# STREAMING PLATFORMS

GROUP 5

RDB Media Analytics Hackathon 2025


# INTRODUCTION

**Overview**

Our group developed a comprehensive data analysis solution for streaming platform media coverage, processing 748 stories from July 2025. The project demonstrates end-to-end data pipeline capabilities from cleaning raw data through advanced analytics to actionable business insights.

**Challenge Scope**

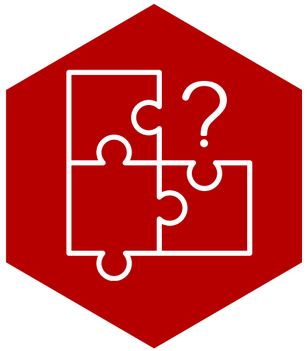
We tackled **eight interconnected challenges:** data cleaning, sentiment analysis, trend visualization, topic identification, word cloud visualization, business analysis, exporting results, and performance optimization. Our solution emphasizes scalability, accuracy, and practical business value.

# CHALLENGE 1
# DATA CLEANING

## Methodology

### Deduplication

- Removed duplicate entries based on URL and Headline
- Reduced dataset size while preserving unique content



### Missing Value Handling

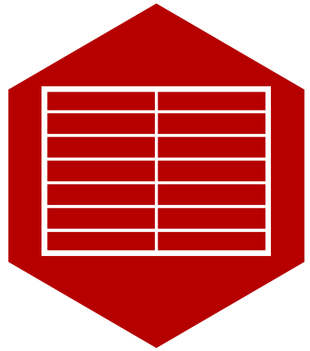
- Identified key fields: Release Date, Headline, Source, Country
- Dropped rows with missing values in these critical fields

### Data Standardization

- Parsed multiple date formats
- Added Year, Month, Day columns for easier analysis
- Filtered out rows with invalid dates

### Source Name Normalization

- Applied standardized naming rules
- Used fuzzywuzzy library for string matching



### Column Organization

- Structured columns logically for analysis
- Added derived columns (e.g., TextBlob Sentiment)
- Exported cleaned data to CSV

# CHALLENGE 2
# SENTIMENT ANALYSIS

## Methodology

### Automated Sentiment Scoring

- Used TextBlob for natural language processing
- Applied polarity thresholds:

> 0.1: Positive
< -0.1: Negative
Otherwise: Neutral

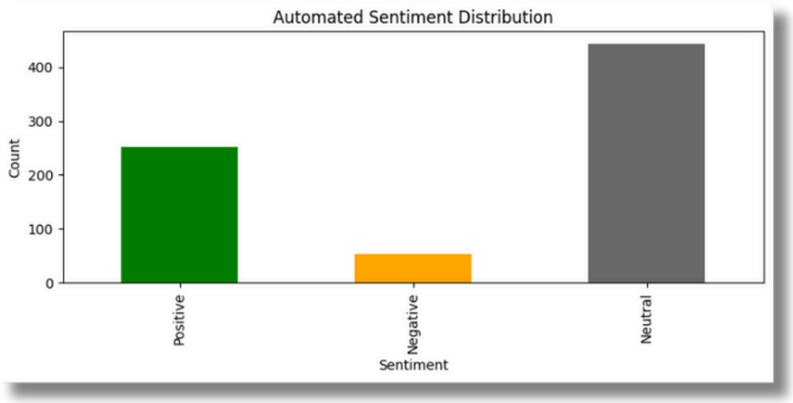
### Validation

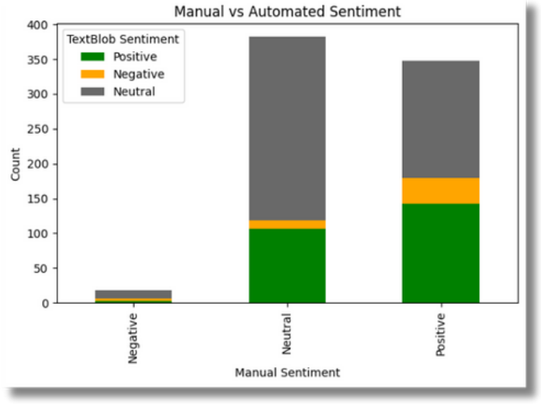
- Compared automated vs manual sentiment (where available)
- Visualized sentiment distributions
- Tracked sentiment trends over time

### Analysis Dimensions

- Temporal: Monthly trends
- Geographic: Country-level sentiment
- Source-based: Platform Comparisons

## Automated vs. Manual Sentiment Distribution

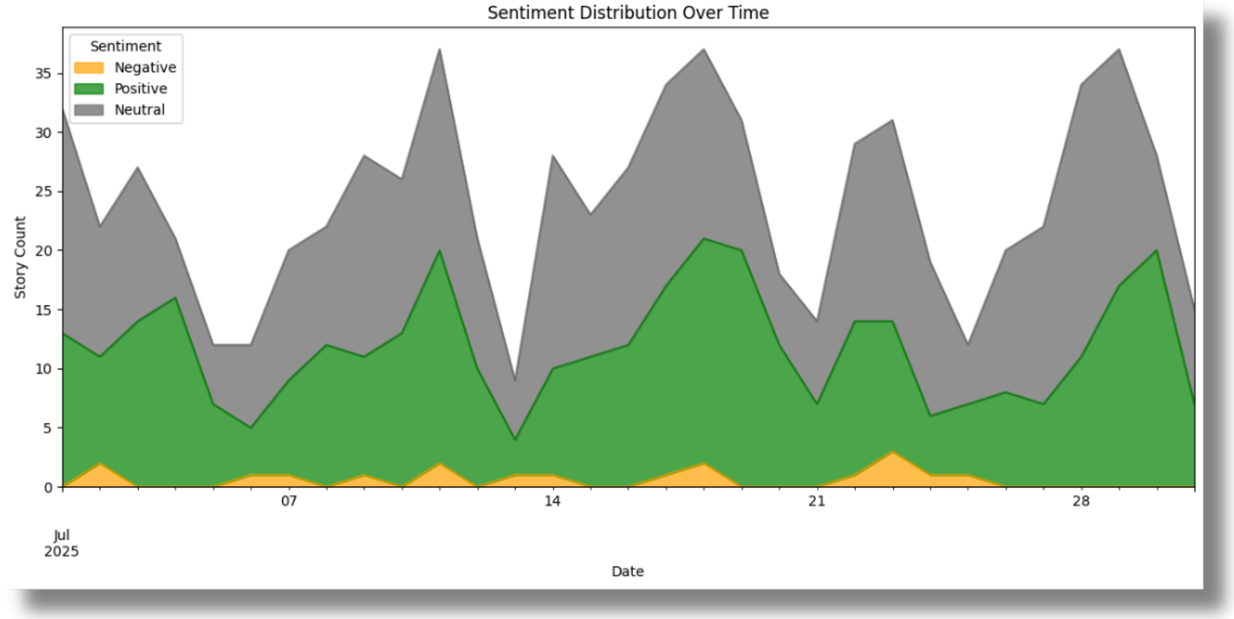



**Distribution Breakdown:** 51.1% neutral, 46.5% positive, only 2.4% **negative coverage**
**Key Findings:** Automated sentiment aligned well with manual

# CHALLENGE 3
# TREND VISUALIZATION

## Sentiment Distribution




**Dominant Neutral Tone:** Neutral sentiment consistently represented the largest portion of coverage throughout the month

**Stable Distribution:** The proportion of positive vs negative vs neutral content remained relatively consistent over time

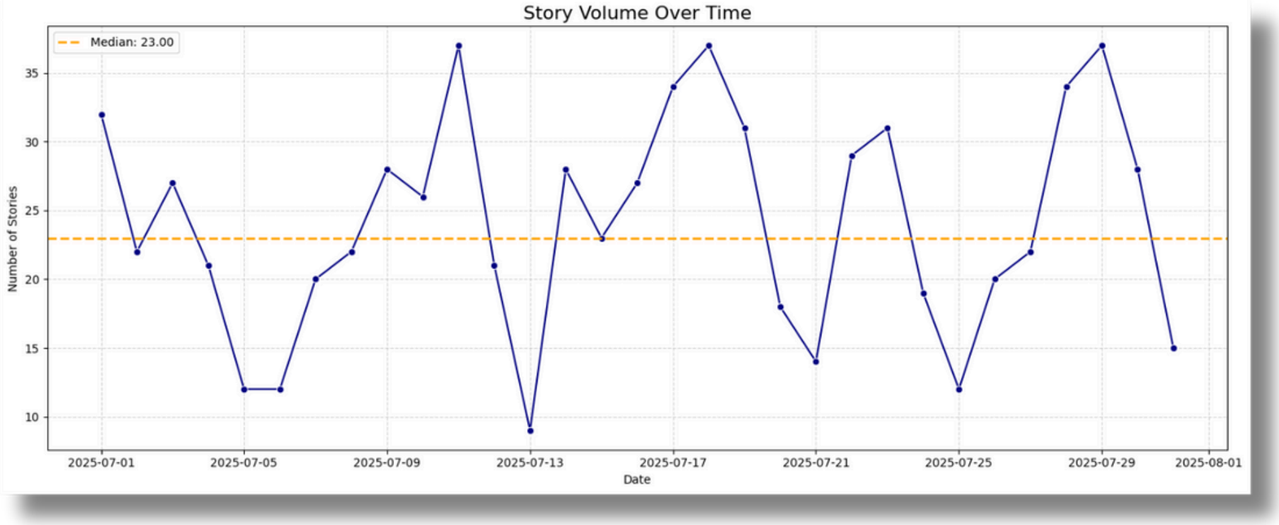
**Limited Negative Coverage:** Negative sentiment stories were consistently the smallest segment (only 2.4% overall)

**No Major Sentiment Shifts:** Unlike news cycles that might show dramatic sentiment swings, streaming platform coverage maintained steady sentiment

# CHALLENGE 3
# TREND VISUALIZATION

## Story Volume

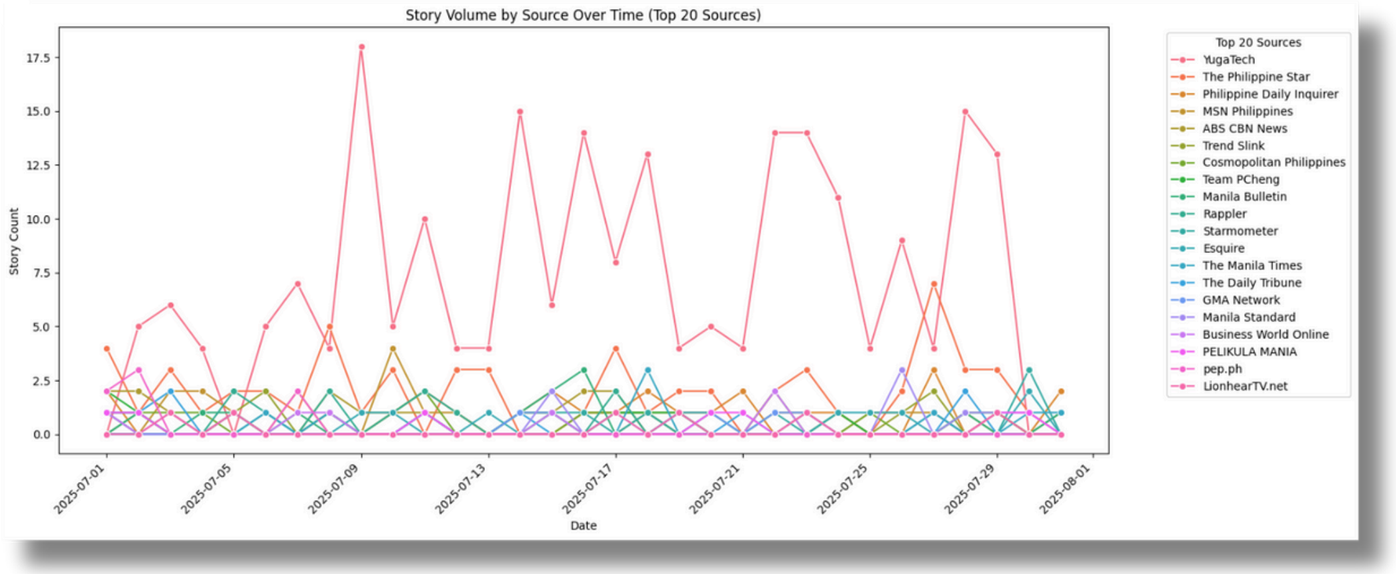
In July 2025, story publishing showed **irregular spikes** with no consistent daily schedule. The median daily story count provided a baseline, with several days exceeding this norm due to major events or announcements. While peak activity suggested coordinated coverage, other periods fell below the median, highlighting **potential opportunities for more consistent content planning.**



# CHALLENGE 3
# TREND VISUALIZATION

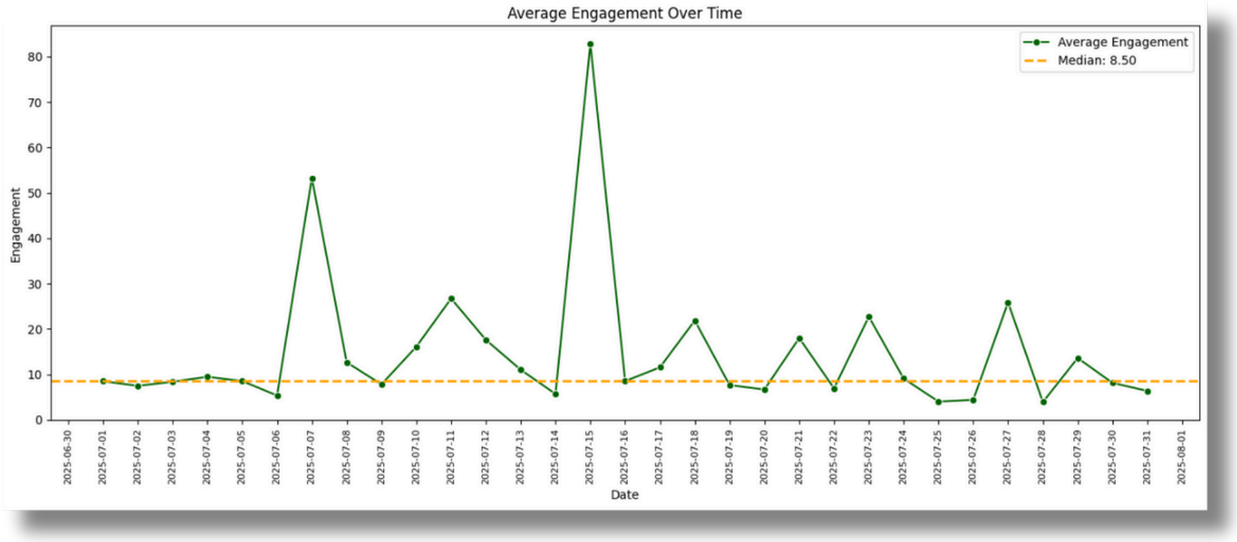
## Story Volume by Source

The source comparison showed that **YugaTech dominated** output with 225 stories, far surpassing other outlets. A clear tier structure emerged, with Philippine Star (61 stories) and Philippine Daily Inquirer (25 stories) leading the mid-level group, while smaller outlets published far less.



# CHALLENGE 3
# TREND VISUALIZATION

## Average Engagement



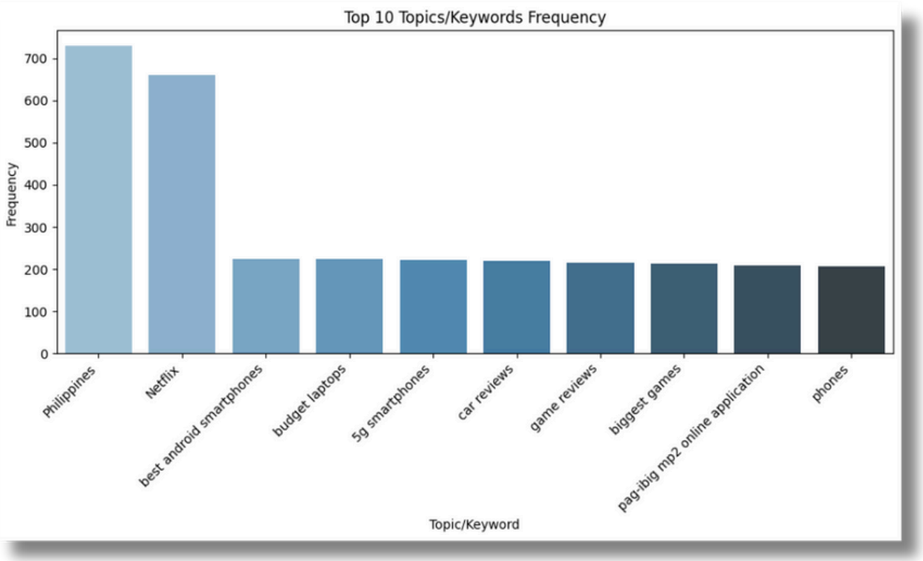

**Low Overall Engagement:** Average daily engagement remained relatively low throughout the observation period

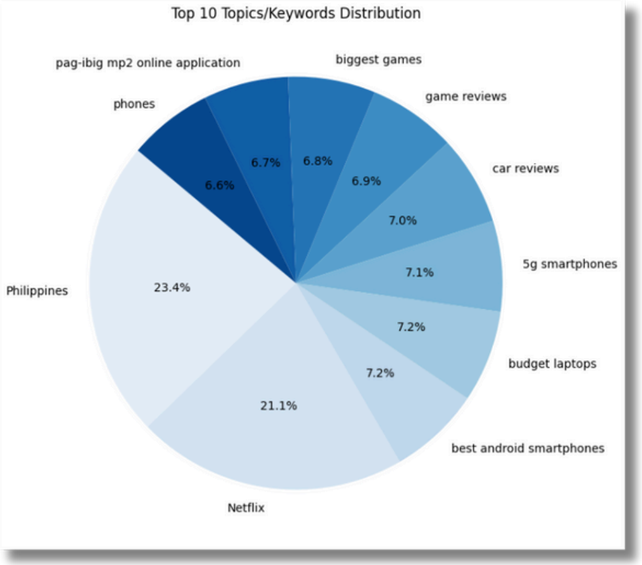
**Sporadic Spikes:** Engagement showed occasional peaks but no sustained high-performance periods

**Median Performance:** Most days fell near or below the median engagement level

**Volume-Engagement Disconnect:** **High story volume days didn't necessarily correspond to high engagement days**, suggesting content quality or timing factors influence audience response more than quantity

# CHALLENGE 4
# TOPIC IDENTIFICATION



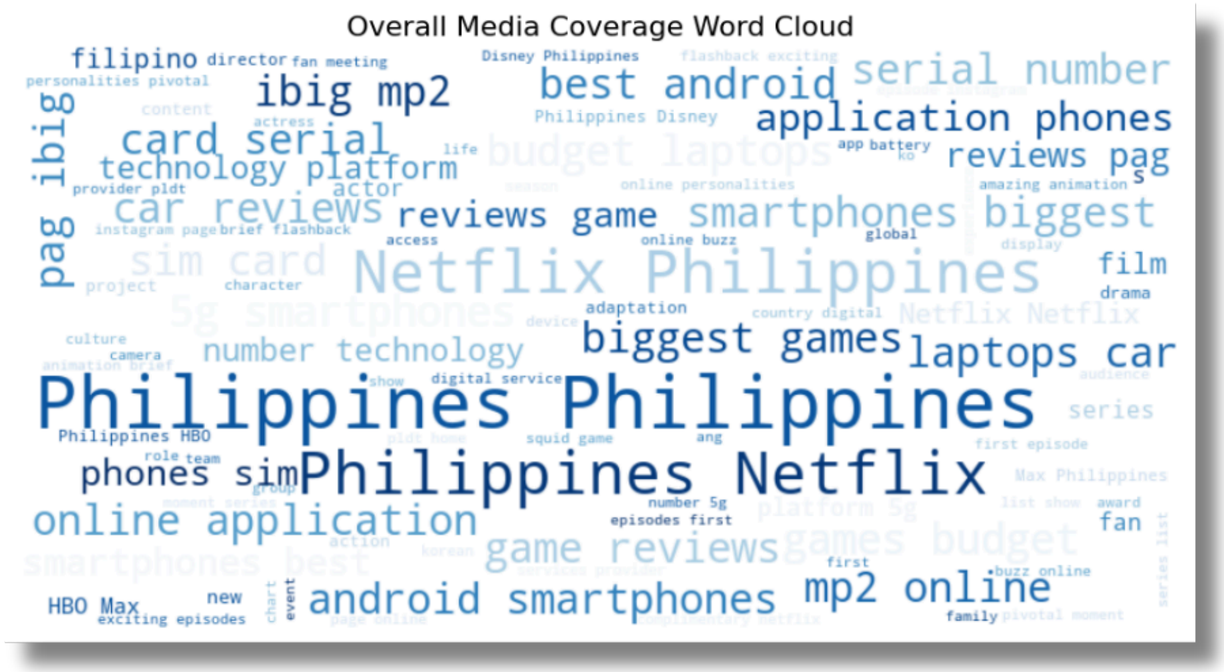




Content analysis reveals that **Philippines** and **Netflix** dominated content mentions. Coverage emphasizes consumer electronics, including **smartphones**, **laptops**, and **gaming**, clustered around **product reviews**, **tech comparisons**, and regional entertainment. The top 10 keywords account for a large share of mentions, reflecting concentrated topic focus, while Netflix leads other streaming platforms in visibility. Overall, the prevalence of product-related terms highlights a substantial commercial and review-oriented content strategy tailored to a localized audience.

# CHALLENGE 5
# WORD CLOUD VISUALIZATION

## Overall Media Coverage

Word cloud analysis showed **“Netflix”** and **“Philippines”** as consistently dominant across all sentiments, with recurring mentions of consumer technology such as **smartphones**, **laptops**, and **gaming**.



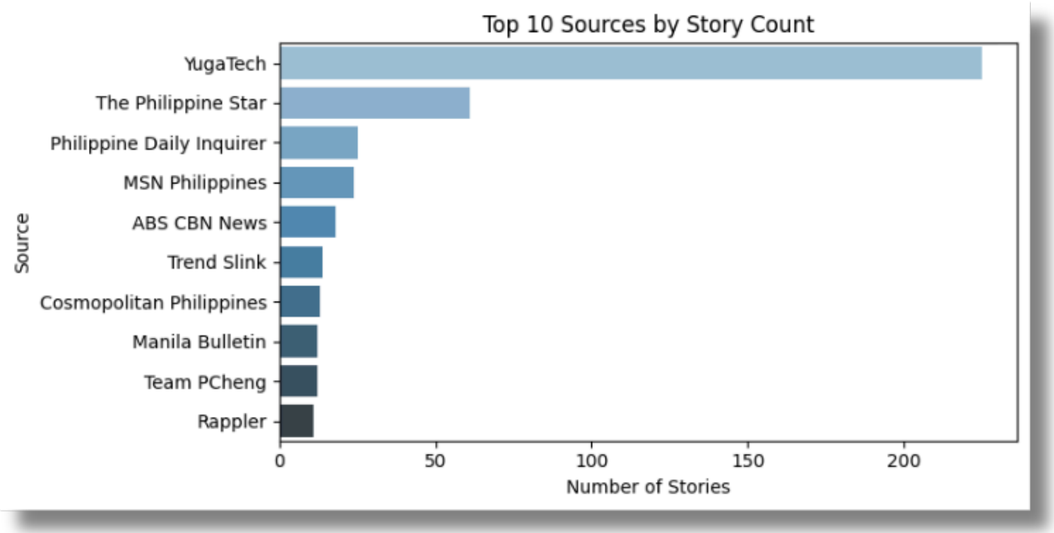
# CHALLENGE 5
# WORD CLOUD VISUALIZATION

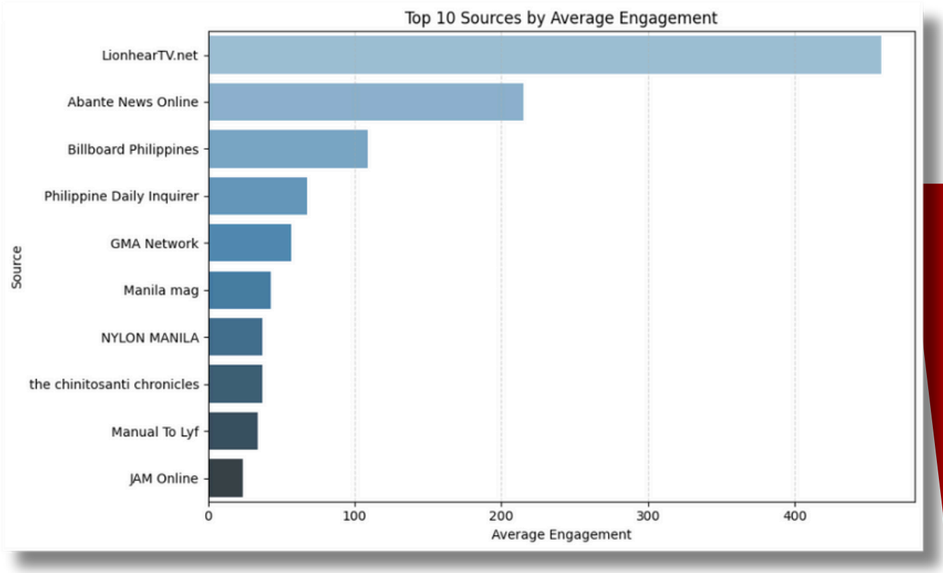
## Word Cloud per Sentiment




Positive, negative, and neutral clouds reflected **similar keyword patterns**, though the negative set had fewer distinct terms due to its smaller sample size (18 stories). Overall, the visuals highlighted **homogenous topic coverage with limited diversity**, heavily focused on streaming platforms and technology brands.

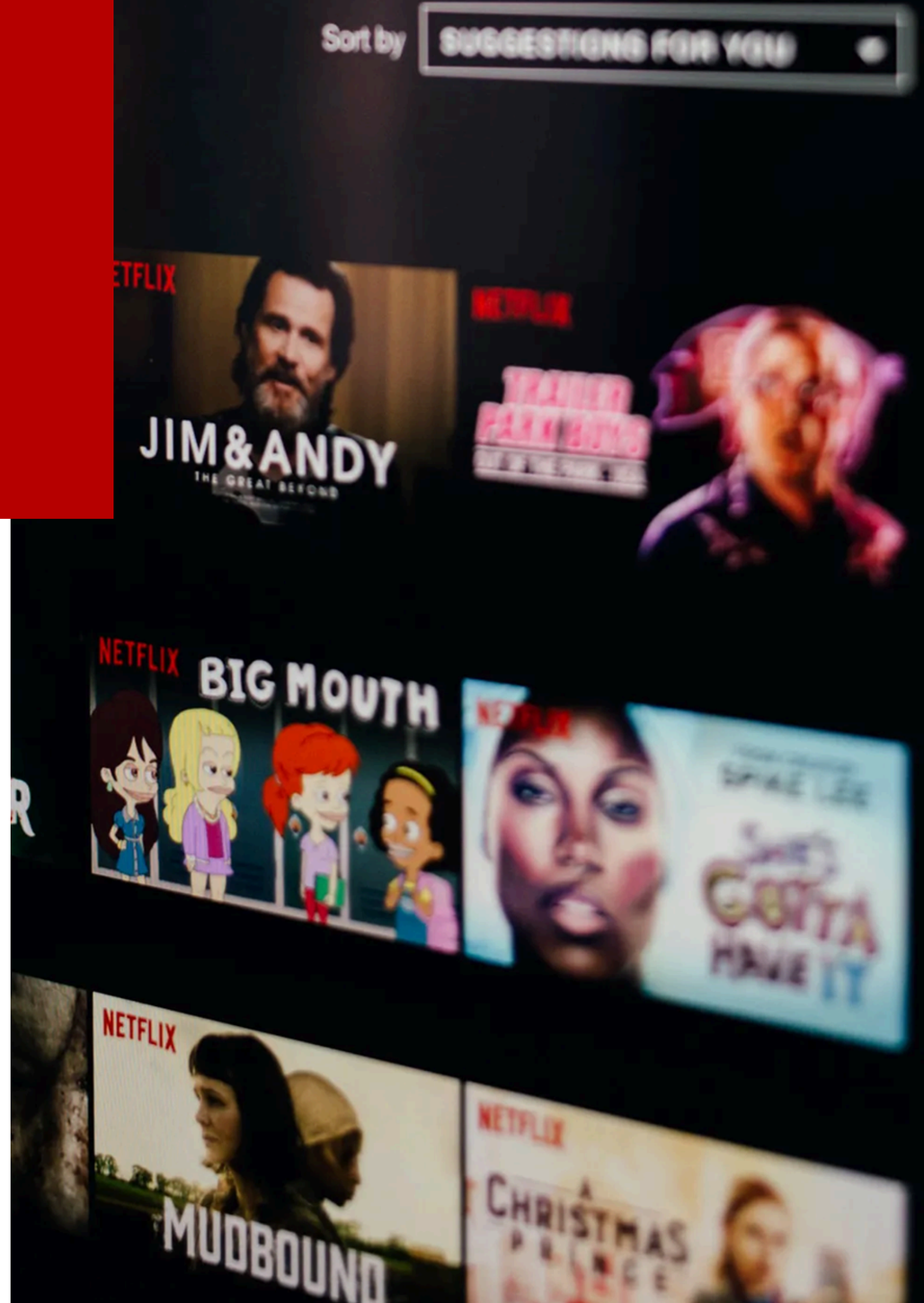
# CHALLENGE 6
# SUMMARY INSIGHTS




**YugaTech** led with 225 stories, accounting for **30% of total coverage**, but volume did not equate to impact as **LionhearTV.net** achieved the **highest engagement** (459 average) **despite lower output.** Coverage was heavily **concentrated in the Philippines**, which also led to positive sentiment, while July 2025 marked the dataset's peak activity. A sharp drop from YugaTech to the second-ranked source (61 stories) underscored market fragmentation, with many high-volume publishers struggling to generate strong engagement, highlighting clear **quality-versus-quantity gaps.**

# CHALLENGE 7
# EXPORTING OF RESULTS



# CHALLENGE 8
# PERFORMANCE OPTIMIZATION

## Efficiency Gains

Reduced dataset from 4,138 to 748 relevant rows through targeted filtering

## Memory Optimizatfion

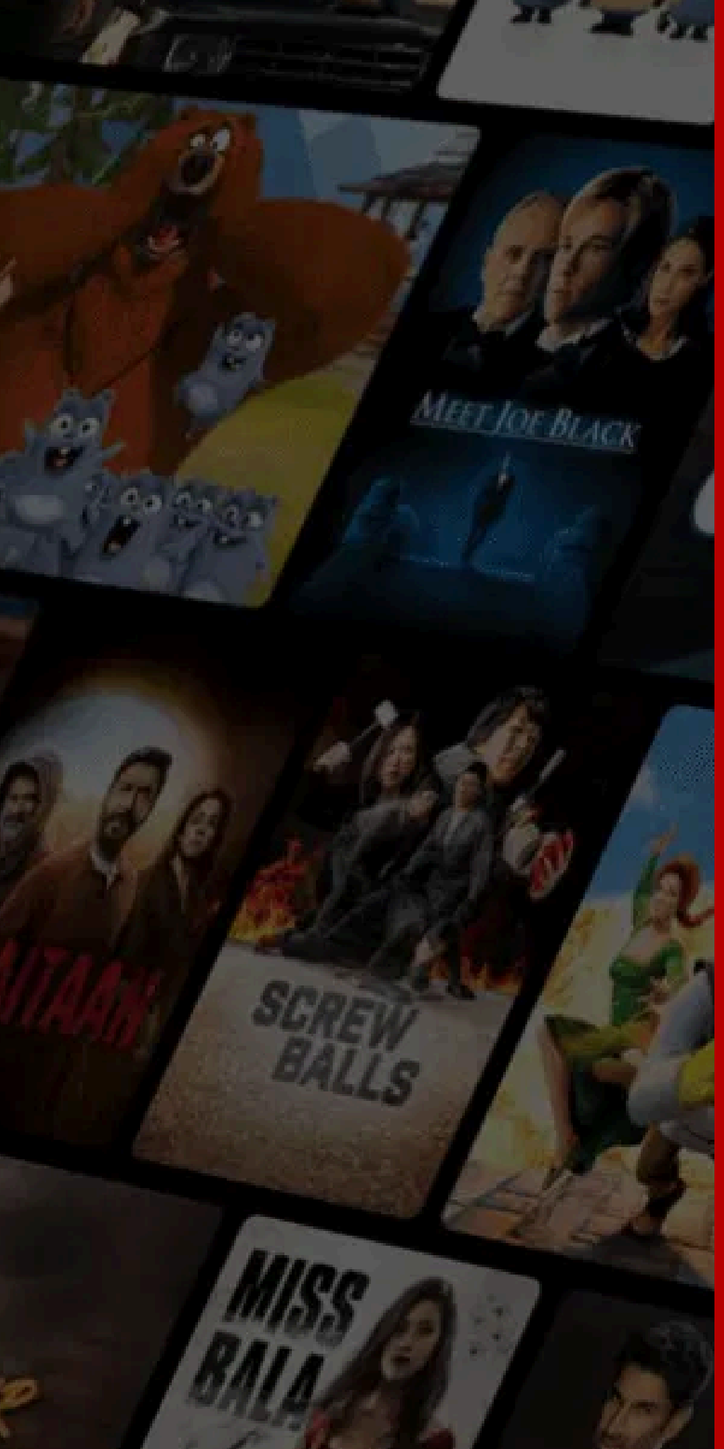
Chunked processing eliminated memory bottlenecks for large file handling

## Processing Speed

Aggregation operations completed in under 0.01 seconds after optimization

## Scalability Proof

Optimization techniques demonstrated readiness for much larger datasets




# RECOMMENDATIONS FOR IMPROVEMENT

## Data Collection

- Add more granular engagement metrics
- Include content categories/tags
- Can include geographical location by city, region, etc.

## Sentiment Data Analysis

- Improve sentiment data and develop more sophisticated sentiment analysis

## Reporting

- Can automate report generation
- Include trend forecasting

# THANK YOU!

RDB Media Analytics Hackathon 2025