Challenge 7 Documentation: Export Results with Enhanced Reporting

Process & Implementation

Step 1 – Data Cleaning

- Removed duplicate rows (based on `URL` if available).

- Dropped empty rows.

- Converted `Date` columns to a consistent datetime format.

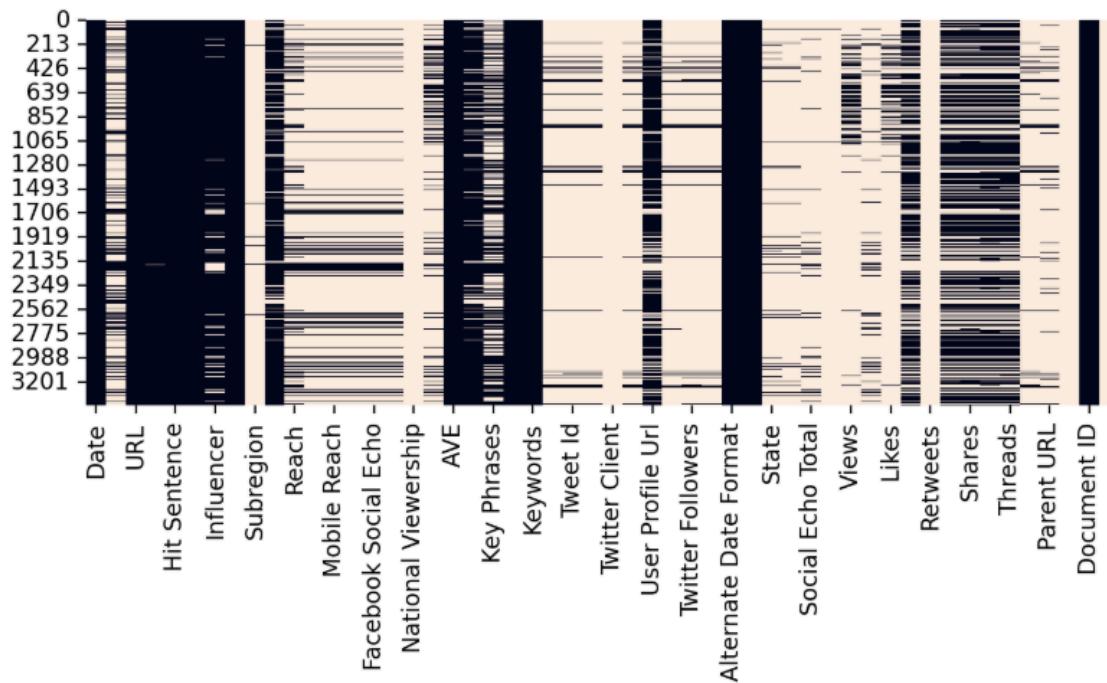- Created a summary report showing rows before and after cleaning.

Step 2 – Data Exploration

- Displayed dataset shape and a sample of rows.

- Visualized missing values using a **heatmap** to identify data quality issues.
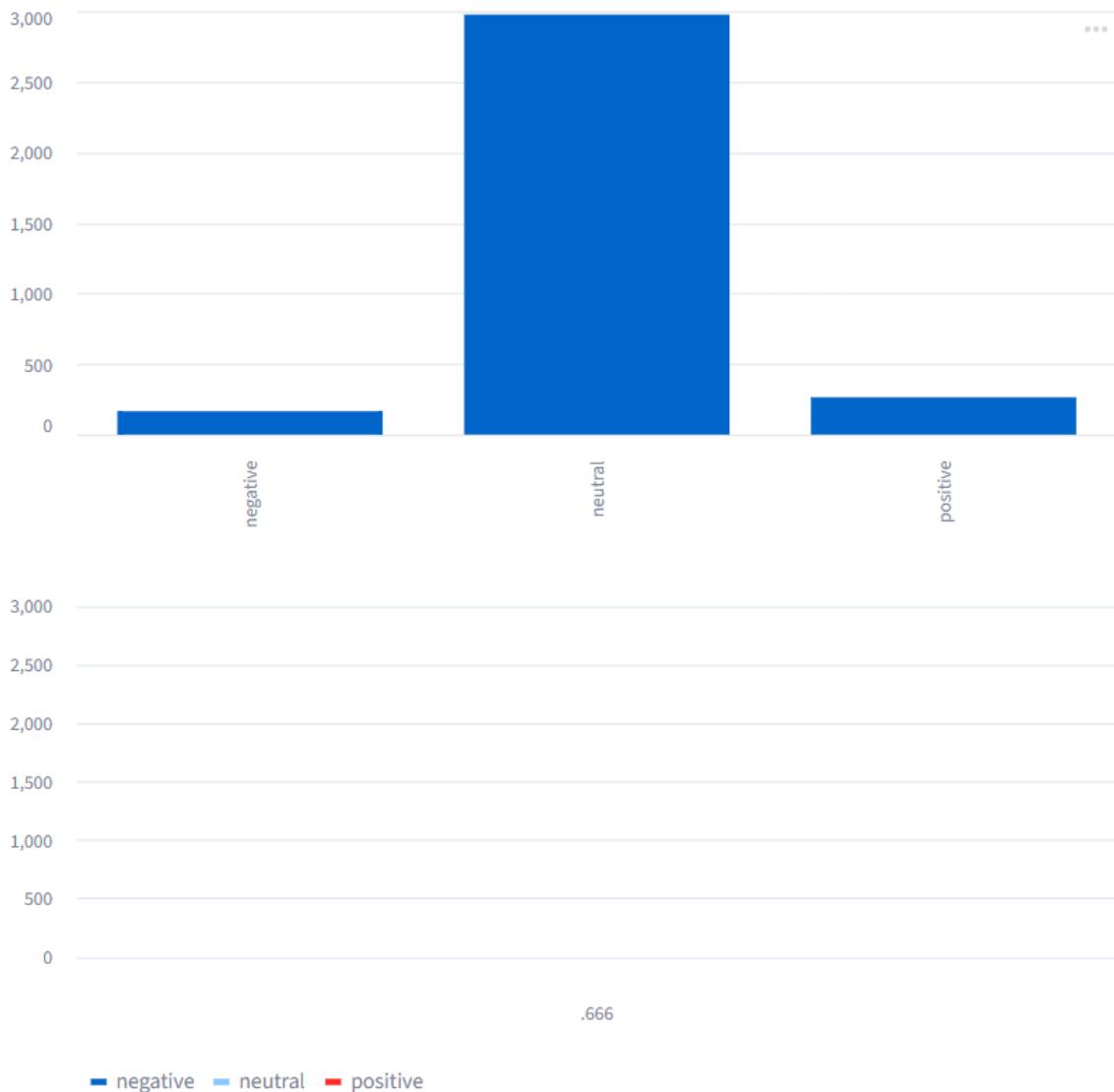
🔍 Step 2: Data Exploration

Shape: `(3405, 52)`

	Date	Headline	URL
0	2025-07-31 23:39:00	None	https://www.tiktok.com/@xirb87/video/75332
1	2025-07-31 23:33:00	Kelra heroics propel ONIC to MSC semis	https://tiebreakertimes.com.ph/tbt/kelra-her
2	2025-07-31 23:26:00	None	https://www.tiktok.com/@nongdarrrrr17/vide
3	2025-07-31 23:25:00	None	https://www.tiktok.com/@ronsfever/video/75
4	2025-07-31 23:07:00	None	https://www.tiktok.com/@aungaunga28/vide

Step 3 – Sentiment Analysis

- Selected text fields (Reviews , Headline , or Opening Text).

- Used **TextBlob** to assign sentiment labels (positive , neutral , negative).

- Displayed results as a bar chart and a timeline chart (sentiment over months).

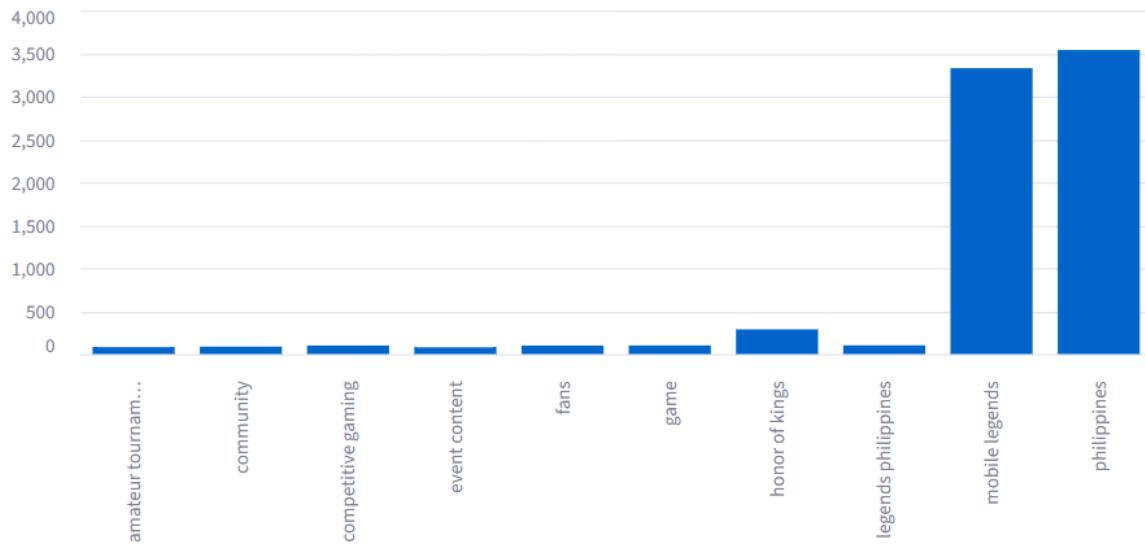## 😊 Step 3: Sentiment Analysis 🔗

Step 4 – Keyword Analysis

- Extracted keywords from the dataset (`Key Phrases` or `Keywords` columns).
- Counted frequency and visualized the **Top 10 keywords** with a bar chart.

🔑 Step 4: Keyword Analysis

Step 5 – Word Cloud

- Combined keyword data into a single text corpus.
- Generated a **Word Cloud** visualization to highlight frequently mentioned terms.

☁️ Step 5: Word Cloud

Step 6 – Business Insights

- Computed quick insights as metrics:
    - 🏆 Top Source
    - 🔑 Most Common Keyword
    - 📅 Busiest Month
    - 📡 Highest Average Reach

📊 Step 6: Business Insights

🏆 Top Source
Tiktok
↑ 2138

🔑 Top Keyword
Philippines
↑ 3201

📅 Busiest Month
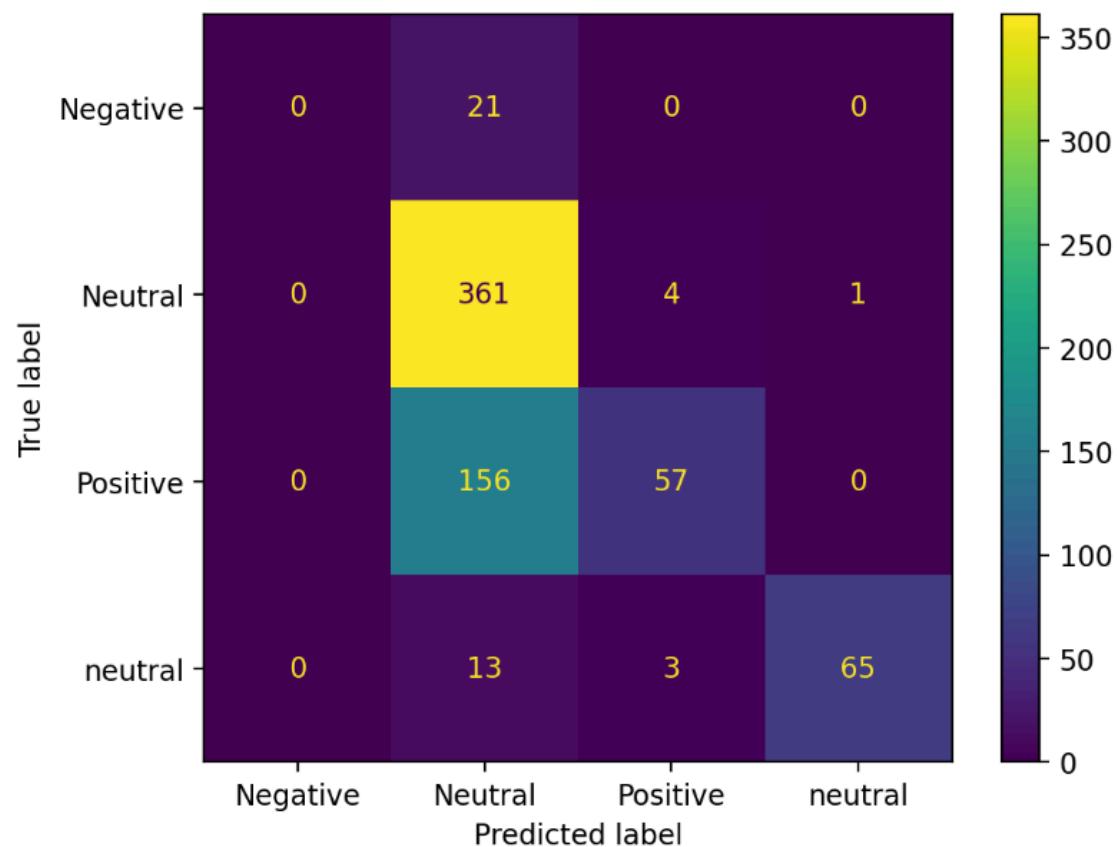**2025-07**
↑ 3405

📡 Highest Avg Reach
MSN.com
↑ 101199920.0

Step 7 – Predictive Modeling

- Built a **Logistic Regression** model using TF-IDF vectorization.
- Trained it to predict sentiment labels from text.
- Evaluated model with a **confusion matrix** and **accuracy score**.
- Enabled interactive testing with custom user input.

⊚ Step 7: Predictive Modeling

Model Accuracy

70.93%

Step 8 – Final Export (Core of Challenge 7)

- Added **multi-format export**:
  - CSV (`filename_cleaned.csv`)
  - Excel (`filename_cleaned.xlsx`)

↓ Step 8: Download Data

	Date	Headline	URL
0	2025-07-31 23:39:00	None	https://www.tiktok.com/@xirb87/video/7533271096259349768?utm_campa
1	2025-07-31 23:33:00	Kelra hero	https://tiebreakertimes.com.ph/tbt/kelra-heroics-propel-onic-to-msc-semis,
2	2025-07-31 23:26:00	None	https://www.tiktok.com/@nongdarrrrr17/video/7533267741797584136?utm
3	2025-07-31 23:25:00	None	https://www.tiktok.com/@ronsfever/video/7533267590412504327?utm_can
4	2025-07-31 23:07:00	None	https://www.tiktok.com/@aungaunga28/video/7533262892074831111?utm_
5	2025-07-31 23:01:00	None	https://www.tiktok.com/@nongdarrrrr17/video/7533261241180081426?utm
6	2025-07-31 22:51:00	None	https://www.tiktok.com/@ronsfever/video/7533258681043995912?utm_can
7	2025-07-31 22:42:00	None	https://www.tiktok.com/@aiesports64/video/7533256400164670727?utm_c.
8	2025-07-31 22:32:00	None	https://www.tiktok.com/@jayyytamayo22/video/7533253898149088530?utn
9	2025-07-31 22:21:00	None	https://www.tiktok.com/@dr.d.999/video/7533251015571787026?utm_camp

[Download Cleaned Data (CSV)]

[Download Excel Report]

- Excel contains **3 sheets**:

  - *Cleaned Data* → final dataset after preprocessing

  - *Summary* → key metrics (row counts, sentiment distribution, top keyword, model accuracy)

---

3. Technical Improvements

- Used `pandas.ExcelWriter` + `openpyxl` to manage **multi-sheet Excel reports**.

- Dynamic filenames: exports use the original uploaded dataset name + `_cleaned`.

  - Example: `sales_data.xlsx` → `sales_data_cleaned.xlsx`

---

4. Outcome

- Stakeholders receive a **single Excel report** containing:

  - Clean, deduplicated dataset

- - Summarized metrics and business insights
- Judges can evaluate the **end-to-end workflow** without touching code.
- This step transforms raw analysis into a **professional deliverable**.

5. Significance

This approach bridges the gap between **technical analysis** and **business usability**:

- Analysts can run interactive exploration in Streamlit.
- Non-technical users can download polished Excel reports.
- Ensures reproducibility, clarity, and accessibility of results.

TRY IT OUT!

If you want to test the application, you may download the code directly from our GitHub repository:

RDB-Hackathon-Team-9. Detailed instructions for installation and running the Streamlit app are provided in the README file.