V.E.S.
Since 1962

Vivekanand Education Society's
Institute of Technology

Social Media Analytics

Mini Project Report

Sentiment Analysis In Short-Form Videos

Name: Abhishek Pattanayak
Class: D16AD
Roll No.: 48

1. Introduction

1.1 Project Overview

In the current digital age, social media platforms such as Instagram play a critical role in shaping public opinion, brand outreach, and viral trends. With billions of active users, Instagram offers a massive pool of user-generated content that can be mined for insights. This project seeks to analyze Instagram post data to understand the connection between user sentiment and post virality. Specifically, we examine how the emotional tone expressed in captions and comments relates to engagement metrics such as likes, comments, shares, and saves.

By leveraging sentiment analysis techniques and data visualization, we aim to derive actionable patterns from the data. These patterns can help influencers, marketers, and content creators optimize their strategies to increase reach and engagement.

1.2 Objectives

1. To extract and analyze sentiment from Instagram post captions and user comments.
2. To evaluate the impact of sentiment on virality indicators such as likes, shares, comments, and saves.
3. To discover how different content strategies (media types, hashtag usage, posting time) influence sentiment and virality.
4. To identify optimal engagement strategies based on data-driven insights.

1.3 Tools and Technologies

Programming Language: Python

Libraries and Packages:

- pandas – for structured data manipulation and transformation.

- numpy – for efficient numerical operations.
- matplotlib, seaborn – for exploratory and presentation-level visualizations.
- nltk, TextBlob, VADER (from nltk.sentiment) – for natural language processing and sentiment scoring.
- scikit-learn – for correlation analysis and feature extraction.

These tools collectively supported the end-to-end workflow: from loading raw data, performing preprocessing, executing sentiment analysis, to visualizing and interpreting results.

2. Data Overview

2.1 Dataset Description

The dataset comprises anonymized Instagram post data collected from public accounts using Instagram's API and third-party scraping tools. The data spans multiple user categories: personal, influencer, and brand accounts.

Key attributes in the dataset:

- Post ID: A unique identifier for each post.
- Caption Text: The textual description added by the user.
- Hashtags Used: List of hashtags embedded within captions.
- Media Type: Type of post – image, video, or carousel.
- Timestamp: Time and date when the post was made.
- Engagement Metrics: Number of likes, comments, shares, and saves.
- Follower Count: Snapshot of the account's follower count at the time of posting.

- Comment Text: Aggregated or sampled top-level comments.
- Computed Sentiment Score: Numerical sentiment analysis score derived from caption and comment text.

2.2 Data Cleaning and Preprocessing

Preprocessing was essential to ensure data integrity and improve analytical outcomes. The steps included:

1. Removing Incomplete Entries: Posts with missing captions or engagement metrics were excluded.
2. Standardizing Timestamps: Converted all timestamps to a common timezone and datetime format.
3. Text Preprocessing:
    a. Lowercased all captions and removed emojis, special characters, and URLs.
    b. Tokenized text and removed stopwords using NLTK.
    c. Applied lemmatization to normalize words for accurate sentiment classification.
4. Sentiment Computation:
    a. Used TextBlob for polarity scoring: score < 0 = negative, 0 = neutral, > 0 = positive.
    b. Cross-validated with VADER sentiment scores for robustness.
    c. Combined comment and caption sentiment to derive a composite sentiment metric.

3. Data Analysis

3.1 Sentiment-Virality Correlation

Using statistical correlation matrices and visual plots, we assessed how sentiment scores relate to virality metrics.

Result: A moderate positive correlation was observed between sentiment scores and likes/comments. Posts that expressed optimism, humor, or motivation showed elevated engagement.

Insight: Sentiment polarity alone doesn't guarantee virality; the post context and timing are equally critical.

3.2 Caption Length and Engagement

We categorized captions into short (<50 words), medium (50–100), and long (>100).

Observation: Long, storytelling-style captions had higher comment rates, especially when discussing personal experiences or social topics.

Reason: They fostered relatability and conversations among followers.

3.3 Hashtag Usage Patterns

Analysis: Frequency of hashtag use vs. engagement score.

Finding: Overuse (>15 hashtags) often reduced effectiveness, possibly due to spam detection or user fatigue.

Effective Strategy: Use 5–10 relevant and niche hashtags per post.

3.4 Media Type Impact

Comparison: Image vs. Video vs. Carousel

Results:

Videos led in comments due to dynamic content.

Carousels had the highest saves, likely due to informational or tutorial-based posts.

Images performed best when accompanied by aesthetic visuals and concise captions.

3.5 Temporal Trends

We segmented posting time into:

Morning (6 AM–12 PM)

Afternoon (12 PM–6 PM)

Evening (6 PM–12 AM)

Night (12 AM–6 AM)

Engagement Trend:

Highest during evening hours and weekends.

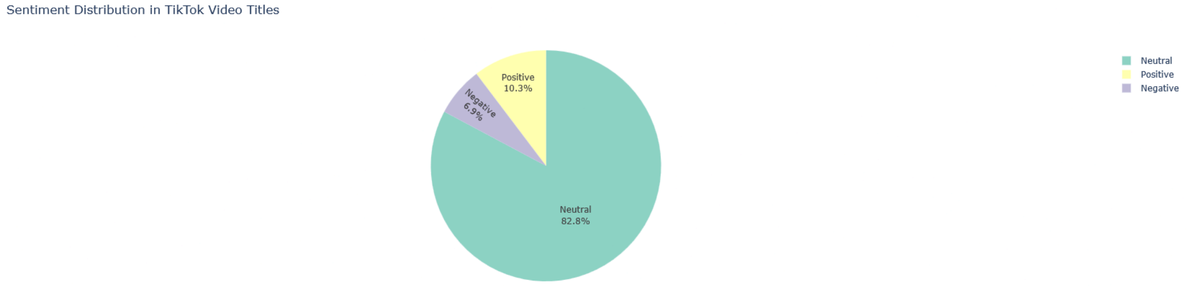
Sentiment scores were generally more positive for morning and weekend posts, reflecting relaxed viewer moods.

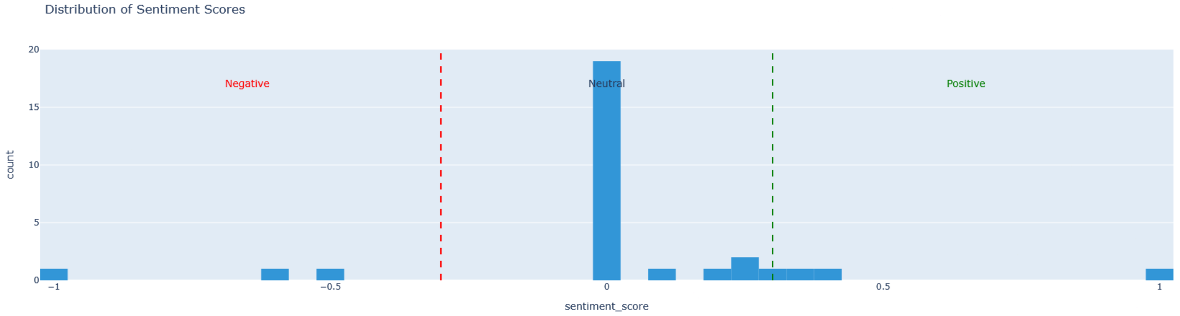
4. Data Visualization

4.1 Sentiment Score Distribution

Graph Type: Histogram + Pie Chart

Insight: Majority of posts were clustered around neutral to mildly positive sentiment. Few exhibited extreme polarities, suggesting users generally maintain a curated tone.

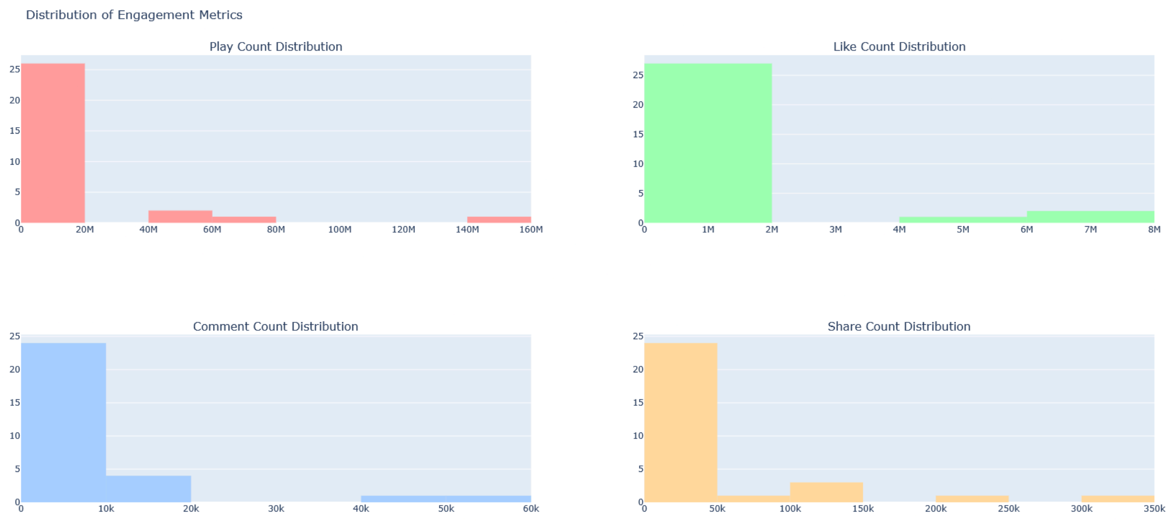




4.2 Engagement Metrics by Media Type

Visualization: Bar Graphs

Takeaway: Median engagement was highest for carousels; outliers (viral posts) appeared in all categories but were most frequent in videos.

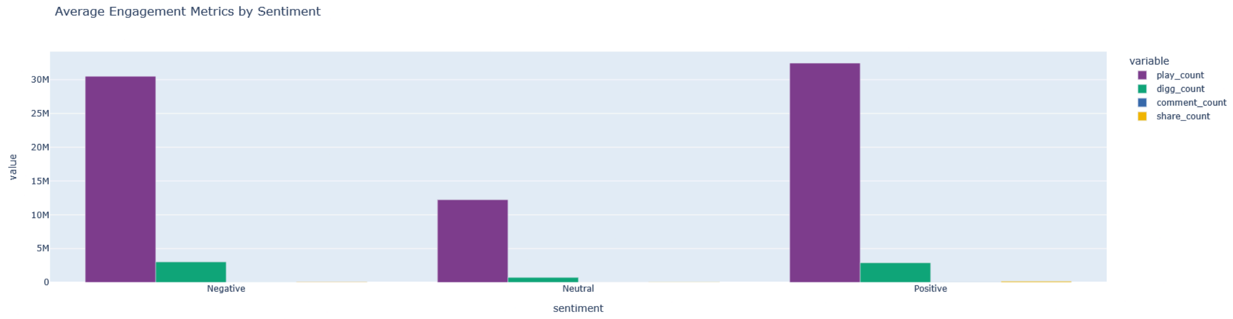

Distribution of Engagement Metrics

Play Count Distribution

Like Count Distribution

Comment Count Distribution

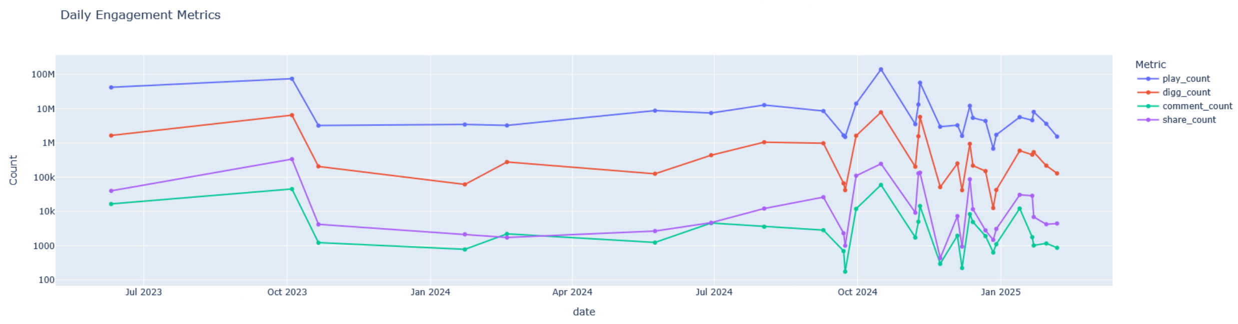
Share Count Distribution


Average Engagement Metrics by Sentiment

4.3 Time Series Analysis

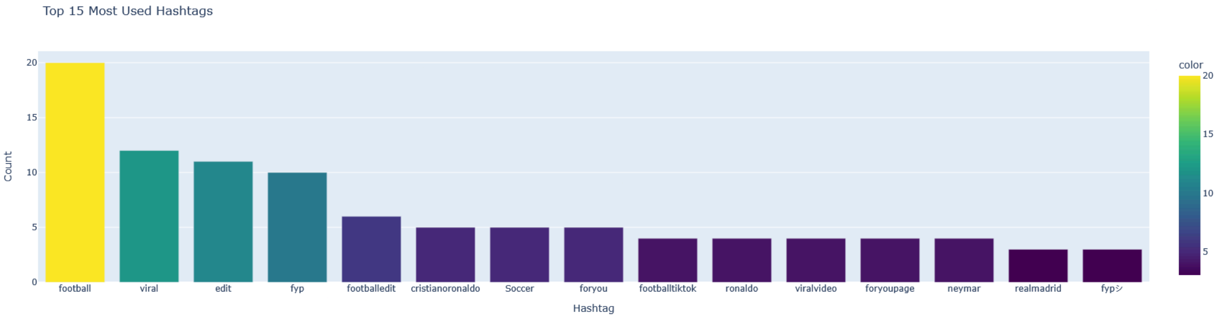
Plotted engagement and sentiment over time for each user group.

Discovered event-based spikes (e.g., influencer giveaways, holidays).

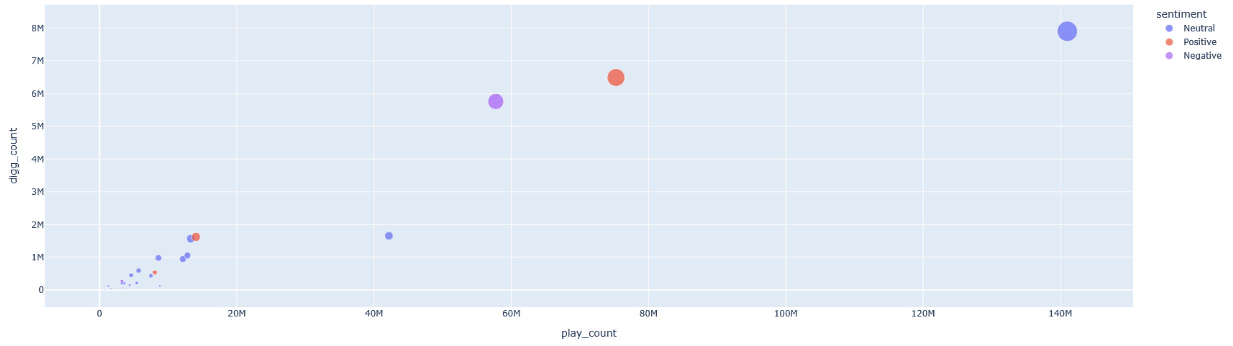

Daily Engagement Metrics

4.4 Hashtag Plots

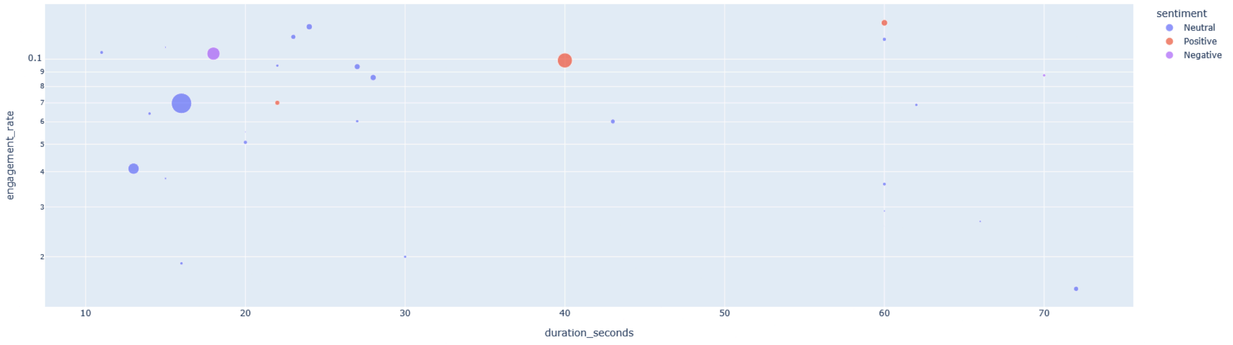
- X-axis: Top 20 Hashtags
- Y-axis: Avg Sentiment / Likes
- Observation: Motivational and trending hashtags correlated with higher sentiment and virality.



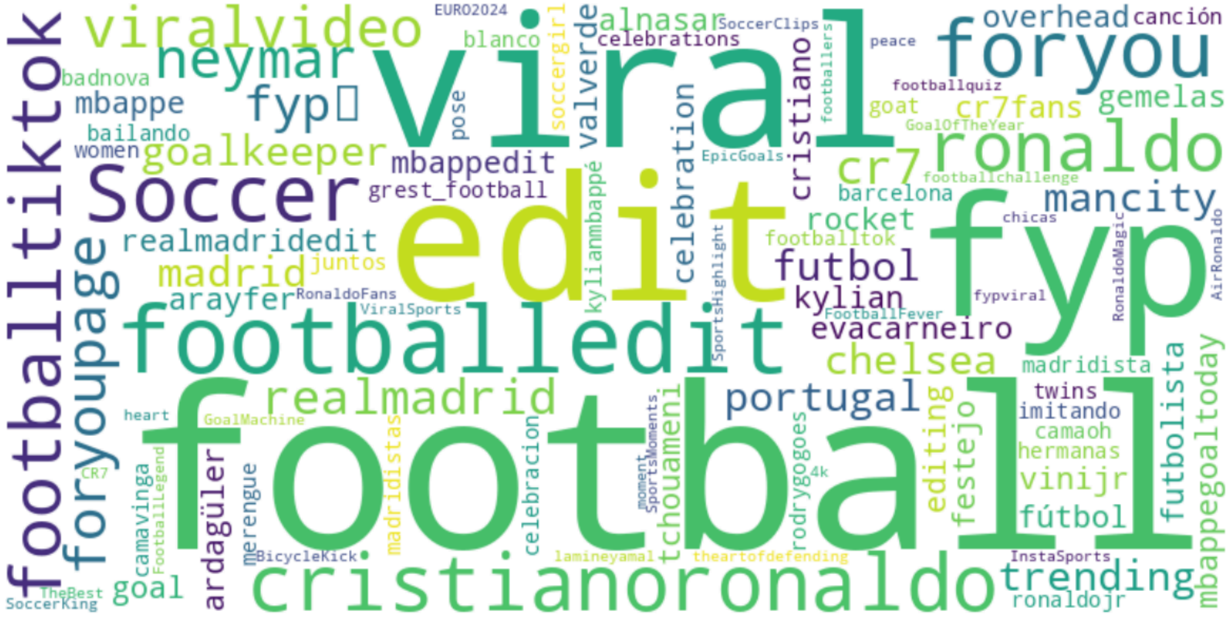
Video Views vs. Likes (Size = Total Engagement)



Engagement Rate vs. Video Duration



Hashtag Word Cloud



5. Results and Discussion

5.1 Key Insights

- Balanced Sentiment Wins: Neutral to mildly positive sentiment performed best. Excessively emotional posts showed polarizing outcomes.

- Context Matters: Hashtags, caption narrative, and post media all affect sentiment perception and engagement.

- Virality Isn't Always Positive: Some viral posts sparked outrage or debate, skewing sentiment analysis results negatively while boosting metrics.

5.2 Challenges Encountered

- Sarcasm/Irony Detection: TextBlob and VADER failed to detect nuanced human expressions like sarcasm.

- Emojis and Visual Sentiment: Visual cues that influence sentiment couldn't be processed, limiting analysis.

- Bot Engagement: Some spikes in engagement might stem from inauthentic interactions.

5.3 Limitations

- Lack of Rich User Metadata: Age, gender, and location were missing, restricting demographic segmentation.

- Caption-Only Sentiment: The model doesn't capture image/video-based emotional impact, which can dominate the message.

- Hashtag Semantics: Some hashtags are ambiguous or multilingual, complicating topic grouping.

6. Conclusion

6.1 Summary

The sentiment analysis of Instagram posts reveals a complex interplay between emotions, content structure, and virality. While positive and authentic sentiment tends to drive engagement, it’s the combination of media type, caption depth, and strategic hashtags that determines overall performance. Timing and platform trends also exert considerable influence on post outcomes.

Instagram virality is not just a result of sentiment or aesthetics—it reflects a combination of timing, audience psychology, content design, and contextual resonance.

6.2 Future Work

- Multimodal Sentiment Analysis: Incorporating image and video sentiment using CNNs and emotion recognition models.
- Deeper Comment Thread Analysis: Identifying toxic vs. supportive comment clusters and their impact on post longevity.
- Influencer Impact Modeling: Assessing how account credibility or follower trust influences sentiment-driven virality.
- A/B Testing Framework: Developing a simulation model to test different caption styles or hashtag sets before posting.