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

FEBRUARY 23, 2023

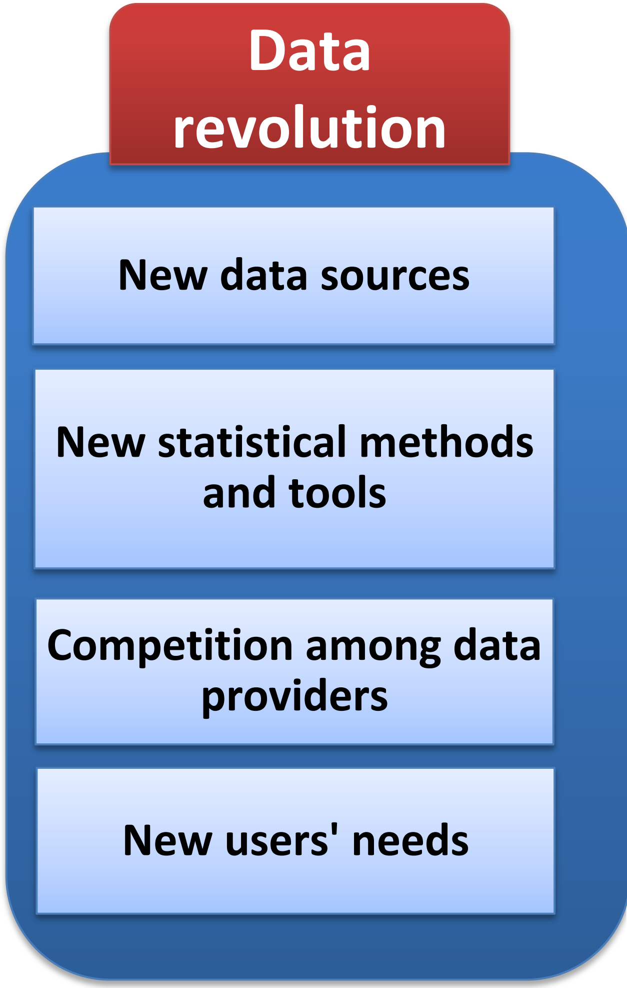
Philippine Institute for Development Studies
Surian sa mga Pag-aaral Pangkaunlaran ng Pilipinas

Outline

1. Introduction
   - Official statistics, digital information, and big data
   - Policy Questions
2. Research Design
3. Empirical Findings
   a) Examining PIDS web download data
   b) Analyzing Twitter and other web scraped data
      - Web scraped news data on violence against women (VAW)
      - Text mining tourism data
4. Recommendations and Ways Forward

1. Introduction

- With the advent of digital transformation, ICT innovations have led to a **data revolution,** i.e., more data captured, produced, stored, accessed, analyzed, archived, and re- analyzed, and at an exponential pace.
  - New data sources, including big data and crowd sourced data, can complement traditional data sources (Albert *et al*. 2019).
  - PIDS should harness use of non-traditional data sources to provide policy insights to decision-makers with near real time information.


Data revolution

New data sources

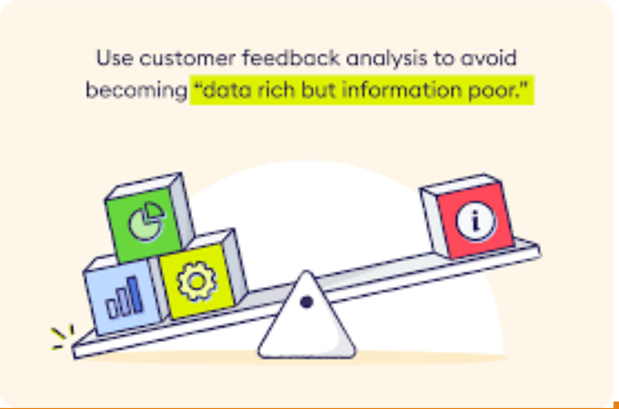
New statistical methods and tools

Competition among data providers

New users' needs

1. Introduction (cont'd)

SUSTAINABLE DEVELOPMENT GOALS
17 GOALS TO TRANSFORM OUR WORLD
1 NO POVERTY
2 ZERO HUNGER
3 GOOD HEALTH AND WELL-BEING
4 QUALITY EDUCATION
5 GENDER EQUALITY
6 CLEAN WATER AND SANITATION
7 AFFORDABLE AND CLEAN ENERGY
8 DECENT WORK AND ECONOMIC GROWTH
9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
10 REDUCED INEQUALITIES
11 SUSTAINABLE CITIES AND COMMUNITIES
12 RESPONSIBLE CONSUMPTION AND PRODUCTION
13 CLIMATE ACTION
14 LIFE BELOW WATER
15 LIFE ON LAND
16 PEACE, JUSTICE AND STRONG INSTITUTIONS
17 PARTNERSHIPS FOR THE GOALS
SUSTAINABLE DEVELOPMENT GOALS

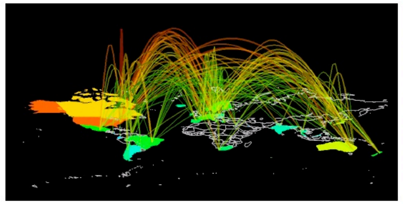

Use customer feedback analysis to avoid becoming "data rich but information poor."

- National dev't priorities identified (in PDP, Ambisyon 2040 and in our commitment to the SDGs), but data gaps persist
- PIDS is accumulating data, but little data analytics are being performed on data holdings.
  - PIDS website download data
  - public sentiments in FB page and on twitter
  - sentiments during public webinars
- This study was designed to examine some new data accumulated at PIDS as well as address selected data gaps on tourism, violence against women (VAW), among others.

1.1. Official statistics and big data

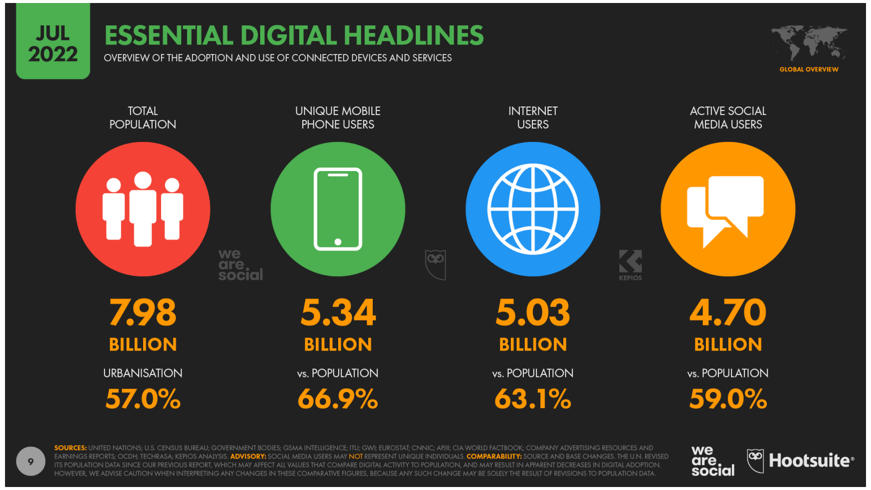
Airplane Flights

Telephone Calls



Internet Traffic

Social Networking



JUL 2022

ESSENTIAL DIGITAL HEADLINES

OVERVIEW OF THE ADOPTION AND USE OF CONNECTED DEVICES AND SERVICES

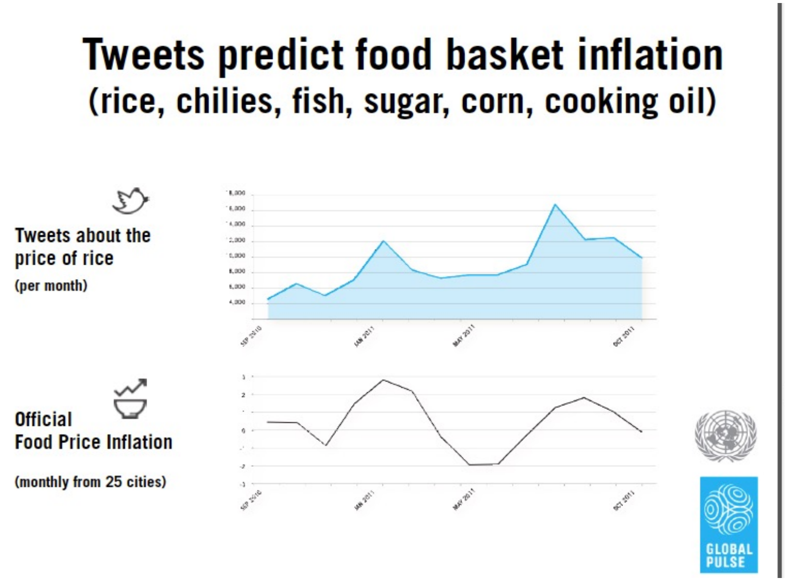
TOTAL POPULATION	UNIQUE MOBILE PHONE USERS	INTERNET USERS	ACTIVE SOCIAL MEDIA USERS
7.98 BILLION	5.34 BILLION	5.03 BILLION	4.70 BILLION
URBANISATION 57.0%	vs. POPULATION 66.9%	vs. POPULATION 63.1%	vs. POPULATION 59.0%

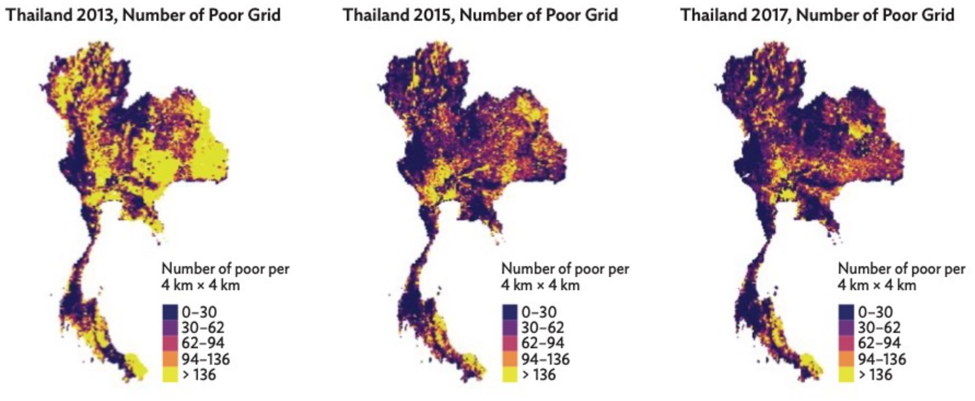
SOURCES: UNITED NATIONS; U.S. CENSUS BUREAU; GOVERNMENT BODIES; GSMA INTELLIGENCE; ITU; GWI; EUROSTAT; CNNIC; APJII; CIA WORLD FACTBOOK; COMPANY ADVERTISING RESOURCES AND EARNINGS REPORTS; OCDH; TECHRASA; KEPIOS ANALYSIS. ADVISORY: SOCIAL MEDIA USERS MAY NOT REPRESENT UNIQUE INDIVIDUALS. COMPARABILITY: SOURCE AND BASE CHANGES. THE U.N. REVISED ITS POPULATION DATA SINCE OUR PREVIOUS REPORT, WHICH MAY AFFECT ALL VALUES THAT COMPARE DIGITAL ACTIVITY TO POPULATION, AND MAY RESULT IN APPARENT DECREASES IN DIGITAL ADOPTION. HOWEVER, WE ADVISE CAUTION WHEN INTERPRETING ANY CHANGES IN THESE COMPARATIVE FIGURES, BECAUSE ANY SUCH CHANGE MAY BE SOLELY THE RESULT OF REVISIONS TO POPULATION DATA.

Data Sources	Pros	Cons
Census	• Complete enumeration • Source of population statistics • Provides area and/or list frame	• Costly • Robust staffing • Restricted periodicity • Longer lag time in producing results
Surveys	• Relatively easy to administer • Cost-effective • Wider scope	• Non-response • Sampling error • Response bias • Need for an adequately trained manpower
Administrative Data	• Low-cost data collection • Timely statistical outputs • Up-to-date (more frequent data) • Reduce burden for respondents • Better data coverage and availability	• Not designed for statistical purposes • Needs strong coordination among statistical agencies, government agencies, and public and private data provider. • Confidentiality issue • Missing data • Different time periods
Other Types of Big Data	• Large volume of data • Wide variety of data types • Timely data • Improves accuracy and granularity of statistics	• Data privacy and security • Accessibility • Challenges in technological infrastructure • Requires new skill sets • Coverage and representativeness

1.1. Official statistics and big data

- Utilizing big data for development:
  - UN Global Pulse (2014) : twitter conversations on food in Jakarta as proxy for food inflation
  - UN Women (2018): Making Gender Data Visible
  - ADB and World Data Lab (2021): satellite imagery integrated with census and survey data for high quality poverty estimates at small area areas in Thailand
- UN Statistical Commission established UN Committee of Experts on Big Data and Data Science for Official Statistics (UN-CEBD) in 2014




Thailand 2013, Number of Poor Grid
Number of poor per 4 km × 4 km
0–30
30–62
62–94
94–136
> 136
Thailand 2015, Number of Poor Grid
Number of poor per 4 km × 4 km
0–30
30–62
62–94
94–136
> 136
Thailand 2017, Number of Poor Grid
Number of poor per 4 km × 4 km
0–30
30–62
62–94
94–136
> 136

km = kilometer.

1.2. Policy Questions

1. How can data from these new data sources be transformed into meaningful insights for development to effect better development outcomes, in some areas such as gender, tourism and traffic management?
2. What strategies can be developed to promote the access, analysis and use and re-use of new data sources (and mitigate risks from abuse of big data analytics)?

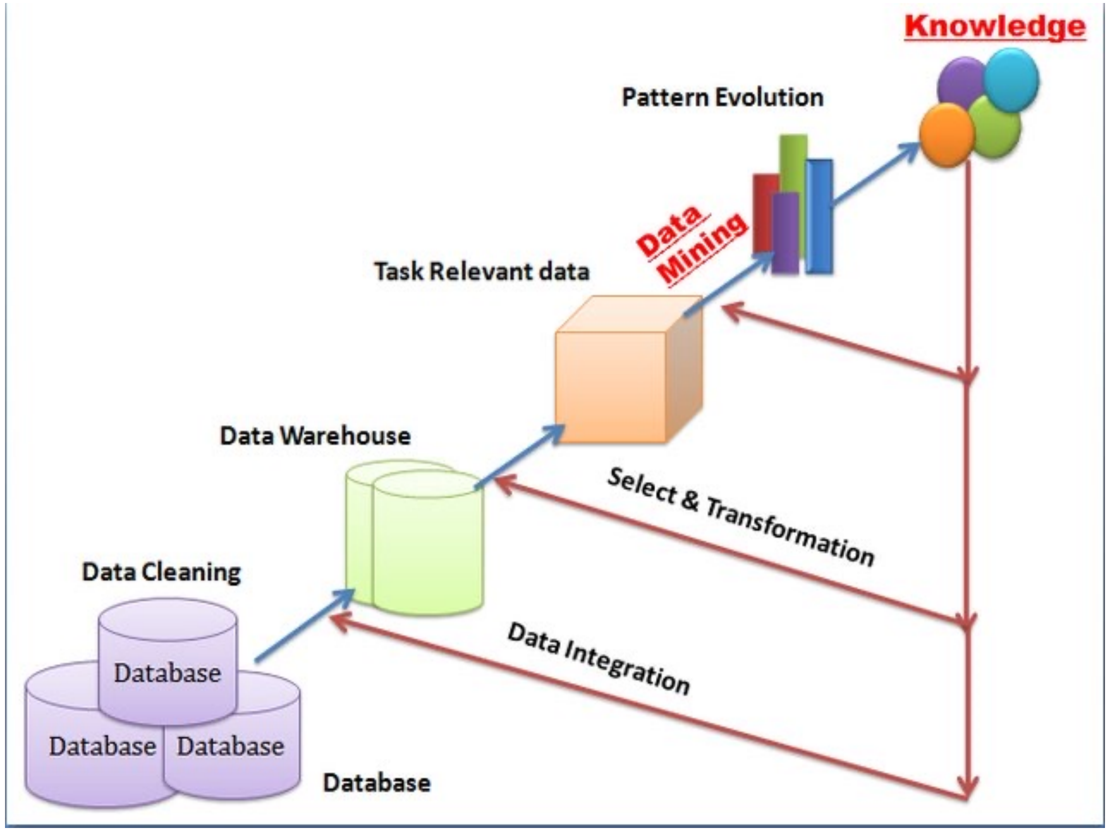
2. Research Design

- ❑ Data sources
  - PIDS web download data
  - Twitter data and web-scraped data
- ❑ Data collection methods/tools
  - Market basket analysis
  - Text mining (sentiment analysis, topic modelling)
  - Social media analysis

2. Research Design (cont’d)

Knowledge Discovery Process (or Data Mining)

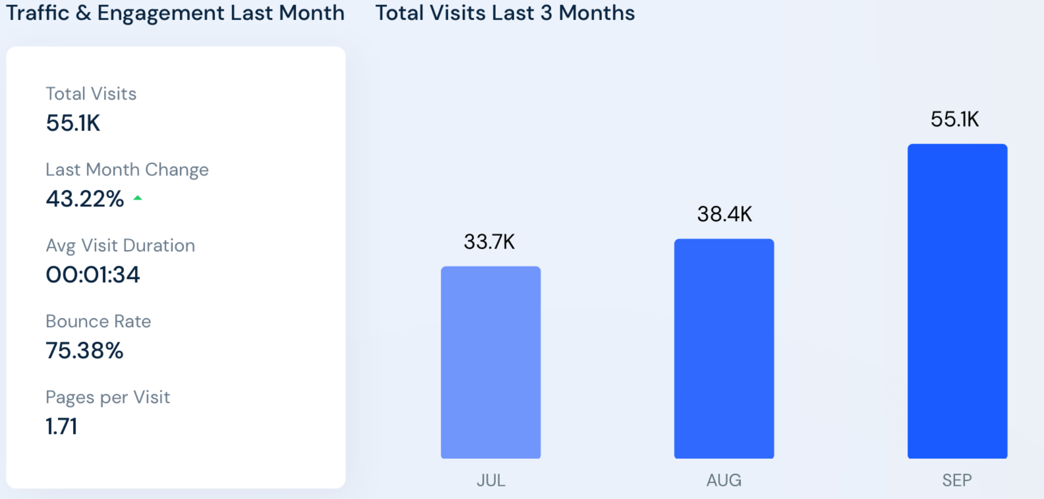
1. Selection: Selecting data relevant to the analysis task from the database
2. Preprocessing: Removing noise and inconsistent data; combining multiple data sources
3. Transformation: Transforming data into appropriate forms to perform data mining
4. Data mining: Choosing a data mining algorithm which is appropriate to pattern in the data; Extracting data patterns
5. Interpretation/Evaluation: Interpreting the patterns into knowledge by removing redundant or irrelevant patterns; Translating the useful patterns into terms that human-understandable



(Source: Fayyad *et al.*, 1996)

3.1. PIDS Web Download data

Some useful data on PIDS website visits from SimilarWeb (September 2022)



Traffic & Engagement Last Month

Total Visits
55.1K

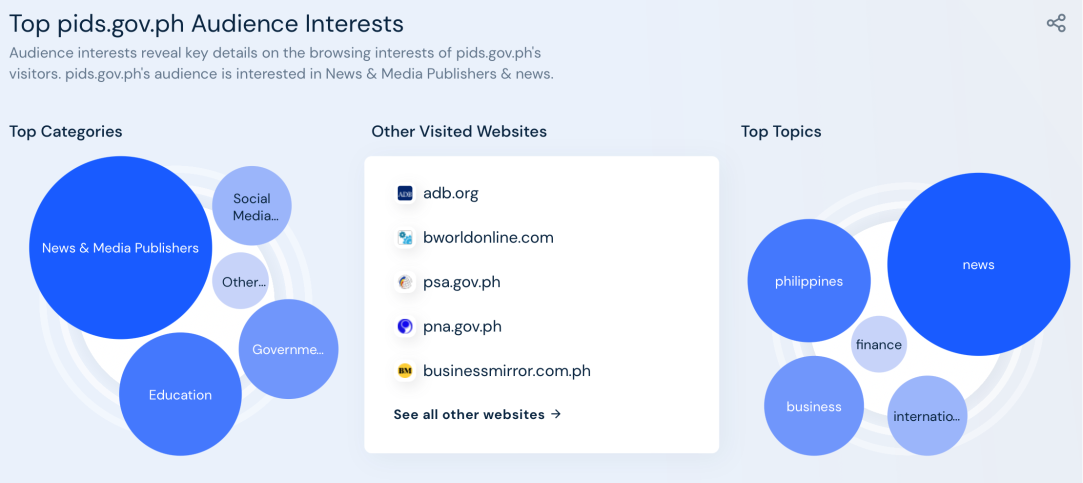
Last Month Change
43.22% ▲

Avg Visit Duration
00:01:34

Bounce Rate
75.38%

Pages per Visit
1.71

Total Visits Last 3 Months



Top pids.gov.ph Audience Interests

Audience interests reveal key details on the browsing interests of pids.gov.ph's visitors. pids.gov.ph's audience is interested in News & Media Publishers & news.

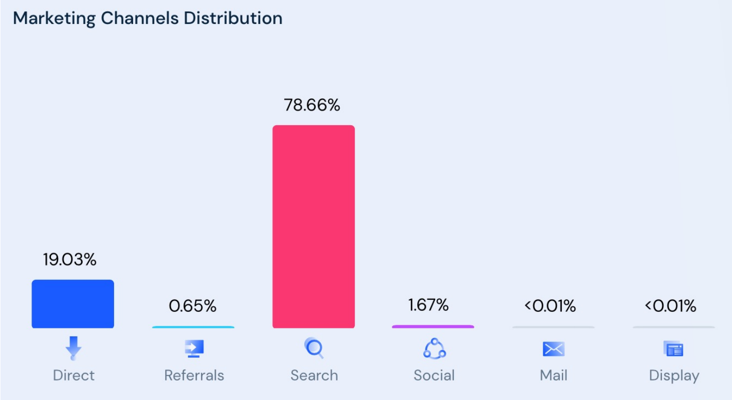
Top Categories

Other Visited Websites

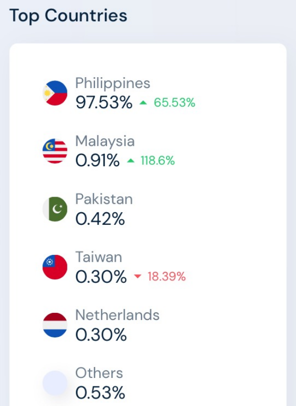
- adb.org
- bworldonline.com
- psa.gov.ph
- pna.gov.ph
- businessmirror.com.ph

See all other websites →

Top Topics

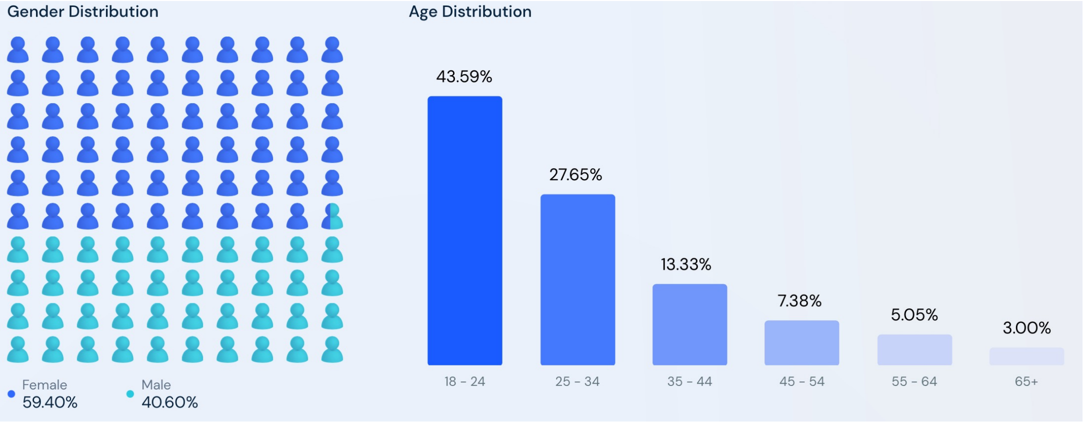


Marketing Channels Distribution



Top Countries

Philippines	97.53%	▲ 65.53%
Malaysia	0.91%	▲ 118.6%
Pakistan	0.42%	
Taiwan	0.30%	▼ 18.39%
Netherlands	0.30%	
Others	0.53%	



Gender Distribution

Age Distribution

Source: SimilarWeb (free version) https://www.similarweb.com/website/pids.gov.ph/#interests

3.1. Visitor Profile

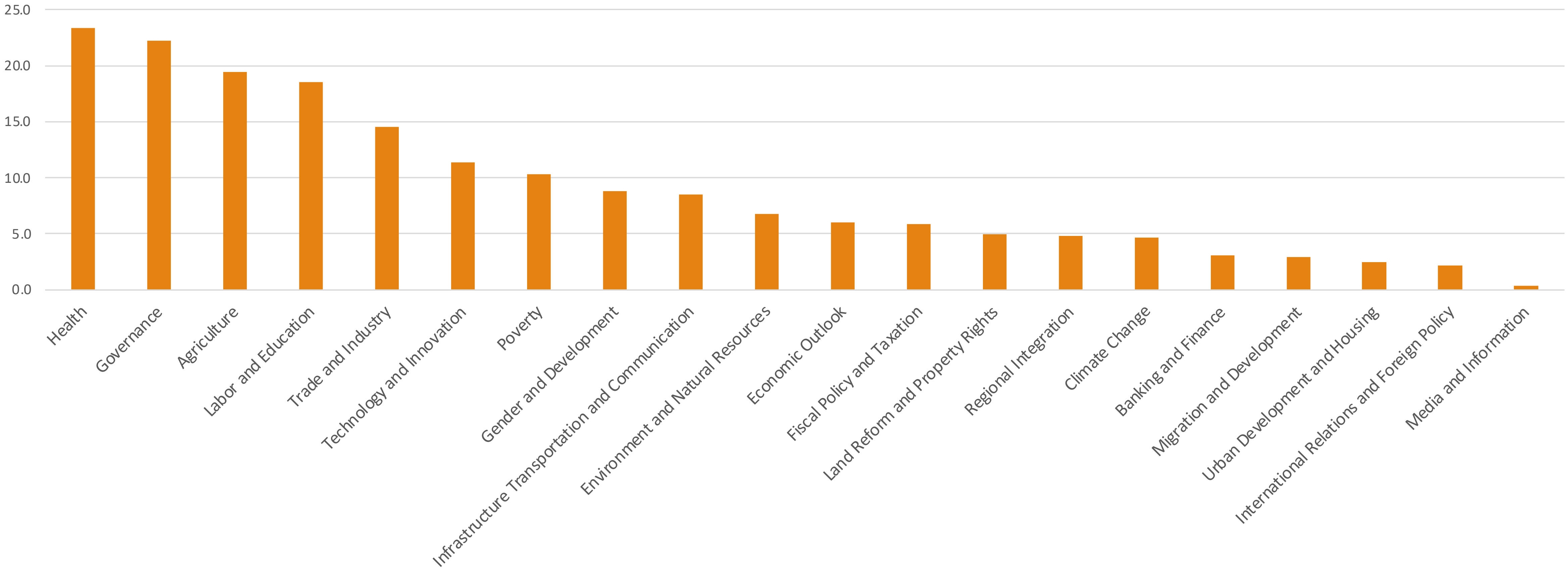
Downloader Profile	Frequency	Distribution (%)
Age		
Below 18	2,037	3.2
19-35	14,188	22.1
36-50	4,961	7.7
51-65	2,403	3.7
66 and above	400	0.6
Missing data	40,218	62.6
Total	**64,207**	**100.0**
Gender		
Female	12,328	19.2
Male	9,906	15.4
Prefer not say	1,067	1.7
Prefer to self-describe	117	0.2
Missing data	40,789	63.5
Total	**64,207**	**100.0**

Downloader Profile	Frequency	Distribution (%)
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
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**

Data Source: PIDS

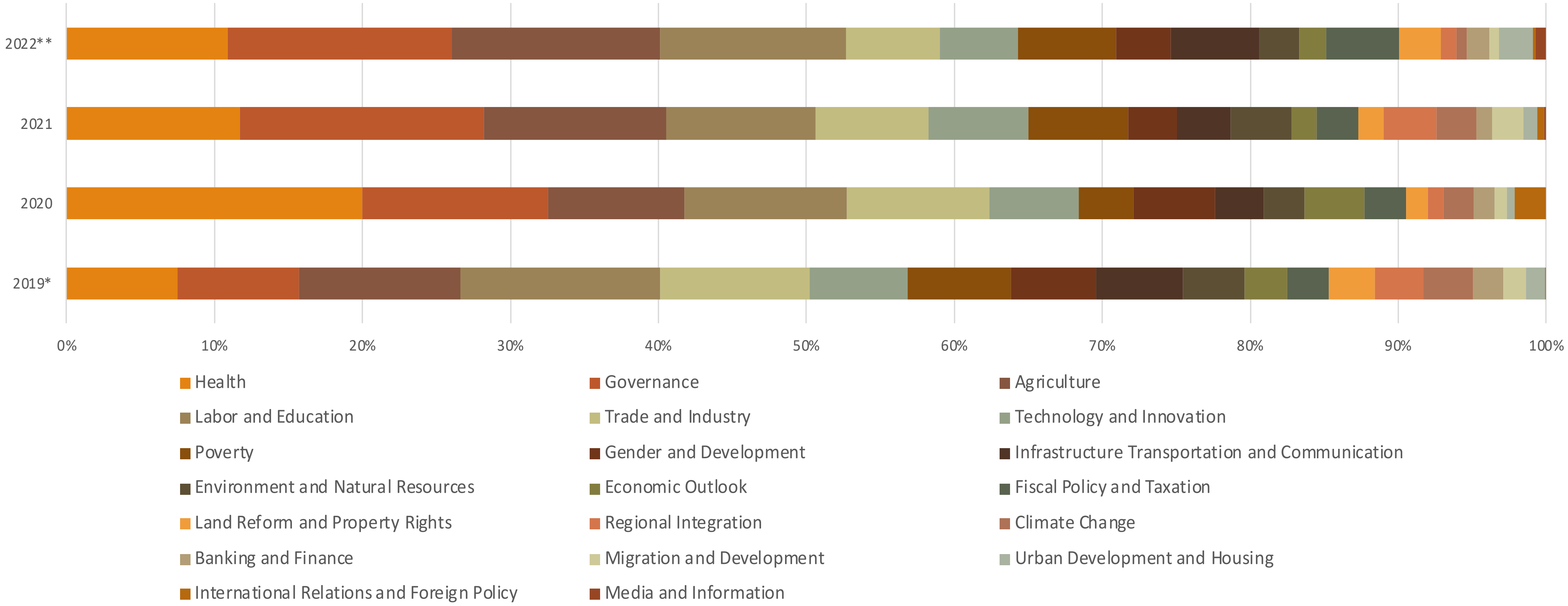
3.1.2. Publication Downloads

PIDS Website Publication Downloads by Theme (%): April 18, 2019 to August 9, 2022



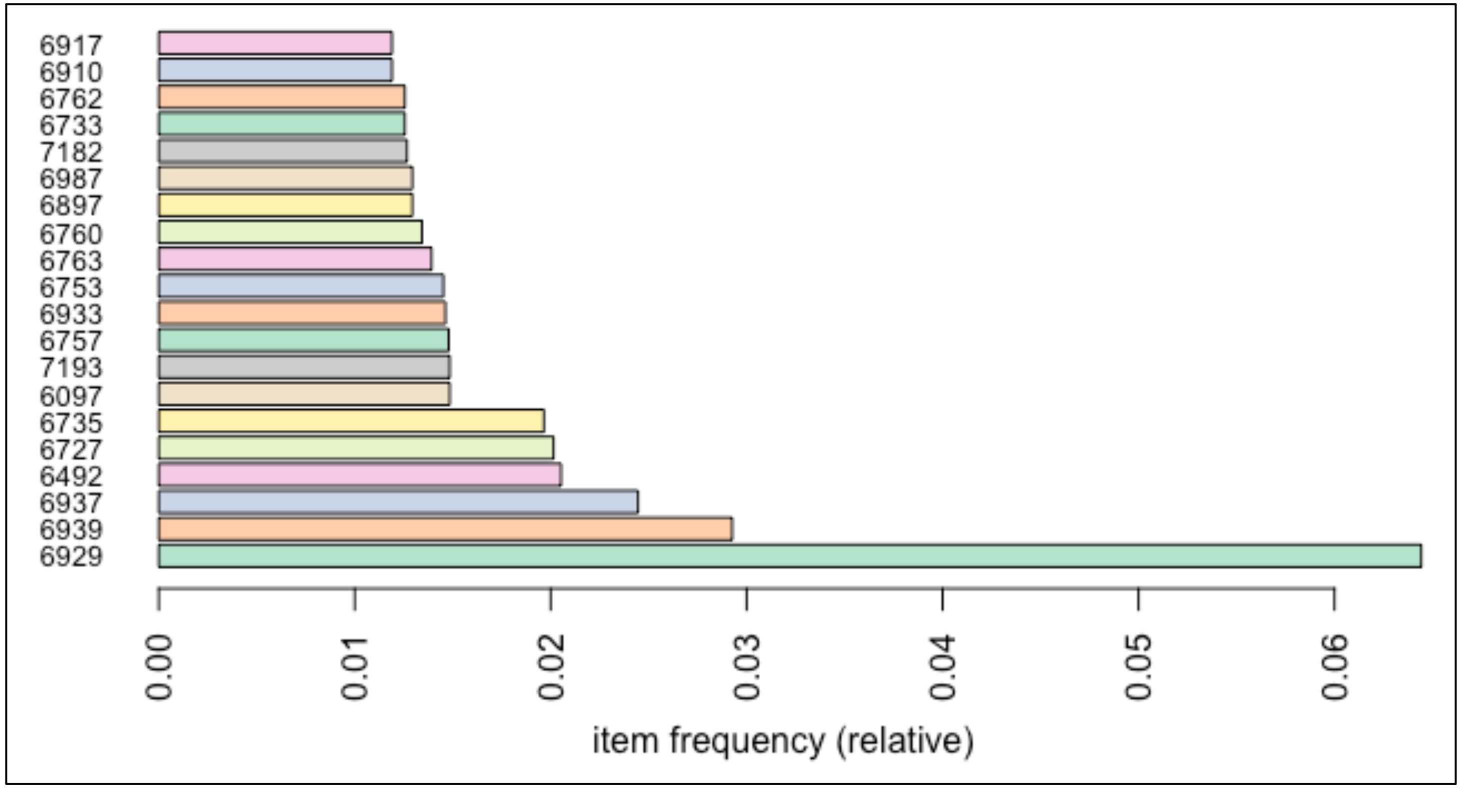
3.1.2. Publication Downloads (cont'd)

PIDS Website Publication Downloads by Theme by Year (%)



3.1.2. Publication Downloads (cont'd)

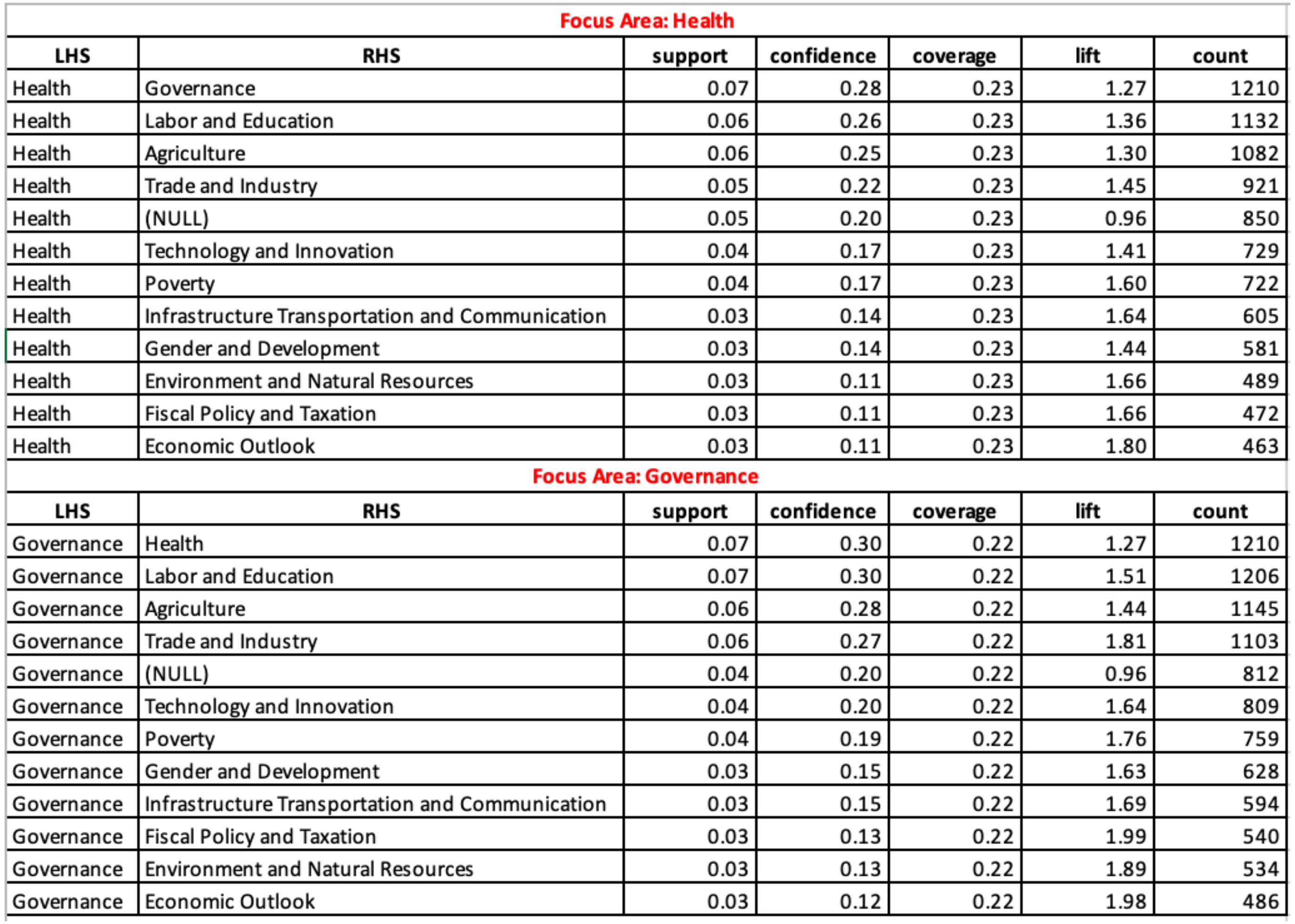
Top 20 most downloaded publications in PIDS website (relative value)



- Top 1: Publication ID **6929**, a DP on the Situation Analysis of ECCD-F1KD Initiatives in Selected UNICEF-KOICA Provinces constitute around **6%** of PIDS website downloads
- Top 2 : Publication ID 6939, a PN on the Issues and Concerns in the Implementation of the PBB at DepEd (2%)
- Top 3: Publication ID 6937, a DP on Expanding Health Insurance for the Elderly of the Philippines (2%)

3.1.2 Market Basket Analysis of PIDS Download Data

- identifying what PIDS products (publications), or groups of products, tend to occur together (are associated) when PIDS “customers” make transactions/downloads (baskets)
- examining the association between different “items”, to find frequent patterns in the PIDS website download transaction database



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

3.1.2 Market Basket Analysis of PIDS Download Data

Left Hand Side	Right Hand Side	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 Rightsizing Government)	6758 (Child Stunting Prevention: The Challenge of Mobilizing Local Governments for National Impact)	0.001420	0.928571	0.001529	147.89	26
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

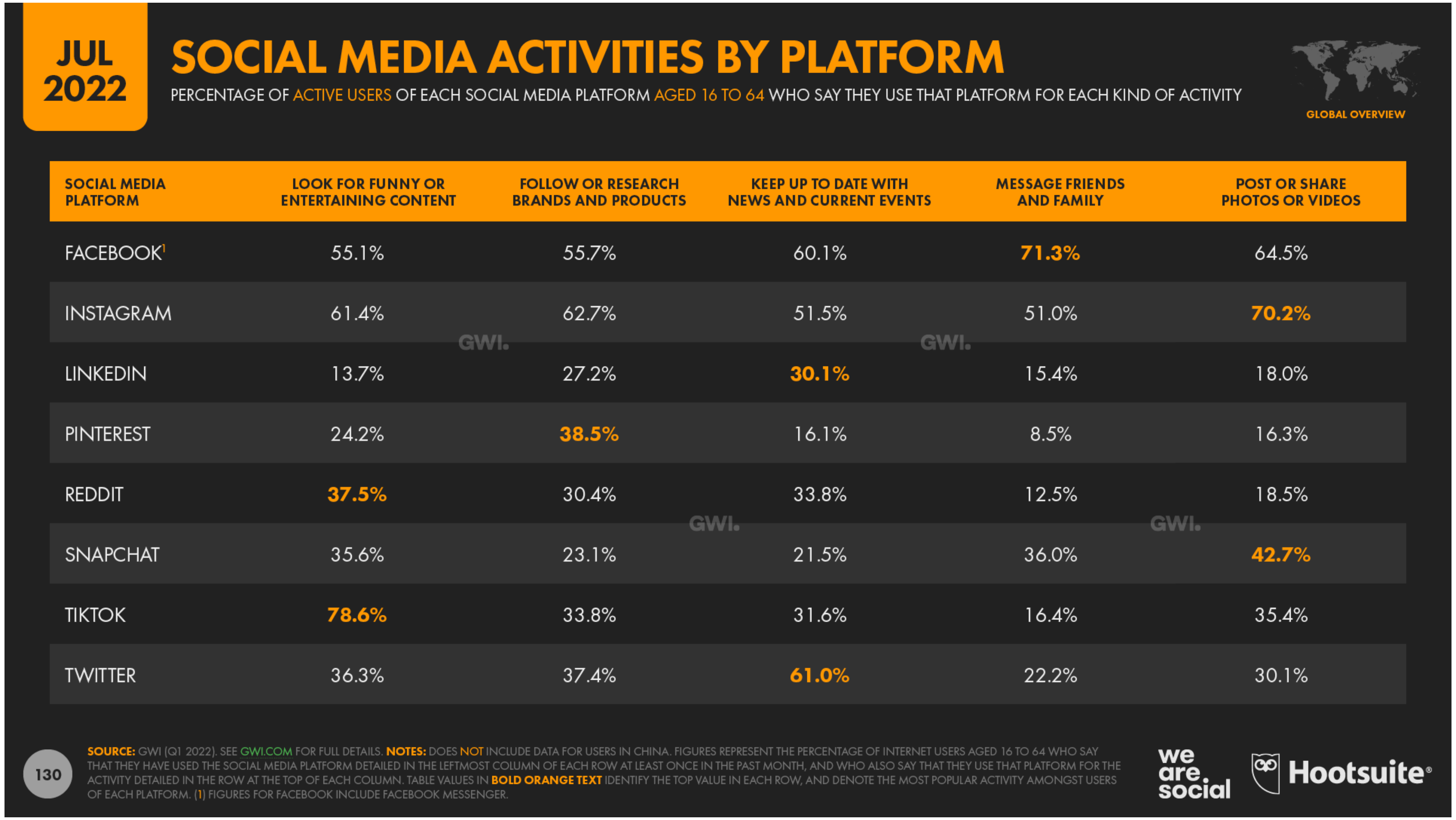
3.2. Twitter and other web scraped data

DataReportal (2022):

- 4.7 B social media users (58% of world population)
- Average daily time on social media: 2h 29mins

Twitter

- 238 M users (world);
- 10.5 M users (PH)
- 7th favorite social media platform



JUL 2022

SOCIAL MEDIA ACTIVITIES BY PLATFORM

PERCENTAGE OF ACTIVE USERS OF EACH SOCIAL MEDIA PLATFORM AGED 16 TO 64 WHO SAY THEY USE THAT PLATFORM FOR EACH KIND OF ACTIVITY

GLOBAL OVERVIEW

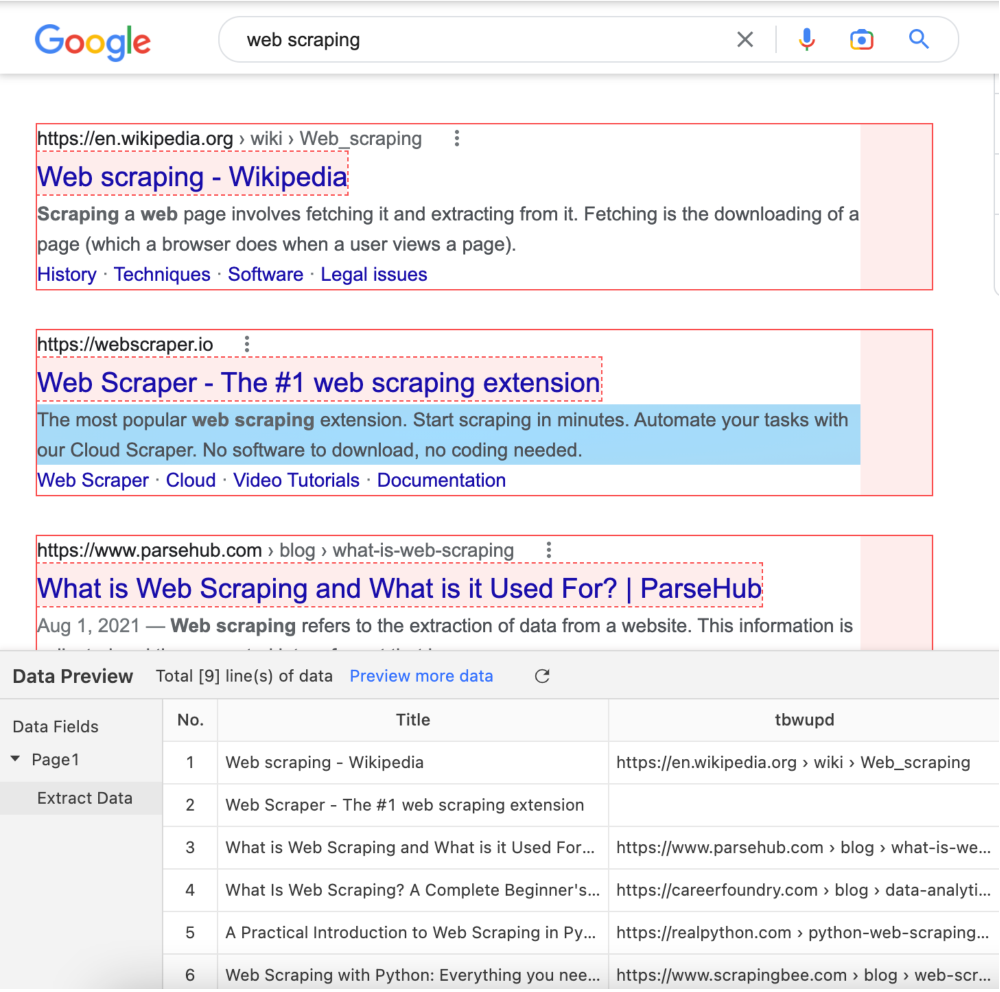
SOCIAL MEDIA PLATFORM	LOOK FOR FUNNY OR ENTERTAINING CONTENT	FOLLOW OR RESEARCH BRANDS AND PRODUCTS	KEEP UP TO DATE WITH NEWS AND CURRENT EVENTS	MESSAGE FRIENDS AND FAMILY	POST OR SHARE PHOTOS OR VIDEOS
FACEBOOK[1]	55.1%	55.7%	60.1%	**71.3%**	64.5%
INSTAGRAM	61.4%	62.7%	51.5%	51.0%	**70.2%**
LINKEDIN	13.7%	27.2%	**30.1%**	15.4%	18.0%
PINTEREST	24.2%	**38.5%**	16.1%	8.5%	16.3%
REDDIT	**37.5%**	30.4%	33.8%	12.5%	18.5%
SNAPCHAT	35.6%	23.1%	21.5%	36.0%	**42.7%**
TIKTOK	**78.6%**	33.8%	31.6%	16.4%	35.4%
TWITTER	36.3%	37.4%	**61.0%**	22.2%	30.1%

SOURCE: GWI (Q1 2022). SEE GWI.COM FOR FULL DETAILS. **NOTES:** DOES NOT INCLUDE DATA FOR USERS IN CHINA. FIGURES REPRESENT THE PERCENTAGE OF INTERNET USERS AGED 16 TO 64 WHO SAY THAT THEY HAVE USED THE SOCIAL MEDIA PLATFORM DETAILED IN THE LEFTMOST COLUMN OF EACH ROW AT LEAST ONCE IN THE PAST MONTH, AND WHO ALSO SAY THAT THEY USE THAT PLATFORM FOR THE ACTIVITY DETAILED IN THE ROW AT THE TOP OF EACH COLUMN. TABLE VALUES IN **BOLD ORANGE TEXT** IDENTIFY THE TOP VALUE IN EACH ROW, AND DENOTE THE MOST POPULAR ACTIVITY AMONGST USERS OF EACH PLATFORM. (1) FIGURES FOR FACEBOOK INCLUDE FACEBOOK MESSENGER.

Source: Digital 2022 July Global Statshot Report, DataReportal.

3.2. Twitter and other web scraped data (cont’d)

- **Web scraping**
  - process of extracting publicly available data from a website
  - can be programmed through programming languages such as Python
  - scraping tweets using the Twitter API platform can provide insights on global to local topics and events, gain information to better profile target audience, and identify trends and important conversations on Twitter
  - Note: important to be familiarized 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



https://en.wikipedia.org › wiki › Web_scraping

Web scraping - Wikipedia

Scraping a **web** page involves fetching it and extracting from it. Fetching is the downloading of a page (which a browser does when a user views a page).

History · Techniques · Software · Legal issues

https://webscraper.io

Web Scraper - The #1 web scraping extension

The most popular **web scraping** extension. Start scraping in minutes. Automate your tasks with our Cloud Scraper. No software to download, no coding needed.

Web Scraper · Cloud · Video Tutorials · Documentation

https://www.parsehub.com › blog › what-is-web-scraping

What is Web Scraping and What is it Used For? | ParseHub

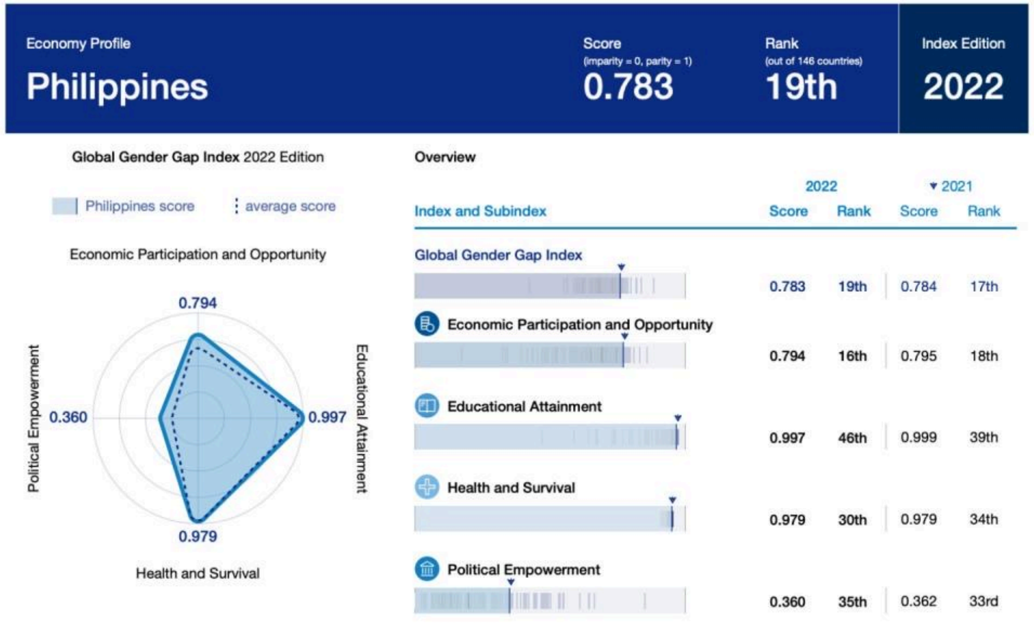
Aug 1, 2021 — **Web scraping** refers to the extraction of data from a website. This information is

Data Preview Total [9] line(s) of data Preview more data

No.	Title	tbwupd
1	Web scraping - Wikipedia	https://en.wikipedia.org › wiki › Web_scraping
2	Web Scraper - The #1 web scraping extension	
3	What is Web Scraping and What is it Used For...	https://www.parsehub.com › blog › what-is-we...
4	What Is Web Scraping? A Complete Beginner's...	https://careerfoundry.com › blog › data-analyti...
5	A Practical Introduction to Web Scraping in Py...	https://realpython.com › python-web-scraping...
6	Web Scraping with Python: Everything you nee...	https://www.scrapingbee.com › blog › web-scr...

3.3. Web scraped news data on VAW

- Gender Gap Index 2022 (WEF): While being the only Asian country in the Top 20, PH fell two spots from 2021 rankings (17th), indicator on political empowerment remains low
- Data on VAW reported cases: "Low incidence of cases does not mean that VAW decreased." - Anna Laurene Del Rosario, Information Officer of the Inter-Agency Council on Violence Against Women and their Children



Economy Profile

Philippines

Score (imparity = 0, parity = 1) 0.783

Rank (out of 146 countries) 19th

Index Edition 2022

Global Gender Gap Index 2022 Edition

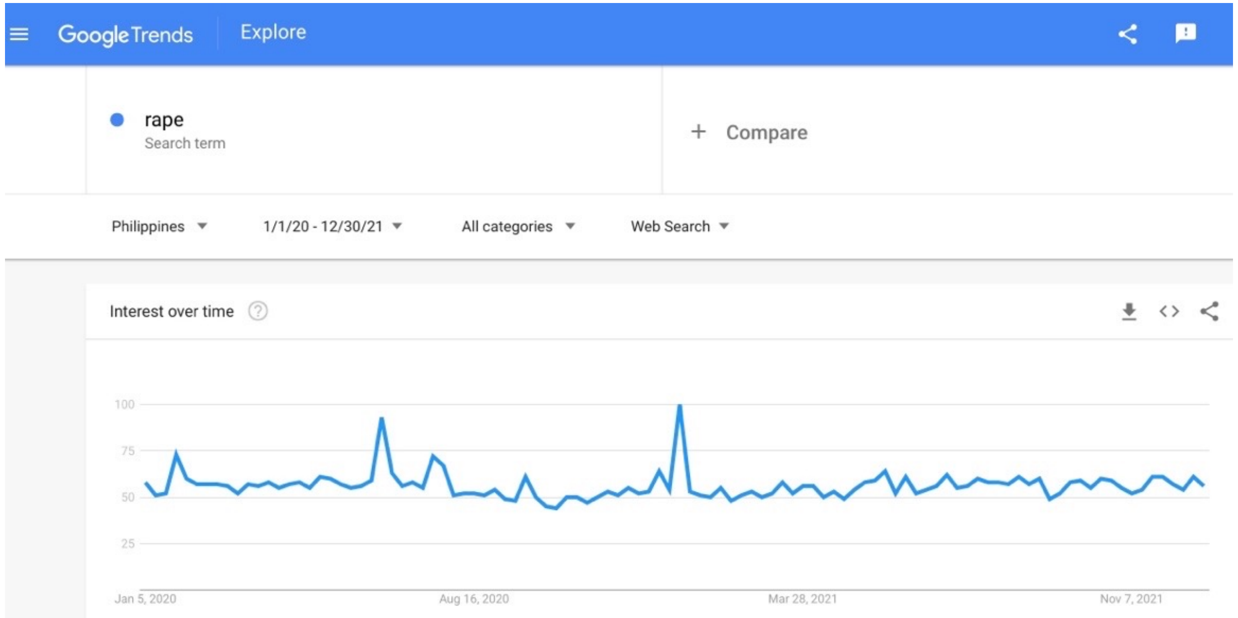
Overview

Index and Subindex	2022 Score	2022 Rank	▾ 2021 Score	▾ 2021 Rank
Global Gender Gap Index	0.783	19th	0.784	17th
Economic Participation and Opportunity	0.794	16th	0.795	18th
Educational Attainment	0.997	46th	0.999	39th
Health and Survival	0.979	30th	0.979	34th
Political Empowerment	0.360	35th	0.362	33rd

Year	No. of Cases Served by DSWD (under RA 9262)	No. of Cases Reported to PNP (under RA 9262)
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

3.3. Web scraped news data on VAW (cont’d)

- DataReportal (2022): **82%** of GWI survey respondents worldwide rely on online channels for news
- Web scraped news related to violence against women in the Philippines: **561 contents** from ABS-CBN, The Philippine Daily Inquirer, Manila Bulletin, The Manila Times, and Rappler (2016-2022)



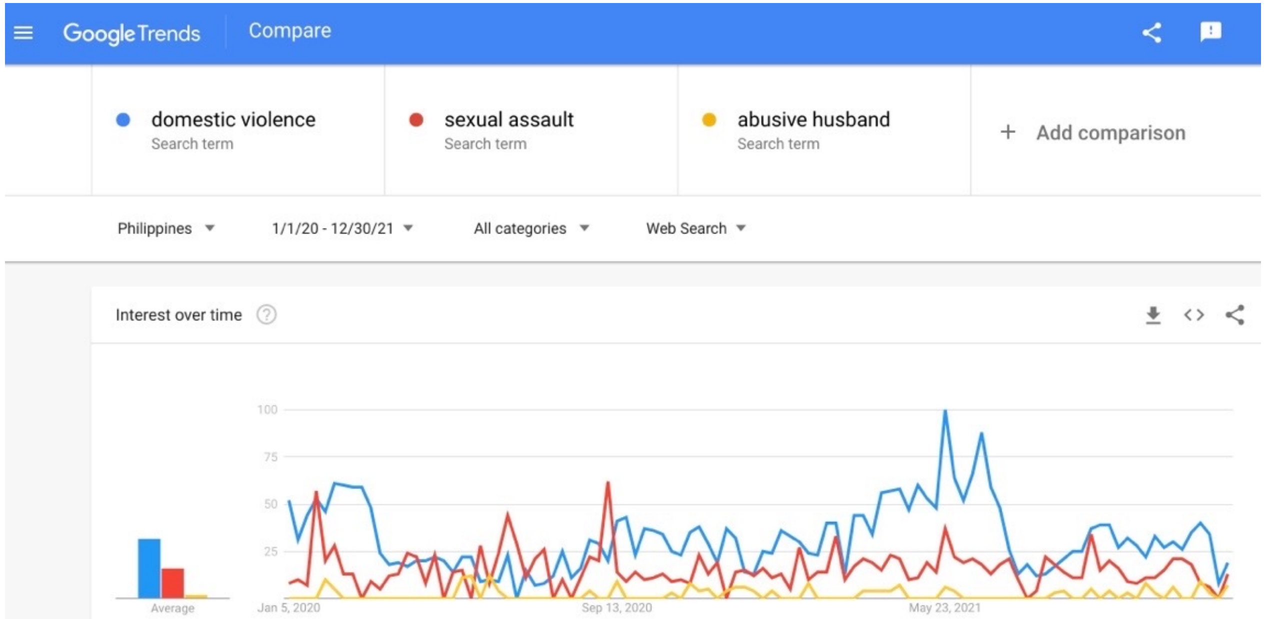
Google Trends Explore

● rape
Search term

+ Compare

Philippines ▾ 1/1/20 - 12/30/21 ▾ All categories ▾ Web Search ▾

Interest over time



Google Trends Compare

● domestic violence
Search term

● sexual assault
Search term

● abusive husband
Search term

+ Add comparison

Philippines ▾ 1/1/20 - 12/30/21 ▾ All categories ▾ Web Search ▾

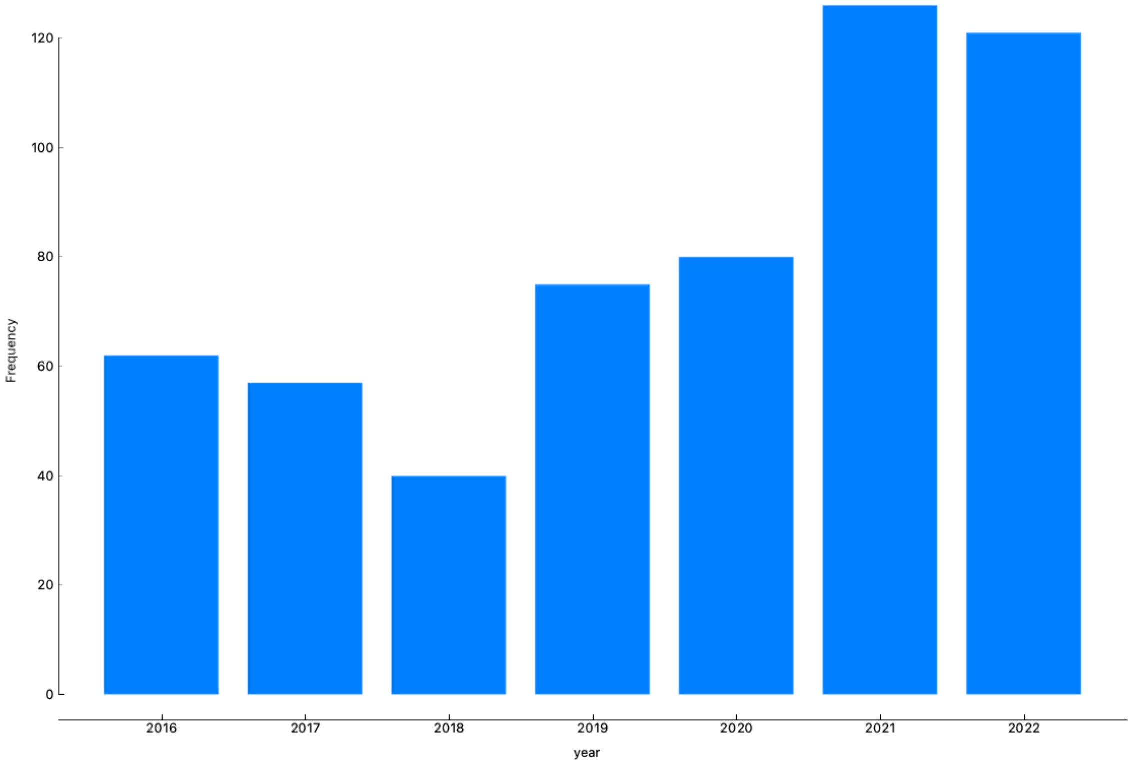
Interest over time

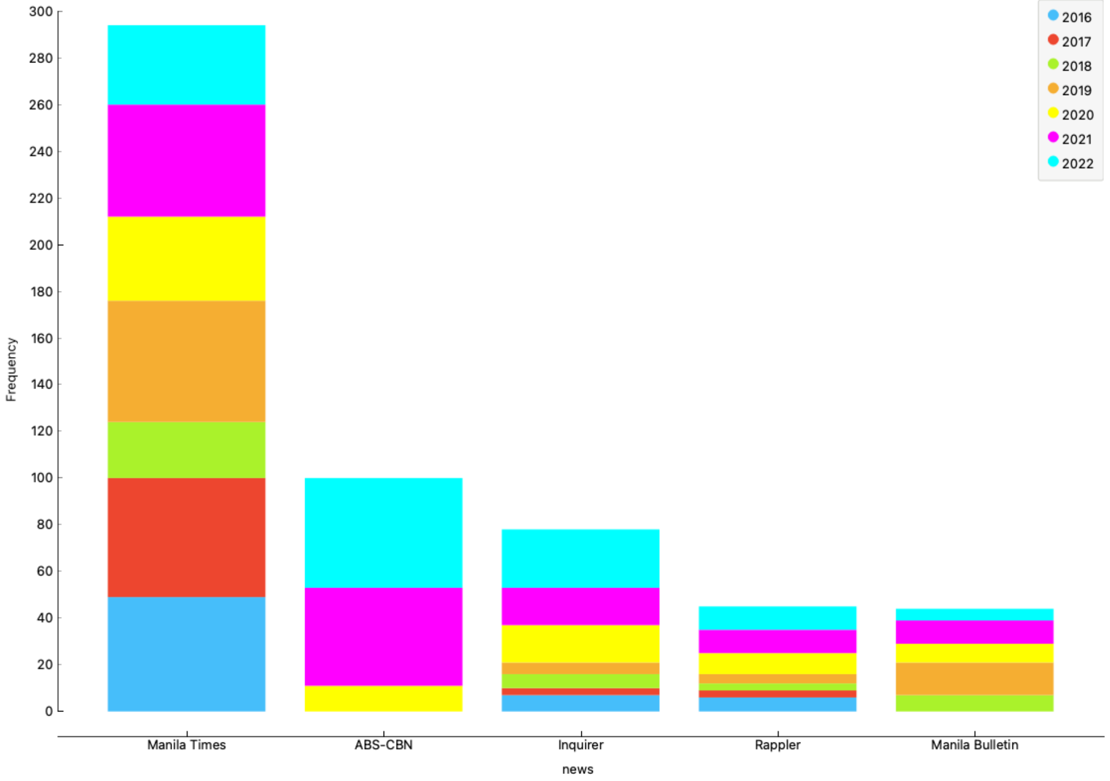
Google Trends in the Philippines on Searches for (a) the term “rape”’, and (b) the terms “domestic violence”, “sexual assault”, “abusive husband”: Jan 2020-Dec 2021

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

3.3. Web scraped news data on VAW (cont'd)

- DataReportal (2022): **82%** of GWI survey respondents worldwide rely on online channels for news
- Web scraped news related to violence against women in the Philippines: **561 contents** from ABS-CBN, The Philippine Daily Inquirer, Manila Bulletin, The Manila Times, and Rappler (2016-2022)





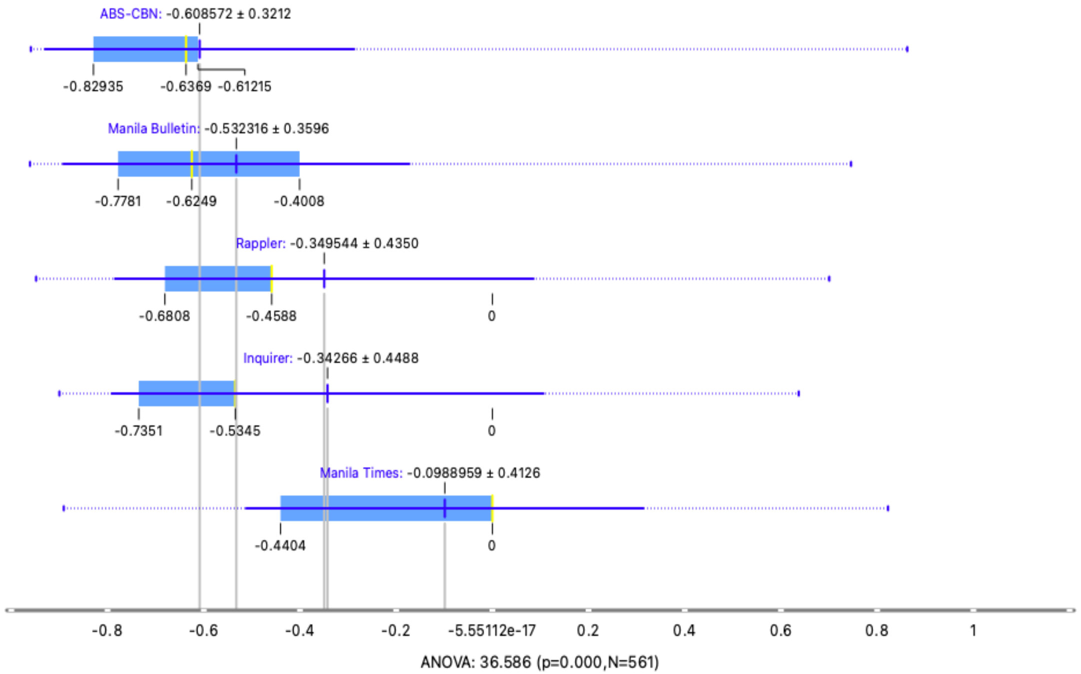
3.3. Web scraped news data on VAW (cont’d)

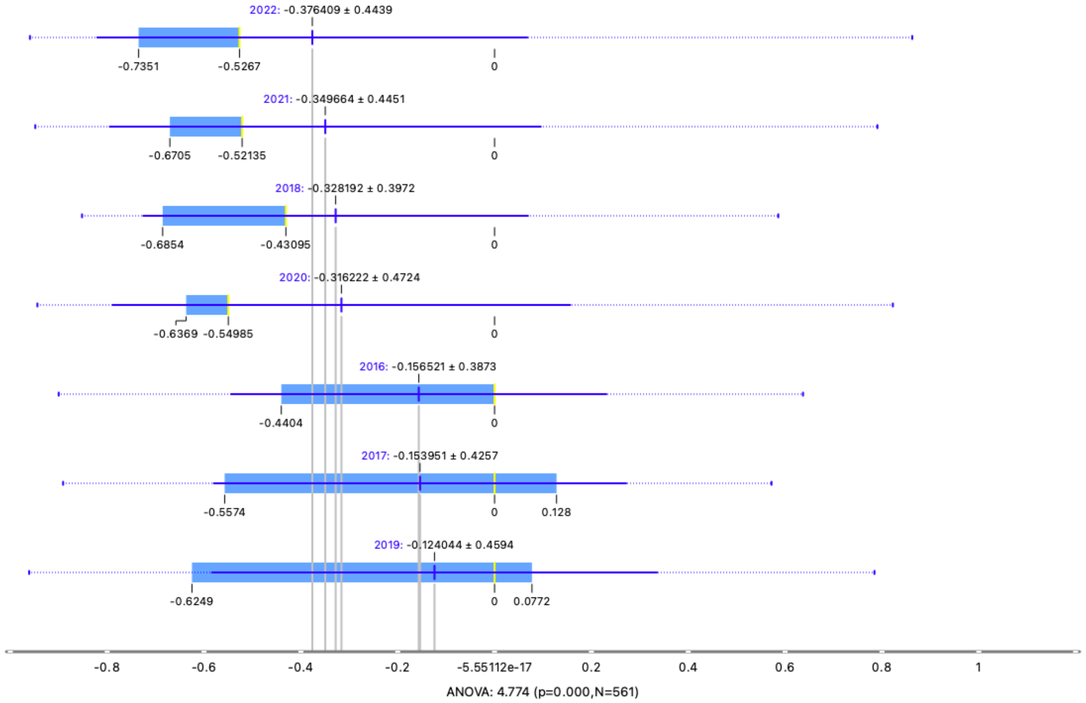
- **Sentiment Analysis**: (or opinion mining) a natural language processing that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions via the computational treatment of subjectivity in text (Hutto and Gilbert, 2014).
  - VADER method: (Valence Aware Dictionary for sEntiment Reasoning).
    - Uses a combination of qualitative and quantitative methods to produce, and empirically validate, a gold-standard sentiment lexicon (i.e., VADER uses a list of positive and negative words with scores depending on intensity)

3.3. Web scraped news data on VAW (cont’d)

Sentiment Analysis (results):

- ❑ By news site: distribution of news content lean more on the negative side across all news sites
- ❑ By year: VAW-related news contents during in the last two years reflect more negative scores
- ❑ However, positive and negative scores provide limited insights to address a policy issue

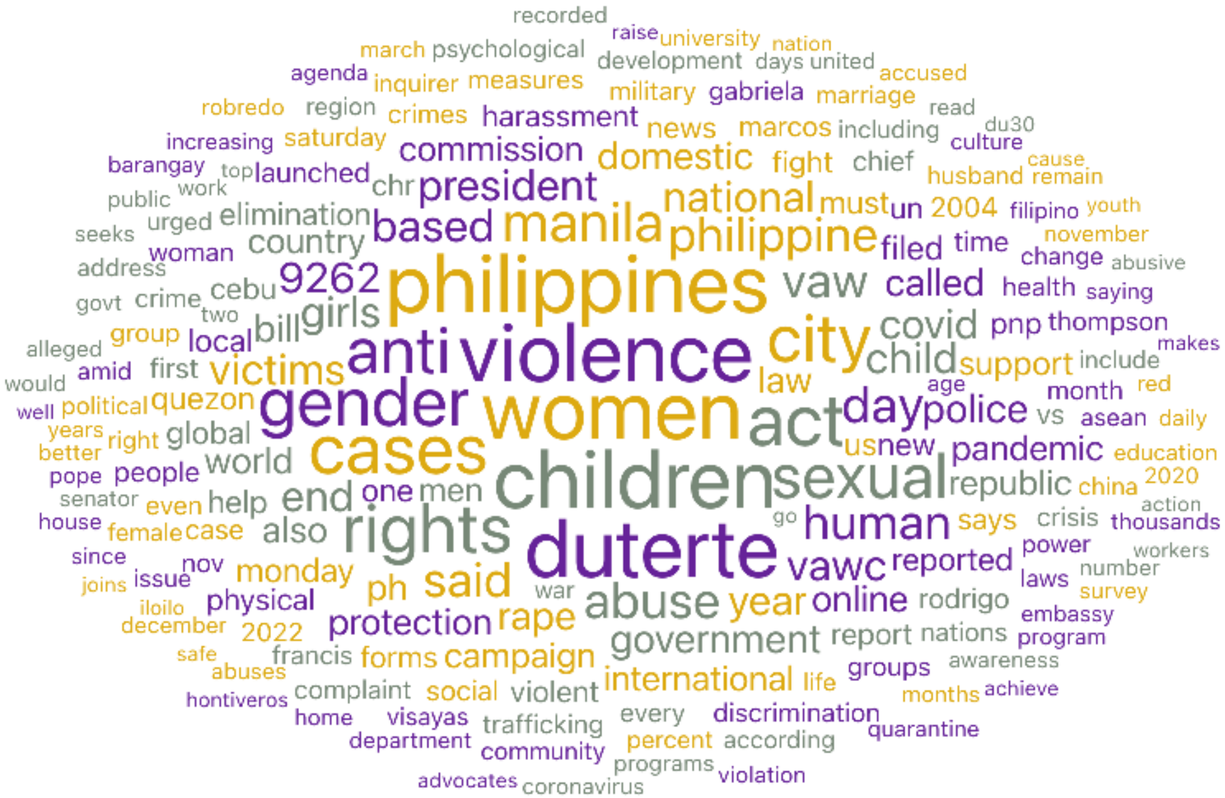


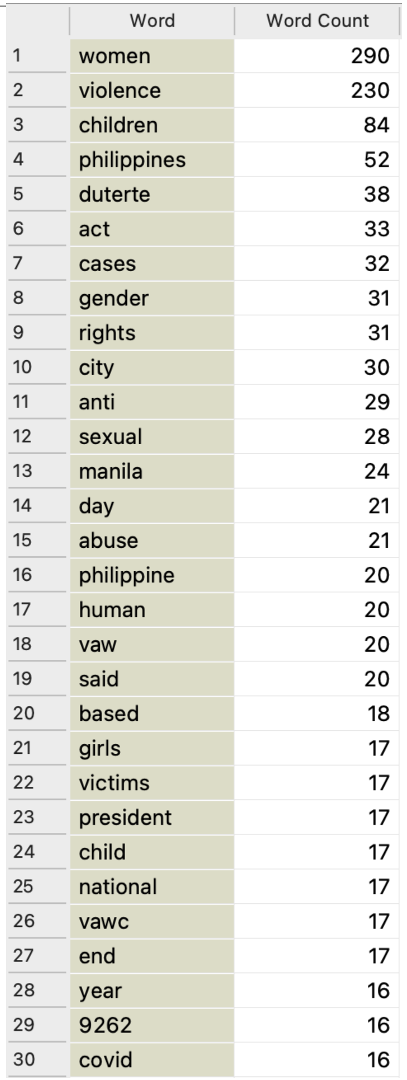


3.3. Web scraped news data on VAW (cont’d)

Word Cloud

- Visual representation of words in a corpus (i.e., collection of documents), with the size of the word reflecting its frequency or importance.





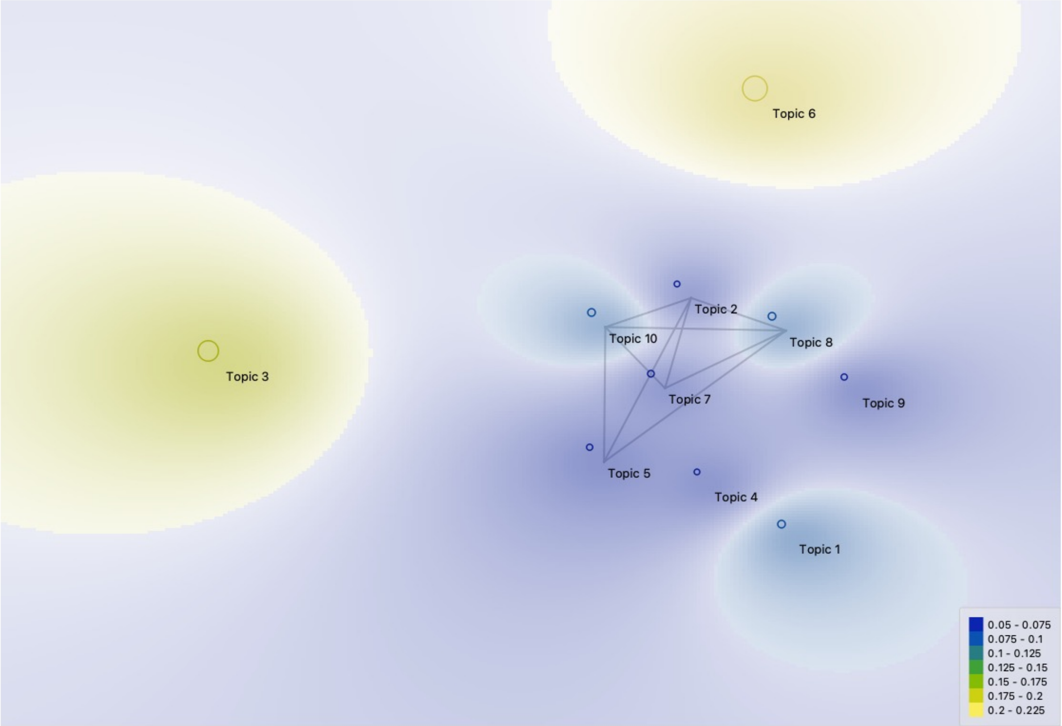
	Word	Word Count
1	women	290
2	violence	230
3	children	84
4	philippines	52
5	duterte	38
6	act	33
7	cases	32
8	gender	31
9	rights	31
10	city	30
11	anti	29
12	sexual	28
13	manila	24
14	day	21
15	abuse	21
16	philippine	20
17	human	20
18	vaw	20
19	said	20
20	based	18
21	girls	17
22	victims	17
23	president	17
24	child	17
25	national	17
26	vawc	17
27	end	17
28	year	16
29	9262	16
30	covid	16

3.3. Web scraped news data on VAW (cont'd)

Topic Modelling

- Statistical modelling that discovers abstract topics in clusters of similar words found in a corpus
- Latent Dirichlet allocation (LDA): documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words

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



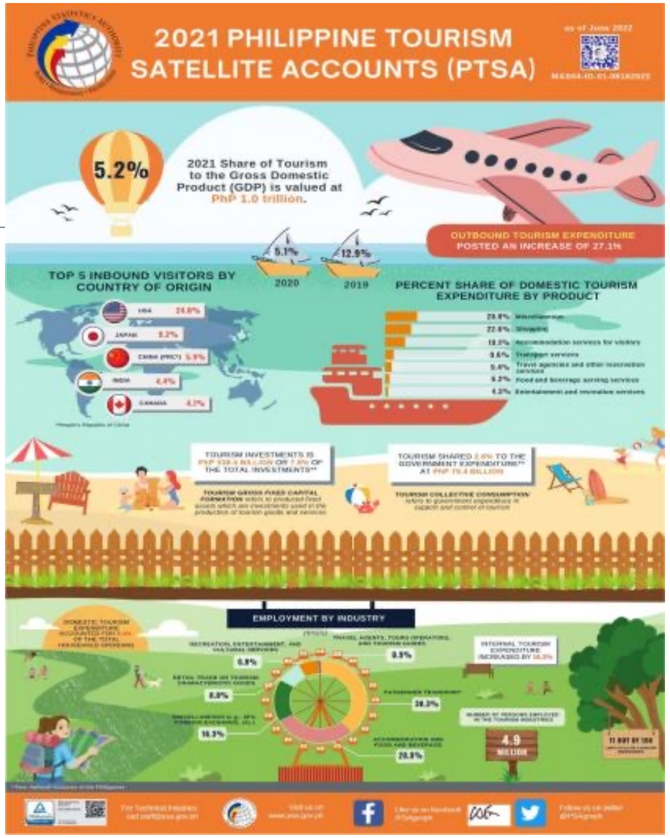
3.4. Text Mining Tourism Data

<u>Official sources of tourism statistics</u>:

- PSA (Tourism Satellite Accounts)
  - E.g., Share of tourism to GDP, total employment
- DOT (Tourism Demand Statistics)
  - E.g., Data on visitor arrivals by country of residence (latest: Sept 2022); regional travelers (latest: 2020 data)

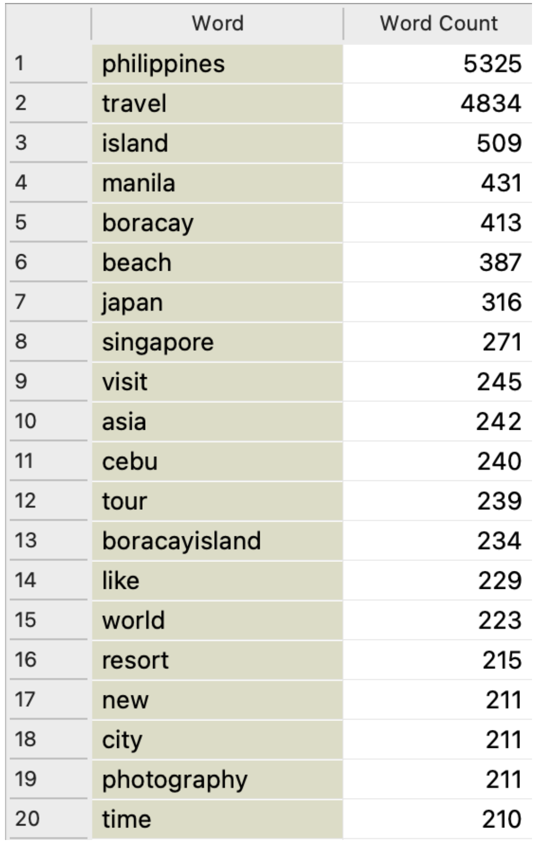
<u>Other data sources:</u>

- LGU data
- Travel reviews (e.g., Conde Nast: Boracay as top island in the world, Palawan at 8th place; PH 30th top country and top 10 friendliest country)
- **Social media (Twitter data)**

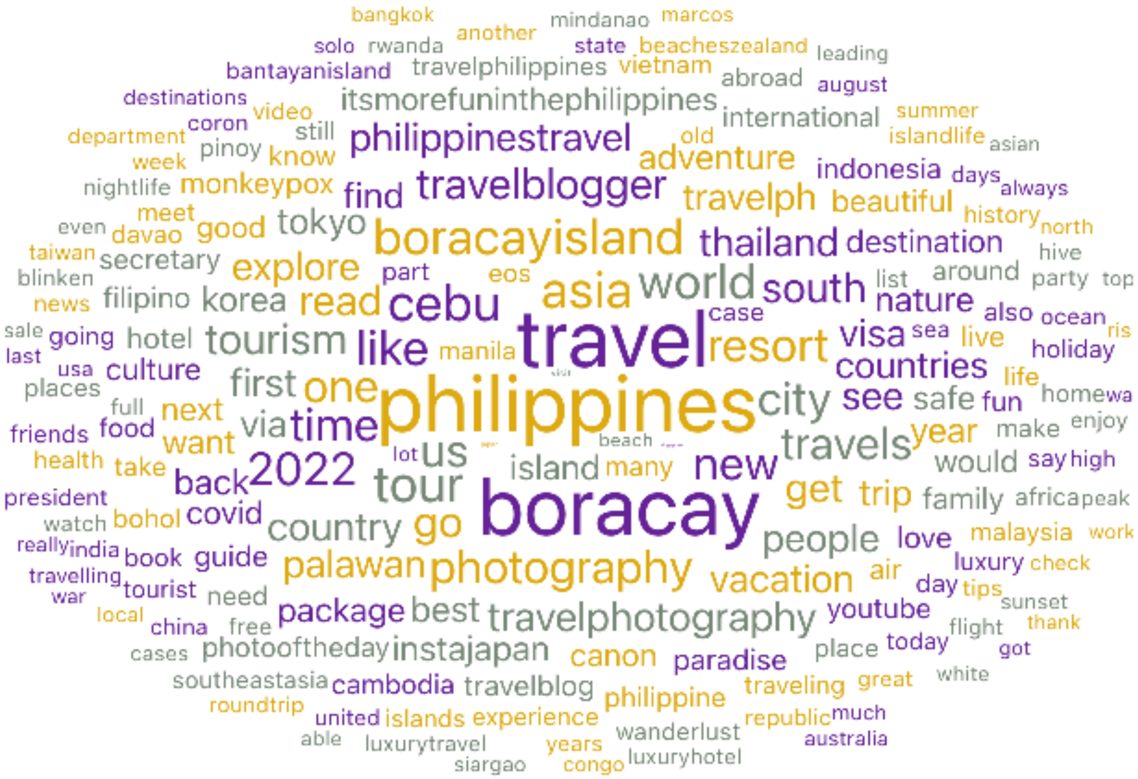

2021 PHILIPPINE TOURISM SATELLITE ACCOUNTS (PTSA)
as of June 2022
5.2%
2021 Share of Tourism to the Gross Domestic Product (GDP) is valued at PhP 1.0 trillion.
OUTBOUND TOURISM EXPENDITURE POSTED AN INCREASE OF 27.1%
5.1%
12.9%
2020
2019
TOP 5 INBOUND VISITORS BY COUNTRY OF ORIGIN
USA
JAPAN
CHINA (PRC*)
INDIA
CANADA
PERCENT SHARE OF DOMESTIC TOURISM EXPENDITURE BY PRODUCT
Miscellaneous
Shopping
Accommodation services for visitors
Transport services
Travel agencies and other reservation services
Food and beverage serving services
Entertainment and recreation services
TOURISM INVESTMENTS IS
OF THE TOTAL INVESTMENTS**
TOURISM SHARED
TO THE GOVERNMENT EXPENDITURE**
TOURISM GROSS FIXED CAPITAL FORMATION
TOURISM COLLECTIVE CONSUMPTION
EMPLOYMENT BY INDUSTRY
INTERNAL TOURISM EXPENDITURE INCREASED BY
4.9 MILLION
11 OUT OF 100

3.4. Text Mining Tourism Data (cont’d)

Word Cloud



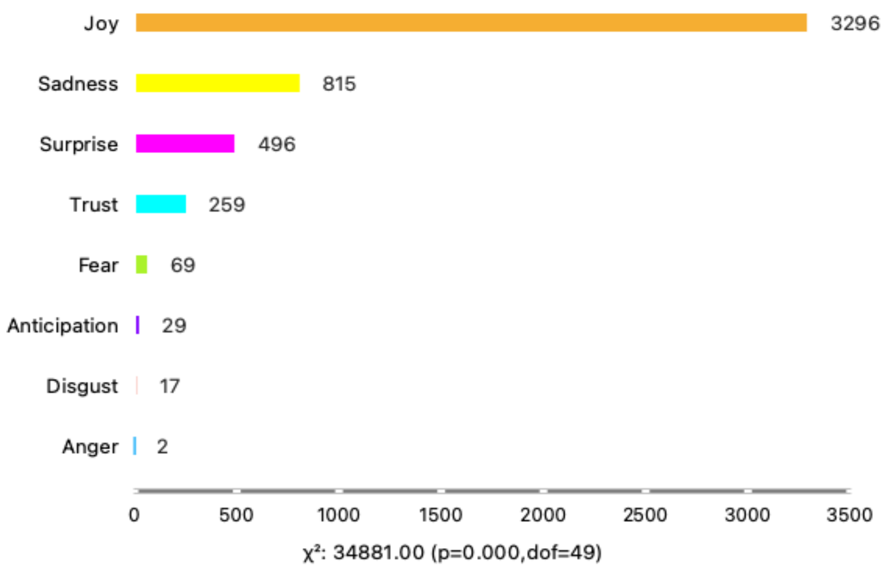
	Word	Word Count
1	philippines	5325
2	travel	4834
3	island	509
4	manila	431
5	boracay	413
6	beach	387
7	japan	316
8	singapore	271
9	visit	245
10	asia	242
11	cebu	240
12	tour	239
13	boracayisland	234
14	like	229
15	world	223
16	resort	215
17	new	211
18	city	211
19	photography	211
20	time	210



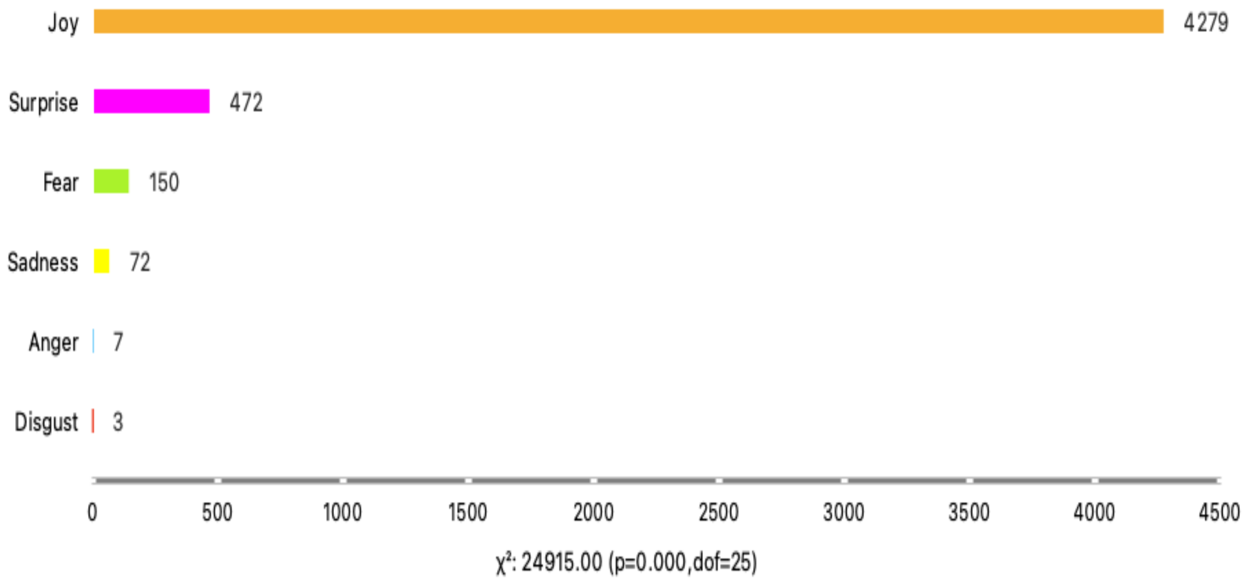
3.4. Text Mining Tourism Data (cont'd)

Sentiment Analysis (Emotions Classification)

- Emotion detection (types of feelings) of tweets using Plutchik and Ekman classifications



Plutchik Classification



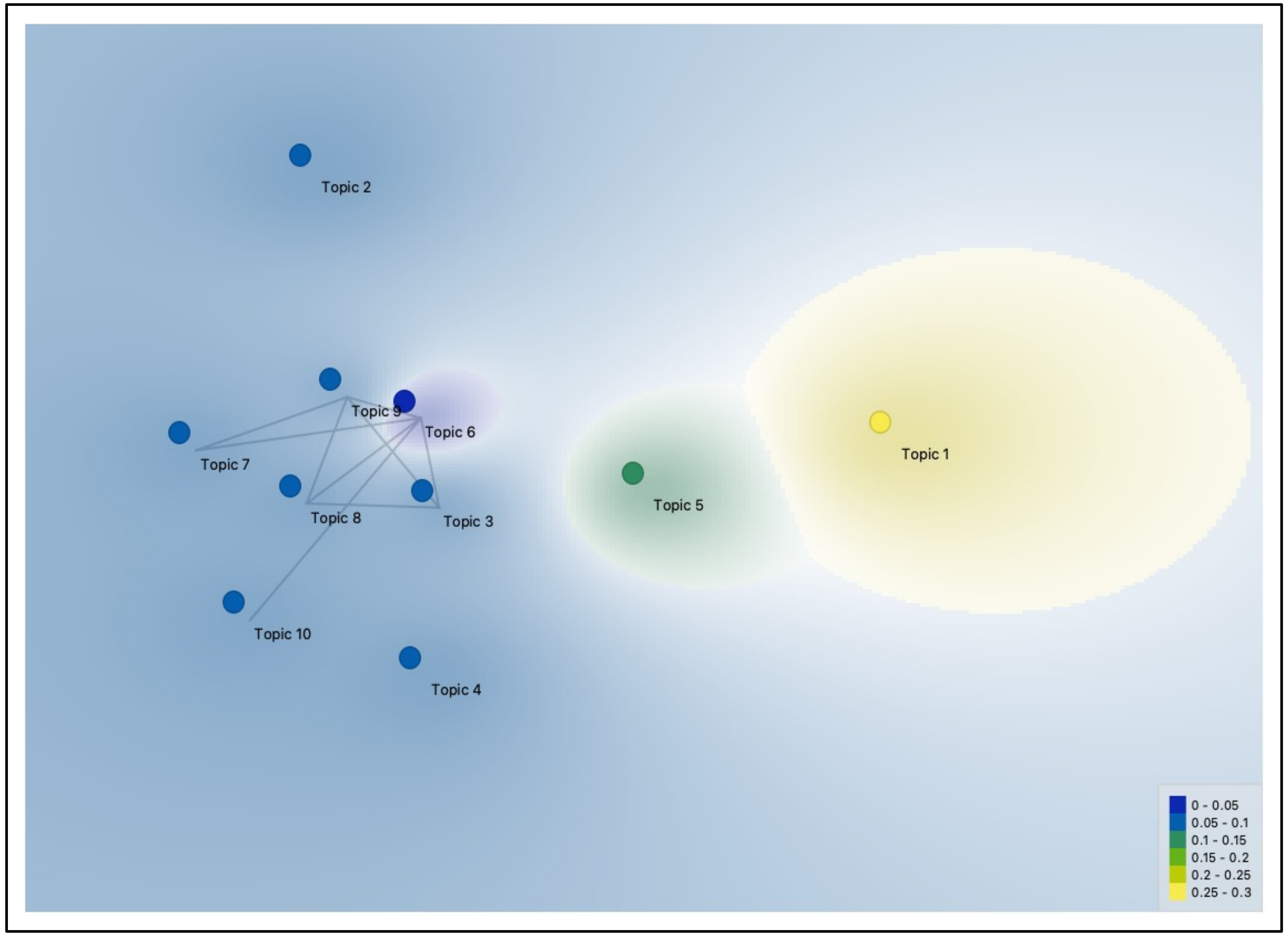
Ekman Classification

3.4. Text Mining Tourism Data (cont'd)

Topic Modelling

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

Multidimensional Scaling



4. Recommendations and Ways Forward

- PIDS research staff (and the institute as whole) should examine new data sources as they complement though they can not replace traditional data sources (i.e., surveys, censuses)
  - PIDS should regularly conduct data analytics on download data (such as market basket analysis) to identify patterns of association beyond themes but also publications to develop targeted marketing campaigns
  - Standards must be set to ensure data are fit for use, e.g., examine dimensions of data quality (Brackstone, 1999) such as relevance, accuracy, timeliness, accessibility, interpretability, and coherence
  - New data sources have many benefits, especially in the context of addressing data gaps and other gaps on disaggregated data for monitoring development outcomes (e.g., gender, tourism). Big data provides a fast and cheap stream of information that can supplement traditional data analyses, enhancing responsiveness to policy issues (Ceron and Negri, 2015)

4. Recommendations and Ways Forward

- Risk assessment and risk mitigation on use of big data and other new data sources since the world of big data and hyper connectivity no longer guarantees irreversible de-identification
  - potential harms posed to individuals and to identifiable groups or populations
  - identifying threshold 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?
  - balance needs to be struck between protecting data privacy and harnessing use of new data sources for safeguarding civil rights, ensuring fairness, and preventing discrimination.

END