

Analysis of Foreign Media Influence on the South Pacific Environment

Derek Lilienthal*[†]

Advisor: Dr. Elizabeth Gooch[‡][§]

February 2021

Abstract

In this research, we quantify foreign actors media events in the South Pacific involving an environmental theme. We use the GDELT (Global Database of Events, Language, and Tone) database to compare the tones of articles produced by Western, Chinese, and South Pacific (Local) media sources that involve an environmental theme and when a great power (United States, China, Australia, New Zealand, Japan, and Russia) is involved as an actor. We found when comparing Western, Chinese, and Local news sources, the average sentimental analysis of Western tones is negative, the average of Local tones is slightly positive, and the average of Chinese tones are very positive. When comparing the difference in means by each set of news sources, we used the Welch's two sample t-test because the distribution of Western, Chinese, and Local tones followed a normal distribution but had unequal variances among the groups. After conducting our statistical analysis, we found there is strong evidence to conclude the difference in means of tones between the three media sources are statistically significant between each pairwise comparison.

Introduction

The GDELT (Global Database of Events, Language and Tone) database was created to provide a platform that collects the world's news from nearly every corner of every country in print, broadcast, and web formats [1]. GDELT provides a big data resource to analyze the world's news that allows researchers to explore trends in media that was not easily accessible before. The data is stored in a relational database that is freely accessible through Google's

*California State University, Monterey Bay, Seaside CA 93955

[†]dlilienthal@csumb.edu

[‡]Naval Postgraduate School, Monterey CA 93943

[§]elizabeth.gooch@nps.edu

BigQuery platform. The accessibility of GDELT through BigQuery allows researchers to use standard SQL (Structured Query Language) to access the data. GDELT additionally updates its database every 15 minutes [2]. Allowing researchers the ability to track eventsⁱ as they are happening in almost real time.

The amount of raw data that GDELT gives researchers is monumental. At the time of this research, there were over five billion unique entries in the GKG (Global Knowledge Graph) table. GDELT's database itself has over 50 tables available through BigQuery. But for the purpose of this research, we will solely be focusing only three of GDELT's tables. The Events, Eventmentions, and GKG.

An example on how these three tables are related to each other: If Fox News is the first to report on about a political scandal, the event will first have an entry in the Events table. After 15 minutes, any follow-up news articles (or if a different media outlet like CNN) that report on the same event will all be separate entries in the Eventmentions table.

When a new event is created in the Events table, it is assigned a unique identifier called the *GLOBALEVENTID*. The *GLOBALEVENTID* is also present in each entry in the Eventmentions table. This way, we can join both the Event and Eventmentions table on each events' unique id. The GKG table uniquely identifies each event by the URL where article originated. This way, we can join the Eventmentions and GKG tables through each events URL because the Eventmentions table also contains the source URL of where the article originated from. Thus, allowing us to join the Events, Eventmentions, and GKG tables collectively to form our entire data set that we will be analyzing throughout this research.

Attributes of the Events, Eventmentions, and GKG Tables

The Events table is where GDELT creates a new entry for each event, the Eventmentions table is where GDELT tracks the life of each event as it spans across more media outlets and as the event continues to develop, and the GKG table is a detailed analysis of every news article itself [3][4]. Each event in the Events table is stored in a CAMEO (Conflict and Mediation Event Observation) format where the two actorsⁱⁱ and the action performed by *Actor1* upon *Actor2* is recorded [4]. The Geographical location where the event took place is also recorded along with the geographical location where *Actor1* and *Actor2* reside from. If the event only involves one or no actors, then these attributes can be left blank [4].

The Eventmentions table is an extension of the Events table. It records all the mentions of each event as it spans across multiple news sources [4]. It tracks each event as it is reported after the first initial recording in the Events table. The Eventmentions table also includes additional details about each event compared that is not in the Events table. Some of the additional attributes are: the articles toneⁱⁱⁱ, called *MentionDocTone*, each actors

ⁱIn this research, we refer to each *event* as a news article captured by GDELT

ⁱⁱIn this research, we refer to each *Actor* as a country or territory.

ⁱⁱⁱEach *tone* is generated through sentiment analysis and is given a value of -100 to 100 but the majority of articles are within the -10 to 10 range [4].

character offset in the article, the level of confidence (from 0 to 100) that the article is related to the *GLOBALEVENTID*, etc [4]. Many of those attributes can be used to filter articles out articles based on length or the importance of an actor based on where it was located within the article. However, in this research the only attribute we are utilizing in the Eventmentions table will be the *MentionDocTone*.

The GKG table “connects every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what’s happening around the world, what its connect is and who’s involved, and how the world is feeling about it, every single day” [3]. In other words, the GKG table gives an additional level of analysis for each event recorded in GDELT. These additional attributes can be used to further filter articles by organizations, persons, themes, tones, locations, and more. For this research, we will be focusing on using the themes, *V2Themes*, and tones, *V2Tone*, attributes of the GKG table.

Purpose of Report

This paper’s main purpose is to demonstrate the methods and data engineering techniques that can highlight geopolitical issues involving great power nations and events involving a specified theme by using GDELT. In this paper, we will focus primarily on Chinese, Western, and Local news sources in South Pacific when a great power is involved as an *actor* and the event involves an environmental theme.

Choosing Great Powers and the South Pacific

We chose the United States, China, and Russia as our great powers because the United States has identified China and Russia as ‘great power competition’ regarding military capabilities, nuclear weapons, allied military capabilities in the Indo-Pacific region, supply chain security, capabilities for countering hybrid warfare and gray-zone tactics, etc [5]. Australia, New Zealand, and Japan were also chosen to be a Great Power because the Department of Defense (DOD) has been ‘strengthen[ing] the military capabilities of U.S. allies in the [Indo-Pacific] region, in particularly Japan and Australia’ and New Zealand to improve the ability of forces from these countries to operate effectively with U.S. forces [5].

Because this Indo-Pacific region includes many of the South Pacific islands, we generalized to analyzing the regions of the South Pacific that are especially affected by any drastic changes in their environment. Islands in the South Pacific are ‘facing devastating impacts of climate change including increasing droughts and water scarcity, coastal flooding and erosion, changes in rainfall that affect ecosystems and food production’ [6]. Due to this, it is essential to these remote locations to have national governments recognize and support these communities [6].

With DOD officials have subsequently identified countering China’s military capabilities as DOD’s top priority, we wanted to analyze how Chinese, Western, and Local media sources

write about environmental issues in the South Pacific when one of the great powers is directly involved [5].

Defining Locations of Interest

For the remainder of the report, we will be focusing on analyzing Western, Chinese, and Local media tones across the South Pacific when a great power is either *Actor1* or *Actor2*, the location of the event is a South Pacific country or territory, and the event involves an environmental theme^{iv}. The great powers of interest in this research are: The United States, China, Australia, New Zealand, Japan, and Russia.

The locations of interest in the South Pacific are: Micronesia, Fiji, Kiribati, Marshall Islands, Nauru, Palau, Papua New Guinea, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu, Cook Islands, Niue, American Samoa, Ashmore Reef, Baker Island, Coral Sea, Easter Island, Galapagos Islands, French Polynesia, Guam, Howland Island, Jarvis Island, Johnston Atoll, Kingman Reef, Midway Island, New Caledonia, Norfolk Island, Northern Mariana Islands, Ogasawaramura Japan, Palmyra Atoll, Papua Indonesia, Pitcairn Islands, Tokelau, Wake Island, Wallis and Futuna, West Papua, and Bonin Islands.^v

Data

In this next section, I will be talking about the methods used to collect our data set. We used SQL to query GDELT and Python to further refine our data.

Using SQL through BigQuery

In order to gain a perspective of how many articles are contained in GDELT and what portion of those are located in the South Pacific, we tabulated the total number of articles in GDELT and the total number of articles located in the South Pacific using two separate SQL scripts.

LISTING 1: Tallying the total number of articles in GDELT

```
1 SELECT
2     count(1)
3 FROM
4     `gdelt-bq.gdeltv2.eventmentions` AS em JOIN `gdelt-bq.gdeltv2.event` AS e
5     ON em.GLOBALEVENTID = e.GLOBALEVENTID JOIN `gdelt-bq.gdeltv2.gkg` AS GKG
6     ON em.MentionIdentifier = GKG.DocumentIdentifier
```

^{iv}There are 21 environmental themes created by GDELT: *ENV_CLIMATECHANGE*, *ENV_OIL*, *ENV_FISHERY*, *ENV_MINING*, *ENV_COAL*, *ENV_GREEN*, *ENV_SOLAR*, *ENV_METALS*, *ENV_POACHING*, *ENV_NATURALGAS*, *ENV_DEFORESTATION*, *ENV_OVERFISH*, *ENV_FORESTRY*, *ENV_NUCLEARPOWER*, *ENV_WATERWAYS*, *ENV_SPECIESENDANGERED*, *ENV_HYDRO*, *ENV_BIOFUEL*, *ENV_GEOTHERMAL*, *ENV_WINDPOWER*, *ENV_CARBONCAPTURE*, and *ENV_SPECIESEXTINCT*.

^vWe did not include Hawaii, New Zealand and Australia in our locations of interest because those three locations accounted for the majority of all the entries in our South Pacific subset of data from GDELT.

LISTING 2: Tallying the total number of articles in the South Pacific

```

1 SELECT
2     count(1)
3 FROM
4     `gdelt-bq.gdeltv2.eventmentions` AS em JOIN `gdelt-bq.gdeltv2.event` AS e
5     ON em.GLOBALEVENTID = e.GLOBALEVENTID JOIN `gdelt-bq.gdeltv2.gkg` AS GKG
6     ON em.MentionIdentifier = GKG.DocumentIdentifier
7 WHERE
8     (ActionGeo_ADM1Code like `FM%` -- Micronesia
9     OR ActionGeo_ADM1Code like `FJ%` -- Fiji
10    OR ActionGeo_ADM1Code like `KR%` -- Kiribati
11    OR ActionGeo_ADM1Code like `RM%` -- Marshall Islands
12    OR ActionGeo_ADM1Code like `NR%` -- Nauru
13    OR ActionGeo_ADM1Code like `PS%` -- Palau
14    OR ActionGeo_ADM1Code like `PP%` -- Papua New Guinea
15    OR ActionGeo_ADM1Code like `WS%` -- Samoa
16    OR ActionGeo_ADM1Code like `BP%` -- Solomon Islands
17    OR ActionGeo_ADM1Code like `TN%` -- Tonga
18    OR ActionGeo_ADM1Code like `TV%` -- Tuvalu
19    OR ActionGeo_ADM1Code like `NH%` -- Vanuatu
20    OR ActionGeo_ADM1Code like `CW%` -- Cook Islands
21    OR ActionGeo_ADM1Code like `NE%` -- Niue
22    OR ActionGeo_ADM1Code like `AQ%` -- American Samoa
23    OR ActionGeo_FullName = `Ashmore Reef, Queensland, Australia`
24    OR ActionGeo_ADM1Code like `FQ%` -- Baker Island
25    OR ActionGeo_FullName = `Coral Sea, Oceans (general), Oceans`
26    OR ActionGeo_FullName like `Easter Island, V%`
27    OR ActionGeo_FullName = `Galapagos, Imbabura, Ecuador`
28    OR ActionGeo_ADM1Code like `FP%` -- French Polynesia
29    OR ActionGeo_ADM1Code like `GQ%` -- Guam
30    OR ActionGeo_ADM1Code like `HQ%` -- Howland Island
31    OR ActionGeo_ADM1Code like `DQ%` -- Jarvis Island
32    OR ActionGeo_ADM1Code like `JQ%` -- Johnston Atoll
33    OR ActionGeo_ADM1Code like `KQ%` -- Kingman Reef
34    OR ActionGeo_FullName = `Midway Island, Western Australia, Australia`
35    OR ActionGeo_ADM1Code like `NC%` -- New Caledonia
36    OR ActionGeo_ADM1Code like `NF%` -- Norfolk Island
37    OR ActionGeo_ADM1Code like `CQ%` -- Norther Mariana Islands
38    OR ActionGeo_FullName = `Ogasawaramura, Tokyo, Japan`
39    OR ActionGeo_ADM1Code like `LQ%` -- Palmyra Atoll
40    OR ActionGeo_ADM1Code = `ID36` -- Papua, Indonesia
41    OR ActionGeo_ADM1Code like `PC%` -- Pitcairn Islands
42    OR ActionGeo_ADM1Code like `TL%` -- Tokelau
43    OR ActionGeo_ADM1Code like `WQ%` -- Wake Island
44    OR ActionGeo_ADM1Code like `WF%` -- Wallis and Futuna
45    OR ActionGeo_ADM1Code = `ID39` -- West Papua, Indonesia
46    OR ActionGeo_FullName = `Bonin Islands, Tokyo, Japan`)

```

With both SQL statements tallying the number of articles captured in all of GDELT and the number of articles where the location of the event in the South Pacific, we found that only 0.08% of articles contained in GDELT were located in the South Pacific [Figure 1].

Next, we ran a third query that returned us a data set containing the attributes *Ac-*

tor1CountryCode, *Actor2CountryCode*, *AvgTone*, *MentionDocTone*, *SourceCommonName*, *V2Themes*, and *V2Tone*. This query is identical to the second query (*Listing2*) we ran except we replaced line 2 with the attributes previously listed. The data set returned from this query contains every article ever reported by GDELT where the location of the event is one the territories or countries in the South Pacific that we have explicitly stated. For the rest of the report, we will be doing the remainder of our data and statistical analysis using Python and R.

Using Python for Tallying

Next, we used the Python libraries, *Pandas* and *Numpy*, to further filter, tally, and visualize our data. Loading in the data into a *Pandas* data frame in Jupyter Notebooks allowed us to manipulate the data further without the need to constantly be running queries through external resources, like BigQuery.

Of all the articles in the South Pacific, only 14.43% of them contain an environmental theme^{vi}. When further filtering the data set by only containing the great powers as either *Actor1* or *Actor2* in each event, we found that of the 624,075 articles that contained an environmental theme, 27.02% (168,594 articles) involved a great power [*Figure 1*].

A further breakdown of the six great powers in the South Pacific, shows that the United States (*USA*) is the top actor with being in over 47% (48,561) of the articles where a great power is present and the article involves an environmental theme. Next is Australia (*AUS*) with over 17% (17,919), China (*CHN*) with over 14% (15,057), New Zealand (*NZL*) with over 11% (11,228), Japan (*JPN*) with over 7% (7,919), and Russia (*RUS*) with over 1% (1,253). [*Figure 2*]

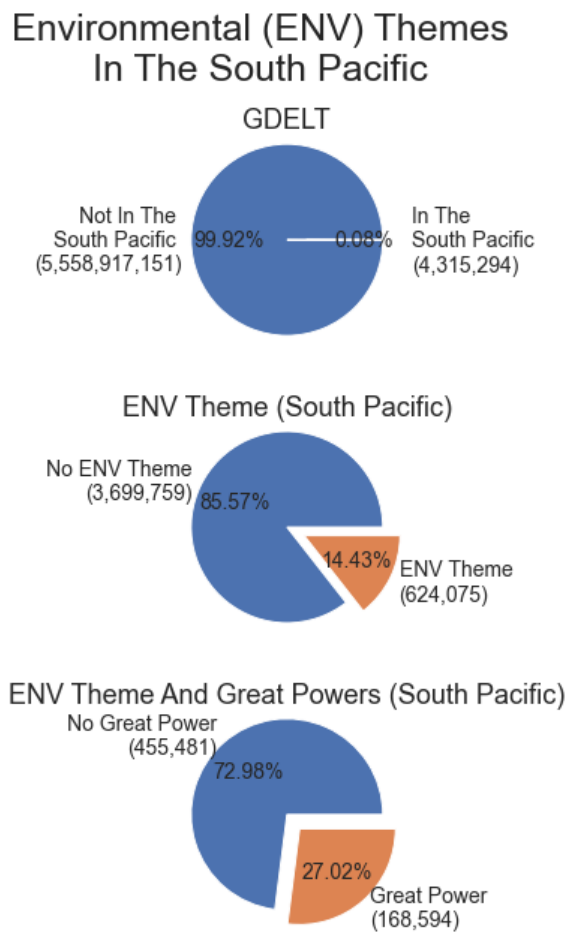


FIGURE 1: Breakdown of GDELT

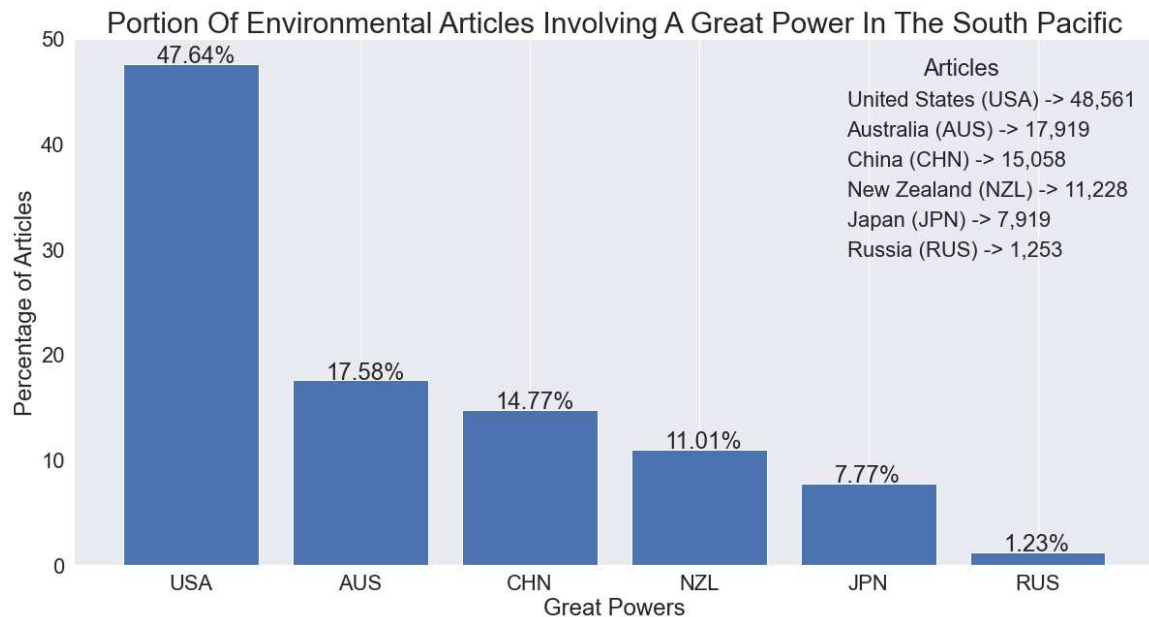


FIGURE 2: Breakdown of Articles by Great Power

Methods

This next section will be discussing how we were able to identify each location of interest within GDELT, which fields we used to specify each location, how we used Python to further filter the data set produced by BigQuery, expand on some of the differences in the fields with similar names and how their features can make a significant difference, how we gathered our media sources, and how we calculated the means of Western, Chinese, and Local tones.

Identifying Locations in the South Pacific

To create a subset of data that only contains events in the South Pacific, we had to identify all the location within the South Pacific and their corresponding FIPS10-4 country codes. GDELT uses the Federal Information Processing Standards (FIPS) codes to identify each events unique location. We looked at Wikipedia to find all the countries and territories in the South Pacific because Wikipedia has a list of names of most of the sovereign states and dependent territories in the South Pacific [7]. We also decided to excluded Australia and New Zealand from our initial list of nations to include in our data set because both of those nations are already well developed and are major influences on the many of the nations in the South Pacific. After we had a list of countries and territories we wanted to include in our analysis, we then needed to find each locations corresponding FIPS code. To find each FIPS code, we ran the query in [Listing 3] through BigQuery:

^{vi}Note, there is an increase in 8,540 articles in the middle pie chart in figure 1 because there was a time difference of a few weeks between running queries.

LISTING 3: Getting The List Of Countries And Territories In GDELТ

```
1  SELECT
2      Actor1Name , Actor1CountryCode , Actor2Name , Actor2CountryCode ,
3      Actor1Geo_FullName Actor1Geo_CountryCode , Actor2Geo_FullName ,
4      Actor2Geo_CountryCode , ActionGeo_FullName , ActionGeo_ADM1Code
5  FROM
6      `gdelt-bq.gdeltv2.events`
```

Next, I manually searched the results in Excel for each of the locations in the South Pacific corresponding two-digit country code (*ActionGeo_CountryCode*). I found this method to be the fastest and most precise way of finding each location compared to other methods like querying for each location in BigQuery, using *Pandas* and *Numpy*, or searching other online resources. We could not trust the FIPS codes found on the internet because there were inconsistencies between what GDELТ labeled certain territories compared to what was found online.

Using this method, I was successful in identifying almost every region except for the locations of Territory of Ashmore and Cartier Islands, Galápagos Islands, Midway Islands, Ogasawara village, Papua Province, and West Papua. Searching for these locations returned some results that did not match the location within the South Pacific or did not return a result at all. I found the names and their country codes within GDELТ after querying GDELТ itself. I did however initially miss-identify West Papua's *ActionGeo_ADM1Code* for a different part of Indonesia. I eventually found the correct *ActionGeo_ADM1Code* for West Papua after a series of queries made against the GDELТ data set on BigQuery and using Google maps to confirm the regions.

ActionGeo_ADM1Code instead of ActionGeo_CountryCode

I decided to use *ActionGeo_ADM1Code* as the main predicate for filtering by location in the queries ran through BigQuery. There are two reasons that made us decide to use *ActionGeo_ADM1Code* instead of *ActionGeo_CountryCode* or *ActionGeo_FullName* in my query. First, I found when filtering by explicitly stating the name of the region using *ActionGeo_FullName*, it produced a result with less data compared to using the *ActionGeo_ADM1Code* predicate. For example, I found that when filtering by *ActionGeo_FullName = 'Hawaii, United States'* compared to using *ActionGeo_ADM1Code = 'USHI'*, there was **346,532**^{vii} additional entries returned using *ActionGeo_ADM1Code = 'USHI'*. The second reason was for consistency in which predicates I was using within the queries. Because every *ActionGeo_ADM1Code* is anywhere between two and four characters and the first two characters of the *ActionGeo_CountryCode* is the also the first two characters of the *ActionGeo_ADM1Code*, I am still capturing the same number of articles using *ActionGeo_ADM1Code* as my filtering predicate as I would if using *ActionGeo_CountryCode* instead.

^{vii}At the time of the initial query

However, some locations within the South Pacific are part of countries that are not entirely considered to be in the South Pacific. For example, Indonesia has two regions that are a part of the South Pacific, but Indonesia itself is not entirely in the South Pacific. To only get those specific regions of Indonesia, I had to explicitly state the whole *ActionGeo_ADMICode* for West Papua and Papua. Also, the locations of Ashmore and Cartier Islands falls underneath an *ActionGeo_ADMICode* that captures more than just that region. For this reason, I had to explicitly state the regions, using the *ActionGeo_FullName*, of Ashmore Reef, Queensland, Australia, Coral Sea Islands, Easter Island, the Galapagos Islands, Midway Islands, Ogasawaramura Islands, and the Bonin Islands to prevent capturing additional articles from their respective countries they belong to.

Python to Filter for Themes

LISTING 4: Filtering the data set using Python

```
# Create boolean masks of each great power
usa_mask = ((df['Actor1CountryCode'] == 'USA') | (df['Actor2CountryCode'] == 'USA'))
chn_mask = ((df['Actor1CountryCode'] == 'CHN') | (df['Actor2CountryCode'] == 'CHN'))
aus_mask = ((df['Actor1CountryCode'] == 'AUS') | (df['Actor2CountryCode'] == 'AUS'))
nzl_mask = ((df['Actor1CountryCode'] == 'NZL') | (df['Actor2CountryCode'] == 'NZL'))
rus_mask = ((df['Actor1CountryCode'] == 'RUS') | (df['Actor2CountryCode'] == 'RUS'))
jpn_mask = ((df['Actor1CountryCode'] == 'JPN') | (df['Actor2CountryCode'] == 'JPN'))

# Filter the data set by only allowing articles that contain an ENV_ theme
df = df[df['V2Themes'].astype(str).str.contains('ENV_')]

# Filter the data set by only allowing a great power being Actor1 or Actor2
df = df[(usa_mask | chn_mask | aus_mask | nzl_mask | rus_mask | jpn_mask)].copy()
```

From the GKG table, the one of the attributes we are interested in for this research are its *Themes*. The *Themes* attribute allowed us to filter our data set of the South Pacific to only contain events that involves the environment. The environmental themes of interest were filtered from the original data set by using *Pandas* by creating boolean masks to apply to the entire data set while also using Python's *str.contains()* method. This allowed us to create a data set that only contained articles where there is an environmental theme associated with it.

GDELDT also has included World Bank themes in their *Themes* field [9]. The World Bank themes offers a greater description in each of their themes compared to themes GDELDT has identified. But for this research, we only are filtering by the themes that begin with 'ENV_' because the World Bank themes do not follow a uniform structure like GDELDT's themes. Therefore, making it exceedingly difficult to only filter by a certain topic using World Bank themes.

V2Themes vs. Themes

The GKG table that has two fields that contain the themes of each article, *Themes* and *V2Themes*. After doing a thorough analysis of both the *Themes* and *V2Themes* fields, we concluded that both produced nearly identical results and choosing one over the other was

not important for this research. The major differences between the *Themes* and *V2Themes* is the *V2Themes* contains the character offset where the theme was identified within the article itself. *V2Themes* also allows for the same theme to appear more than once if the theme itself appears more than once in the article. The *Themes* is only a list of each unique theme as it appears in the event. For this research, I decided to use *V2Themes* instead of *Themes* when filtering by articles involving the environment. However, we could have used *Themes* instead and gotten the same results for this research.

Tones

The Events, Eventmentions, and GKG tables all contain a tones' field: *AvgTone*, *MentionDocTone*, and *V2Tone* respectively. Each of the tones presented in each table are calculated slightly differently from one another. The tone in the Events table, *AvgTone*, only represents the average tone of the all of the articles that reported the event within the first 15 minutes the event was created [3]. The tone in the Eventmentions table, *MentionDocTone*, is calculated the same way as the *AvgTone*. Just for each particular article and not all the articles for a specific event [4]. Finally, the *V2Tone* is calculated by subtracting the positive and negative score^{viii} of the whole document [4].

Only analyzing the *AvgTone* can lead to bias results because those tones are only generated from when the event first appears. As a major story develops, the tones of the event can change. The Eventmentions tone, *MentionDocTone*, and the GKG tone, *V2Tone*, are less susceptible to bias because their tones are generated on an individual article level.

The tones, and positive and negative scores are automatically generated by GDELT using sentiment analysis. The exact algorithms or techniques are not mentioned in the documentation but GDELT promises that it is an advance Natural Language Processing (NLP) algorithm that generates these scores [4][8].

Great Powers

There are two attributes available in the Events table we can use to filter for our great powers, *ActorCountryCodes*^{ix} or *ActorGeo_CountryCode*^x. The *ActorGeo_CountryCode* method uses the FIPS country codes to specify who the actors are, the attribute *ActorCountryCode* uses a 3-character CAMEO^{xi} code for the country affiliations. There are resources online to find the CAMEO codes of each of the great powers we are looking for. Instead however,

^{viii}**Positive Score** is the percentage of all words in the article that were found to have a positive emotional connotation. Ranges from 0 to +100. **Negative Score** is the percentage of all words in the article that were found to have a positive emotional connotation. Ranges from 0 to +100. [4]

^{ix}*ActorCountryCodes* is generalized. The actual names of the attributes are *Actor1CountryCodes* and *Actor2CountryCodes*

^x*ActorCountryCodes* is generalized. The actual names of the attributes are *ActorGeo1_CountryCodes* and *ActorGeo2_CountryCodes*

^{xi}The three digit CAMEO codes used in GDELT for the United States, Australia, China, New Zealand, Japan, and Russia are: *USA*, *AUS*, *CHN*, *NZL*, *JPN*, and *RUS* respectively

I searched the existing data sets using *Pandas* and *Numpy*. This way, I was able to get the exact CAMEO code used for each of the great powers in GDELT.

I decided to use the *ActorCountryCode* instead of *ActorGeo_CountryCode* for filtering the data set to only contain the great powers. I chose this because the results of using *ActorGeo_CountryCode* as a filter produced a data set with less than half of the articles compared to using the *ActorCountryCode*. The Events table in GDELT only specifies two actors and a location. Because of this, if multiple actors are present, GDELT will either leave the field blank or will choose which actors take precedence in the article and assign those actors to those attributes [4]. But the *ActorCountryCode* does not always line up with the *ActorGeo_CountryCode*. For instance, "if the text refers to 'French Assistant Minister Smith was in Moscow', the CountryCode field will list France in the CountryCode field, while the geographic fields may list Moscow as the location" [4]. While only using *ActorCountryCodes* as our filter leads to potentially missing some articles involving certain actors, if we were to look for actors by also including their location, *ActorGeo_CountryCode*, we may be including articles where we have the actors country location but they may not have been an actor in the article itself. We did experiment with specifying the *ActorCountryCode* and *ActorGeo_CountryCode* equivalently matching. But this produced a data set that only contained a fraction of the number of events (ten's of thousands compared to hundred's of thousands) compared to only using *ActorCountryCode* and allowing *ActorGeo_CountryCode* to be any value.

Media Sources

When researching the Local media sources of the South Pacific, I created a function in Python to output every unique media source that has involves a great power as an actor. From there, my research mentor, *Dr. Elizabeth Gooch*, went through the list of sources for each individual location and cross checked each source to verify if it was a Local media source^{xii} to the South Pacific. The Western news sources^{xiii} were selected because they give a range of liberal, conservative, and neutral media outlets. The Chinese news sources^{xiv} were select because each one of those news sources are either state-owned, state-ran, or authorized by the Chinese government.

Calculating Tones

^{xii}**Local media sources:** samoanews.com, samoaoobserver.ws, solomonstarnews.com, solomon-times.com, islandsun.com.sb, saipantribune.com, fijisun.com.fj, fjitimes.com, fijivillage.com, fiji.gov.fj, fiji-one.tv, thejetnewspaper.com, islandsbusiness.com, kpress.info, presidency.pf, ladepeche.pf, guampdn.com, kuam.com, pacificnewscenter.com, mvguam.com, postguam.com, pncguam.com, dailypost.vu, nouvellecaledonie.lalere.fr, postcourier.com.pg, emtv.com.pg, thenational.com.pg, marshallislandsjournal.com, tonga-broadcasting.net, matangitonga.to, tongadailynews.to, parliament.gov.to, mic.gov.to, and planet-tonga.com

^{xiii}**Western news sources:** yahoo.com, bbc.co.uk, foxnews.com, go.com, cbsnews.com, nytimes.com, msn.com, usatoday.com, and npr.org.

^{xiv}**Chinese news sources:** xinhuanet.com, chinadaily.com, china.org.cn, ecns.cn, and peopledail.com.cn

Once we had the geographical locations, the CAMEO codes for each great power, and the proper filtering methods for creating subsets of data, I then created multiple Python functions to compute the average weighted tone for each news source in the South Pacific [Figure 3].

Results

When comparing each of the three media sources (Chinese, Local, and Western) side-by-side, Chinese news sources have an average tone of around 3.2, Local news sources have an average tone of around 0.7, and Western news sources have an average tone of 1.8 to 1.9. The tones in GDELT can have a score of up to 100 or as low as -100, but most of the tone scores are between +10 and -10 (where -10 is an article written with a very negative sentiment, +10 is written with a very positive sentiment, and 0 indicating neutral) [Figure 4] [4]. Even throughout our experimentation's, we rarely saw any tones go above 10 or below -10.

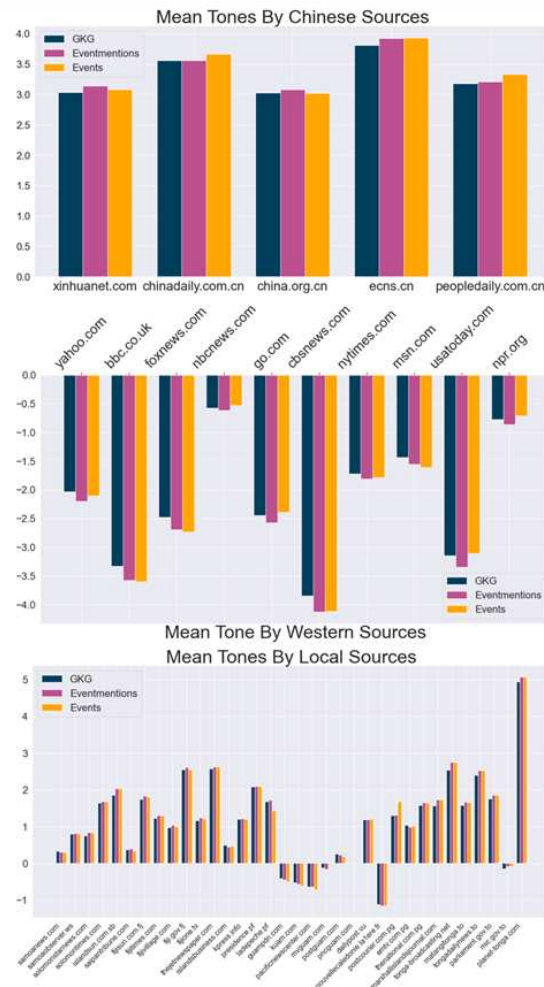


FIGURE 3: Tones by Sources

Statistical Analysis

In order to prove each of the population (tone) means are statistically significance, I conducted a difference in means tests for each of our three sources. I conducted the statistical analysis using R instead of Python. I did this for a few reasons. One, R has better interpretability with the results. In Python, similar functions used to conducted statistical analysis using the *SciPy* library usually produce only single number results. To show the results in a presentable manor, I would need to create additional functions with print statements explaining the results of each test. Instead of doing this, I decided R would be the better choice. Two, the statistical visualizations produced in R using *ggplot* and *base R* produce interpretable graphics with minimal amount of coding compared to Python. I still however used Python for all the data aggregation, pre-processing data, and for the pie and bar plots. But I imported the same data into R for the statistical visualizations and analysis.

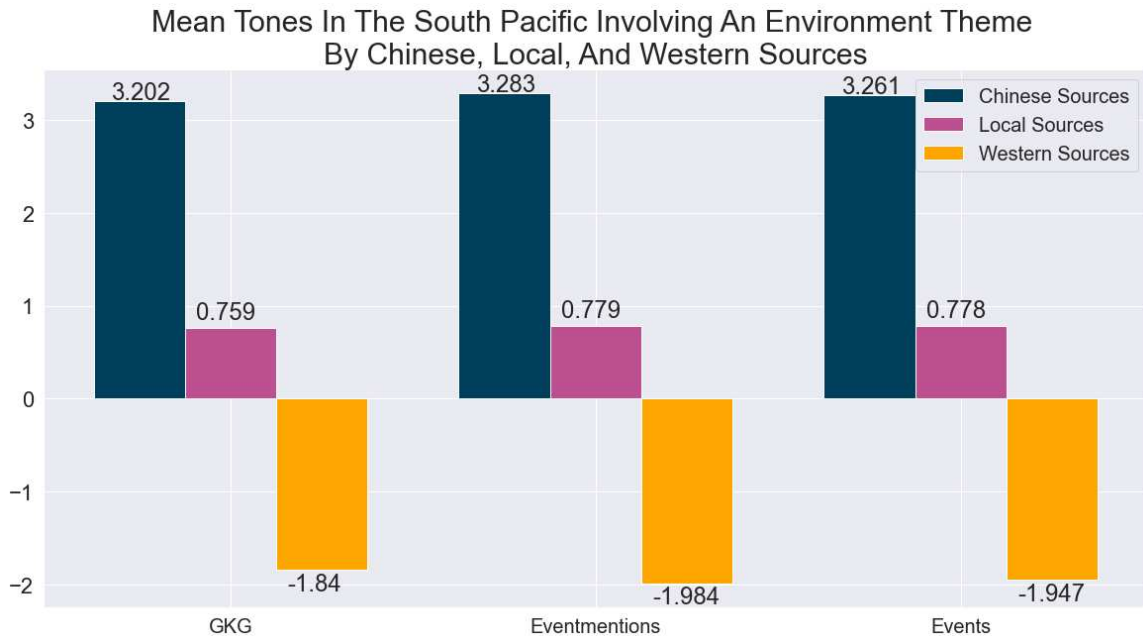


FIGURE 4: Mean Tones By Source

In order to prepare the data for statistical analysis using R, I filtered the data set one more time to only contain the *GKG_Tone* and added a new field to indicated if the *Source* was a Chinese source, Local source, or Western source.

Difference in Means

In order to properly conduct a difference in means test, we must acknowledge the assumptions to conduct this statistical analysis. First, we will assume that each observation from each of the groups are independent of each other group. Next, I checked the normality by creating three distribution plots and a Q-Q plot [Figure 6, 5]. This showed that the data overall follows a normal distribution^{xv}. Finally, when checking for equal variances in the Residuals vs Fitted plot, the variances appeared to not be equal among the three groups [Figure 6]. Because of this inconsistency, I conducted a Levene Test to test the homogeneity of variances of the residuals.

LISTING 5: Levene's Test for Homogeneity of Variance

```
> leveneTest(GKG_Tone ~ Source, data=tones)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value    Pr(>F)
group  2  7.5008 0.0005548 ***
14650
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

^{xv}Because the number of observations is large (Chinese = 1,013, Local = 10,292, Western = 3,348), violating the normality assumption is okay.

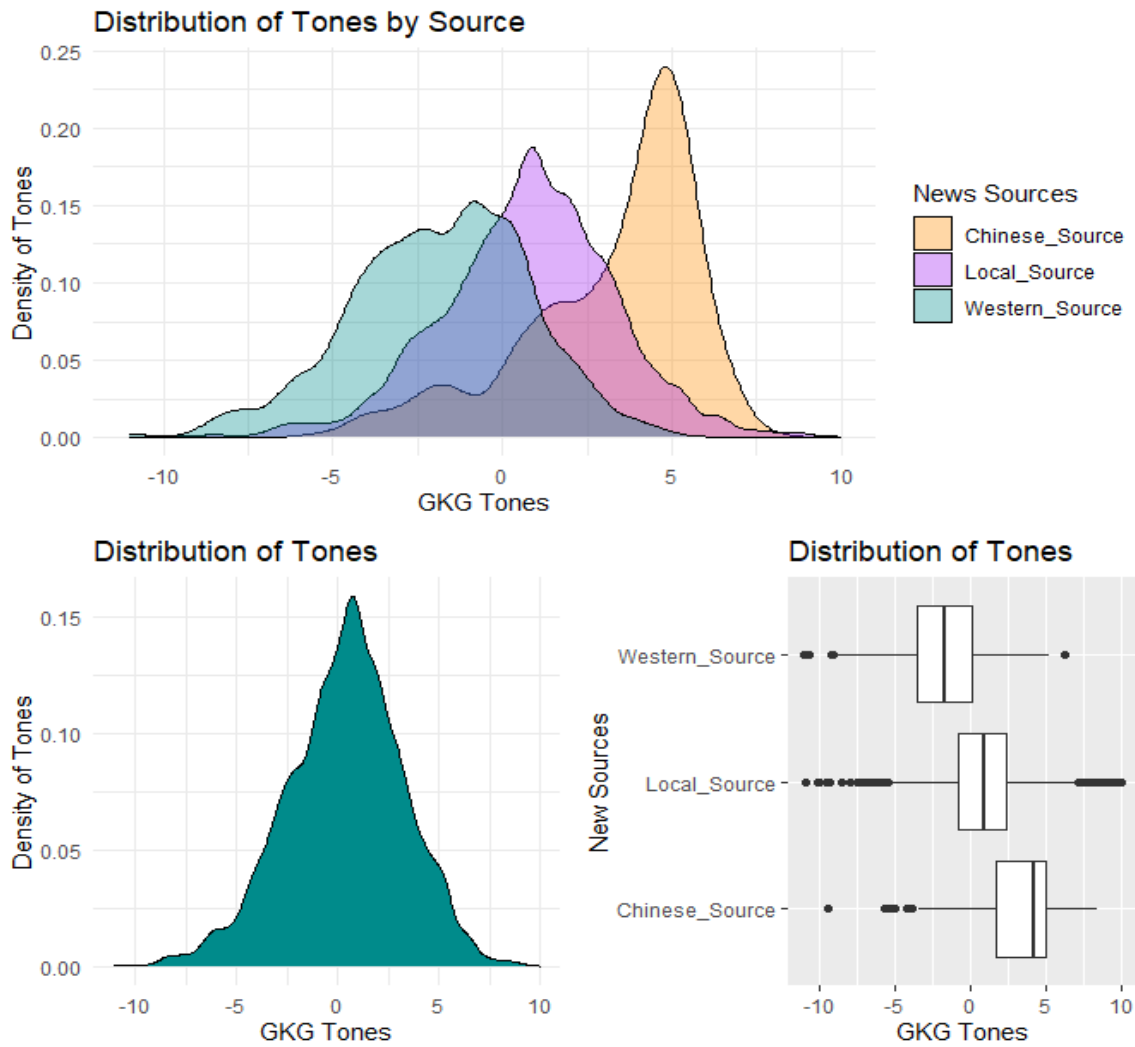


FIGURE 5: Normality of Tones

To interpret the results of the Levene’s Test for Homogeneity of Variance, at a confidence level of $\alpha = 0.05$ and p-value of 0.0005 we can conclude the population variances are not equal. Because the population variances are not equal, we have to conduct three Welch’s t-test’s for each of the pairwise groups to prove statistical significance.

Interpreting Welch’s two-sample t-test

After conducting these tests in R, the results showed that the difference in means between each pairwise group is statistically significant [Table1]. To interpret these results: at a significance level of $\alpha = 0.05$, we fail to reject the null between each of the three pairwise groups and prove that the difference in tones observed is statistically significant between each pairwise group.

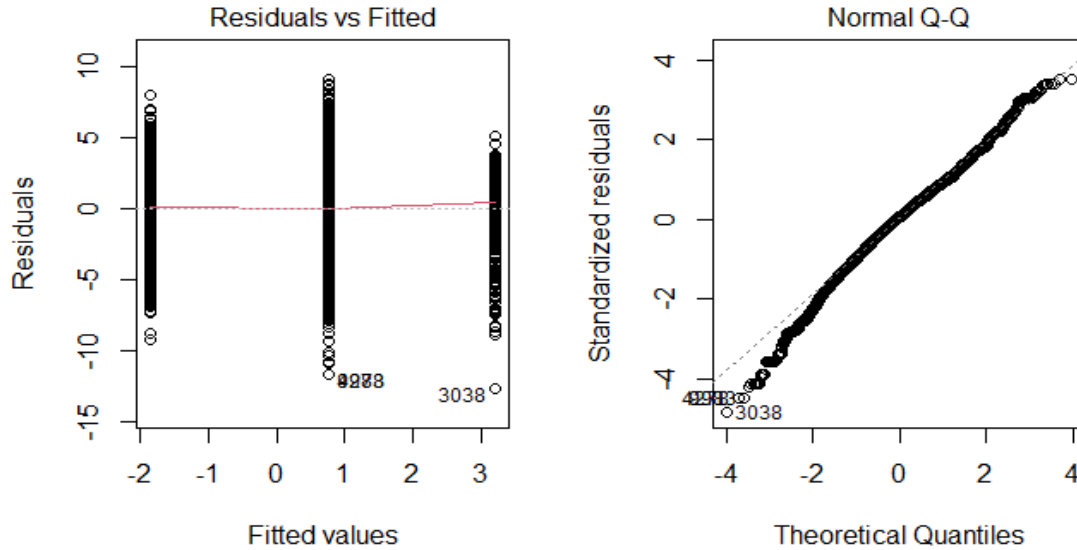


FIGURE 6: Normality and Variance of Tones

Pairwise Groups	T-statistic	Degrees of freedom	P-value
Local - Chinese	-28.536	1218.7	2.2e-16
Western - Chinese	-53.947	1684.2	2.2e-16
Western - Local	-49.899	5624.3	2.2e-16

TABLE 1: Results of Welch's two-sample t-test

Conclusion

In this report, I introduce GDELT and how GDELT can be used to track events in regions of the world. I talked about three of the tables (Events, Eventmentions, GKG) we used in this research, how they are related to one another, and which attributes we used for this research. I showed how we found the names of actors, great powers, and locations of interests within GDELT by using Python, SQL, and Wikipedia. I mentioned which countries and territories we included in our region of the South Pacific and some that we did not. I showed all the methods I used to develop our final data set. I gave explanations on which attributes I decided to chose for our filtering and how I used Python for all of my data wrangling, exploration, and most of the visualizations. I explained why we chose our media sources and how I calculated the average tones for each source. I then showed that Chinese sources have an overall very positive tone, Local news sources have a positive to neutral tone, and Western sources have a negative tone in the South Pacific when the event itself involves the environment and a great power an actor. And finally, I used Welch's two-sample t-test (using R) to show the mean tones of Chinese, Local, and Western new sources are statistically different.

Acknowledgment

I would like to thank Dr. Elizabeth Gooch from the Naval Postgraduate School for mentoring me and the Undergraduate Research Opportunity Center (UROC) Program at California State University, Monterey Bay for providing me a scholarship for this research.

References

- [1] <https://www.gdeltproject.org/about.html>
- [2] <https://www.gdeltproject.org/>
- [3] http://data.gdeltproject.org/documentation/GDELGlobal_Knowledge_Graph_Codebook-V2.1.pdf
- [4] http://data.gdeltproject.org/documentation/GDEL-Event_Codebook-V2.0.pdf
- [5] <https://fas.org/sgp/crs/natsec/R43838.pdf>
- [6] <https://www.frontiersin.org/articles/10.3389/fmars.2019.00289/full>
- [7] <https://en.wikipedia.org/wiki/Oceania>
- [8] <https://blog.gdeltproject.org/introducing-gkg-2-0-the-next-generation-of-the-gdelt-global-knowledge-graph/>
- [9] <https://www.worldbank.org/>