undergo regular processes of data curation to maintain data quality in terms of relevance, accuracy, timeliness, accessibility, interpretability, and coherence. Big Data and other new data sources provide a fast and cheap stream of information thus enhancing responsiveness to socio economic development problems being addressed in the policy cycle. Despite the limitations of new data sources, generating statistics from them can also be promising since costs are generally little, if not none, and data is timely, in fact, in near real-time (compared to the much slower pace of production of statistics from traditional data sources). Governments, such as the European Union, have growing interest in exploring ICT-based methods of exploiting political contents in various web sites and social media accounts of EU citizens by employing opinion mining and sentiment analysis "to obtain a better understanding of the needs and problems of society, and also the perceptions and feelings of the citizens, and to formulate effective public policies" (Charalabidis *et al.* 2015, p. 158). Meanwhile National Statistics Offices (NSOs) collect monthly prices of data to produce statistics on inflation, but twitter conversations, e.g. on the price of rice in Jakarta, have provided a near-real time and innovative way to monitor actual prices (UN Global Pulse 2012). Further, big data provide a way of increasing granularity. ADB, in cooperation with NSO of Thailand and World Data Lab, has worked on integrating satellite imagery with census and survey data to yield high quality poverty statistics at small area areas (ADB 2021b).

5.4. Mitigating risks of using personal data in new data sources

There are, however, weeds among the wheat in the use of new data sources. Risk assessment and risk mitigation on use of big data and other non-traditional data sources is necessary since the world of big data and hyperconnectivity no longer guarantees irreversible de-identification that can yield potential harms posed to individuals and to identifiable groups or populations. Further examination is needed to identify the thresholds at which deidentified data is no longer personal: is it feasible (and practical) to seek consent in situations of emergency, development response when data is de-identified? There needs to be a balance between protecting data privacy and harnessing use of new data sources for safeguarding civil rights, ensuring fairness, and preventing discrimination.

6. References

Albert, J. R. G. 2021. Towards Measuring the Platform Economy: Concepts, Indicators, and Issues. Chapter 2 of Asian Development Bank (ADB) December 2021 Publication "Managing The Development Of Digital Marketplaces In Asia" Edited By Cyn-Young Park, James Villafuerte, and Josef T. Yap. Pasig City, Philippines: ADB. https://www.adb.org/sites/default/files/publication/761016/managing-development-digital-marketplaces-asia.pdf (accessed on November 12, 2022).

Albert, J.R.G, Martinez, A., C. De Dios, I. Sebastian-Sameniego, K.G. Miradora, and J.A. Lapuz. 2019. *Readiness of National Statistical Systems in Asia and the Pacific for Leveraging Big Data to Monitor the SDGs.* ADB Briefs No. 106. March 2019. Asian Development Bank. Pasig City, Philippines: ADB. http://dx.doi.org/10.22617/BRF190026-2 (https://www.adb.org/publications/national-statistical-systems-big-data-sdgs) (accessed on February 18, 2022)

Albert, J.R.G. and Martinez, A. 2018. *The Future of Data Today*. Development Asia. Asian Development Bank. Pasig City, Philippines: ADB. https://development.asia/explainer/future-data-today (accessed on February 18, 2022)

Ang, M. 2022. *Manila eighth most traffic-congested city in the world*. Yahoo! News. https://ph.news.yahoo.com/manila-eighth-most-traffic-congested-city-in-the-world-011555951.html (accessed on October 24, 2022)

Asian Development Bank (ADB). 2021a. Practical Guidebook on Data Disaggregation for The Sustainable Development Goals. Pasig City, Philippines: ADB. https://www.adb.org/sites/default/files/publication/698116/guidebook-data-disaggregation-sdgs.pdf (accessed on November 12, 2022)

ADB. 2021b. *Mapping the Spatial Distribution of Poverty Using Satellite Imagery in Thailand*. http://dx.doi.org/10.22617/TCS210112-2 Pasig City, Philippines: ADB. (accessed on February 23, 2022)

Blei, D.M. A.Y. Ng, M.I. Jordan. 2003. Lafferty, John (ed.). *Latent Dirichlet Allocation*. Journal of Machine Learning Research. 3 (4–5): pp. 993–1022. Cambridge, Massachusetts, USA: MIT Press. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed on October 17, 2022)

Boiy, E. and M.F. Moens. 2008. A machine learning approach to sentiment analysis in multilingual Web texts. Information Retrieval 12, 526–558. MA, United States: Kluwer Academic Publishers. https://doi.org/10.1007/s10791-008-9070-z (https://www.semanticscholar.org/paper/A-machine-learning-approach-to-sentiment-analysis-Boiy-Moens/462754e479123ef2167c3946fd4cd04a242c2767) (accessed on July 4, 2022)

Boquet, Y. 2013. *Battling congestion in Manila*. Transport and Communications Bulletin for Asia and the Pacific No. 82. Bangkok, Thailand: United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP) . https://www.unescap.org/sites/default/files/bulletin82_Article-4.pdf (accessed on October 22, 2022)

Brachman, R. and T. Anand. 1996. *The process of knowledge discovery in databases: A human-centered approach, ' in Advances in Knowledge Discovery and Data Mining*, U. Fayyad, G. Piatetsky-Shapiro, S. P.Amith, and R. Uthurusamy, Eds. Cambridge, MA: MIT Press 1996, pp. 37-58.

Brackstone, G. 1999. *Managing Data Quality in a Statistical Agency*. Survey Methodology, December 1999 Vol. 25, No. 2, pp. 139-149 Ottawa, Canada: Statistics Canada. https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1999002/article/4877-eng.pdf?st=NKWj1vK0 (accessed on October 20, 2022)

Ceron, A. & F. Negri. 2015. *Public policy and social media: How sentiment analysis can support policy-makers across the policy cycle*. 10. 309-338. 10.1483/81600. Vaulx-en-Velin, France: International Public Policy Association. (https://www.ippapublicpolicy.org/file/paper/1435660874.pdf) (accessed on October 24, 2022)

Charalabidis, Y., M. Maragoudakis, and E. Loukis. 2015. *Opinion Mining and Sentiment Analysis in Policy Formulation Initiatives: The EU-Community Approach*. In: , et al. Electronic Participation. ePart 2015. Lecture Notes in Computer Science, vol 9249. Springer, Cham. https://doi.org/10.1007/978-3-319-22500-5_12 (https://link.springer.com/content/pdf/10.1007/978-3-319-22500-5_12.pdf) (accessed on October 24, 2022)

Clark, D. E. & B.M. Cushing. 2004. *Rural and urban traffic fatalities, vehicle miles, and population density*. Accident Analysis Prevention. 36(6), 967–972. Amsterdam, Netherlands: Elsevier. https://doi.org/10.1016/j.aap.2003.10.006

Colneric, N. 2019. *Emotion Recognition on Twitter Using Neural Networks*. Faculty of Computer and Information Science, University of Ljubljana. Ljubljana, Slovenia: University of Ljubljana. http://eprints.fri.uni-lj.si/4432/1/63080007-Niko_Colnerič-Prepoznavanje_čustev_na_Twitterju_z_uporabo_nevronskih_mrež.pdf) (accessed on October 16, 2022)

Cox, D.R., C. Kartsonaki, and R.H. Keogh. 2018. *Big data: Some statistical issues*. Statistics & Probability Letters Volume 136. Pages 111-115. ISSN 0167-7152. Amsterdam, Netherlands: Elsevier. https://doi.org/10.1016/j.spl.2018.02.015. (https://www.sciencedirect.com/science/article/pii/S0167715218300609) (accessed on February 23, 2022)

Data Privacy Act of 2012. (Phil) Republic Act 10173. Manila, Philippines: Republic of the Philippines. https://www.privacy.gov.ph/data-privacy-act/ (accessed on October 27, 2022)

DataReportal. 2022a. *Digital 2022 July Global Statshot Report*. https://datareportal.com/reports/digital-2022-july-global-statshot (accessed on October 17, 2022)

DataReportal. 2022b. *Digital 2022: The Philippines*. https://datareportal.com/reports/digital-2022-philippines (accessed on October 17, 2022)

Diwate, Rahul. 2014. *Data Mining Techniques in Association Rule : A Review*. International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 227 - 229. Chennai, Tamil Nadu, India: AIRCC Publishing. https://www.researchgate.net/publication/282651888 (accessed on October 22, 2022)

Gomez, J.P. & A.K. Robredillo. 2021, June 17. *'Fewer violence vs. women cases, but more unreported'*. Manila Standard. https://manilastandard.net/news/national/357417/fewer-violence-vs-women-cases-but-more-unreported.html (accessed on October 17, 2022)

Hackernoon. 2020. *Why do we use hexagons and not squares to aggregate location data*. Why Do We Use Hexagons And Not Sqaures to Aggregate Location Data | HackerNoon (accessed on October 24, 2022)

Hoehner, C. M., C.E. Barlow, P. Allen, and M. Schootman. 2012. Commuting distance, cardiorespiratory fitness, and Metabolic Risk. American Journal of Preventive Medicine, 42(6), 571–578. Washington, D.C., USA: The American College of Preventive Medicine. https://doi.org/10.1016/j.amepre.2012.02.020

Hu, M. and B. Liu. 2004.. Mining opinion features in customer reviews. In Proceedings of AAAI Conference on Artificial Intelligence, vol. 4, pp. 755–760. July 25–29, 2004. San Jose, California, USA. https://www.aaai.org/Papers/AAAI/2004/AAAI04-119.pdf (accessed on October 27, 2022)

Hutto, C.J. & E.E. Gilbert. 2014. *VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text*. Eighth International Conference on Weblogs and Social Media (ICWSM-14). June 1-4, 2014. Ann Arbor, Michigan, USA. https://doi.org/10.1609/icwsm.v8i1.14550 (https://ojs.aaai.org/index.php/ICWSM/article/view/14550/14399) (accessed on October 24, 2022)

Interaksyon. 2017, October 11. '*ON DATA PRIVACY | Can you use public information on social media freely?*' Philstar.com. https://interaksyon.philstar.com/national/2017/10/11/102741/on-data-privacy-can-you-use-public-information-on-social-media-freely/ (accessed on October 26, 2022)

Jakarta Government. *Jakarta Smart City: Beranda*. http://smartcity.jakarta.go.id (accessed on February 19, 2022)

Kudarvalli, H., & Fiaidhi, J. 2020. Detecting *Fake News using Machine Learning Algorithms* (Version 1). TechRxiv. https://doi.org/10.36227/techrxiv.12089133.v1 (https://www.techrxiv.org/ndownloader/files/22227396) (accessed on December 28, 2022)

Litman, T. 2013. *Smarter Congestion Relief in Asian Cities*. Transport and Communications Bulletin for Asia and the Pacific. 82(13), pp. 1-15. Bangkok, Thailand: UNESCAP. https://www.unescap.org/sites/default/files/bulletin82_Article-1.pdf (accessed on October 31, 2022)

Liu, B. M. Hu, and J. Cheng. 2005. *Opinion Observer: Analyzing and Comparing Opinions on the Web*. Proceedings of the 14th International World Wide Web conference (WWW-2005). May 10-14, 2005. Chiba, Japan. https://www.cs.uic.edu/~liub/publications/www05-p536.pdf (accessed on October 22, 2022)

Martinez, A., J.R.G. Albert, I. Sebastian-Sameniego, K.G. Miradora, and J.A. Lapuz. 2018. *Big Data Can Transform SDG Performance. Here's How.* Asian Development Blog. Asian Development Bank. Pasig City, Philippines: ADB. https://blogs.adb.org/blog/big-data-can-transform-sdg-performance-here-s-how (accessed on February 18, 2022)

National Economic Development Authority & Japan International Cooperation Agency. 2014. Roadmap for transport infrastructure development for metro manila and its surrounding areas (region III & region IV-A) in the Republic of the Philippines: Final report3–1-3–7. Tokyo; Japan International Cooperation Agency.

Philippine Statistics Authority (PSA). *Women and Men in the Philippines: Women and Men Factsheet* (2014-2022). Quezon City, Philippines: PSA. https://psa.gov.ph/gender-stat/wmf (accessed on October 18, 2022)

PSA. 2022, April 28. *Highlights on the Economic Performance of Regional Economies for 2021*. Philippine Statistics Authority. Quezon City, Philippines: PSA. https://psa.gov.ph/grdp/highlights-id/167340 (accessed on October 24, 2022)

PSA. 2021. *Philippine Tourism Satellite Accounts (PTSA) Report*. Philippine Statistics Authority. Quezon City, Philippines: PSA. https://psa.gov.ph/system/files/Publication_PTSA_2021.pdf (accessed on October 18, 2022)

Red Dot Foundation. *What is Safecity?* Safecity India. https://www.safecity.in (accessed on February 19, 2022)

Renniger, A et al. 2018. *WAZE: Congestion Predictive Study*. Master of Urban Spatial Analytics 801. WAZE: Congestion Predictive Study (pennmusa.github.io) (accessed on October 24, 2022)

Silver, N. 2012. The signal and the noise: Why so many predictions fail--but some don't. New York: Penguin Press.

Subingsubing, K. 2020. *Metro commuters lost 257 hours to traffic last year*. Philippine Daily Inquirer. https://newsinfo.inquirer.net/1222065/metro-commuters-lost-257-hours-to-traffic-last-year. (accessed on October 31, 2022)

Twitter. *Getting Started: About the Twitter API*. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api (accessed on October 17, 2022)

Twitter, 2022. *Investor Earnings Report for Q2 2022* (published July 2022). https://s22.q4cdn.com/826641620/files/doc_financials/2022/q2/Final_Q2'22_Earnings_Release.pdf (accessed on October 17, 2022)

Twitter. 2022. *Terms of Service* (effective June 10, 2022). https://twitter.com/en/tos (accessed on October 17, 2022)

UN Department of Economic and Social Affairs. 2021. UN *Global Platform Regional Hubs*. UN Global Platform. New York, USA: UN DESA. https://unstats.un.org/bigdata/regional-hubs.cshtml#china (accessed on February 20, 2022)

UN Global Pulse. 2012. *Big Data for Development: Opportunities and Challenges*. New York, USA: UN Global Pulse. https://www.unglobalpulse.org/document/big-data-for-development-opportunities-and-challenges-white-paper/ (accessed on February 20, 2022)

UN Women. 2018. *Gender Equality and Big Data*. New York, USA: UN Women. https://unsdg.un.org/sites/default/files/Gender-equality-and-big-data-en.pdf (accessed on December 28, 2022)

UN Women. 2021. *Measuring the Shadow Pandemic: Violence Against Women During COVID-19*. New York, USA: UN Women. https://data.unwomen.org/sites/default/files/documents/Publications/Measuring-shadow-pandemic.pdf (accessed on November 12, 2022)

UN Women. 2021. UN Fund for Population Activities (UNFPA) and Quilt.ai. *COVID-19 and Violence Against Women: The Evidence Behind The Talk* (Insights from big data analysis in Asian countries). New York, USA: UN Women.
https://data.unwomen.org/sites/default/files/documents/Publications/COVID-19%20and%20VAW_Insights%20from%20big%20data%20analysis_final.pdf (accessed on November 12, 2022)

World Economic Forum. 2022. *Global Gender Gap Report 2022*. Cologny, Switzerland: WEF. https://www3.weforum.org/docs/WEF_GGGR_2022.pdf (accessed on October 14, 2022)

Zhang, K and Batterman, S. 2013. *Air pollution and health risks due to vehicle traffic.* Sci Total Environ. Amsterdam, Netherlands: Elsevier. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4243514/ (accessed on October 25, 2022)

Appendix Table 1. PIDS Publication by ID, Publication Type, and Focus Area

Publication ID	Publication Title	Publication Type	Focus Area	Frequency	Percentage
6929	Situation Analysis of ECCD-F1KD Initiatives in Selected UNICEF-KOICA Provinces	Discussion Papers	Health	1,180	0.0644
6939	Issues and Concerns in the Implementation of PBB at DepEd	Policy Notes	Governance	536	0.0293
6937	Expanding Health Insurance for the Elderly of the Philippines	Discussion Papers	Health	448	0.0245
6492	The Comprehensive Agrarian Reform Program After 30 Years: Accomplishments and Forward Options	Research Paper Series	Land Reform and Property Rights	376	0.0205
6727	Out-of-School Children: Changing Landscape of School Attendance and Barriers to Completion	Discussion Papers	Labor and Education	369	0.0201
6735	Exploring Neural Network Models in Understanding Bilateral Trade in APEC: A Review of History and Concepts	Discussion Papers	Regional Integration	360	0.0197
6097	What Drives Filipino Firms to Innovate	Policy Notes	Technology and Innovation	272	0.0149
7193	Policy, Institutional, and Expenditure Review of Bottom-up Approach Disaster Risk Reduction and Management	Discussion Papers	Governance	272	0.0149

6757	Impact Evaluation D a esign for the CHED K-to-12 Transition Program	Discussion Papers	Labor and Education	271	0.0148
6933	Perception Survey in the Philippines on APEC	Discussion Papers	International Relations and Foreign Policy	268	0.0146
6753	Regulatory Measures Affecting Services Trade and Investment: Financial Services	Discussion Papers	Trade and Industry	266	0.0145
6763	Exploring the Presence of Cognitive Social Capital in Philippine Communities	Discussion Papers	Governance	255	0.0139
6760	Benefit-Cost Analysis of the Resurgent Irrigation System Program of the Philippines	Discussion Papers	Agriculture	246	0.0134
6897	Social Protection and Demand for Health Care among Children in the Philippines	Philippine Journal of Development	Health	237	0.0129
6987	Assessing the Resurgent Irrigation Development Program of the Philippines - Communal Irrigation Systems Component	Discussion Papers	Governance	237	0.0129
7182	Deepening the Narrative: Qualitative Follow-up Study on the Third Impact Evaluation of Pantawid Pamilya	Discussion Papers	Health	232	0.0127
6733	A Public Expenditure Review of Social Protection Programs in the Philippines	Discussion Papers	Poverty	230	0.0126

6762	Senior High School and the Labor Market: Perspectives of Grade 12 Students and Human Resource Officers	Discussion Papers	Labor and Education	230	0.0126
6910	Assessment of the Philippine Local Government Planning and Budgeting Framework	Discussion Papers	Governance	218	0.0119
6917	Regional Analysis of the Philippine Services Sector	Discussion Papers	Trade and Industry	218	0.0119

Appendix Table 2. Association of PIDS publications by Focus Area

Health

lhs	rhs	support	confidence	coverage	lift	count
Health	Governance	0.07	0.28	0.23	1.27	1210
Health	Labor and Education	0.06	0.26	0.23	1.36	1132
Health	Agriculture	0.06	0.25	0.23	1.30	1082
Health	Trade and Industry	0.05	0.22	0.23	1.45	921
Health	(NULL)	0.05	0.20	0.23	0.96	850
Health	Technology and Innovation	0.04	0.17	0.23	1.41	729
Health	Poverty	0.04	0.17	0.23	1.60	722
Health	Infrastructure Transportation and Communication	0.03	0.14	0.23	1.64	605
Health	Gender and Development	0.03	0.14	0.23	1.44	581
Health	Environment and Natural Resources	0.03	0.11	0.23	1.66	489
Health	Fiscal Policy and Taxation	0.03	0.11	0.23	1.66	472
Health	Economic Outlook	0.03	0.11	0.23	1.80	463

Governance

lhs	rhs	support	confidence	coverage	lift	count
Governance	Health	0.07	0.30	0.22	1.27	1210
Governance	Labor and Education	0.07	0.30	0.22	1.51	1206
Governance	Agriculture	0.06	0.28	0.22	1.44	1145
Governance	Trade and Industry	0.06	0.27	0.22	1.81	1103
Governance	(NULL)	0.04	0.20	0.22	0.96	812
Governance	Technology and Innovation	0.04	0.20	0.22	1.64	809
Governance	Poverty	0.04	0.19	0.22	1.76	759
Governance	Gender and Development	0.03	0.15	0.22	1.63	628
Governance	Infrastructure Transportation and Communication	0.03	0.15	0.22	1.69	594
Governance	Fiscal Policy and Taxation	0.03	0.13	0.22	1.99	540

Governance	Environment and Natural Resources	0.03	0.13	0.22	1.89	534
Governance	Economic Outlook	0.03	0.12	0.22	1.98	486

(NULL)

lhs	rhs	support	confidence	coverage	lift	count
(NULL)	Health	0.05	0.23	0.21	0.96	850
(NULL)	Labor and Education	0.04	0.22	0.21	1.12	824
(NULL)	Governance	0.04	0.21	0.21	0.96	812
(NULL)	Agriculture	0.04	0.21	0.21	1.09	803
(NULL)	Trade and Industry	0.04	0.20	0.21	1.34	753
(NULL)	Infrastructure Transportation and Communication	0.03	0.16	0.21	1.80	586
(NULL)	Technology and Innovation	0.03	0.13	0.21	1.10	505
(NULL)	Poverty	0.03	0.13	0.21	1.23	490
(NULL)	Gender and Development	0.02	0.11	0.21	1.17	418

Agriculture

lhs	rhs	support	confidence	coverage	lift	count
Agriculture	Governance	0.06	0.32	0.19	1.44	1145
Agriculture	Health	0.06	0.30	0.19	1.30	1082
Agriculture	Labor and Education	0.06	0.30	0.19	1.55	1076
Agriculture	Trade and Industry	0.05	0.25	0.19	1.71	903
Agriculture	(NULL)	0.04	0.23	0.19	1.09	803
Agriculture	Technology and Innovation	0.04	0.20	0.19	1.66	714
Agriculture	Poverty	0.04	0.18	0.19	1.74	653
Agriculture	Gender and Development	0.04	0.18	0.19	1.93	648
Agriculture	Infrastructure Transportation and Communication	0.03	0.17	0.19	1.93	590
Agriculture	Environment and Natural Resources	0.03	0.14	0.19	1.99	489
Agriculture	Fiscal Policy and Taxation	0.02	0.12	0.19	1.83	434
Agriculture	Economic Outlook	0.02	0.11	0.19	1.90	406
Agriculture	Climate Change	0.02	0.11	0.19	2.26	377
Agriculture	Regional Integration	0.02	0.10	0.19	2.10	361

Labor and Education

lhs	rhs	support	confidence	coverage	lift	count
Labor and Education	Governance	0.07	0.34	0.19	1.51	1206
Labor and Education	Health	0.06	0.32	0.19	1.36	1132
Labor and Education	Agriculture	0.06	0.30	0.19	1.55	1076
Labor and Education	Trade and Industry	0.05	0.28	0.19	1.85	983
Labor and Education	(NULL)	0.04	0.23	0.19	1.12	824
Labor and Education	Poverty	0.04	0.22	0.19	2.06	777
Labor and Education	Technology and Innovation	0.04	0.21	0.19	1.77	763

Labor and Education	Gender and Development	0.04	0.19	0.19	1.97	662
Labor and Education	Infrastructure Transportation and Communication	0.03	0.17	0.19	1.94	595
Labor and Education	Economic Outlook	0.03	0.14	0.19	2.39	512
Labor and Education	Environment and Natural Resources	0.03	0.14	0.19	1.99	489
Labor and Education	Fiscal Policy and Taxation	0.03	0.13	0.19	1.95	462
Labor and Education	Climate Change	0.02	0.11	0.19	2.34	391
Labor and Education	Regional Integration	0.02	0.11	0.19	2.18	375

Trade and Industry

lhs	rhs	support	confidence	coverage	lift	count
Trade and Industry	Governance	0.06	0.40	0.15	1.81	1103
Trade and Industry	Labor and Education	0.05	0.36	0.15	1.85	983
Trade and Industry	Health	0.05	0.34	0.15	1.45	921
Trade and Industry	Agriculture	0.05	0.33	0.15	1.71	903
Trade and Industry	(NULL)	0.04	0.28	0.15	1.34	753
Trade and Industry	Technology and Innovation	0.04	0.24	0.15	2.00	660
Trade and Industry	Poverty	0.04	0.24	0.15	2.29	658
Trade and Industry	Infrastructure Transportation and Communication	0.03	0.19	0.15	2.19	514
Trade and Industry	Gender and Development	0.03	0.18	0.15	1.88	484
Trade and Industry	Environment and Natural Resources	0.02	0.16	0.15	2.39	448
Trade and Industry	Economic Outlook	0.02	0.15	0.15	2.43	398
Trade and Industry	Fiscal Policy and Taxation	0.02	0.14	0.15	2.06	372
Trade and Industry	Climate Change	0.02	0.13	0.15	2.76	352
Trade and Industry	Regional Integration	0.02	0.12	0.15	2.41	316

Technology and Innovation

lhs	rhs	support	confidence	coverage	lift	count
Technology and Innovation	Governance	0.04	0.36	0.12	1.64	809
Technology and Innovation	Labor and Education	0.04	0.34	0.12	1.77	763
Technology and Innovation	Health	0.04	0.33	0.12	1.41	729
Technology and Innovation	Agriculture	0.04	0.32	0.12	1.66	714
Technology and Innovation	Trade and Industry	0.04	0.30	0.12	2.00	660
Technology and Innovation	Poverty	0.03	0.23	0.12	2.21	517
Technology and Innovation	(NULL)	0.03	0.23	0.12	1.10	505

Technology and Innovation	Fiscal Policy and Taxation	0.02	0.20	0.12	2.99	440
Technology and Innovation	Economic Outlook	0.02	0.20	0.12	3.29	439
Technology and Innovation	Infrastructure Transportation and Communication	0.02	0.20	0.12	2.29	436
Technology and Innovation	Gender and Development	0.02	0.19	0.12	1.97	413
Technology and Innovation	Environment and Natural Resources	0.02	0.16	0.12	2.34	357
Technology and Innovation	Climate Change	0.02	0.15	0.12	3.18	330
Technology and Innovation	Regional Integration	0.01	0.11	0.12	2.26	242
Technology and Innovation	Land Reform and Property Rights	0.01	0.11	0.12	2.15	235

Poverty

lhs	rhs	support	confidence	coverage	lift	count
Poverty	Labor and Education	0.04	0.40	0.11	2.06	777
Poverty	Governance	0.04	0.39	0.11	1.76	759
Poverty	Health	0.04	0.37	0.11	1.60	722
Poverty	Trade and Industry	0.04	0.34	0.11	2.29	658
Poverty	Agriculture	0.04	0.34	0.11	1.74	653
Poverty	Technology and Innovation	0.03	0.27	0.11	2.21	517
Poverty	(NULL)	0.03	0.25	0.11	1.23	490
Poverty	Gender and Development	0.02	0.21	0.11	2.21	403
Poverty	Infrastructure Transportation and Communication	0.02	0.21	0.11	2.41	400
Poverty	Fiscal Policy and Taxation	0.02	0.18	0.11	2.65	340
Poverty	Environment and Natural Resources	0.02	0.17	0.11	2.44	325
Poverty	Economic Outlook	0.02	0.17	0.11	2.75	320
Poverty	Climate Change	0.02	0.15	0.11	3.19	288
Poverty	Regional Integration	0.01	0.13	0.11	2.75	256
Poverty	Land Reform and Property Rights	0.01	0.12	0.11	2.34	223

Gender and Development

lhs	rhs	support	confidence	coverage	lift	count
Gender and Development	Labor and Education	0.04	0.38	0.09	1.97	662
Gender and Development	Agriculture	0.04	0.38	0.09	1.93	648
Gender and Development	Governance	0.03	0.36	0.09	1.63	628
Gender and Development	Health	0.03	0.34	0.09	1.44	581
Gender and Development	Trade and Industry	0.03	0.28	0.09	1.88	484
Gender and Development	(NULL)	0.02	0.24	0.09	1.17	418
Gender and Development	Technology and Innovation	0.02	0.24	0.09	1.97	413
Gender and Development	Poverty	0.02	0.23	0.09	2.21	403
Gender and Development	Infrastructure Transportation and Communication	0.02	0.19	0.09	2.23	331

Gender and Development	Fiscal Policy and Taxation	0.02	0.18	0.09	2.67	306
Gender and Development	Environment and Natural Resources	0.02	0.17	0.09	2.52	300
Gender and Development	Economic Outlook	0.01	0.14	0.09	2.40	249
Gender and Development	Climate Change	0.01	0.13	0.09	2.70	218
Gender and Development	Regional Integration	0.01	0.12	0.09	2.48	207
Gender and Development	Land Reform and Property Rights	0.01	0.11	0.09	2.20	187

Infrastructure Transportation and Communication

lhs	rhs	support	confidence	coverage	lift	count
Infrastructure Transportation and Communication	Health	0.03	0.38	0.09	1.64	605
Infrastructure Transportation and Communication	Labor and Education	0.03	0.38	0.09	1.94	595
Infrastructure Transportation and Communication	Governance	0.03	0.38	0.09	1.69	594
Infrastructure Transportation and Communication	Agriculture	0.03	0.37	0.09	1.93	590
Infrastructure Transportation and Communication	(NULL)	0.03	0.37	0.09	1.80	586
Infrastructure Transportation and Communication	Trade and Industry	0.03	0.33	0.09	2.19	514
Infrastructure Transportation and Communication	Technology and Innovation	0.02	0.28	0.09	2.29	436
Infrastructure Transportation and Communication	Poverty	0.02	0.25	0.09	2.41	400
Infrastructure Transportation and Communication	Gender and Development	0.02	0.21	0.09	2.23	331
Infrastructure Transportation and Communication	Environment and Natural Resources	0.02	0.18	0.09	2.59	281
Infrastructure Transportation and Communication	Fiscal Policy and Taxation	0.01	0.17	0.09	2.55	267
Infrastructure Transportation and Communication	Land Reform and Property Rights	0.01	0.15	0.09	3.15	244
Infrastructure Transportation	Economic Outlook	0.01	0.15	0.09	2.58	244

and Communication						
Infrastructure Transportation and Communication	Climate Change	0.01	0.15	0.09	3.24	239
Infrastructure Transportation and Communication	Banking and Finance	0.01	0.15	0.09	4.81	233
Infrastructure Transportation and Communication	Regional Integration	0.01	0.13	0.09	2.63	200

Environment and Natural Resources

lhs	rhs	support	confidence	coverage	lift	count
Environment and Natural Resources	Governance	0.03	0.42	0.07	1.89	534
Environment and Natural Resources	Agriculture	0.03	0.39	0.07	1.99	489
Environment and Natural Resources	Labor and Education	0.03	0.39	0.07	1.99	489
Environment and Natural Resources	Health	0.03	0.39	0.07	1.66	489
Environment and Natural Resources	Trade and Industry	0.02	0.35	0.07	2.39	448
Environment and Natural Resources	Technology and Innovation	0.02	0.28	0.07	2.34	357
Environment and Natural Resources	(NULL)	0.02	0.27	0.07	1.33	346
Environment and Natural Resources	Poverty	0.02	0.26	0.07	2.44	325
Environment and Natural Resources	Gender and Development	0.02	0.24	0.07	2.52	300
Environment and Natural Resources	Infrastructure Transportation and Communication	0.02	0.22	0.07	2.59	281
Environment and Natural Resources	Fiscal Policy and Taxation	0.01	0.19	0.07	2.80	235
Environment and Natural Resources	Economic Outlook	0.01	0.18	0.07	3.01	229
Environment and Natural Resources	Climate Change	0.01	0.18	0.07	3.86	228

Economic Outlook

lhs	s	support	confidence	coverage	lift	count
Economic Outlook	Labor and Education	0.03	0.46	0.06	2.39	512
Economic Outlook	Governance	0.03	0.44	0.06	1.98	486
Economic Outlook	Health	0.03	0.42	0.06	1.80	463
Economic Outlook	Technology and Innovation	0.02	0.40	0.06	3.29	439
Economic Outlook	Agriculture	0.02	0.37	0.06	1.90	406
Economic Outlook	Trade and Industry	0.02	0.36	0.06	2.43	398
Economic Outlook	Fiscal Policy and Taxation	0.02	0.32	0.06	4.79	351
Economic Outlook	Poverty	0.02	0.29	0.06	2.75	320
Economic Outlook	(NULL)	0.02	0.28	0.06	1.35	306
Economic Outlook	Gender and Development	0.01	0.23	0.06	2.40	249
Economic Outlook	Infrastructure Transportation and Communication	0.01	0.22	0.06	2.58	244
Economic Outlook	Environment and Natural Resources	0.01	0.21	0.06	3.01	229
Economic Outlook	Climate Change	0.01	0.17	0.06	3.72	192

Fiscal Policy and Taxation

lhs	rhs	support	confidence	coverage	lift	count
Fiscal Policy and Taxation	Governance	0.03	0.44	0.07	1.99	540
Fiscal Policy and Taxation	Health	0.03	0.39	0.07	1.66	472
Fiscal Policy and Taxation	Labor and Education	0.03	0.38	0.07	1.95	462
Fiscal Policy and Taxation	Technology and Innovation	0.02	0.36	0.07	2.99	440
Fiscal Policy and Taxation	Agriculture	0.02	0.36	0.07	1.83	434
Fiscal Policy and Taxation	Trade and Industry	0.02	0.31	0.07	2.06	372
Fiscal Policy and Taxation	Economic Outlook	0.02	0.29	0.07	4.79	351
Fiscal Policy and Taxation	Poverty	0.02	0.28	0.07	2.65	340
Fiscal Policy and Taxation	Gender and Development	0.02	0.25	0.07	2.67	306
Fiscal Policy and Taxation	(NULL)	0.02	0.24	0.07	1.15	288
Fiscal Policy and Taxation	Infrastructure Transportation and Communication	0.01	0.22	0.07	2.55	267
Fiscal Policy and Taxation	Environment and Natural Resources	0.01	0.19	0.07	2.80	235

Land Reform and Property Rights

lhs	rhs	support	confidence	coverage	lift	count
Land Reform and Property Rights	Labor and Education	0.02	0.39	0.05	1.98	348
Land Reform and Property Rights	Agriculture	0.02	0.38	0.05	1.97	345
Land Reform and Property Rights	Health	0.02	0.35	0.05	1.49	315
Land Reform and Property Rights	Governance	0.02	0.34	0.05	1.51	304
Land Reform and Property Rights	Trade and Industry	0.01	0.29	0.05	1.92	258
Land Reform and Property Rights	Infrastructure Transportation and Communication	0.01	0.27	0.05	3.15	244
Land Reform and Property Rights	(NULL)	0.01	0.27	0.05	1.30	242
Land Reform and Property Rights	Technology and Innovation	0.01	0.26	0.05	2.15	235
Land Reform and Property Rights	Poverty	0.01	0.25	0.05	2.34	223
Land Reform and Property Rights	Gender and Development	0.01	0.21	0.05	2.20	187

Regional Integration

lhs	rhs	support	confidence	coverage	lift	count
Regional Integration	Labor and Education	0.02	0.42	0.05	2.18	375
Regional Integration	Governance	0.02	0.41	0.05	1.84	363
Regional Integration	Agriculture	0.02	0.41	0.05	2.10	361
Regional Integration	Health	0.02	0.40	0.05	1.69	349
Regional Integration	Trade and Industry	0.02	0.36	0.05	2.41	316
Regional Integration	Poverty	0.01	0.29	0.05	2.75	256
Regional Integration	Technology and Innovation	0.01	0.27	0.05	2.26	242
Regional Integration	(NULL)	0.01	0.26	0.05	1.25	227
Regional Integration	Gender and Development	0.01	0.23	0.05	2.48	207
Regional Integration	Infrastructure Transportation and Communication	0.01	0.23	0.05	2.63	200

Climate Change

lhs	rhs	support	confidence	coverage	lift	count
Climate Change	Labor and Education	0.02	0.46	0.05	2.34	391
Climate Change	Governance	0.02	0.46	0.05	2.04	391
Climate Change	Health	0.02	0.45	0.05	1.91	383
Climate Change	Agriculture	0.02	0.44	0.05	2.26	377
Climate Change	Trade and Industry	0.02	0.41	0.05	2.76	352
Climate Change	Technology and Innovation	0.02	0.39	0.05	3.18	330
Climate Change	(NULL)	0.02	0.35	0.05	1.68	297
Climate Change	Poverty	0.02	0.34	0.05	3.19	288
Climate Change	Infrastructure Transportation and Communication	0.01	0.28	0.05	3.24	239
Climate Change	Environment and Natural Resources	0.01	0.27	0.05	3.86	228
Climate Change	Gender and Development	0.01	0.25	0.05	2.70	218
Climate Change	Economic Outlook	0.01	0.22	0.05	3.72	192

Banking and Finance

lhs	rhs	support	confidence	coverage	lift	count
Banking and Finance	(NULL)	0.02	0.57	0.03	2.76	320
Banking and Finance	Governance	0.02	0.50	0.03	2.24	281
Banking and Finance	Health	0.02	0.49	0.03	2.11	277
Banking and Finance	Labor and Education	0.02	0.49	0.03	2.52	276
Banking and Finance	Agriculture	0.01	0.48	0.03	2.46	269
Banking and Finance	Trade and Industry	0.01	0.46	0.03	3.07	257
Banking and Finance	Infrastructure Transportation and Communication	0.01	0.41	0.03	4.81	233
Banking and Finance	Technology and Innovation	0.01	0.35	0.03	2.89	197

Migration and Development

lhs	rhs	support	confidence	coverage	lift	count
Migration and Development	Labor and Education	0.02	0.51	0.03	2.60	280
Migration and Development	Governance	0.01	0.49	0.03	2.20	271
Migration and Development	Agriculture	0.01	0.46	0.03	2.34	252
Migration and Development	Health	0.01	0.45	0.03	1.92	248
Migration and Development	Trade and Industry	0.01	0.43	0.03	2.92	240
Migration and Development	(NULL)	0.01	0.35	0.03	1.71	195
Migration and Development	Technology and Innovation	0.01	0.35	0.03	2.88	193
Migration and Development	Poverty	0.01	0.34	0.03	3.21	187

Urban Development and Housing

lhs	rhs	support	confidence	coverage	lift	count
Urban Development and Housing	Governance	0.01	0.47	0.02	2.11	215
Urban Development and Housing	Health	0.01	0.45	0.02	1.94	207
Urban Development and Housing	Agriculture	0.01	0.43	0.02	2.23	198
Urban Development and Housing	Labor and Education	0.01	0.43	0.02	2.18	194

International Relations and Foreign Policy

lhs	rhs	support	confidence	coverage	lift	count
International Relations and Foreign Policy	Governance	0.01	0.48	0.02	2.17	189