Untitled - TextEdit

File Edit View Help

Every problem has a solution,
it may sometimes just need another
PERSPECTIVE.

PRESIDENTIAL COMMUNICATIONS OFFICE

GROUP 7'S HACKATHON OUTPUT

PROBLEM

How do different sources (MSM and SM) portray the Presidential Communications Office?

01

Which are the top 5 sources by number of stories?

02

Which country has the most positive sentiment stories?

03

What's the most common keyword/topic?

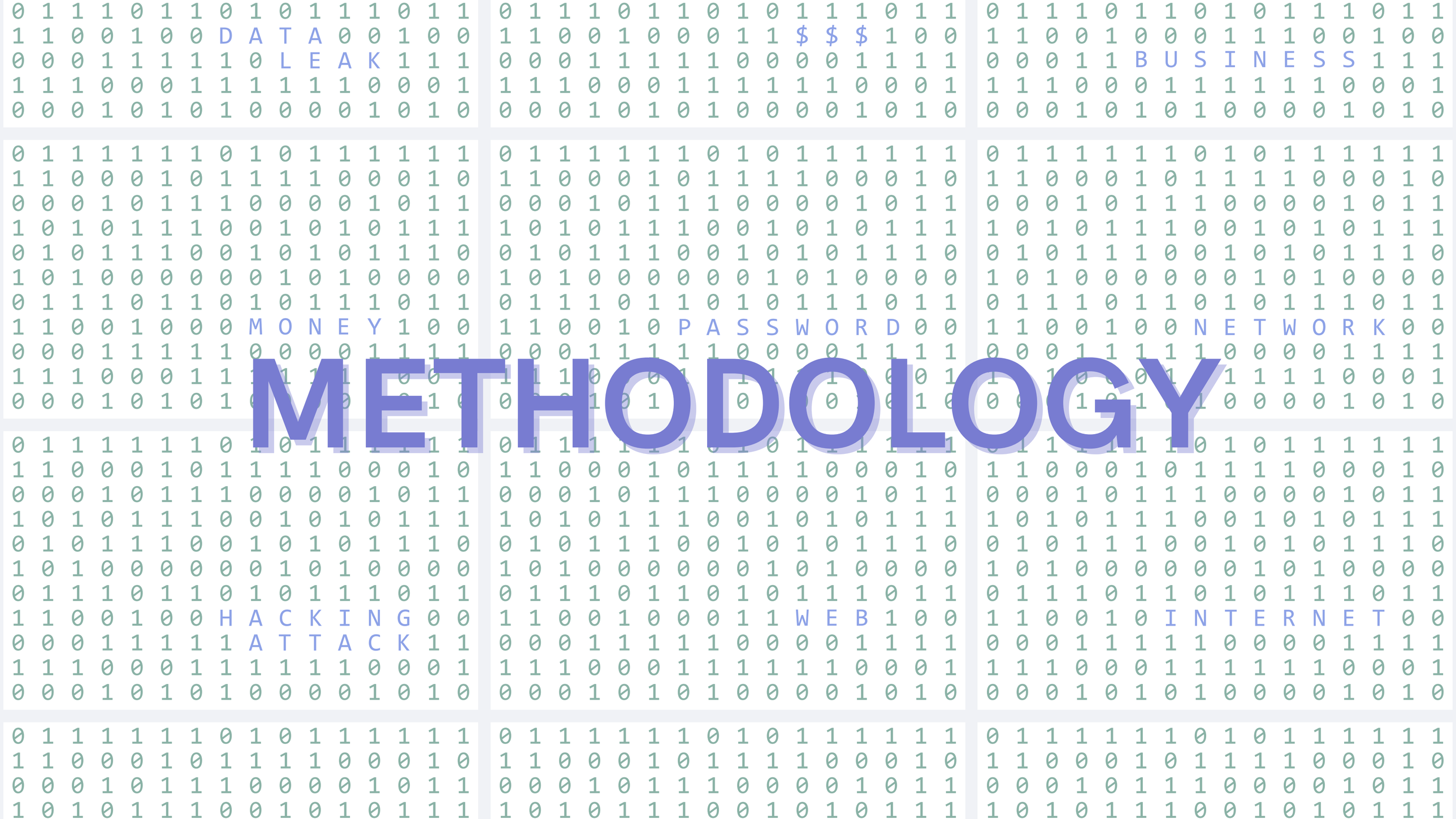
04

Which week had the most stories?

05

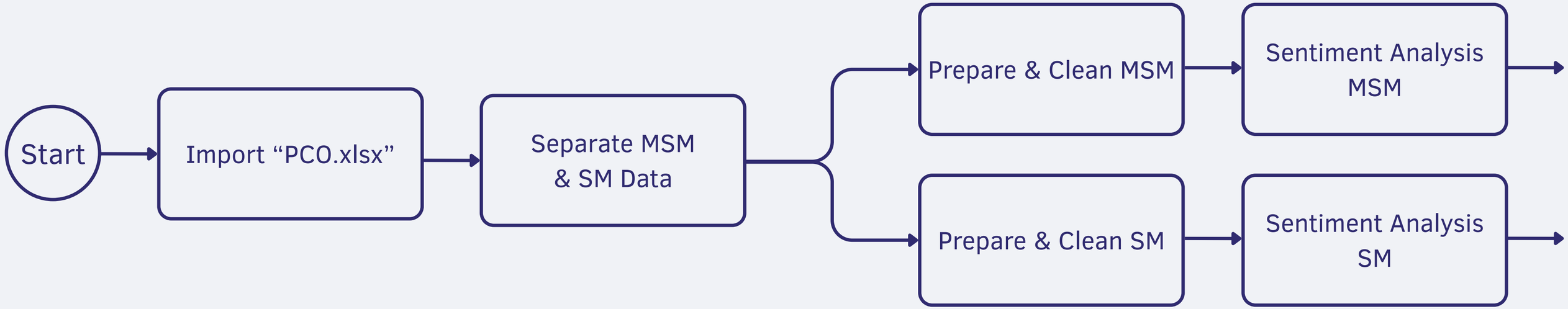
Which source has the highest average reach?

MSM vs SM


METHODOLOGY

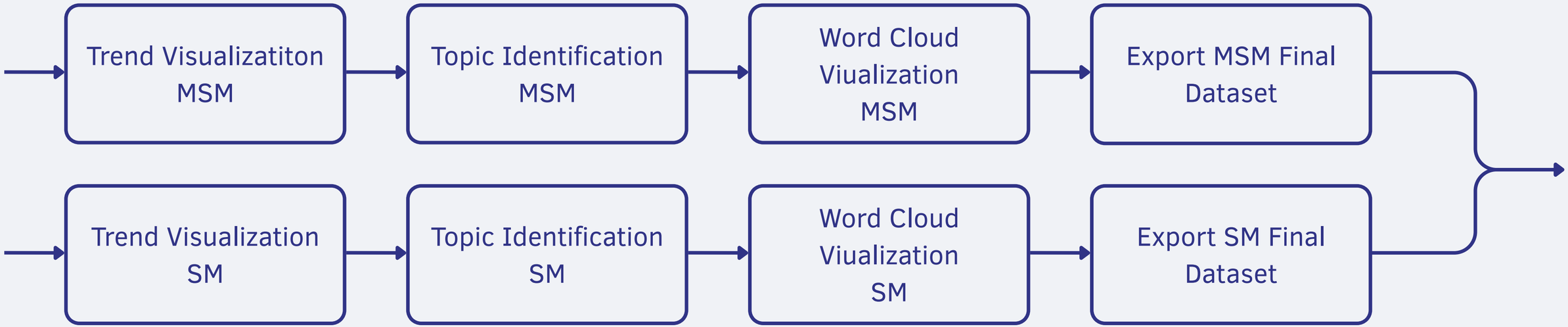
METHODOLOGY

Data Analysis Process



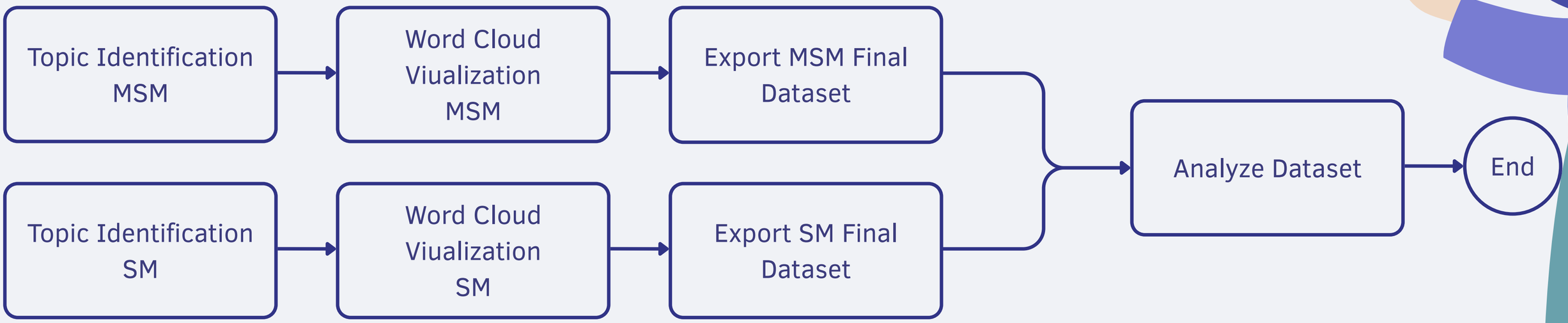
METHODOLOGY

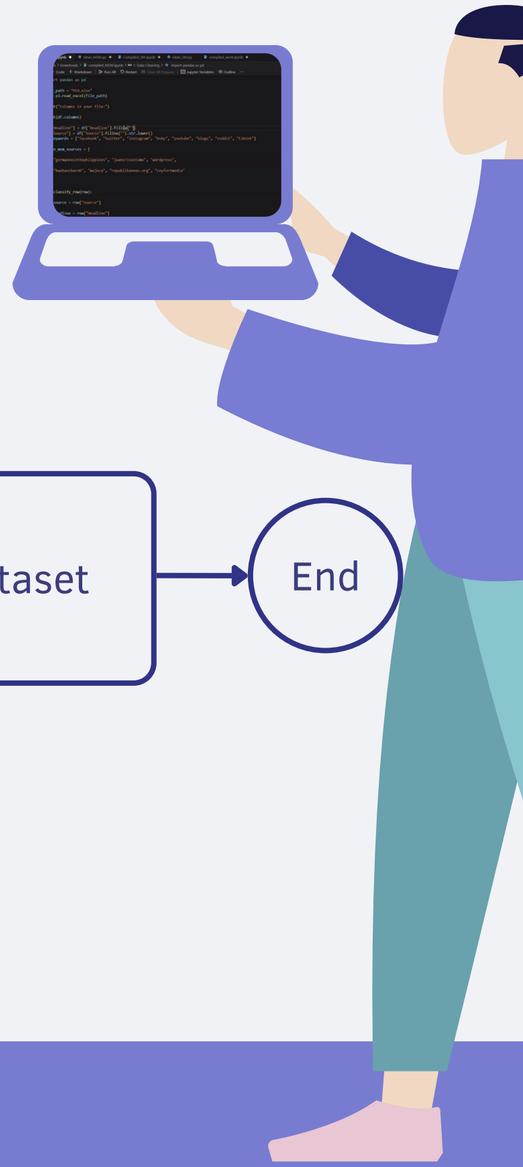
Data Analysis Process



METHODOLOGY

Data Analysis Process

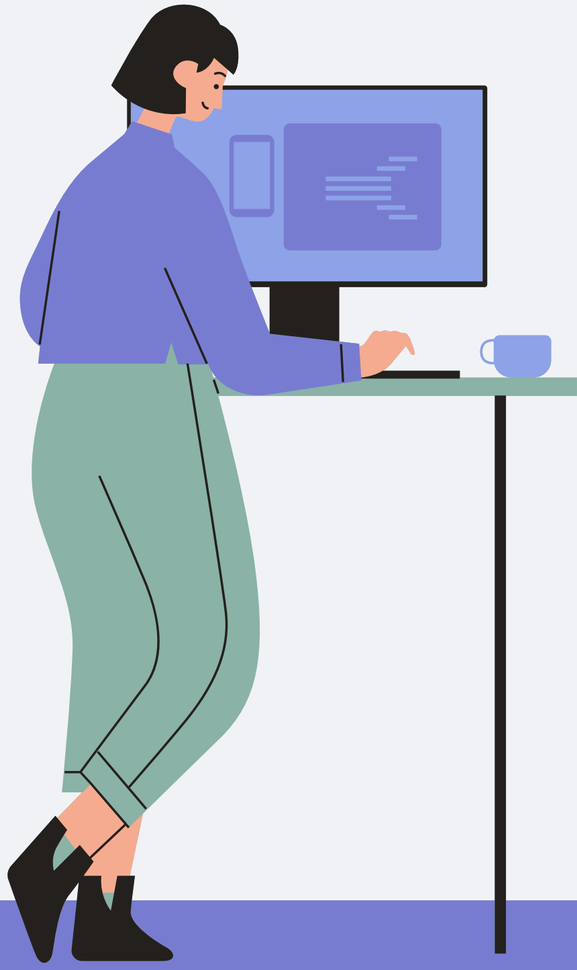




METHODOLOGY

Limitations

- Results depend on **data quality & completeness**
- **Sample bias** possible in source data
- **Rule-based cleaning** cannot resolve all anomalies
- **Library constraints** (performance & interpretability limits)
- Automated checks **may miss context-specific errors**

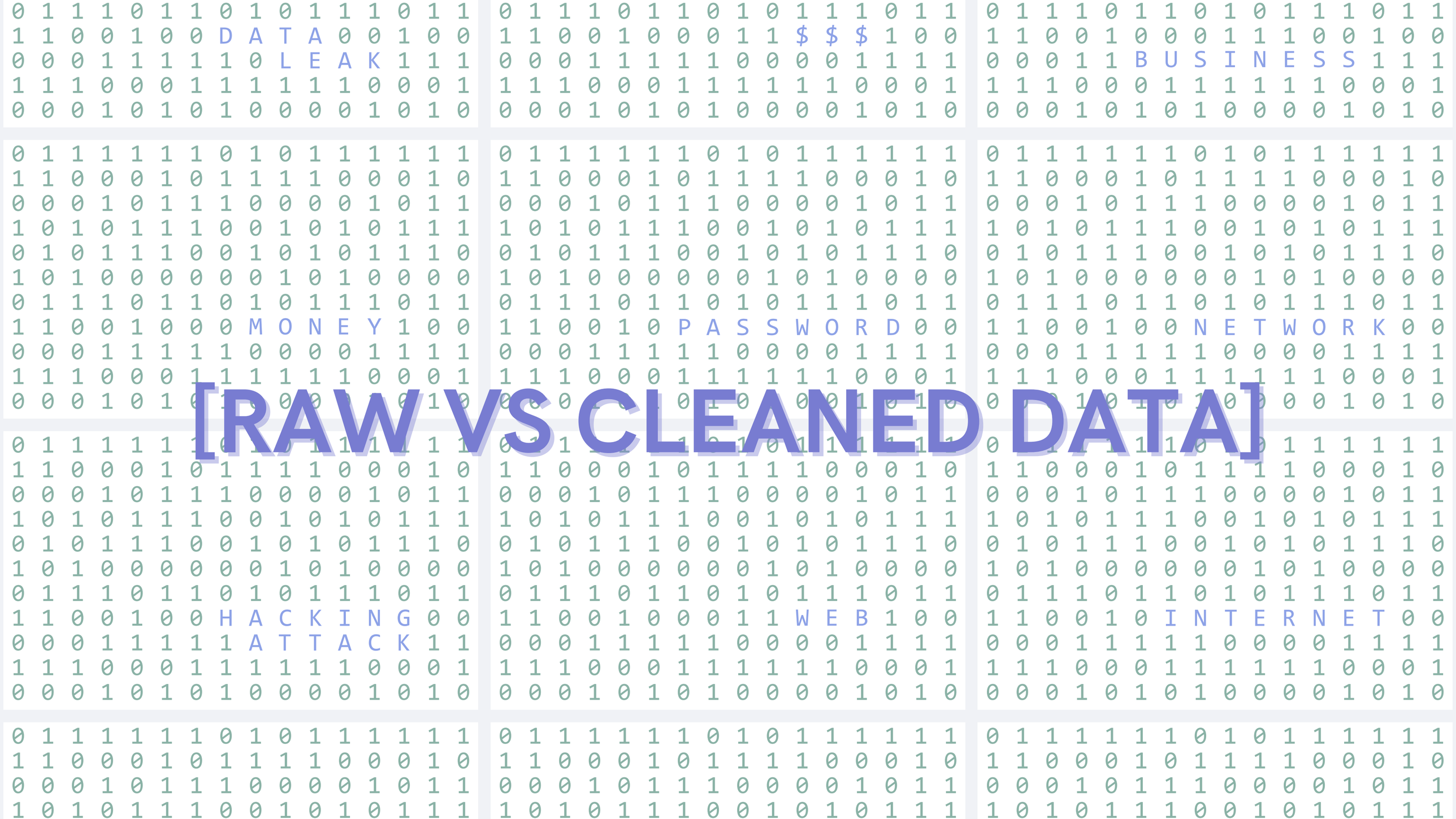


METHODOLOGY

Caveats

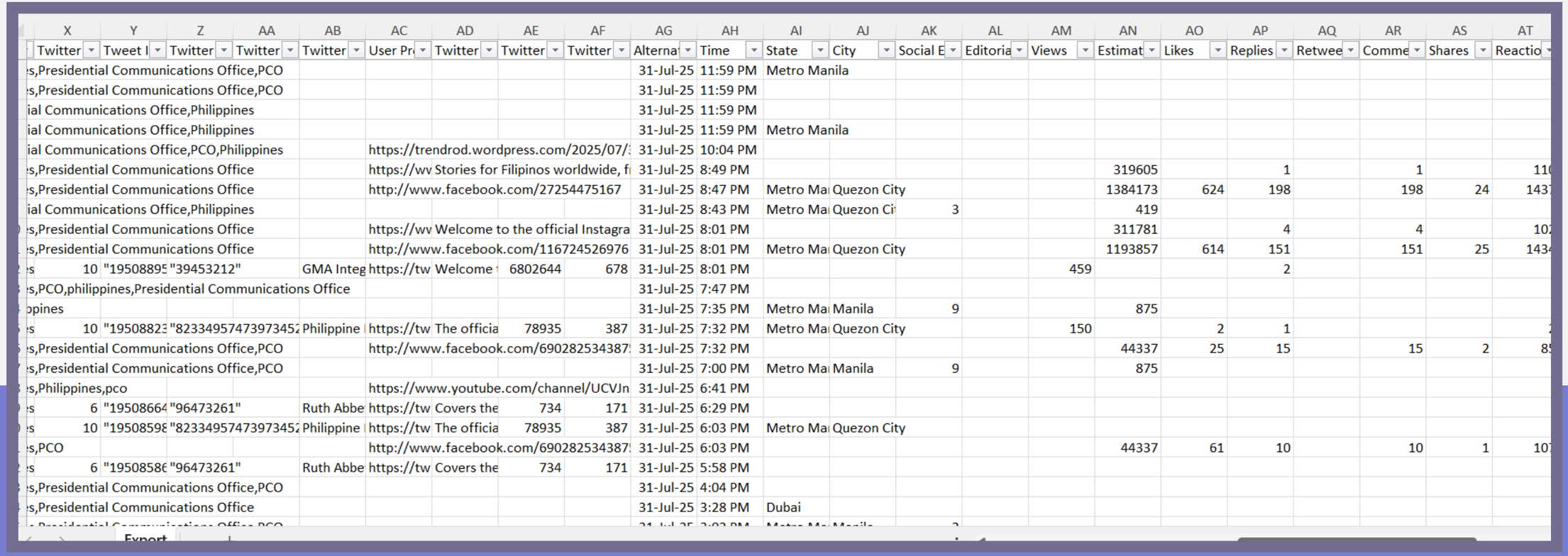
- Some records may remain **incomplete or duplicated**
- **Time sensitivity:** datasets are snapshots, not live updates
- **Textual ambiguity** in headlines/keywords affects sentiment scoring
- Scaling up may face **performance or memory constraints**


DATA


[RAW VS CLEANED DATA]

Raw Data

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
Date	Headlin	Source	Opening	Hit Sent	Source	Influenc	Country	Subregi	Languag	Reach	Desktop	Mobile	Twitter	Faceboo	Reddit S	Nationa	Engagem	AVE	Sentime	Key Phr	Input N	Keywor
31-Jul-202	PBBM Ack	https://wv	Manila: Pr	... Jr. on Th	The Philippines News		Philippines		English	0	0	0	0	0	0			0	Neutral	agency,am	Philippines	Philippines,P
31-Jul-202	PBBM Ack	https://phi	Manila: Pr	... Jr. on Th	Philippines News Dige		Philippines		English	0	0	0	0	0	0			0	Neutral	agency,am	Philippines	Philippines,P
31-Jul-202	PBBM's Su	https://phi	Manila: Th	... palay (u	Philippines News Dige		Philippines		English	0	0	0	0	0	0			0	Positive	agricultura	Philippines	Presidential
31-Jul-202	PBBM's Su	https://wv	Manila: Th	... palay (u	The Philippines News		Philippines		English	0	0	0	0	0	0			0	Positive	agricultura	Philippines	Presidential
31-Jul-202	President I	https://tre	Strengther	nations. Pe	Blogs	trendrod.v	Unknown		English									0	Neutral	75th anniv	Philippines	Presidential
31-Jul-2025 08:49PM	https://wv	LOOK: Pres	LOOK: Pres	Instagram	@abscbnn	Unknown			English	2130705							111	19709.02	Neutral	cooperatic	Philippines	Philippines,P
31-Jul-2025 08:47PM	https://wv	LOOK: Pres	LOOK: Pres	Facebook	abscbnnev	Philippines			English	19773913							1659	182908.7	Neutral	cooperatic	Philippines	Philippines,P
31-Jul-202	Marcos ur	https://wv	NEWLY IN	... G. Nafa	Business World Onlin		Philippines		English	309551	82167	227384	3	0	0		3	2863.35	Neutral	army,chan	Philippines	Presidential
31-Jul-2025 08:01PM	https://wv	President I	President I	Instagram	@gmanew	Unknown			English	2078540							106	19226.5	Neutral	cooperatic	Philippines	Philippines,P
31-Jul-2025 08:01PM	https://wv	President I	President I	Facebook	gmanews	Philippines			English	17055106							1610	157759.7	Neutral	cooperatic	Philippines	Philippines,P
31-Jul-2025 08:01PM	https://tw	President I	President I	Twitter	@gmanew	Philippines			English	6802644							2	62924.46	Neutral	cooperatic	Philippines	Philippines
31-Jul-202	Lies, Trolls	http://drjo	... * Dr. Ro	& Hootsuit	Blogs	noreply@E	Unknown		English									0	Negative	adversaria	Philippines	Philippines,P
31-Jul-202	Marcos ta	https://mb	President I	... again sc	Manila Bulletin		Philippines		English	1085043	225993	859050	3	6	0		9	10036.65	Neutral	cooperatic	Philippines	PCO,Philippir
31-Jul-2025 07:32PM	https://tw	LOOK: Pres	R. Marcos	Twitter	@pnagovp	Philippines			English	78935							3	730.15	Neutral	cooperatic	Philippines	Philippines
31-Jul-2025 07:32PM	https://wv	LOOK: Pres	LOOK: Pres	Facebook	pnagovph	Unknown			English	633398							102	5858.93	Neutral	cooperatic	Philippines	Philippines,P
31-Jul-202	Marcos th	https://mb	President I	... during t	Manila Bulletin		Philippines		English	1085043	225993	859050	3	6	0		9	10036.65	Positive	contributic	Philippines	Philippines,P
31-Jul-202	Did Dutert	https://wv	From 2016	. Was the f	Youtube	Civic Storn	Unknown											0			Philippines	philippines,P
31-Jul-2025 06:29PM	https://tw	LOOK: Pres	Ferdinand	Twitter	@ruthabb	Unknown			English	734								6.79	Neutral	cooperatic	Philippines	Philippines
31-Jul-2025 06:03PM	https://tw	LOOK: Pres	LOOK: Pres	Twitter	@pnagovp	Philippines			English	78935								730.15	Neutral	outgoing is	Philippines	Philippines
31-Jul-2025 06:03PM	https://wv	LOOK: Pres	LOOK: Pres	Facebook	pnagovph	Unknown			English	633398							118	5858.93	Neutral	outgoing is	Philippines	Philippines,P
31-Jul-2025 05:58PM	https://tw	LOOK: Pres	LOOK: Pres	Twitter	@ruthabb	Unknown			English	734								6.79	Neutral	outgoing is	Philippines	Philippines
31-Jul-202	Marcos to	https://wv	Marcos to	... visit aim	MSN Philip	Catherine	Philippines		English	244913	113119	131794	0	0	0			2265.45	Neutral	event,filipi	Philippines	Philippines,P
31-Jul-202	President I	https://the	President I	... of India	TheGlobalF	ilipinoMag	United Arab Emirates		English	54769	4350	50419	0	0	0			506.61	Neutral	filipino cor	Philippines	Philippines,P

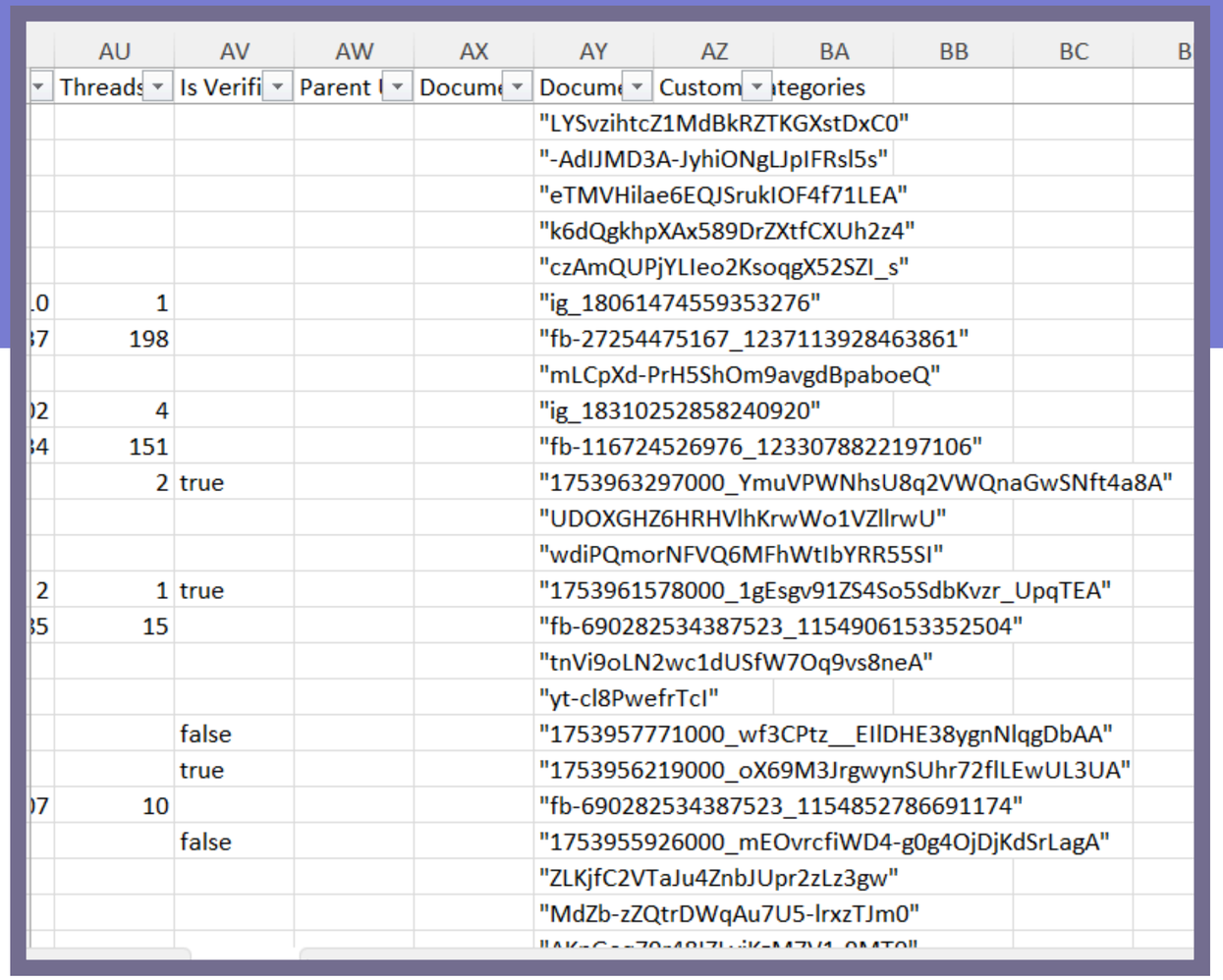


X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT
Twitter	Tweet I	Twitter	Twitter	Twitter	User Pr	Twitter	Twitter	Twitter	Alterna	Time	State	City	Social E	Editoria	Views	Estimat	Likes	Replies	Retwee	Comme	Shares	Reactio
s,Presidential Communications Office,PCO									31-Jul-25	11:59 PM	Metro Manila											
s,Presidential Communications Office,PCO									31-Jul-25	11:59 PM												
ial Communications Office,Philippines									31-Jul-25	11:59 PM												
ial Communications Office,Philippines									31-Jul-25	11:59 PM	Metro Manila											
ial Communications Office,PCO,Philippines					https://trendrod.wordpress.com/2025/07/				31-Jul-25	10:04 PM												
s,Presidential Communications Office					https://wv	Stories for Filipinos worldwide, f			31-Jul-25	8:49 PM						319605		1		1		11
s,Presidential Communications Office					http://www.facebook.com/27254475167				31-Jul-25	8:47 PM	Metro Ma	Quezon City				1384173	624	198		198	24	143
ial Communications Office,Philippines									31-Jul-25	8:43 PM	Metro Ma	Quezon Ci	3			419						
s,Presidential Communications Office					https://wv	Welcome to the official Instagra			31-Jul-25	8:01 PM						311781		4		4		10
s,Presidential Communications Office					http://www.facebook.com/116724526976				31-Jul-25	8:01 PM	Metro Ma	Quezon City				1193857	614	151		151	25	143
s	10	"19508895	"39453212"	GMA Integ	https://tw	Welcome	6802644	678	31-Jul-25	8:01 PM					459			2				
s,PCO,philippines,Presidential Communications Office									31-Jul-25	7:47 PM												
ppines									31-Jul-25	7:35 PM	Metro Ma	Manila	9			875						
s	10	"19508823	"82334957473973452	Philippine	https://tw	The officia	78935	387	31-Jul-25	7:32 PM	Metro Ma	Quezon City			150		2	1				
s,Presidential Communications Office,PCO					http://www.facebook.com/690282534387				31-Jul-25	7:32 PM						44337	25	15		15	2	8
s,Presidential Communications Office,PCO									31-Jul-25	7:00 PM	Metro Ma	Manila	9			875						
s,Philippines,pco					https://www.youtube.com/channel/UCVJn				31-Jul-25	6:41 PM												
s	6	"19508664	"96473261"	Ruth Abbe	https://tw	Covers the	734	171	31-Jul-25	6:29 PM												
s	10	"19508598	"82334957473973452	Philippine	https://tw	The officia	78935	387	31-Jul-25	6:03 PM	Metro Ma	Quezon City										
s,PCO					http://www.facebook.com/690282534387				31-Jul-25	6:03 PM						44337	61	10		10	1	10
s	6	"19508586	"96473261"	Ruth Abbe	https://tw	Covers the	734	171	31-Jul-25	5:58 PM												
s,Presidential Communications Office,PCO									31-Jul-25	4:04 PM												
s,Presidential Communications Office									31-Jul-25	3:28 PM	Dubai											

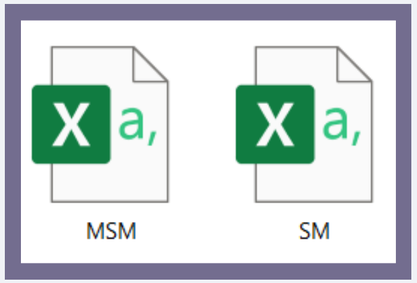
Export

Raw Data

Raw data



	AU	AV	AW	AX	AY	AZ	BA	BB	BC	B
	Threads	Is Verifi	Parent I	Docume	Docume	Custom	tegories			
					"LYSvzihtcZ1MdBkRZTKGXstDxC0"					
					"-AdIJMD3A-JyhiONgLJpIFRsl5s"					
					"eTMVHilae6EQJSrukIOF4f71LEA"					
					"k6dQgkhpXAx589DrZXtfCXUh2z4"					
					"czAmQUPjYLIeo2KsoqgX52SZI_s"					
.0	1				"ig_18061474559353276"					
37	198				"fb-27254475167_1237113928463861"					
					"mLCpXd-PrH5ShOm9avgdBpaboeQ"					
02	4				"ig_18310252858240920"					
34	151				"fb-116724526976_1233078822197106"					
	2	true			"1753963297000_YmuVPWNhsU8q2VWQnaGwSNft4a8A"					
					"UDOXGHZ6HRHVlhKrwWo1VZllrwU"					
					"wdiPQmorNFVQ6MFhWtIbYRR55SI"					
2	1	true			"1753961578000_1gEsgv91ZS4So5SdbKvzr_UpqTEA"					
35	15				"fb-690282534387523_1154906153352504"					
					"tnVi9oLN2wc1dUSfW7Oq9vs8neA"					
					"yt-cl8PwefrTcI"					
		false			"1753957771000_wf3CPtz__EIlDHE38ygnNlqgDbAA"					
		true			"1753956219000_oX69M3JrgwynSUhr72flLEwUL3UA"					
07	10				"fb-690282534387523_1154852786691174"					
		false			"1753955926000_mEOvrcfiWD4-g0g4OjDjKdSrLagA"					
					"ZLKjfC2VTaJu4ZnbJUpr2zLz3gw"					
					"MdZb-zZQtrDWqAu7U5-lrxzTJm0"					



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Clean Date	Date	Headline	Source Link	pening Te	it Sentenc	Source	Influencer	Country	Language	Reach	ter Social	ook Social	dit Social E	Sentiment	Key Phrase	nput Name	Keywords	ial Echo To	Category
2	07-31-202	31-Jul-202	President I	https://tre	Strengther	Nations. p	Blogs	trendrod.v	Unknown	English	0	0	0	0	Neutral	75th anniv	Philippines	Presidentia	0	SM
3	07-31-202	31-Jul-202	Look: Pres	https://wv	Look: pres	Look: pres	Instagram	@abscbnn	Unknown	English	2130705	0	0	0	Neutral	cooperatic	Philippines	Philippines	0	SM
4	07-31-202	31-Jul-202	President I	https://wv	President I	President I	Instagram	@gmanew	Unknown	English	2078540	0	0	0	Neutral	cooperatic	Philippines	Philippines	0	SM
5	07-31-202	31-Jul-202	President I	https://tw	President I	President I	Twitter	@gmanew	Philippines	English	6802644	0	0	0	Neutral	cooperatic	Philippines	Philippines	0	SM
6	07-31-202	31-Jul-202	Lies, Trolls	http://drjc	... * dr. roc	& hootsuit	Blogs	noreply@b	Unknown	English	0	0	0	0	Negative	adversaria	Philippines	Philippines	0	SM
7	07-31-202	31-Jul-202	Look: Pres	https://tw	Look: pres	R. marcos	Twitter	@pnagovp	Philippines	English	78935	0	0	0	Neutral	cooperatic	Philippines	Philippines	0	SM
8	07-31-202	31-Jul-202	Look: Pres	https://wv	Look: pres	Look: pres	Facebook	pnagovph	Unknown	English	633398	0	0	0	Neutral	cooperatic	Philippines	Philippines	0	SM
9	07-31-202	31-Jul-202	Did Dutert	https://wv	From 2016	. was the f	Youtube	Civic Storn	Unknown	Unknown	0	0	0	0	Unknown	Unknown	Philippines	philippines	0	SM
10	07-31-202	31-Jul-202	Look: Pres	https://tw	Look: pres	Look: pres	Twitter	@pnagovp	Philippines	English	78935	0	0	0	Neutral	outgoing is	Philippines	Philippines	0	SM
11	07-31-202	31-Jul-202	Look: Pres	https://wv	Look: pres	Look: pres	Facebook	pnagovph	Unknown	English	633398	0	0	0	Neutral	outgoing is	Philippines	Philippines	0	SM
12	07-31-202	31-Jul-202	Pbbm, Bibi	https://tw	Pbbm, bibi	Sa new del	Twitter	@pilipinas	Philippines	Tagalog	4103	0	0	0	Neutral	august,cou	Philippines	Philippines	0	SM
13	07-31-202	31-Jul-202	Just In: Pre	https://wv	Just in: pre	Aug. 4 to 8	Facebook	onenewsp	Philippines	English	421527	0	0	0	Neutral	75th anniv	Philippines	Presidentia	0	SM
14	07-31-202	31-Jul-202	Isasama N	http://ww	Senate of	Senate of	Facebook	untvnewsr	Unknown	English	0	0	0	0	Neutral	Unknown	Philippines	Philippines	0	SM
15	07-31-202	31-Jul-202	I Strongly I	https://wv	I strongly l	I strongly l	Youtube	Comment	Unknown	Unknown	0	0	0	0	Unknown	Unknown	Philippines	Philippines	0	SM
16	07-30-202	30-Jul-202	Philippines	https://wv	Sona 2025	Sona 2025	Youtube	TIMES NO	Unknown	Unknown	0	0	0	0	Unknown	Unknown	Philippines	Philippines	0	SM
17	07-30-202	30-Jul-202	Should Kei	https://wv	Its still aliv	They have	Forums	ill_young_	Unknown	English	76469	0	0	0	Neutral	governmer	Philippines	pco,Philipp	0	SM
18	07-30-202	30-Jul-202	Phillippine	http://ww	Yuck chea	Of foreign	Facebook	sunstarda	Philippines	English	0	0	0	0	Neutral	minded,pn	Philippines	Philippines	0	SM
19	07-29-202	29-Jul-202	Pinanguna	https://tw	Pinanguna	Ni pangulo	Twitter	@net25tv	Philippines	Tagalog	32781	0	0	0	Neutral	taong	Philippines	Philippines	0	SM
20	07-29-202	29-Jul-202	Philippines	https://wv	Sona 2025	The secon	Youtube	CNBC-TV1	Unknown	Unknown	0	0	0	0	Unknown	Unknown	Philippines	philippines	0	SM
21	07-29-202	29-Jul-202	Patients O	http://ww	Jeralyn lari	Jeralyn lari	Facebook	philippines	Philippines	English	0	0	0	0	Neutral	doh hospit	Philippines	PCO,Philip	0	SM
22	07-29-202	29-Jul-202	A Closer Lo	https://jua	By juan cri	Tesda ann	Blogs	indioako	Unknown	English	0	0	0	0	Neutral	administra	Philippines	Presidentia	0	SM
23	07-29-202	29-Jul-202	Rt @Manil	https://tw	Rt @manil	Rt @manil	Twitter	@godismy	United Sta	English	142	0	0	0	Neutral	economic	Philippines	PCO,Philip	0	SM
24	07-28-202	28-Jul-202	Philippines	https://wv	Philippines	Philippines	Youtube	CNBC Awa	Unknown	Unknown	0	0	0	0	Unknown	Unknown	Philippines	philippines	0	SM

Data Cleaning


DATA CLEANING
[SEPARATING THE DATA]

```python
import pandas as pd

file_path = "PCO.xlsx"
df = pd.read_excel(file_path)

print("Columns in your file:")

print(df.columns)
```

Importing necessary library and importing the file

```python
df["Headline"] = df["Headline"].fillna("")
df["Source"] = df["Source"].fillna("").str.lower()
sm_keywords = ["facebook", "twitter", "instagram", "bsky", "youtube", "blogs", "reddit", "tiktok"]

force_msm_sources = [

    "germanexinthephilppines", "juancrisostomo", "wordpress",

    "kwebanibarok", "majeca", "republikanews.org", "reyfortmedia"
```

Replacing the missing values Headline and Source column

```python
def classify_row(row):
    source = row["Source"]
    headline = row["Headline"]
    if any(keyword in source for keyword in force_msm_sources):
        return "MSM"
    if headline.strip() == "":
        return "SM"
    if any(keyword in source for keyword in sm_keywords):
        return "SM"
    return "MSM"
```

Conditional code to identify which data category is appropriate: SM or MSM

```python
df["Category"] = df.apply(classify_row, axis=1)

msm_df = df[df["Category"] == "MSM"]

sm_df = df[df["Category"] == "SM"]
```

Creating new column category name for MSM and SM

```
msm_df.to_csv("MSM.csv", index=False, encoding="utf-8-sig")

sm_df.to_csv("SM.csv", index=False, encoding="utf-8-sig")
```

Exporting the two separated files into SM and MSM CSV files


DATA CLEANING
[MSM]

```python
import pandas as pd
import re
import string

file_path = "MSM.csv"
df = pd.read_csv(file_path)
```

Importing necessary libraries and importing the MSM CSV file

```
df_clean = df.drop_duplicates(subset=["Source Link"], keep="first")
df_clean = df_clean.dropna(subset=["Date", "Headline", "Source", "Country"])
```

Dropping the null values or missing values

```python
def parse_date(row):
    date_str = row["Date"] if pd.notnull(row["Date"]) else row.get("Alternate Date Format", None)
    try:
        parsed = pd.to_datetime(date_str, errors="coerce")
        return parsed.date() if pd.notnull(parsed) else pd.NaT
    except:
        return pd.NaT

df_clean.loc[:, "Clean Date"] = df_clean.apply(parse_date, axis=1)
df_clean = df_clean.dropna(subset=["Clean Date"])
df_clean.loc[:, "Clean Date"] = pd.to_datetime(df_clean["Clean Date"]).dt.strftime("%m-%d-%Y")
```

Standardization of date format

```
def normalize_source(source):
    s = str(source).strip().lower()

    s = re.sub(r"www\.", "", s)
    s = re.sub(r"\.com|\.ph|\.net|\.org", "", s)
    s = re.sub(r"\s+", " ", s)

    words = s.split()
    result_words = []
    for w in words:
        if w == "abs":
            result_words.append("ABS")
        elif w == "cbn":
            result_words.append("CBN")
        elif w == "msn":
            result_words.append("MSN")
        elif w == "ptv":
            result_words.append("PTV")
        else:
            result_words.append(w.title())
    return " ".join(result_words)
df_clean.loc[:, "Source"] = df_clean["Source"].apply(normalize_source)
```

Removing prefixes, normalizing the source column, capitalizing the acronyms

```python
def normalize_headline(text):
    if pd.isnull(text):
        return ""
    text = text.lower()
    text = re.sub(f"[{re.escape(string.punctuation)}]", "", text)
    text = re.sub(r"\s+", " ", text).strip()
    return text
```

Normalizing the headlines, removing punctuation, trimming spaces

```
df_clean["headline_key"] = df_clean["Headline"].apply(normalize_headline)
df_clean = df_clean.drop_duplicates(subset=["headline_key"], keep="first")
df_clean = df_clean.drop(columns=["headline_key"])
```

Removing duplicate headlines

```python
for col in ["Opening Text", "Influencer", "Key Phrases"]:
    if col in df_clean.columns:
        df_clean[col] = df_clean[col].fillna("Unknown")
        df_clean[col] = df_clean[col].replace(r"^\s*$", "Unknown", regex=True)
```

Replacing the empty cells with "Unknown"

```
social_echo_cols = ["Twitter Social Echo", "Facebook Social Echo", "Reddit Social Echo"]
for col in social_echo_cols:
    if col in df_clean.columns:
        df_clean[col] = pd.to_numeric(df_clean[col], errors="coerce").fillna(0).astype(int)
```

Data imputation, converting social echo as integer, replacing empty cells with 0 for numerical columns

```python
if all(col in df_clean.columns for col in social_echo_cols):
    df_clean["Social Echo Total"] = (
        df_clean["Twitter Social Echo"]
        + df_clean["Facebook Social Echo"]
        + df_clean["Reddit Social Echo"]
    ).astype(int)
elif "Social Echo Total" in df_clean.columns:
    df_clean["Social Echo Total"] = pd.to_numeric(df_clean["Social Echo Total"], errors="coerce").fillna(0).astype(int)
```

Recalculating the social echo by summing the 3 columns

```
df_clean = df_clean[~df_clean["Headline"].str.contains("content from this publisher", case=False, na=False)]
df_clean = df_clean[~df_clean["Source Link"].str.contains("proquest", case=False, na=False)]
df_clean = df_clean[~df_clean["Headline"].str.contains("test", case=False, na=False)]
df_clean = df_clean[df_clean["Source Link"].str.strip() != ""]
```

Removing these hit keywords from headline, source link, and headline

```
if "Hit Sentence" in df_clean.columns:
    df_clean = df_clean[~df_clean["Hit Sentence"].str.contains(r"\[Courtesy:|\[Photo courtesy\]", case=False, na=False)]
```

Removing this keyword from hit sentence column

```
df_clean = df_clean.drop(columns=[c for c in cols_to_drop if c in df_clean.columns])

if "Country" in df_clean.columns:
    df_clean["Country"] = df_clean["Country"].astype(str).str.strip().str.title()

if "Headline" in df_clean.columns:
    df_clean["Headline"] = df_clean["Headline"].astype(str).str.strip()
    df_clean["Headline"] = df_clean["Headline"].replace(r"\s+", " ", regex=True)
    df_clean["Headline"] = df_clean["Headline"].str.title()

text_cols = ["Opening Text", "Hit Sentence"]
for col in text_cols:
    if col in df_clean.columns:
        df_clean[col] = df_clean[col].astype(str).str.strip()
        df_clean[col] = df_clean[col].replace(r"\s+", " ", regex=True)
        df_clean[col] = df_clean[col].str.capitalize()
```

Normalizing the headline, country, opening text, and hit sentence column (Capitalizing)

```python
if "Clean Date" in df_clean.columns:
    cols = ["Clean Date"] + [c for c in df_clean.columns if c != "Clean Date"]
    df_clean = df_clean[cols]
```

Moving the new cleaned date to the first column

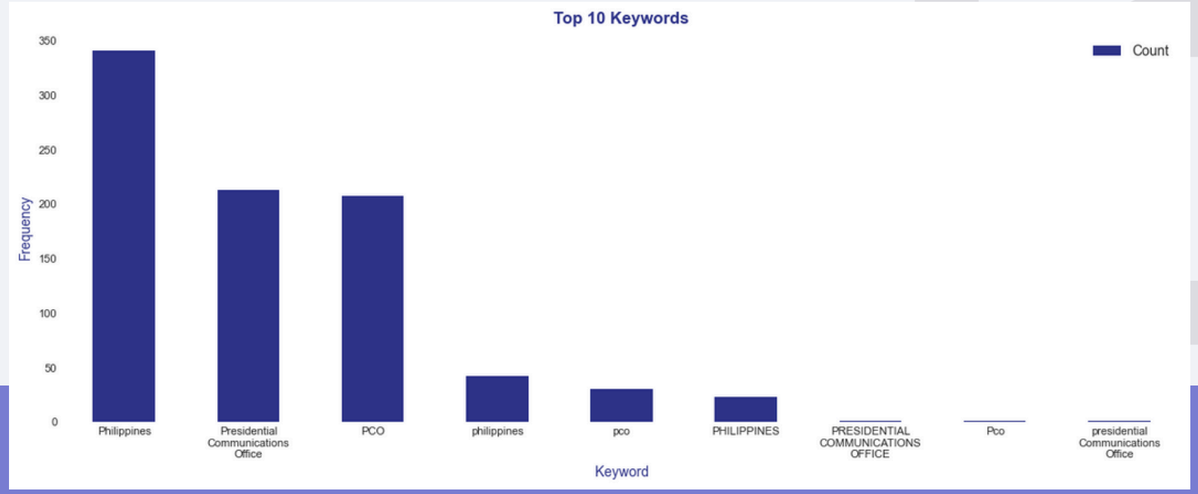
```
if "Keywords" in df_clean.columns:
    df_clean["Keywords"] = df_clean["Keywords"].apply(standardize_keyphrases)
```

Standardization of Keywords Column

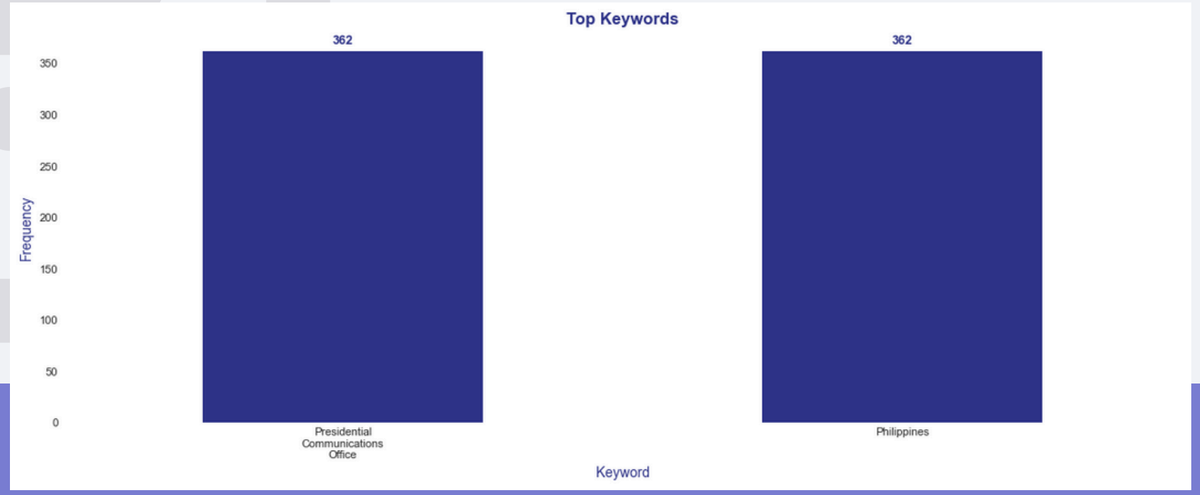
```python
def standardize_keyphrases(text):
    if pd.isna(text) or str(text).strip() == "":
        return "Unknown"
    text = str(text)
    parts = re.split(r'[;,|/•]+', text)
    normalized = []
    for p in parts:
        p = p.strip()
        if p == "":
            continue
        low = p.lower()
        if "pco" in low or "presidential communications" in low:
            normalized.append("Presidential Communications Office")
        elif "philippines" in low or "pilipinas" in low or "pilipino" in low:
            normalized.append("Philippines")
        else:
            normalized.append(p.title())
    seen = set()
    final_list = []
    for item in normalized:
        if item not in seen:
            final_list.append(item)
            seen.add(item)
    return ", ".join(final_list) if final_list else "Unknown"

if "Key Phrases" in df_clean.columns:
    df_clean["Key Phrases"] = df_clean["Key Phrases"].apply(standardize_keyphrases)
```

Standardization of Key Phrases Column

Before



After



Result of standardization of key phrases and keywords column

```python
def detect_language_for_country(text):
    if pd.isna(text) or str(text).strip() == "":
        return None
    try:
        return detect(text)
    except LangDetectException:
        return None

def map_language_to_country(lang_code):
    mapping = {
        "tl": "Philippines",
        "fil": "Philippines",
        "en": "United States",
        "ja": "Japan",
        "ko": "South Korea",
        "zh": "China",
        "fr": "France",
        "es": "Spain",
        "de": "Germany",
        "ms": "Malaysia",
        "id": "Indonesia",
    }
    return mapping.get(lang_code, None)
```

Fill missing Country based on Hit Sentence language

```python
if "Country" in df_clean.columns and "Hit Sentence" in df_clean.columns:
    def fill_country(row):
        country = str(row["Country"]).strip().lower()
        if country == "" or country == "unknown" or country == "nan":
            lang_code = detect_language_for_country(row["Hit Sentence"])
            inferred_country = map_language_to_country(lang_code)
            return inferred_country if inferred_country else "Unknown"
        return row["Country"]

    df_clean["Country"] = df_clean.apply(fill_country, axis=1)
```

Fill missing Country based on Hit Sentence language

```python
def detect_tl_or_en(text):
    if pd.isna(text) or str(text).strip() == "":
        return "English"
    try:
        lang_code = detect(str(text))
    except LangDetectException:
        return "English"

    if lang_code in ["tl", "fil"]:
        return "Tagalog"
    else:
        return "English"

if "Language" in df_clean.columns and "Hit Sentence" in df_clean.columns:
    def fill_language(row):
        lang = str(row["Language"]).strip().lower()
        if lang == "" or lang == "unknown" or lang == "nan":
            return detect_tl_or_en(row["Hit Sentence"])
        return row["Language"]

    df_clean["Language"] = df_clean.apply(fill_language, axis=1)
```

Detect Language based on Hit Sentence - Tagalog or English only

```
cols_to_drop = [
    "Subregion", "Desktop Reach", "Mobile Reach",
    "National Viewership", "Engagement", "AVE", "Twitter Authority",
    "Tweet Id", "Twitter Id", "Twitter Client", "Twitter Screen Name",
    "User Profile Url", "Twitter Bio", "Twitter Followers", "Twitter Following",
    "Alternate Date Format", "Time", "State", "City", "Editorial Echo",
    "Views", "Estimated Views", "Likes", "Replies", "Retweets",
    "Comments", "Shares", "Reactions", "Threads", "Is Verified",
    "Parent URL", "Document Tags", "Document ID", "Custom Categories"
]
```

Dropping the irrelevant columns

```
output_path = "MSM_cleaned.xlsx"
df_clean.to_excel(output_path, index=False, engine="openpyxl")
```

Exporting the new cleaned MSM file


DATA CLEANING
[SM]

```
df_clean = df.drop_duplicates(subset=["Source Link"], keep="first")

df_clean = df_clean.dropna(subset=["Date", "Source", "Country"])
def fill_headline(row):
```

Remove duplicates from the source link and delete rows with missing values in date, source, or country

```
    if pd.isna(row["Headline"]) or str(row["Headline"]).strip() == "":
        if pd.notna(row.get("Opening Text")) and str(row.get("Opening Text")).strip() != "":
            return row["Opening Text"].strip()
    return row["Headline"]

df_clean["Headline"] = df_clean.apply(fill_headline, axis=1)
```

Fill the missing cells in the headline column with the opening text value to prevent Python from dropping Twitter data

```
def normalize_source(source):

    s = str(source).strip().lower()

    s = re.sub(r"www\.", "", s)
    s = re.sub(r"\.com|\.ph|\.net|\.org", "", s)
    s = re.sub(r"\s+", " ", s)
    if "reddit" in s:

        return "Forums"

    return s.title()

df_clean.loc[:, "Source"] = df_clean["Source"].apply(normalize_source)
```

Replacing the Reddit source name as Forums

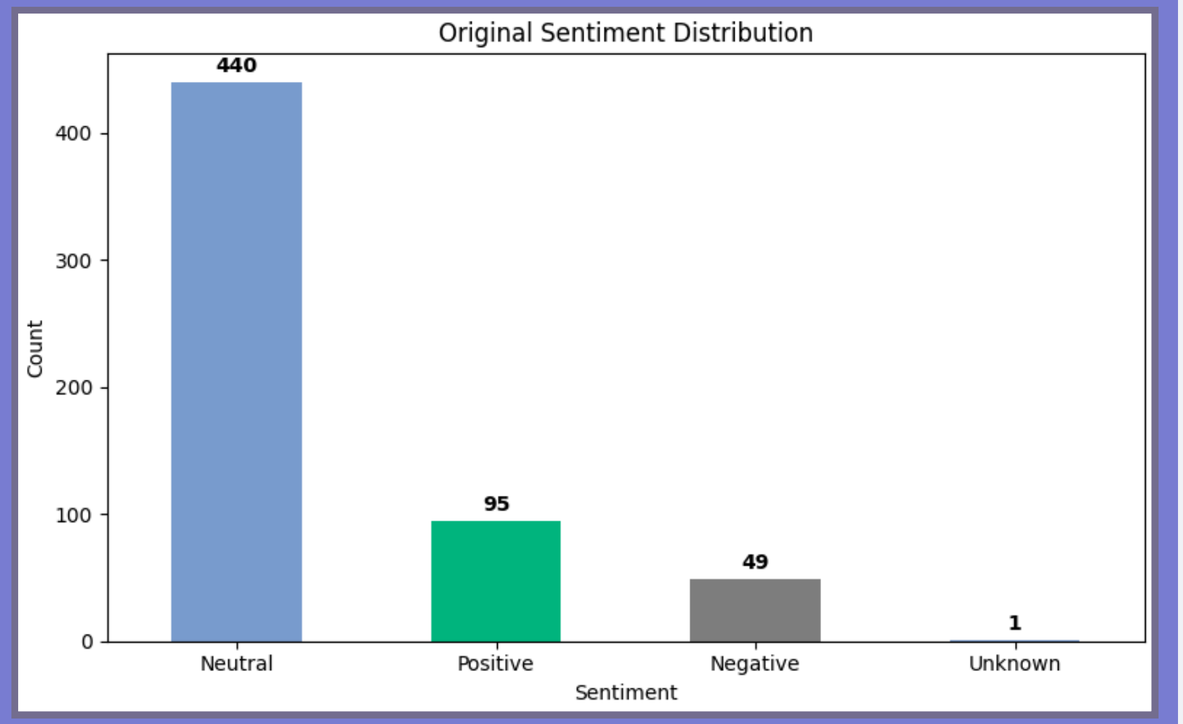
```python
def normalize_headline(text):

    if pd.isnull(text):

        return ""

    text = text.lower()

    text = re.sub(f"[{re.escape(string.punctuation)}]", "", text)

    text = re.sub(r"\s+", " ", text).strip()

    return text

df_clean["headline_key"] = df_clean["Headline"].apply(normalize_headline)

df_clean = df_clean.drop_duplicates(subset=["headline_key"], keep="first")

df_clean = df_clean.drop(columns=["headline_key"])
```

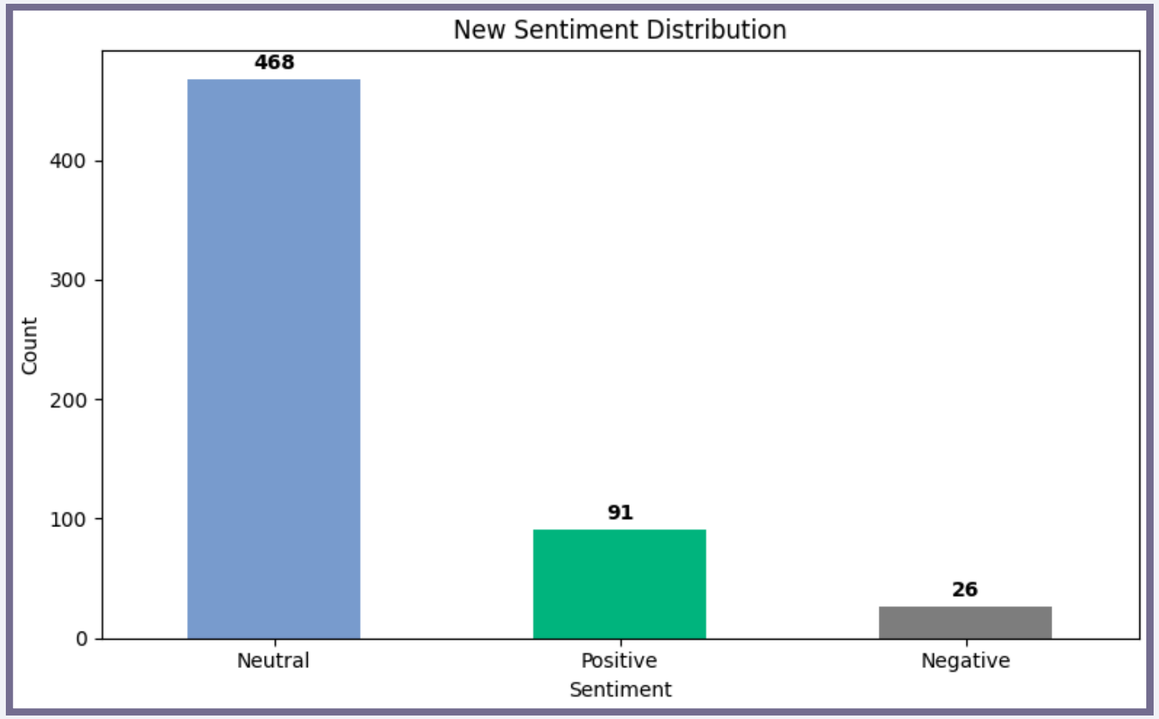
Standardise the headline column by removing punctuation, trimming spaces, and eliminating duplicate entries


MSM RESULTS

Sentiment Analysis

Original Sentiment vs New Sentiment



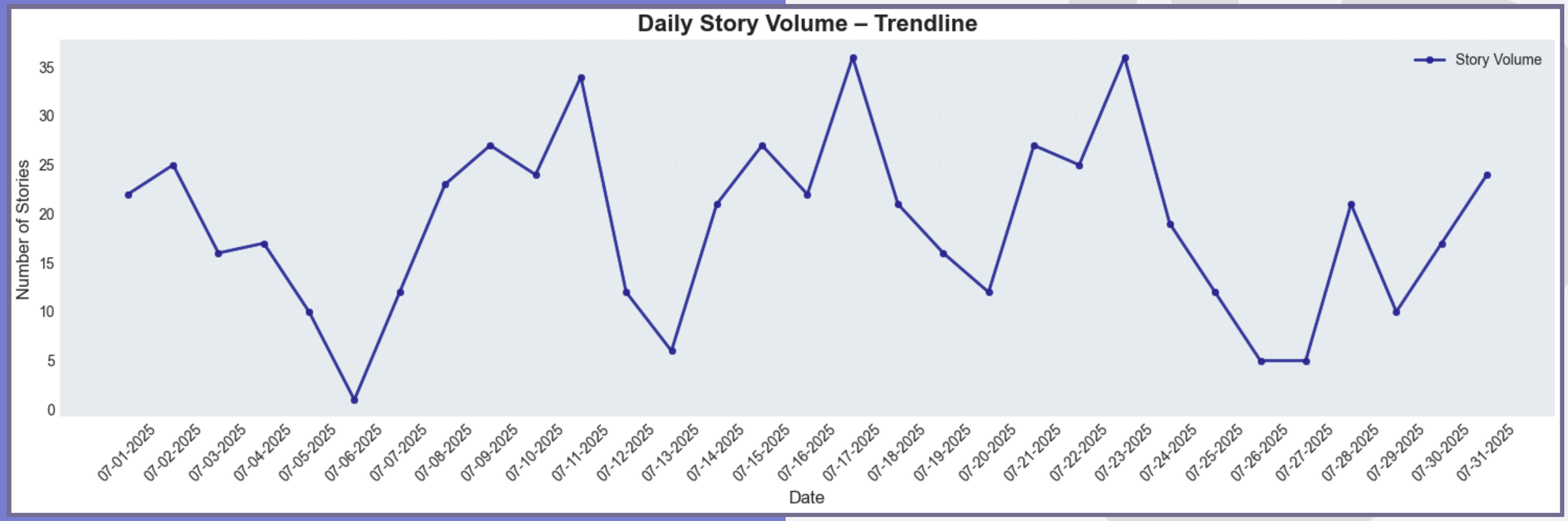


Sentiment Analysis

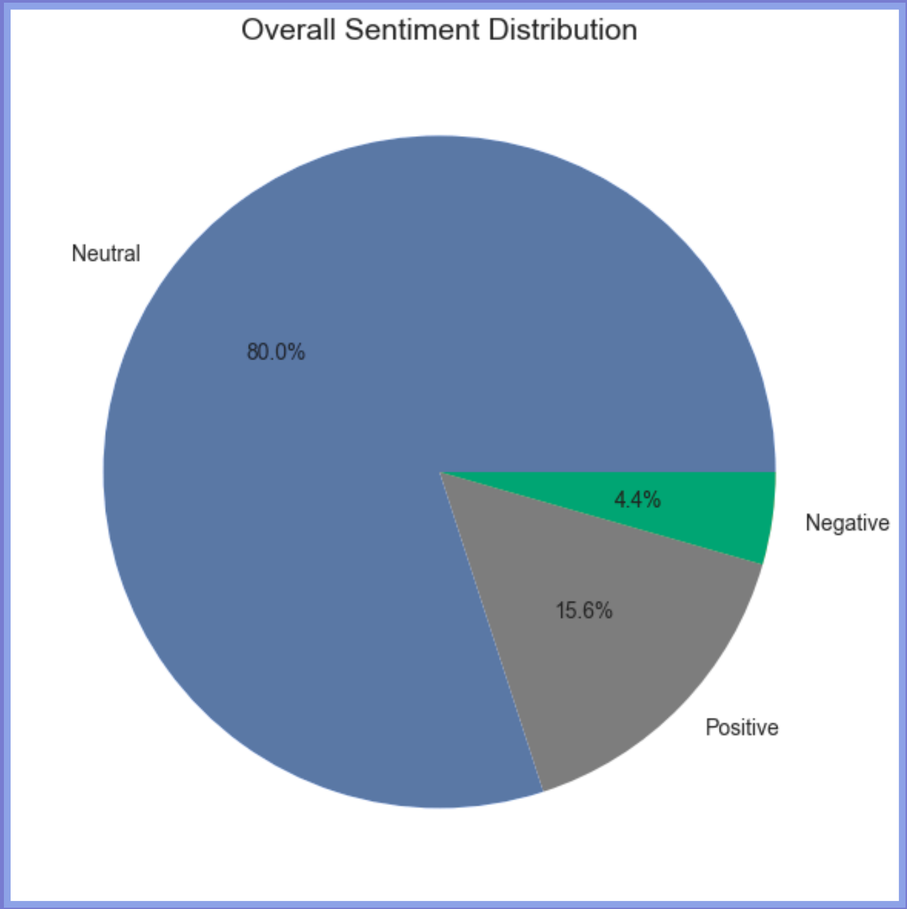
Sample Differences

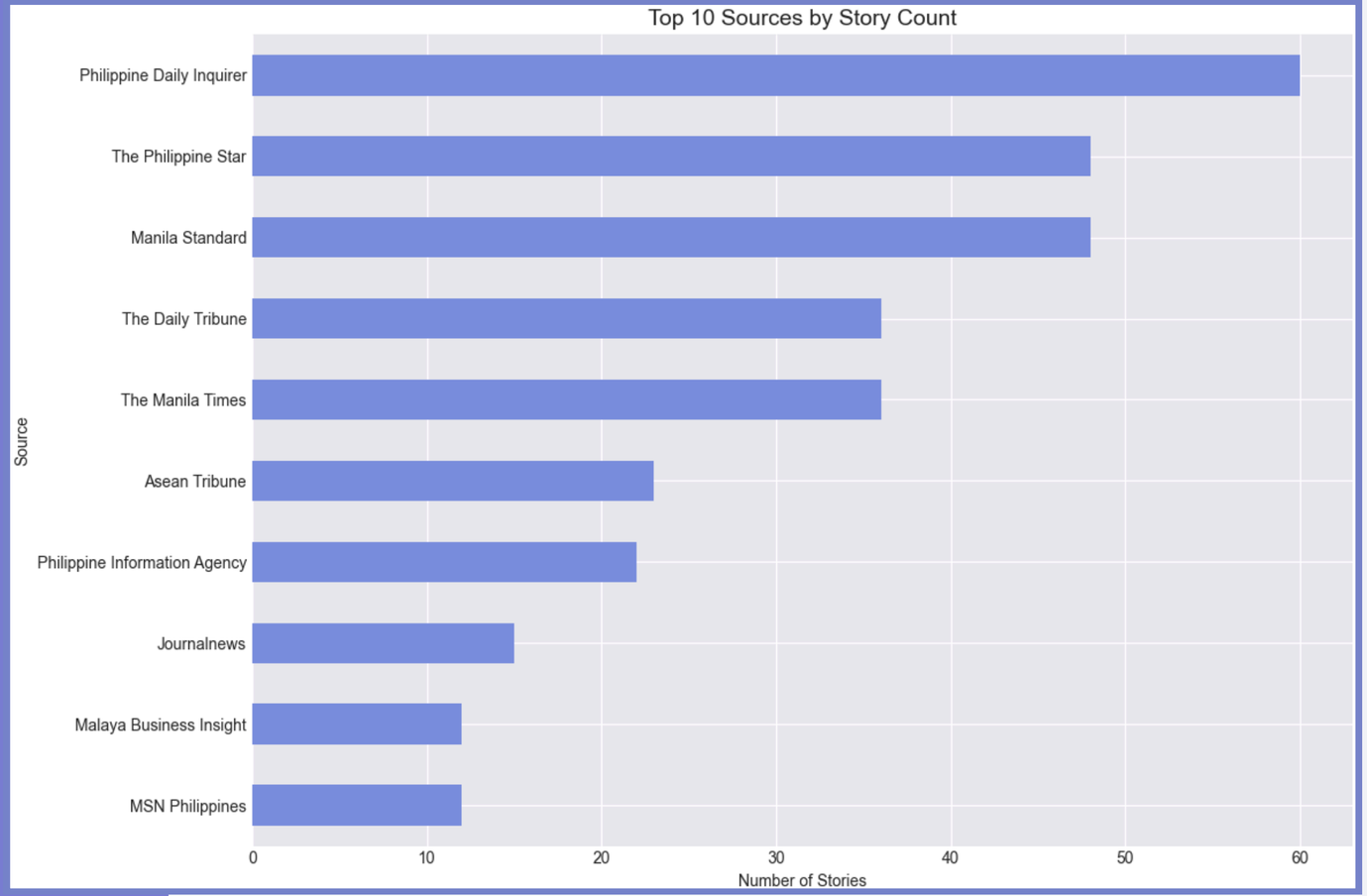
```
🧪 Showing 5 examples where the sentiment changed:

Headline: Paglabag Sa Protocol Sa Presidential Coverage, ‘Di Dapat Maulit - Ruiz
Original Sentiment: Negative
New Sentiment: Neutral
------------------------------------------------------------
Headline: Here'S The Full Text Of President Bongbong Marcos' Fourth State Of The Nation Address
Original Sentiment: Neutral
New Sentiment: Positive
------------------------------------------------------------
Headline: Philippines’ Marcos To Meet Trump Hoping To Secure Trade Deal
Original Sentiment: Neutral
New Sentiment: Positive
------------------------------------------------------------
Headline: Discount Express (From July 16): 50% Off On Manila Trains For Seniors, People Of Determination
Original Sentiment: Positive
New Sentiment: Neutral
------------------------------------------------------------
Headline: Gov’T Extends National Fiber Backbone Project To More Regions
Original Sentiment: Neutral
New Sentiment: Positive
------------------------------------------------------------
```

Trend Visualization

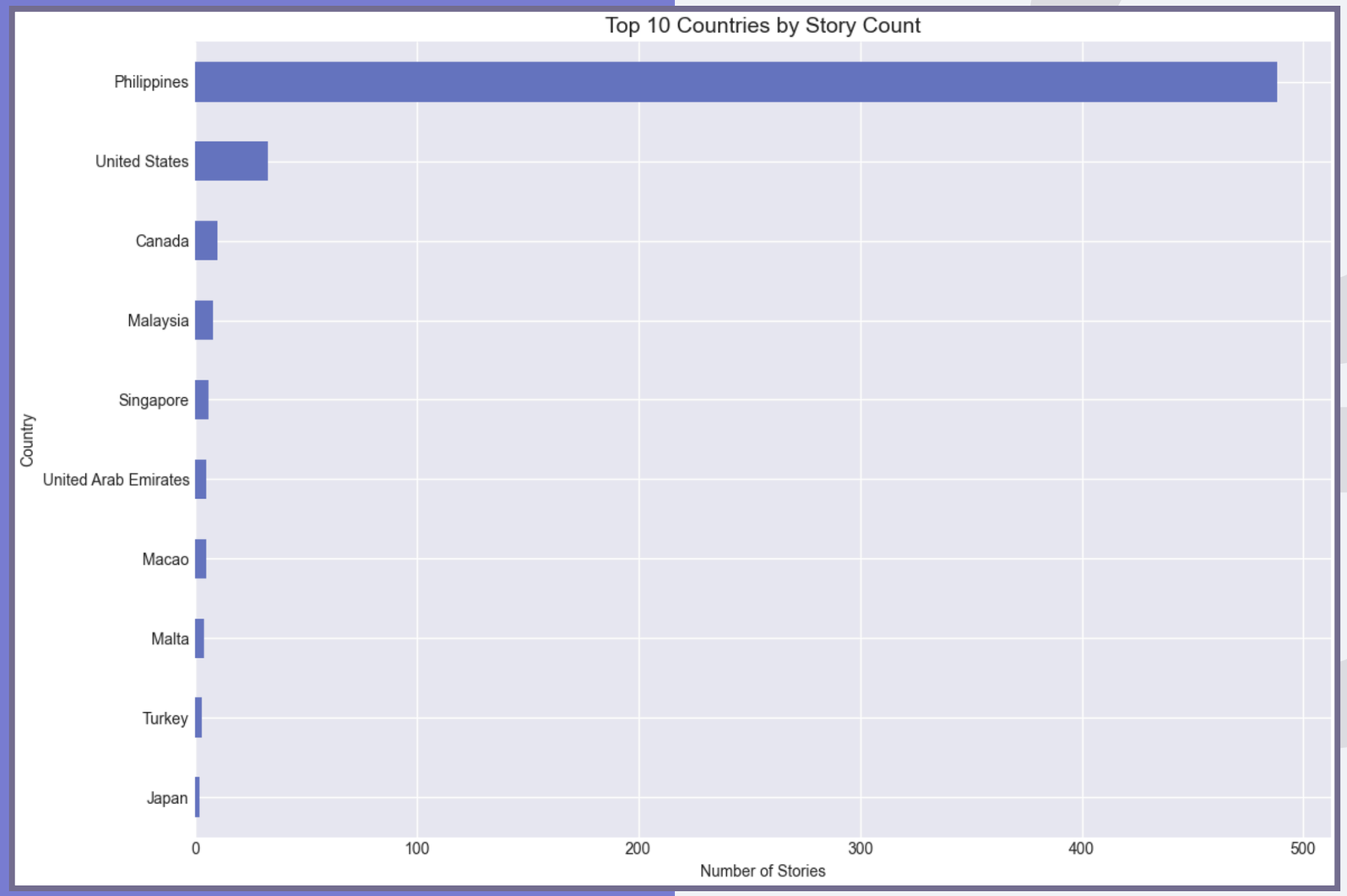


Trend Visualization

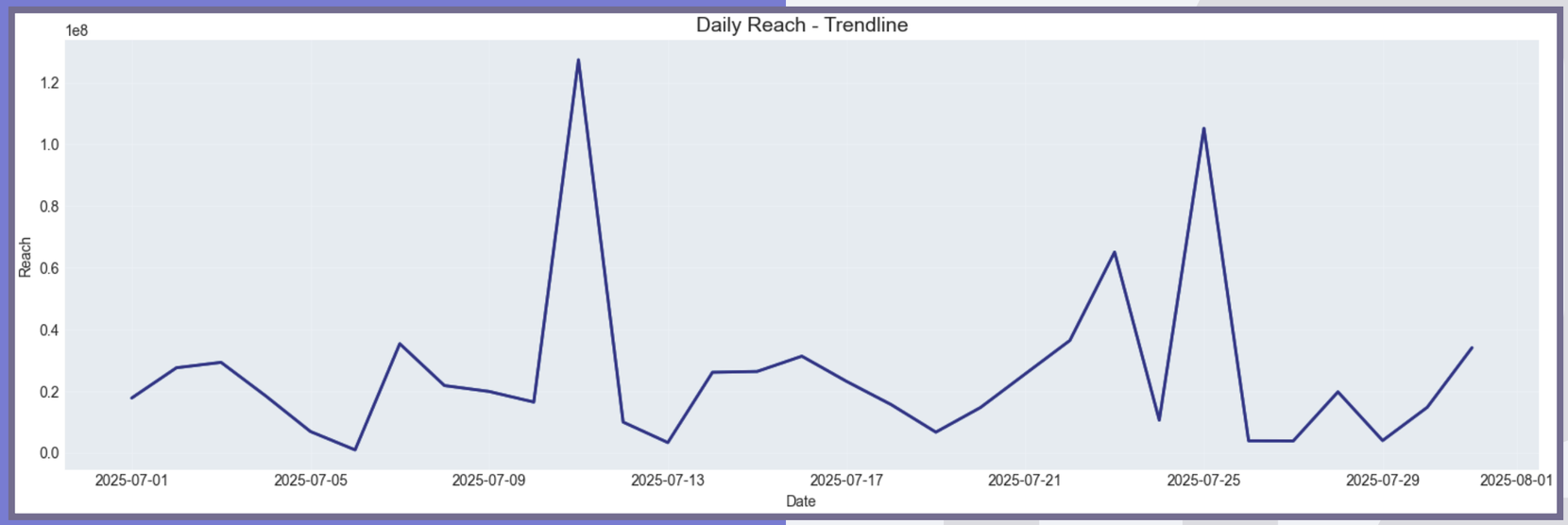


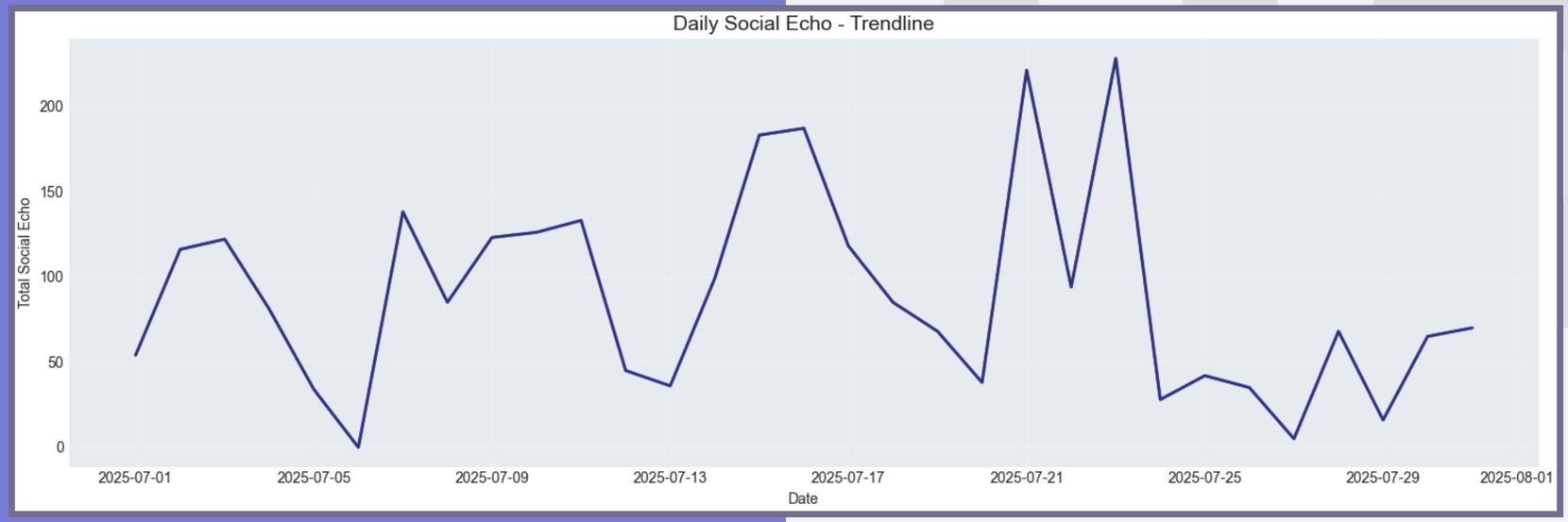


Trend Visualization



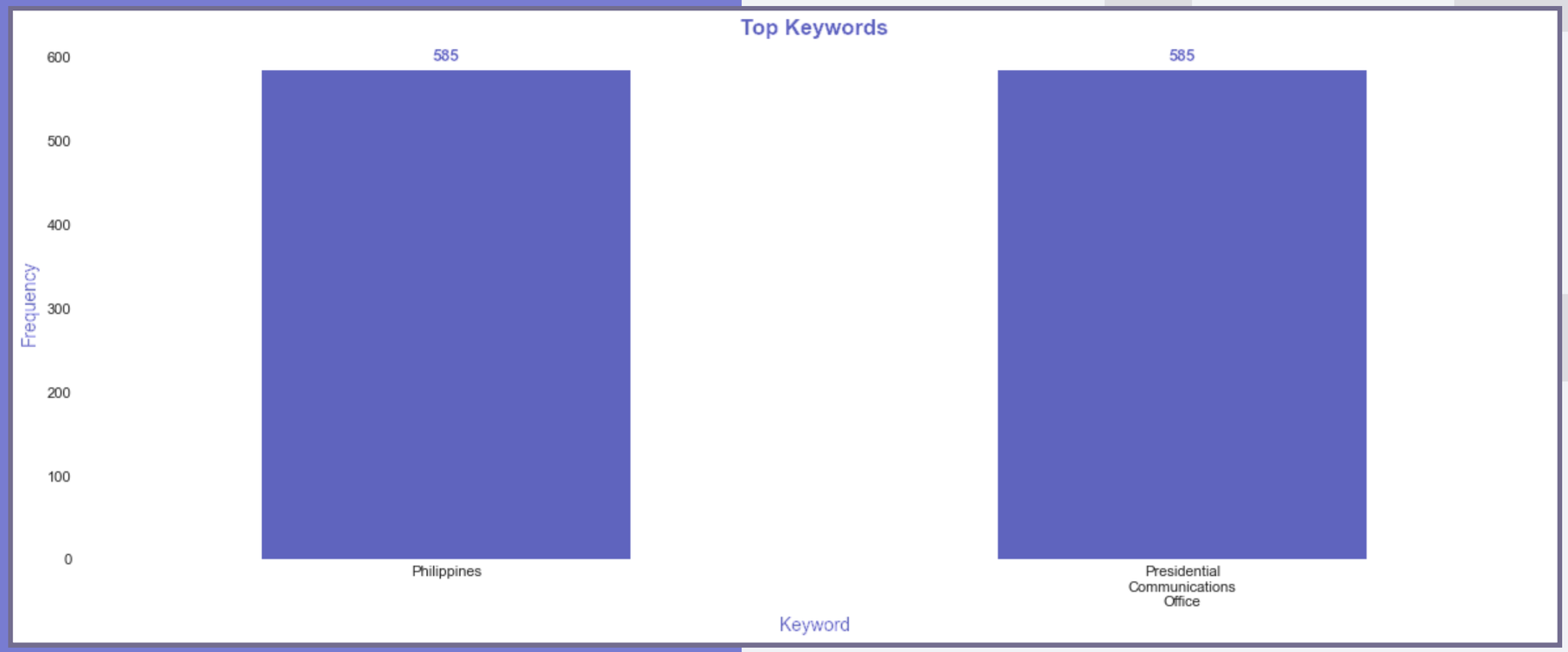
Trend Visualization





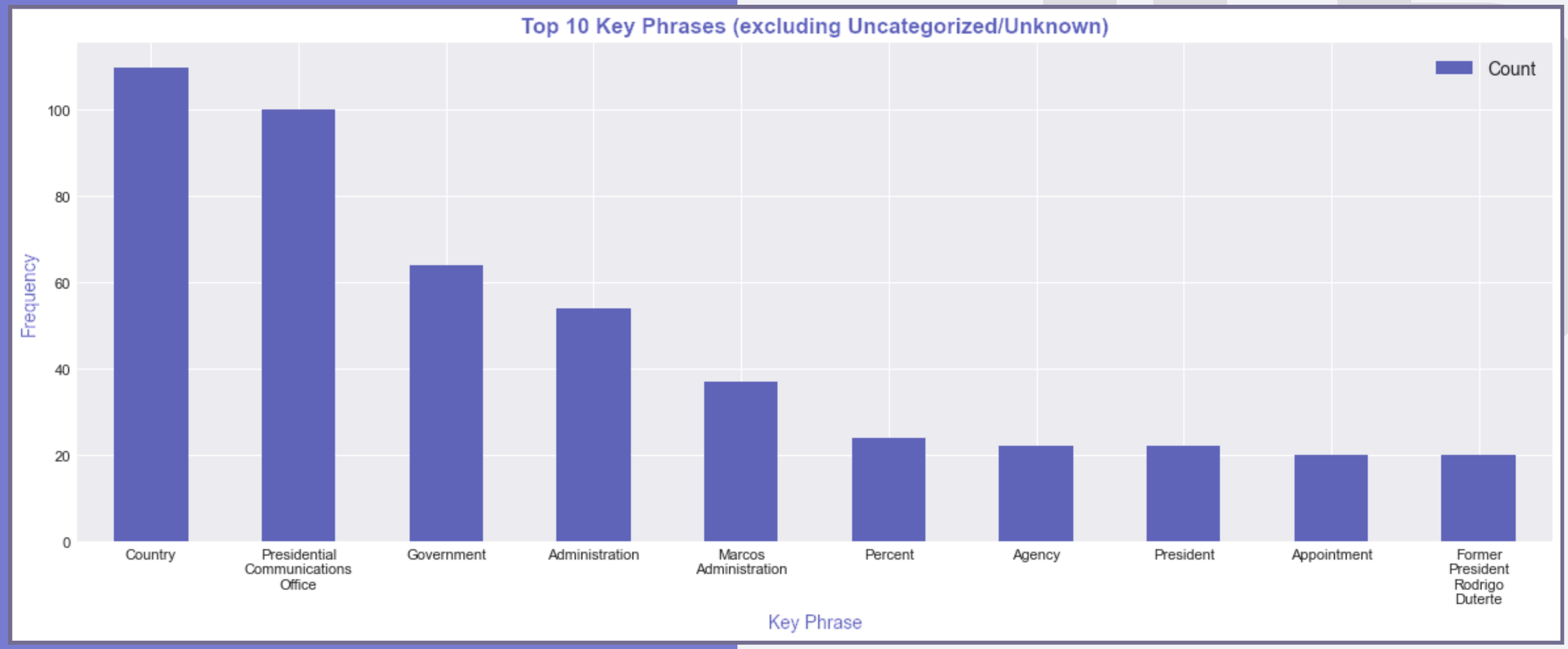
Topic Identification

Content Analysis - Top Common Keywords



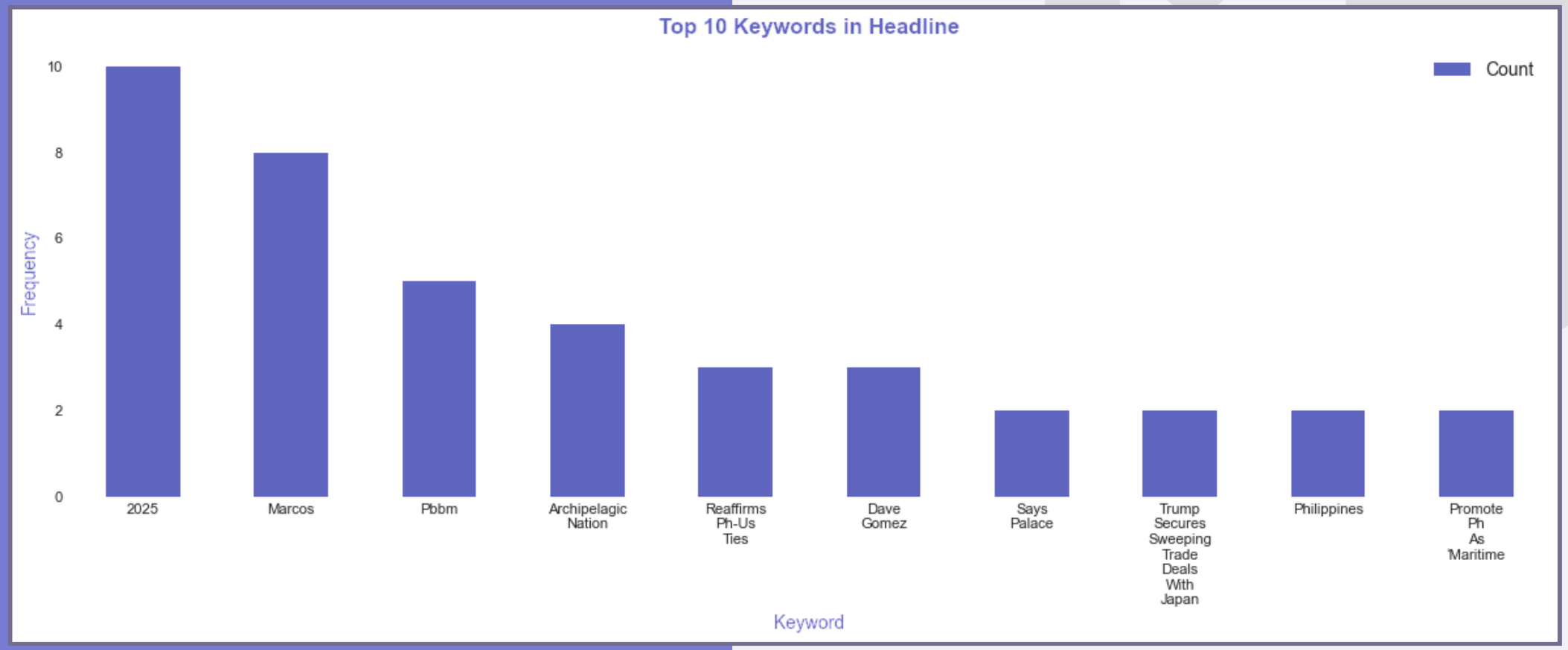
Topic Identification

Content Analysis - Top 10 Key Phrases



Topic Identification

Content Analysis - Top 10 Keywords in Headline



Word Cloud Visualization

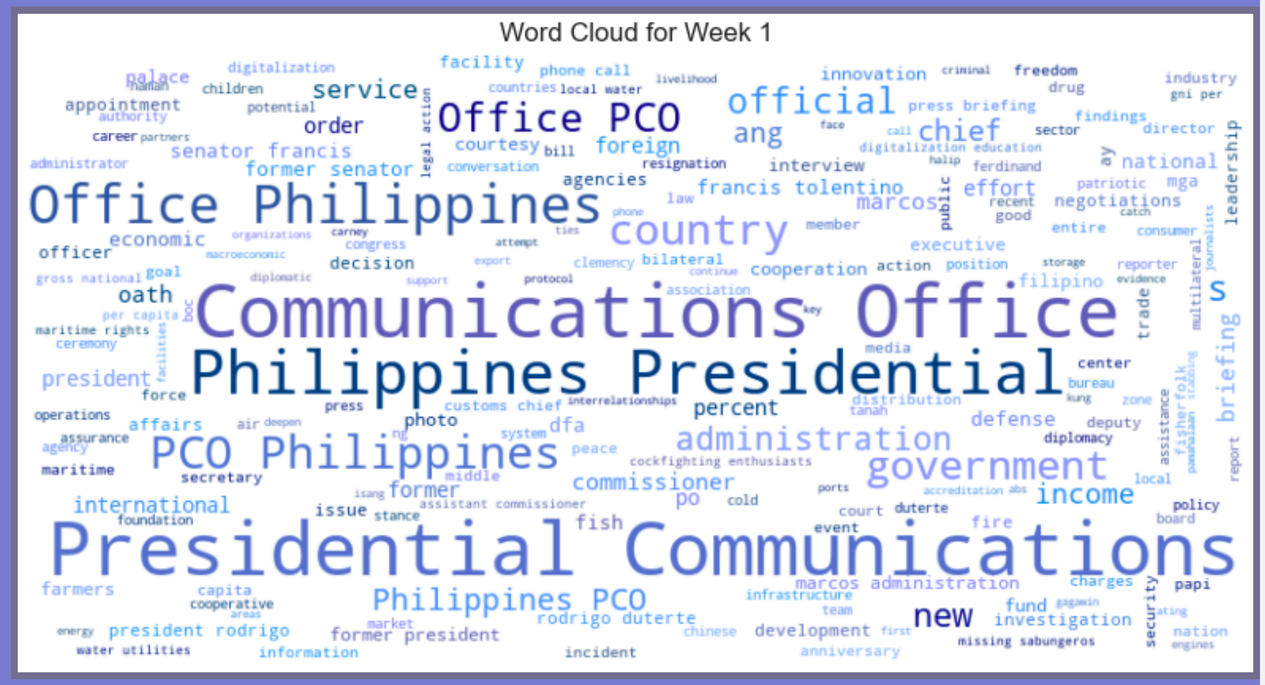
Visual Analysis - All Keywords and Phrases

Word Cloud Visualization

Visual Analysis - All Sentiment Categories

Word Cloud Visualization

Visual Analysis - Week 1 vs Week 2



Word Cloud Visualization

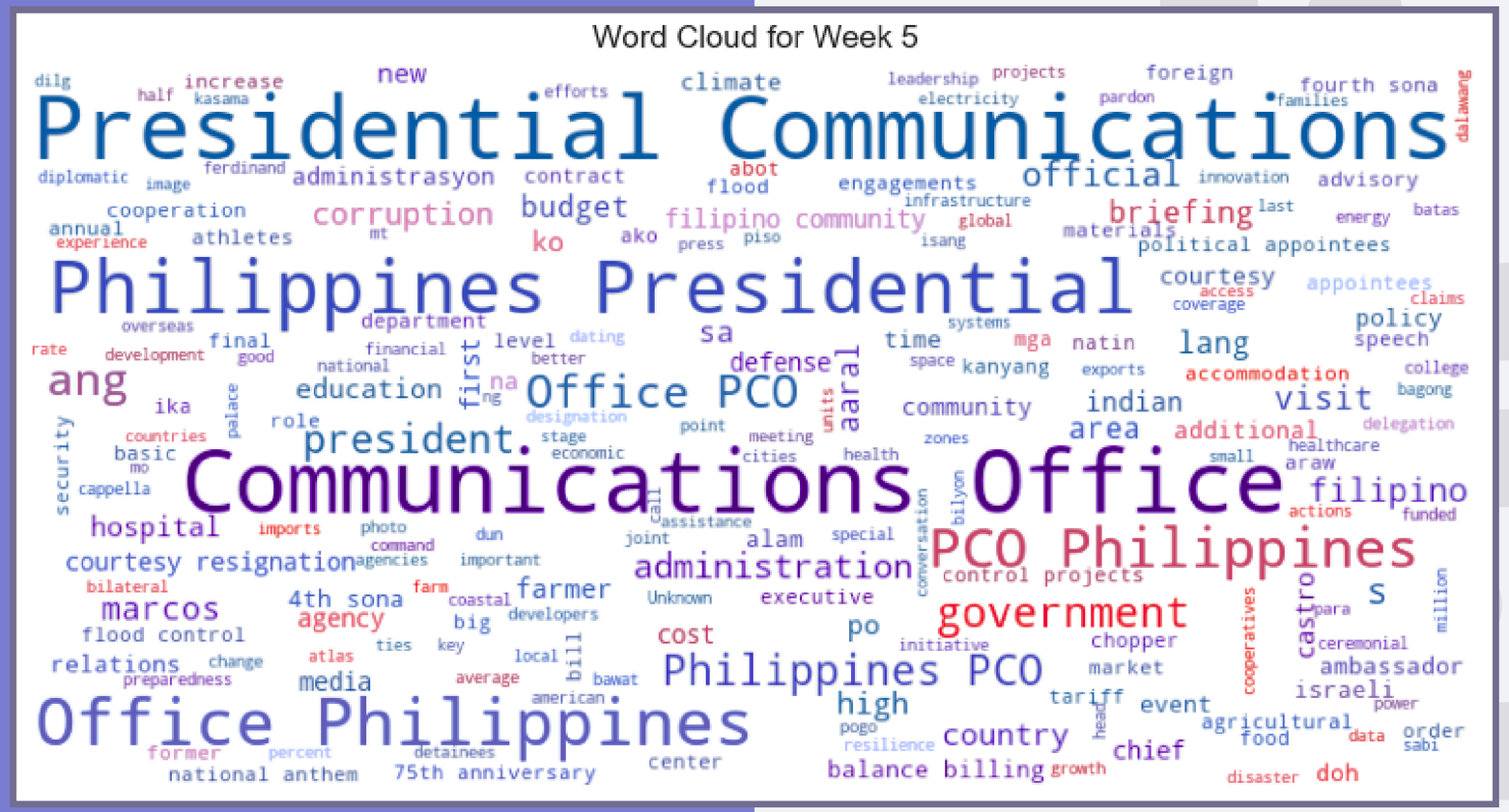
Visual Analysis - Week 3 vs Week 4

Week 3

Week 4

Word Cloud Visualization

Visual Analysis - Week 5



Week 5

Findings

Which are the top 5 sources by number of stories?

Source	Volume
1. Philippine Daily Inquirer	60
2. Manila Standard 2. The Philippine Star	48
3. The Daily Tribune 3. The Manila Times	36
4. Asean Tribune	23
5. Philippine Information Agency	22

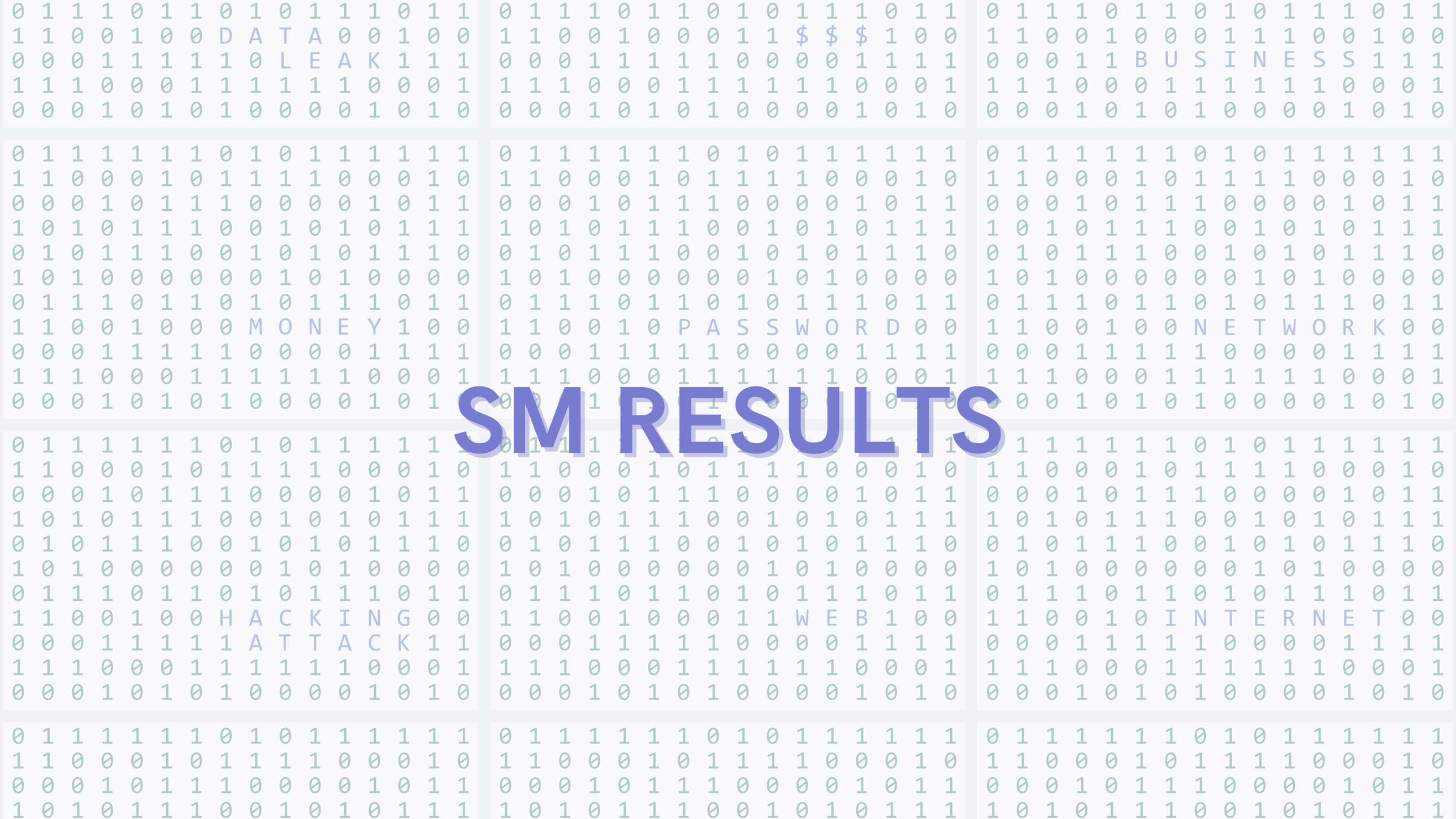
Findings

Which country has the most positive sentiment stories?

Country	Volume
1. Philippines	77
2. United States	4
3. Singapore	3
4. Malaysia	2

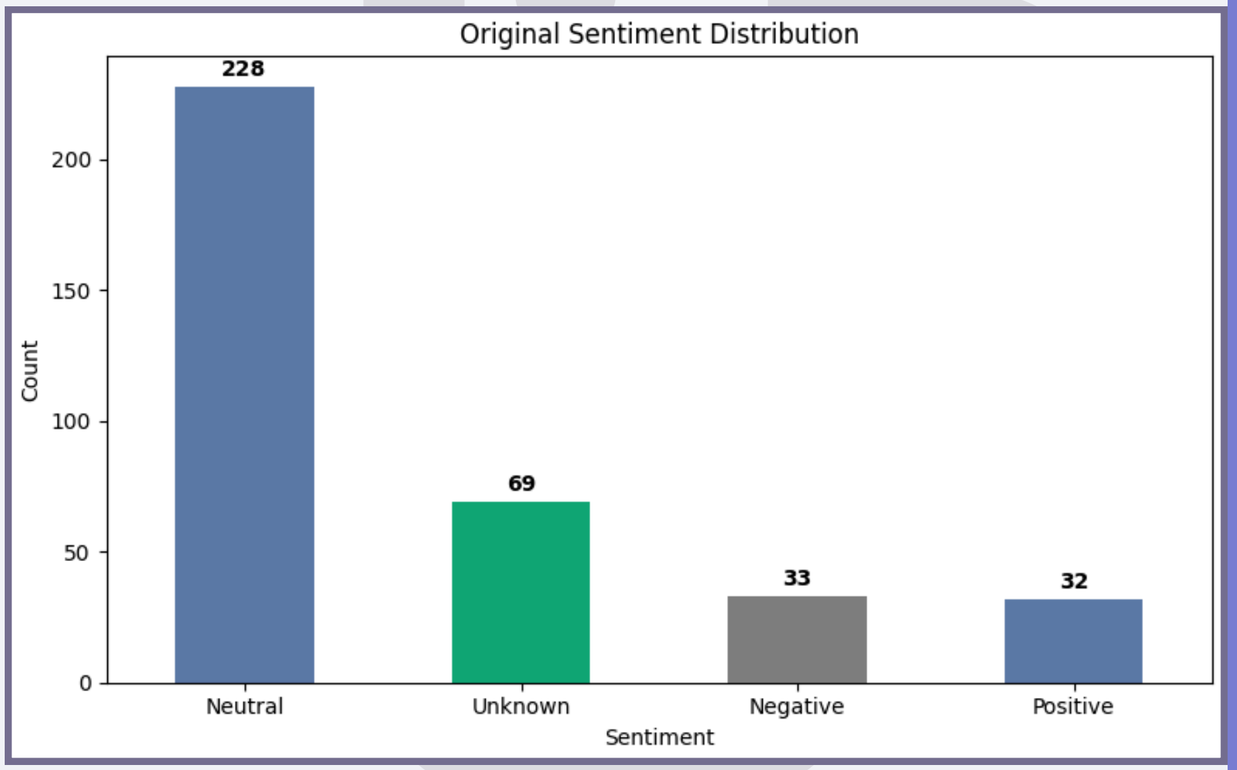
Findings

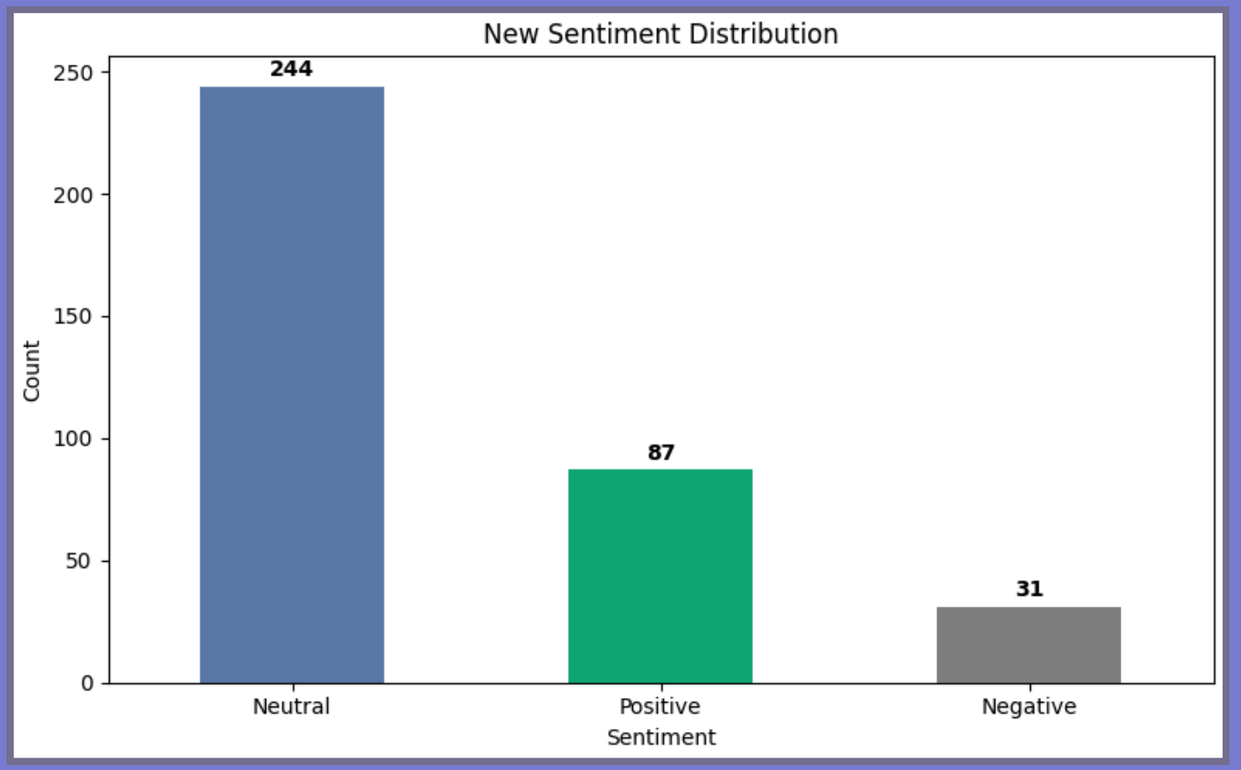
What's the most common keyword/topic?	Volume
Philippines, Presidential Communications Office	585
Which source has the highest average reach?	**Reach**
MSN	101,199,920


SM RESULTS

Sentiment Analysis

Original Sentiment vs New Sentiment



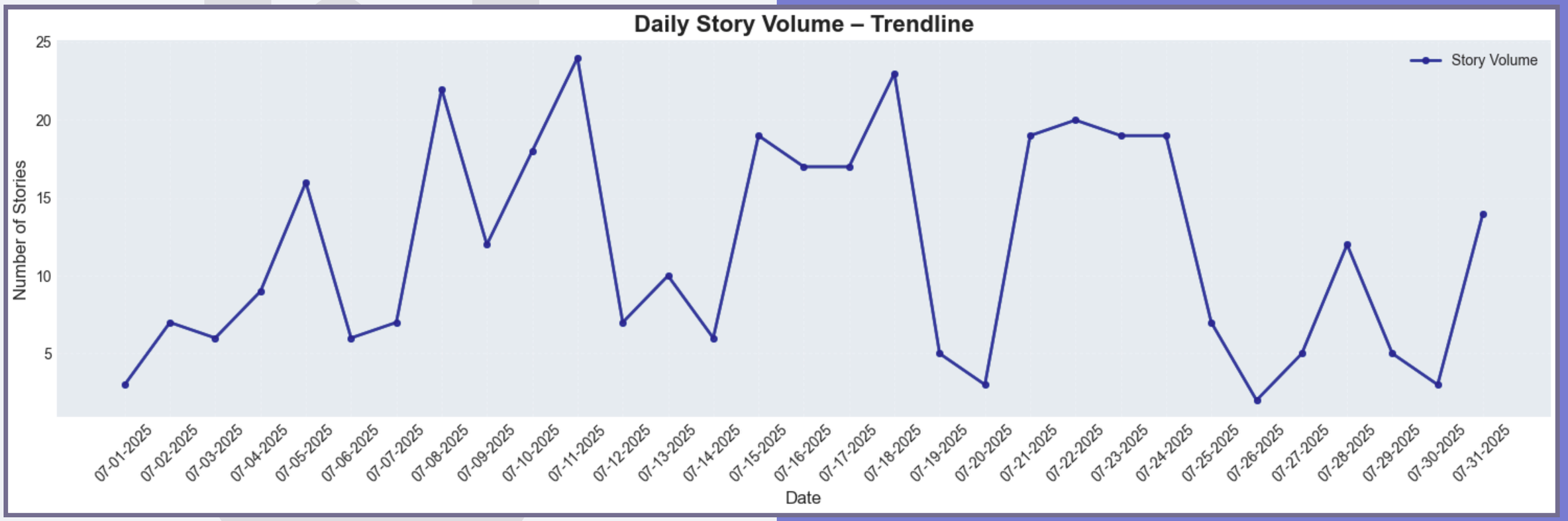


Sentiment Analysis

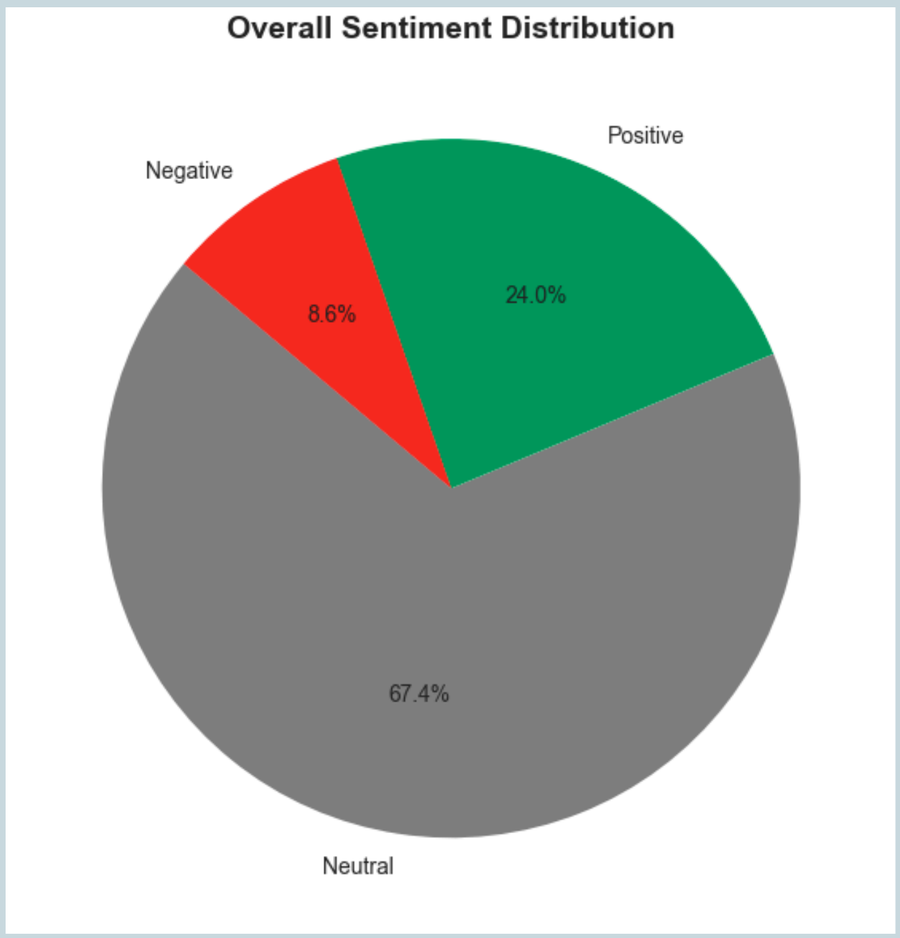
Sample Differences

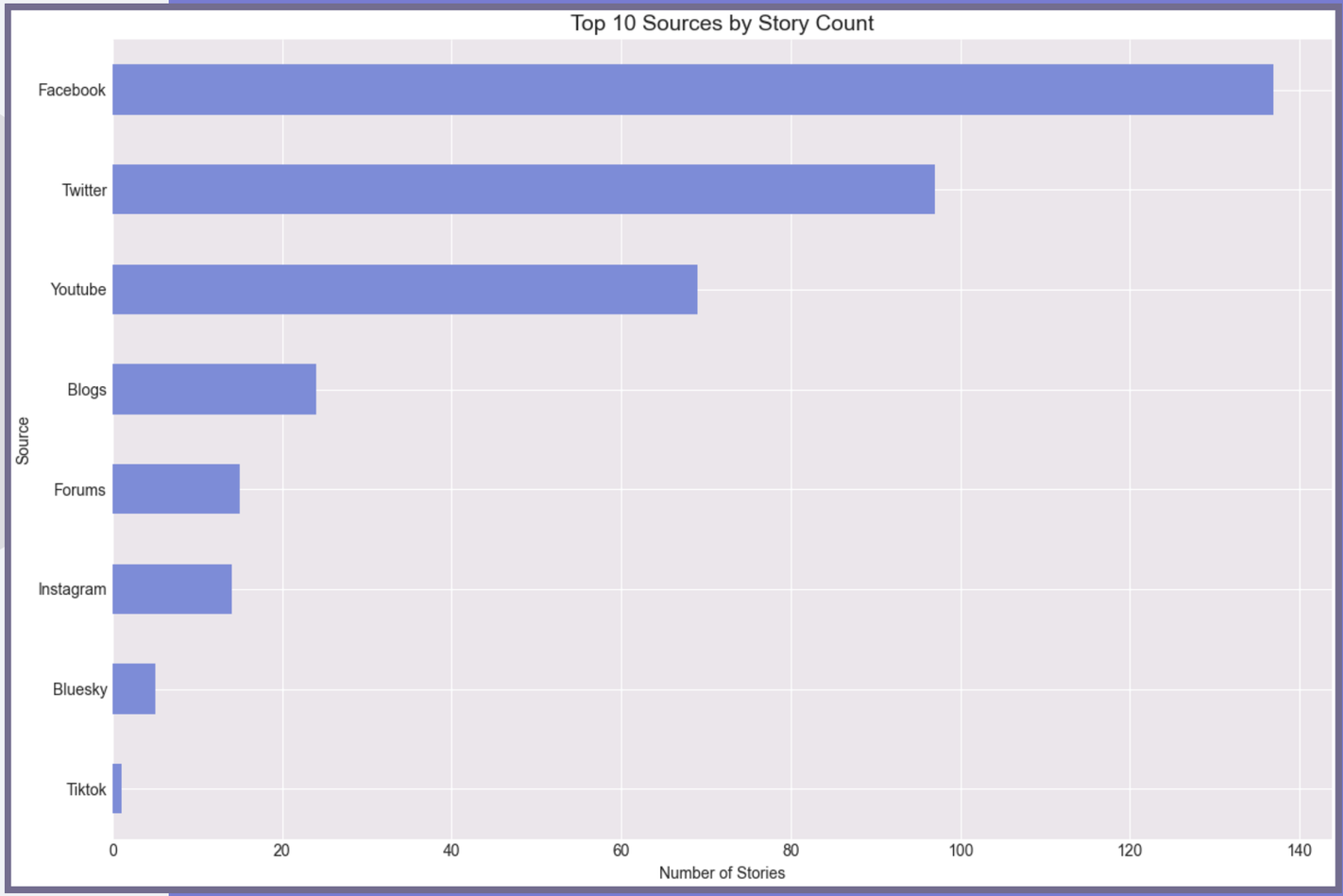
```
Headline: French President Emmanuel Macron Said His Country Would Formally Recognize A Palestinian State During A Un Meeting In September, The Mos
Original Sentiment: Neutral
New Sentiment: Positive
------------------------------------------------------------
Headline: Nearly 500K Families Affected As ‘Crising,’ Habagat Batter Philippines
Original Sentiment: Negative
New Sentiment: Neutral
------------------------------------------------------------
Headline: Pco Approves Removal Of Net 25 Reporter Eden Santos Over Protocol Violations
Original Sentiment: Negative
New Sentiment: Neutral
------------------------------------------------------------
Headline: 🔴Kakapasok Lang Atty. Acosta Ipalit Sa Pco Malacanang Claire Castro Vs Eden Santos Pinakita Palpakan
Original Sentiment: Unknown
New Sentiment: Neutral
------------------------------------------------------------
Headline: Pco Claire Binuking Ni Eden Sa Listahan Ng Drug Free Barangay❗ #Duterte #Bbm #Philippines #Ofwlife
Original Sentiment: Unknown
New Sentiment: Positive
------------------------------------------------------------
```

Trend Visualization

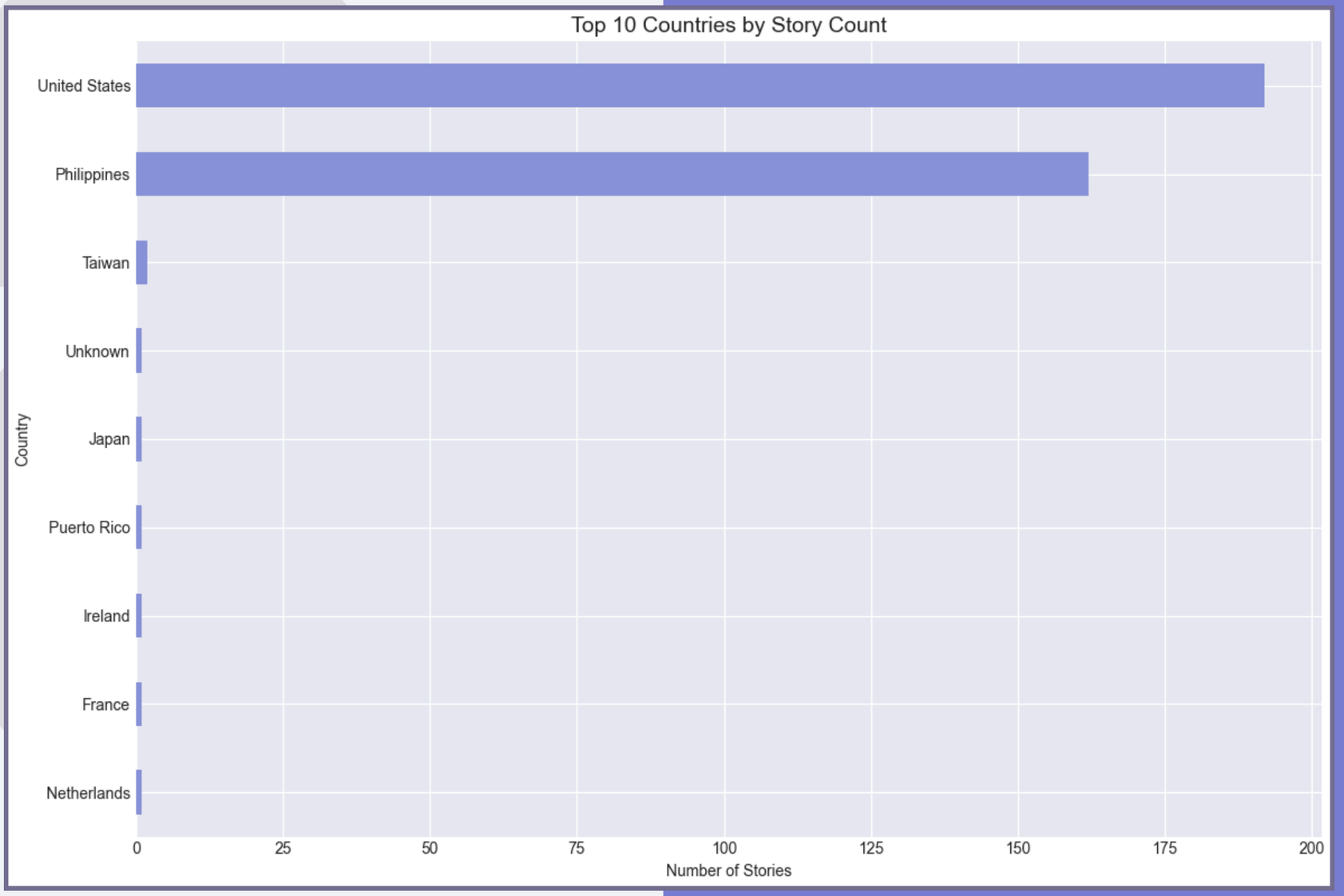


Trend Visualization

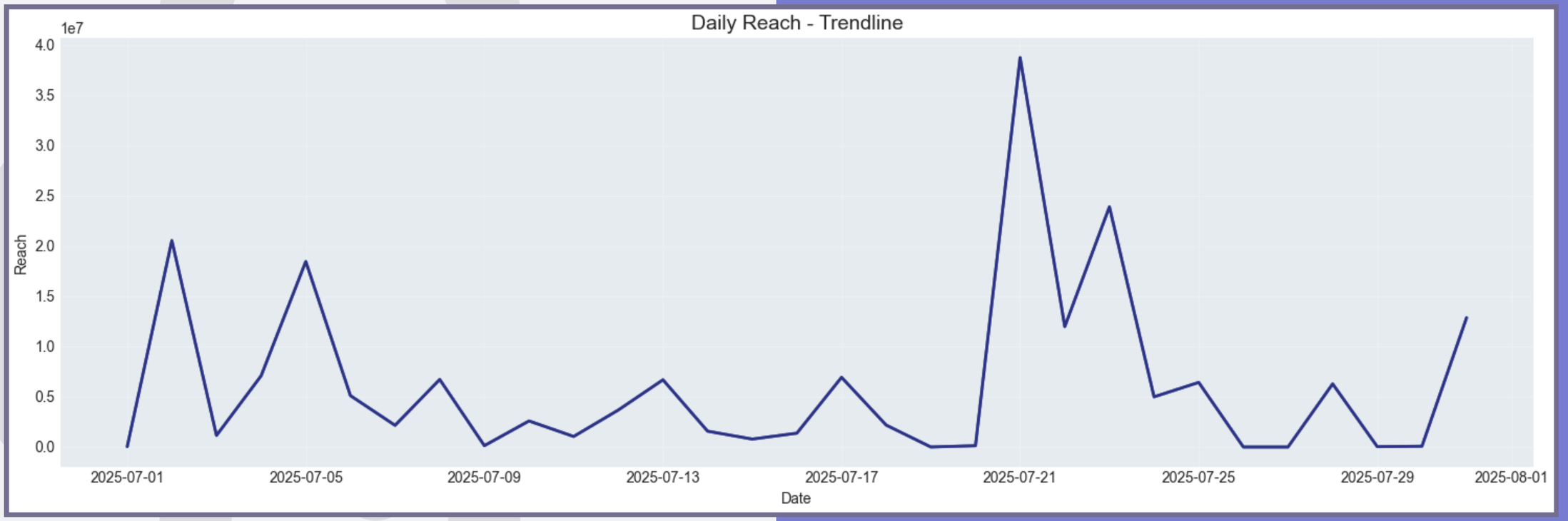




Trend Visualization



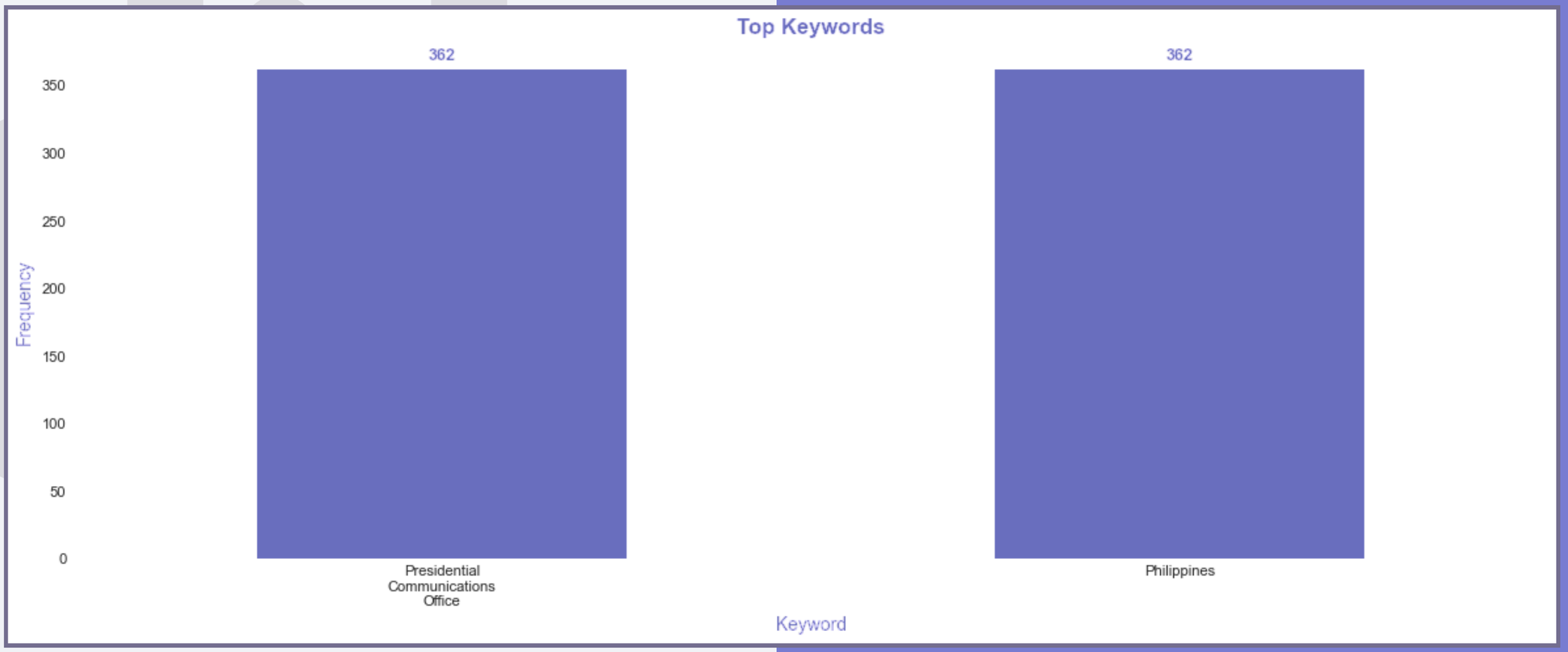
Trend Visualization



*no value for social echo

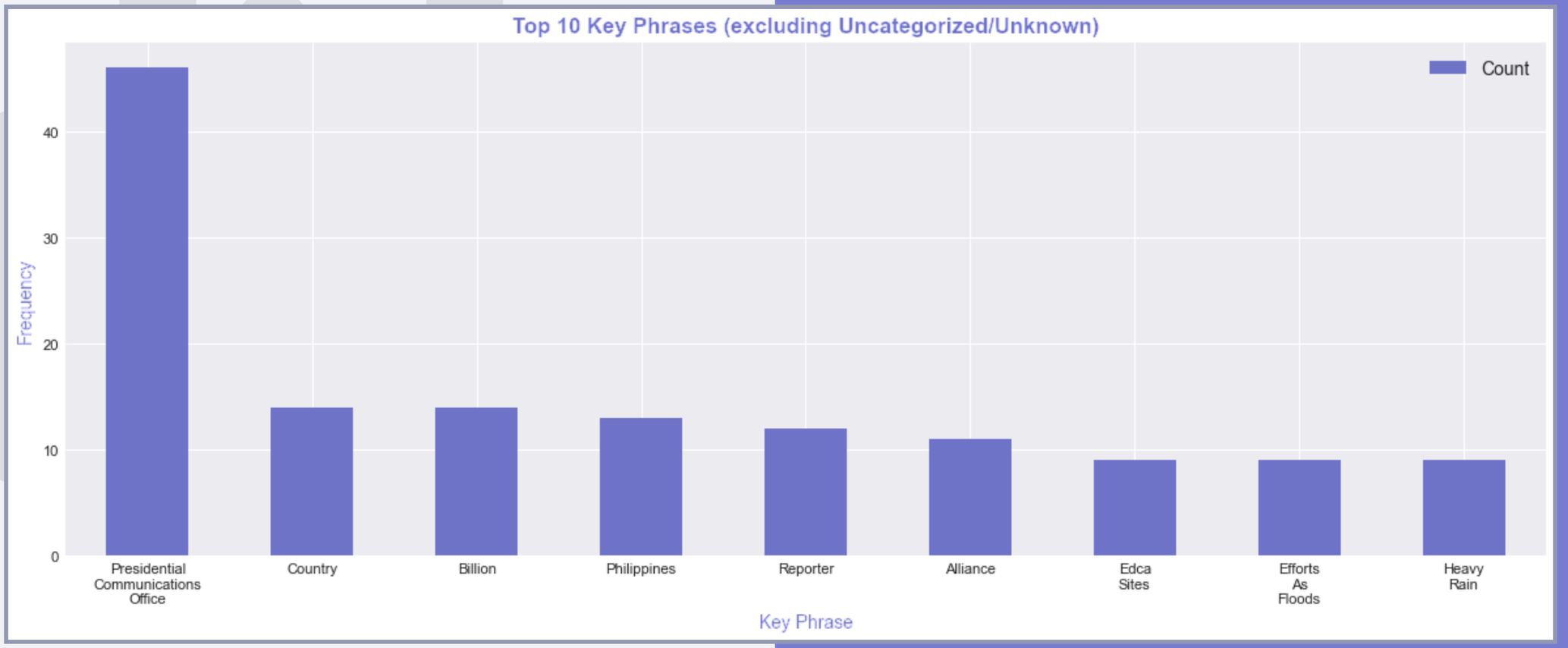
Topic Identification

Content Analysis - Top 10 Common Keywords



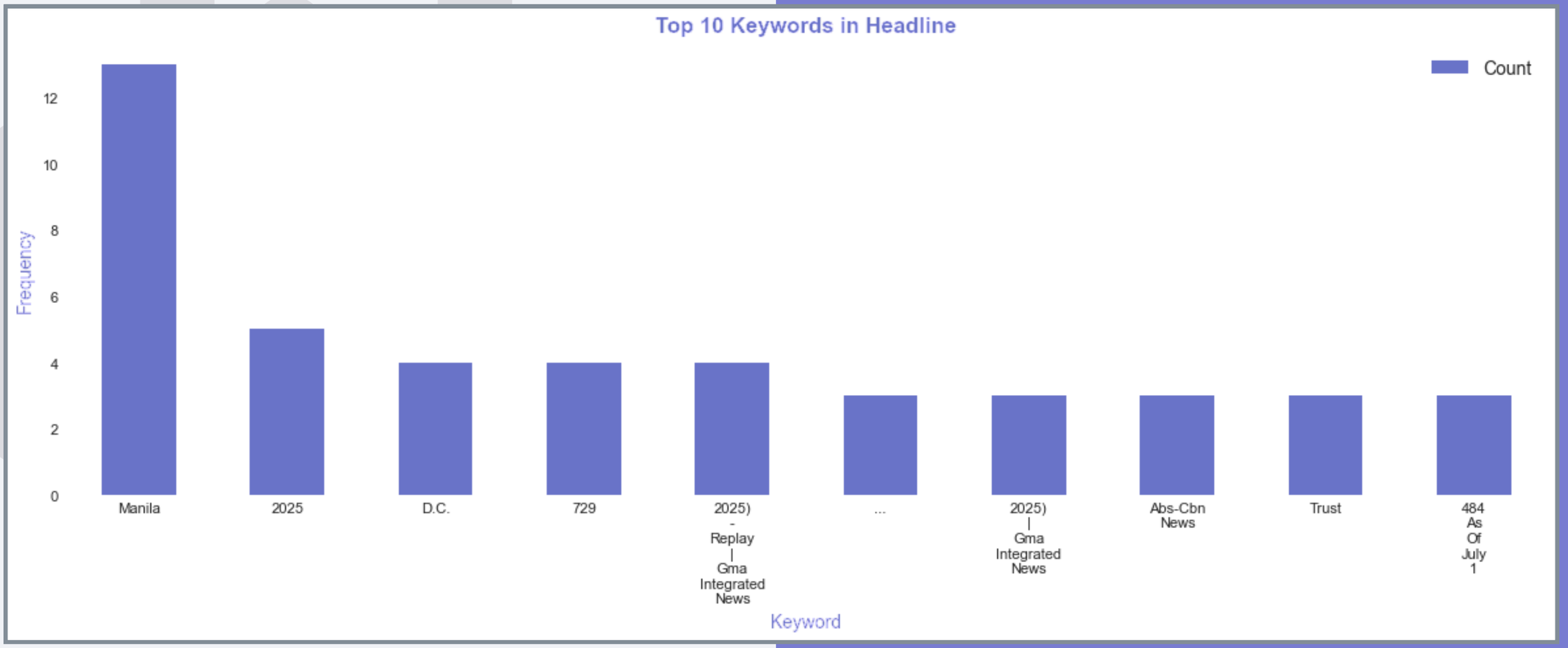
Topic Identification

Content Analysis - Top 10 Key Phrases



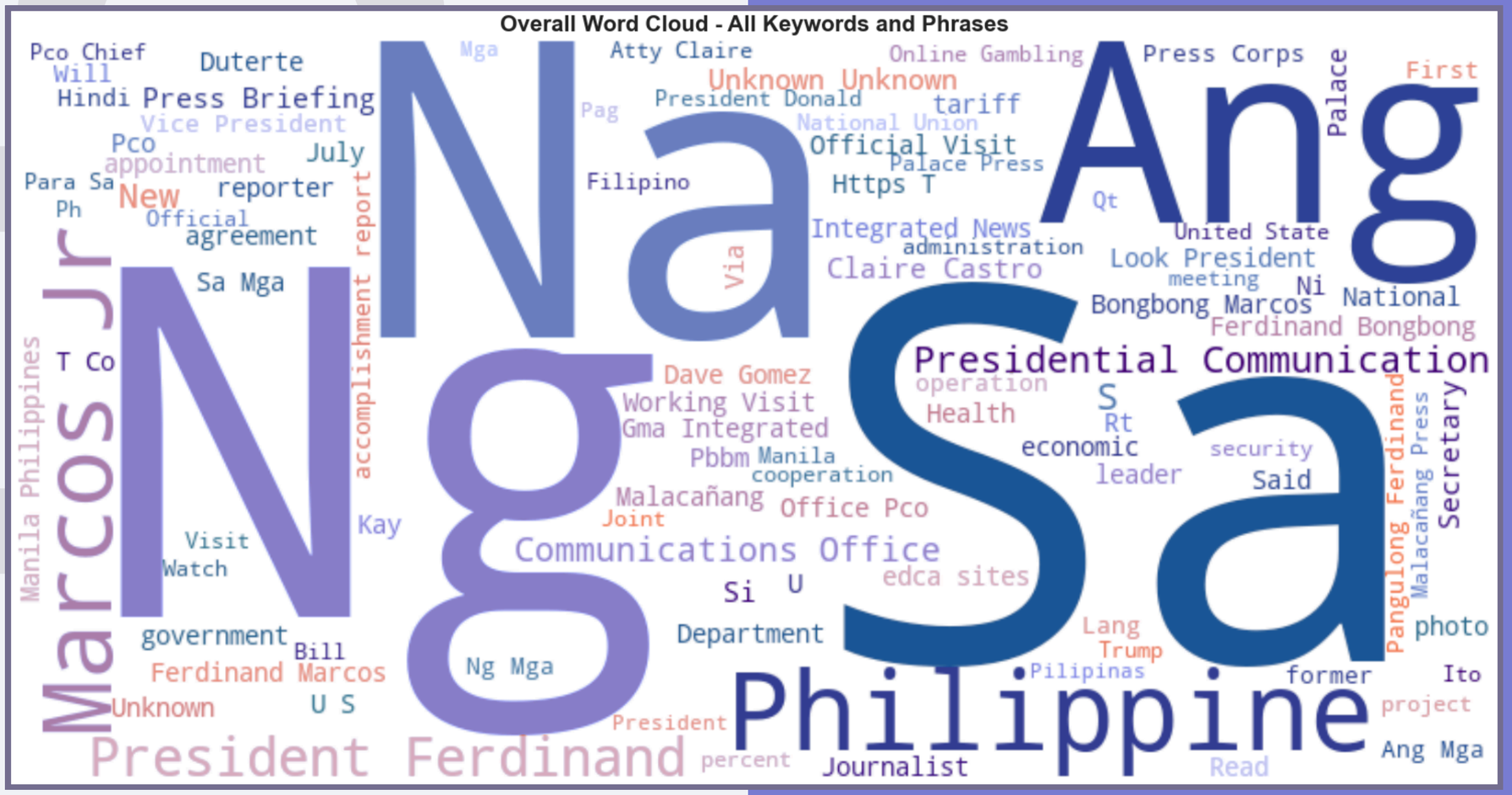
Topic Identification

Content Analysis - Top 10 Keywords in Headline



Word Cloud Visualization

Visual Analysis - All Keywords and Phrases



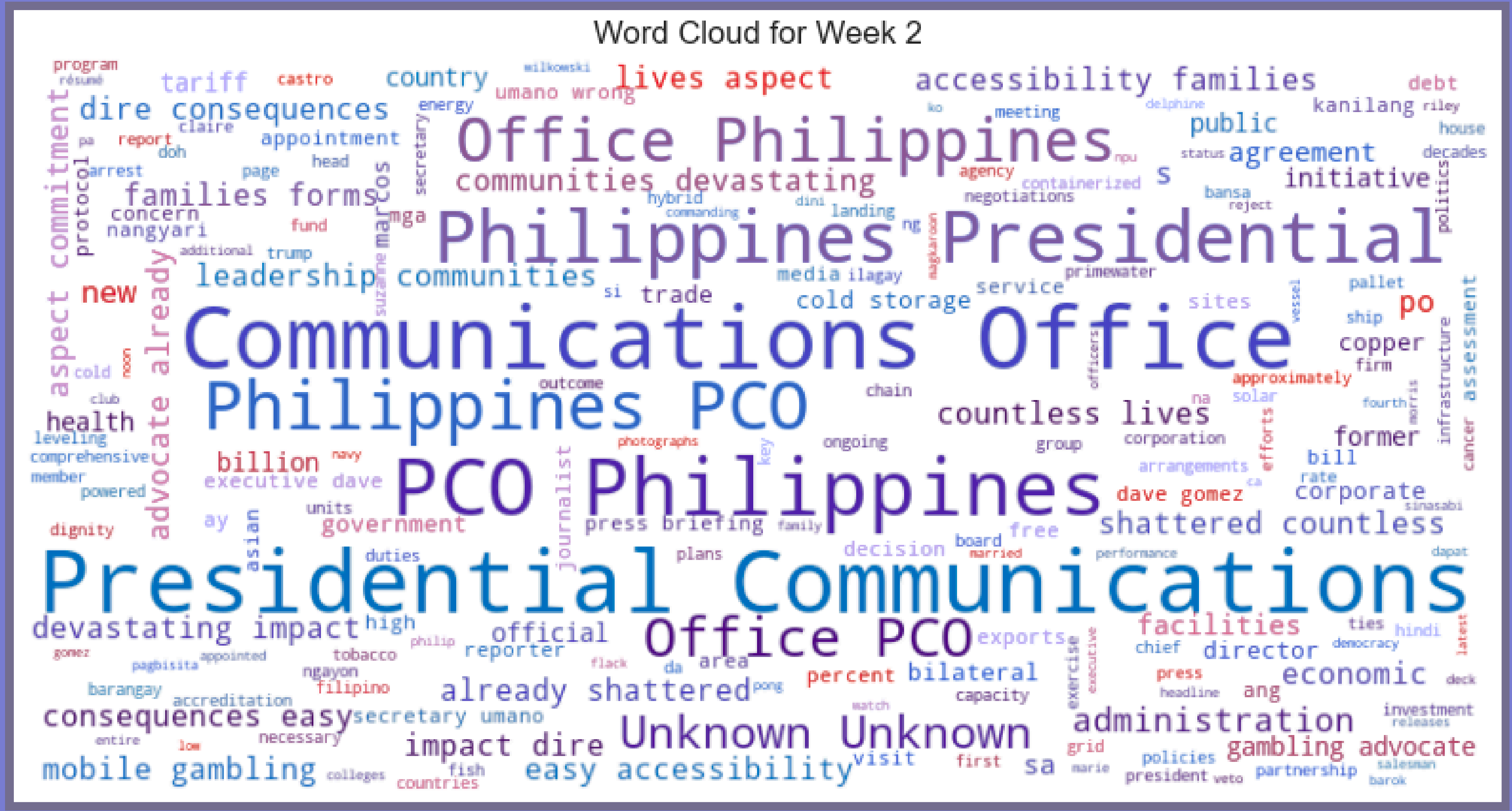
Word Cloud Visualization

Visual Analysis - All Sentiment Categories

Word Cloud Visualization

Visual Analysis - Week 1 vs Week 2

Week 1

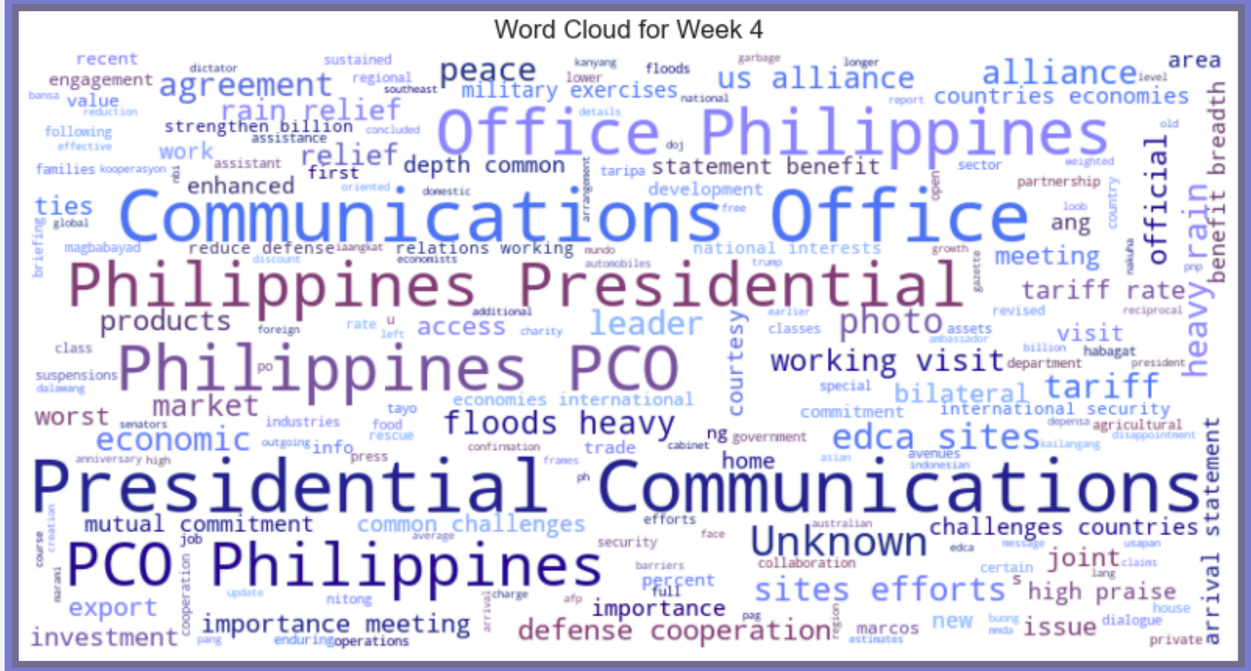


Week 2

Word Cloud Visualization

Visual Analysis - Week 3 vs Week 4

Week 3



Week 4

Word Cloud Visualization

Visual Analysis - Week 5

Week 5

Findings

Which are the top 5 sources by number of stories?

Source	Volume
1. Facebook	137
2. Twitter/X	97
3. YouTube	69
4. Blogs	24
5. Forums	15

Findings

Which country has the most positive sentiment stories?

Country	Volume
1. United States	50
2. Philippines	35
3. Taiwan 3. Netherlands	1

Findings

What's the most common keyword/topic?	Volume
Philippines, Presidential Communications Office	362
Which source has the highest average reach?	**Reach**
Forums	1,224,157
Facebook	1,024,418

SUMMARY INSIGHTS

The Presidential Communications Office (PCO) generated 585 mainstream media (MSM) and 362 social media (SM) coverages in July 2025. Among MSM outlets, Philippine Daily Inquirer accounted for the largest share at 10% (60 articles). On social media, Facebook emerged as the dominant platform with 38% (137 posts), driving much of the critical discourse—particularly surrounding the President's silence on the issue of renewing the Philippines' membership in the ICC.

The Philippines recorded the highest number of positive sentiment stories across both MSM and SM, driven largely by coverage of Communications Secretary Dave Gomez's directive to conduct a performance audit of PCO officials—a move seen as strengthening the government's communication strategy and rebuilding public trust.

SUMMARY INSIGHTS

“Philippines” stood out as the most frequently discussed topic across both media types. Coverage peaked in Week 3 for MSM and in Week 2 for SM, reflecting differences in reporting timelines and public engagement patterns.

In terms of reach, MSN led MSM platforms, with notable coverage of Undersecretary Claire Castro’s call to respect the Supreme Court’s decision to withhold the Vice President’s impeachment trial, while underscoring that a new impeachment complaint could still be filed starting February next year. On social media, Facebook again registered the highest average reach, driven by discourse on the President’s continued silence over ICC membership renewal.

DATA CLEANING DIFFERENCE

DATA CLEANING STEPS	MSM	SM	DIFFERENCE
Load and Deduplicate	Drops duplicates by Source Link, removes rows missing, Date, Headline, Source, Country	Drops duplicates by Source Link, removes rows missing Date, Source, Country	MSM requires Headline to exist; SM allows it to be filled later.
Fix Missing Headline	Keeps rows only if Headline exists	Uses Opening Text as Headline if original is missing/blank	SM is more flexible in filling missing headlines
Date Standardization	Uses Date or Alternate Date Format, cleans into MM-DDYYYY format	Same logic	No major difference

DATA CLEANING DIFFERENCE

DATA CLEANING STEPS	MSM	SM	DIFFERENCE
Source Normalization	Removes www, .com/.ph/.org/.net; maps known acronyms (ABS, CBN, MSN, PTV)	Same cleaning, but: if contains "reddit" → grouped as "Forums"	SM groups Reddit differently, MSM keeps acronyms.
Headline Normalization	Lowercase → remove punctuation → trim → remove duplicate headlines	Same logic	No major difference
Fill Missing Text Fields	Fills Opening Text, Influencer, Key Phrases with "Unknown"	Fills Opening Text, Influencer, Key Phrases, Language, Sentiment with "Unknown"	SM has extra fields (Language, Sentiment)

DATA CLEANING DIFFERENCE

DATA CLEANING STEPS	MSM	SM	DIFFERENCE
Numeric Fields	Converts Twitter/Facebook/Reddit Social Echo to integers, computes Social Echo Total	Same, but also cleans Reach column	SM includes Reach cleaning
Filtering Unwanted Rows	Removes rows with "content from this publisher", "test", "proquest", empty links, and [Courtesy:]	Same filtering rules	No major difference
Drop Columns	Removes metadata columns (social stats, IDs, tags, etc.)	Same logic	Almost identical

DATA CLEANING DIFFERENCE

DATA CLEANING STEPS	MSM	SM	DIFFERENCE
Final Text Cleaning	Country → title case. - Headline → title case. - Opening Text/Hit Sentence → capitalized	Same process	No major difference
Reorder Columns	Moves Clean Date to the first column	Same process	No major difference
Key Phrases and Keyword Column Standardization	Combine the values in the Key Phrases and Keywords columns into a single standardized word, treating them the same regardless of whether they’re in lowercase, uppercase, or mixed case.	Same logic	No major difference

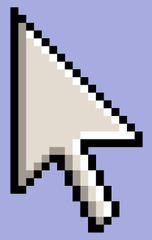
DATA CLEANING DIFFERENCE

DATA CLEANING STEPS	MSM	SM	DIFFERENCE
Country and Language Column Text Data Imputation	Fill in only the blank cells in the Country and Language columns using the language detected from the Hit Sentence, without changing the existing values.	Same logic.	No major difference
Output	Saves as MSM_cleaned.xlsx.	Same logic	No major difference


QUALITY CONTROL
CHECKLIST

Sampling and Spot Checks

Check rows for formatting, alignment, and key details.



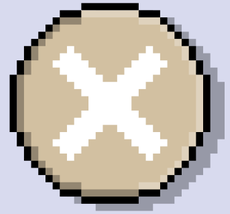
Edge Cases

Ensure missing or duplicate data is handled, dates are valid, and special characters display correctly.

Consistency

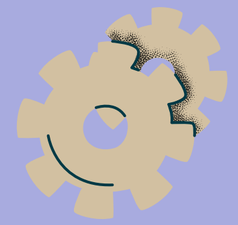
Ensure data structure, validity, and consistency are maintained.

Error Handling



Logs errors and warnings clearly, ensures no rows are skipped silently, and captures exceptions with sufficient detail.

Automation and Reproducibility



Script produces consistent results, edge cases pass, and data-handling assumptions are documented.

Reporting Issues

Log issues with examples, categorize by severity, and escalate unresolved ones before sign-off.

Limitations

Accuracy of Results

Results depend on the quality of raw data; gaps, errors, and bias may still remain despite cleaning.

Rule-Based Constraints

Rule-based cleaning may miss anomalies, and Python libraries have performance and interpretability limits

Text Ambiguity

Text data is ambiguous; sentiment analysis cannot fully interpret tone or sarcasm.

Scalability Challenges

Scaling to larger datasets may pose performance challenges without further optimization.

Recommendations

1

Prioritise Vectorised Operations

Prioritise vectorised operations in Pandas, as they replace Python loops with optimised C-based processes, greatly speeding up repetitive data-cleaning tasks and improving scalability (Compile N Run, n.d.).

2

Employ Data Type Optimisation

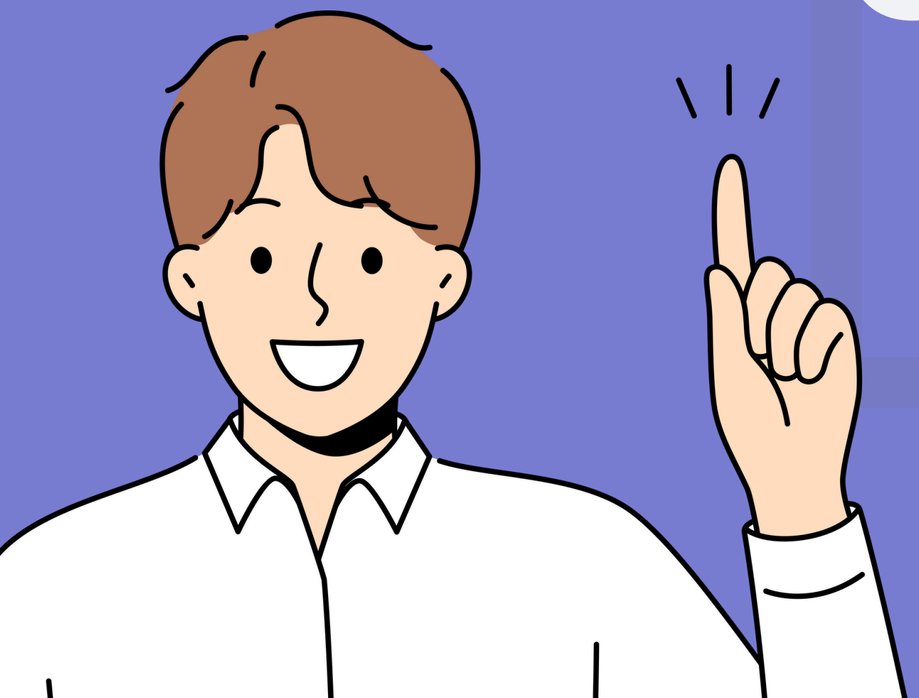
Downcasting numerical columns (e.g. float64 to float32) and converting repeated-value object columns to categorical types can reduce memory usage by up to 80% without affecting accuracy (GeeksforGeeks, 2023a).

3

Utilise Chunk-Based Processing for Large Datasets

Use chunk-based processing (e.g. pd.read_csv(..., chunksize=100000)) to handle large datasets efficiently, reducing RAM use and enabling scalable cleaning of millions of records (GeeksforGeeks, 2023b).

Conclusion



This project successfully demonstrated the capability of the team’s data pipeline to process, clean, and analyze information from multiple sources. The code performed reliably, executing the workflow from data ingestion to visualization with consistency. It effectively managed duplicate removal, standardization, and dataset separation, while also producing sentiment analyses and keyword visualizations. Although limitations remain, particularly with textual ambiguity and scalability for larger datasets, the test confirms that the methodology is **sound** and **provides a strong foundation for future refinement and broader application**.

Team Members

Python Development Team

Chester Mikhail De Guzman
SG-Reports

Patricia Marie Dizon
SG-Reports

Coreen Jois Manay
SG-Reports

Hazel Grace Juaton
SG-Reports

Python Development Team

Ann Relado
SG Government

Sophia Loren Hernandez
M2

Julia Soriano
AU-Newsletter

Visualisation/ Deck Team

Visualisation/ Deck Team

Jan Borce

AU-Newsletter

Lea Murai Sebastian

SG Government

The Presenters

Martin Luigi de Vera
SG Government

Trisha Mae Gubalane
SG Government

Research Team

Isaac Mae Regular

EMEA NL

Arnel Oclinaria

SG Government

The Coordinators

Jan Borce

AU-Newsletter

Lea Murai Sebastian

SG Government


THANK
YOU
LORD!!