YouTube Comments and AMD News Analysis

Bansil Patel
Artificial Intelligence and Machine Learning
Lambton College
Toronto, Canada
c0902873@mylambton.ca

Harsh Mohile
Artificial Intelligence and Machine Learning
Lambton College
Toronto, Canada
c0912872@mylambton.ca

Meet Patel
Artificial Intelligence and Machine Learning
Lambton College
Toronto, Canada
c0910378@mylambton.ca

Rachit Bhatt
Artificial Intelligence and Machine Learning
Lambton College
Toronto, Canada
c0902810@mylambton.ca

***Abstract*—This article demonstrates the analysis on YouTube comments and news of AMD, analyzing the social media platforms to explore the hidden insights and patterns recognized using AI and ML tools.**

Keywords—youtube, amd, comments, news, analysis, sentiment analysis, social media analysis, topic modeling, user engagement, lda, data visualization.

I. Introduction

This open-source project aims to analyze the content created on various social media platforms, deriving insightful patterns, and visualizing them into user engaging charts, elaborating sentiments and user engagements. Using a dataset which has a few metrics for YouTube applied data processing, sentiment analysis, and Latent Dirichlet Allocation (LDA) [1] for topic modeling. The hidden patterns are demonstrated through interactive visualizations including but not limited to bar charts [2], scatter plots [2], and word cloud [3], each unique for respective data. The results are fascinating and the deployment became successful [4] by hosting it to recommend articles on given the keyword.

II. Methodology

The approach of this open-source project has been split into several steps as listed below: data collection, data cleaning, sentiment analysis, topic modeling using LDA, and data visualization.

A. Data Collection

The data was collected using the Google API [5] and Yahoo Finance API [6], focusing on fetching the data from YouTube and Yahoo Finance respectively. The data contains title, description, content, viewCount, likeCount, commentCount, and sentimentScore. The data was exported for version control [7].

B. Data Cleaning

The removal of records where the columns title, description, and content contained null values, and removing outliers from viewCount. The main agenda was to ensure the quality of the data which would be complete and relevant for the analysis, affecting the results of the sentiment analysis and the predictions of the topic modelling.

C. Sentiment Analysis

A pre-trained NLP model [8] was used on the content field. The score of sentiment extracted from the data could be classified into the categories of positive, negative, and neutral. However, the data was not categorized as it requires a threshold point, and making a decision for the actual threshold would be a little challenging at moment because the end results of this project does not depend on the accuracy of the sentiments. The model ran for 10 topics [1], generating a word cloud to visualize the strength of each term within the identified topics [1]. This activity concluded the extracting the thematic content of the videos.

D. Data Visualization

A plethora of interactive and static visualizations were designed to explore the relationships between various metrics:

1. **Bar Chart**: Displaying bivariate relation.
2. **Stacked Bar Chart**: Displaying multivariate relation.
3. **Word Cloud**: Topic distribution and derived terms' strength by LDA model [1].
4. **Scatter Plot**: Alterative to visualize data for word cloud based on occurrence of the words.
5. **Histogram**: Used bar charts to demonstrate the distribution of the polarity score of various records with sentiment scores.

III. Data Collection

The collection process is divided into two parts, one for YouTube and another one for the news.

A. YouTube

- With the help of Google API, the various keys are created and stored in a JSON file.
- Loading that JSON file brings all the keys in action.
- Constructed a modular approach for extracting video data using API.
- The data given by the API was fixed a little during the process of fetching, transforming JSON data into a DataFrame.
- Applied data cleaning and further analysis after loading it into DataFrame.

B. News

- The API key created is stored and fetched from a config file.

Identify applicable funding agency here. If none, delete this text box.

- With the help of the NLTK library, we fetched the data from News API website passing the API key and the topic of interest.
- Cleaned the text by removing the stopwords and tokenizing the content.
- Created and exported filtered data in a DataFrame.

IV. Data Cleaning

Similar to the Data Collection process, the Data Cleaning process is also divided into two parts, one for YouTube and another one for the news.

A. *YouTube*

- When exploring the dataset, the viewCount was found 0 for a few videos.
- After evaluating, the video count was non-zero but was given as 0 by the API.
- It was vital to remove such outliers; hence, dropped them.

B. *News*

- The news data which was not found due to any exception was stored as [Removed] in the title, description, and content columns.
- Hence, such rows were dropped as the [Removed] was like a None or NULL value.

V. Results and Discussion

This analysis presents numerous interesting insightful hidden patterns into user engagement on the platform and the sentiment trending across the video content and news.

A. *Top-10 vidoes based on total views*

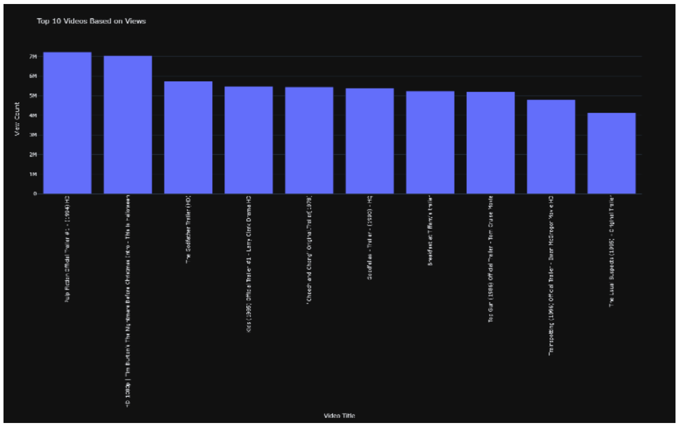


Fig. 1. Top-10 vidoes based on total views

- This visual represents the most viewed videos.
- They ranges between 4.1 million to 7.2 million views so far.
- The trailer of Pulp Fiction from 1994 sits on the top with 7.23 million views while Tim Burton's Nightmare Before Christmas Intro resides closely at 7.04 million views.
- Top Guns being a popular movie from 1980s has 5 million views.
- The movies at or before 2000 had more viewed trailers.

B. *Bottom-10 videos based on total views*

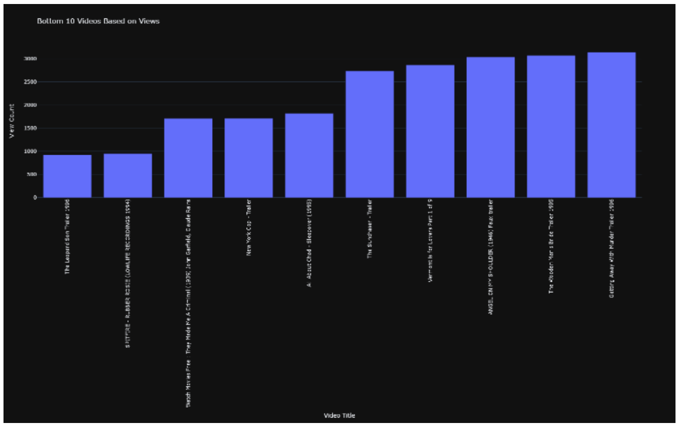
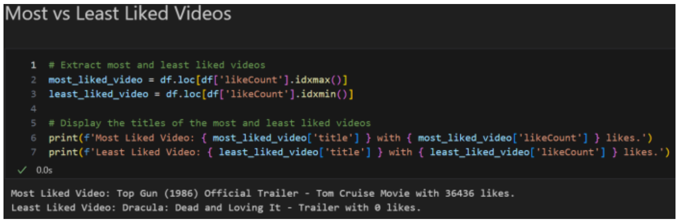


Fig. 2. Bottom-10 videos based on total views

- The most movie trailers released in the 90s have the least popularity.
- Starting with The Leopard Son and SPITFIRE not having views reaching 1K.
- The murder mystery remained at the top of the bottom-10 trailers, Getting Away with Murder in 1996, having 3K viewers.

C. *Most and least liked video title*



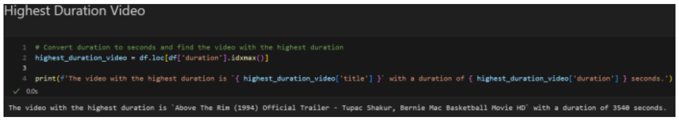
```
# Extract most and least liked videos
most_liked_video = df.loc[df['likeCount'].idxmax()]
least_liked_video = df.loc[df['likeCount'].idxmin()]

# Display the titles of the most and least liked videos
print(f'Most Liked Video: { most_liked_video['title'] } with { most_liked_video['likeCount'] } likes.')
print(f'Least Liked Video: { least_liked_video['title'] } with { least_liked_video['likeCount'] } likes.')
```

```
Most Liked Video: Top Gun (1986) Official Trailer - Tom Cruise Movie with 36436 likes.
Least Liked Video: Dracula: Dead and Loving It - Trailer with 0 likes.
```

Fig. 3. Most and least liked videos

- The most loved movie trailer winning many hearts is a Tom Cruise movie, the Top Gun with 36K likes.
- The movie trailer Dracula: Dead and Loving It sits ideal at 0 likes, disappointing the creators and makers of the movie.

D. *Longest clip*



```
# Convert duration to seconds and find the video with the highest duration
highest_duration_video = df.loc[df['duration'].idxmax()]

print(f'The video with the highest duration is `{ highest_duration_video['title'] }` with a duration of { highest_duration_video['duration'] } seconds.')
```

```
The video with the highest duration is `Above The Rim (1994) Official Trailer - Tupac Shakur, Bernie Mac Basketball Movie HD` with a duration of 3540 seconds.
```

Fig. 4. Highest Duration Video

- A lengthy movie trailer was Above the Rim in 1994 that was almost an hour long.

- Despite the fact being the most lengthiest trailer, from the top-10 highest duration videos, it has the most likes as shown in the following figure:

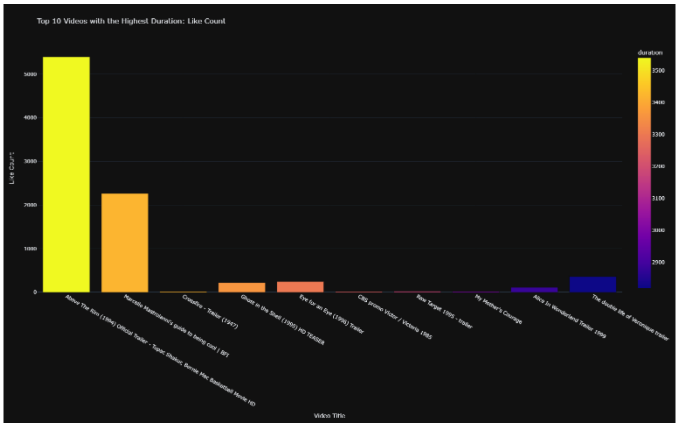


Fig. 5. Highest liked video despite being longest (5396 Likes)

E. Top-10 vidoes with highest positive comments

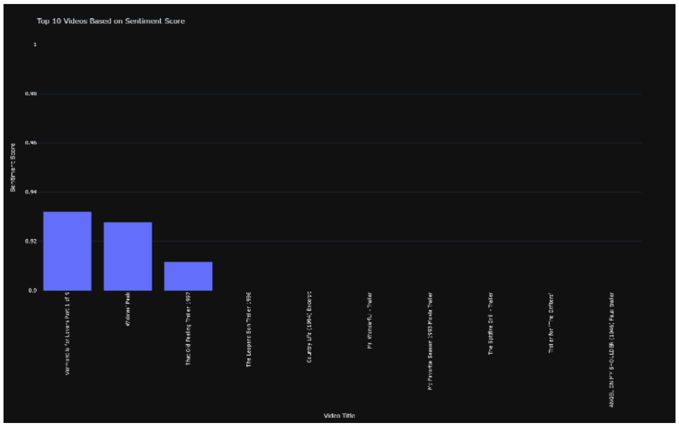


Fig. 6. Top-10 vidoes with highest positive comments

- Due to the limitation of the APIs per day, only three records were able to categorized in this visual.
- Vermont is for Lovers Part 1 of 9, a sequel was viewed and commented well by the audience.
- Just below that, the Windows' Peak leads with 92% positiveness in the comments.
- That Old Feeling Trailer from 1997, one of the most loved movies of all time also perceived positive sentiments in the comments section scoring 91% positiveness overall.

F. Bar diagram showcasing top-5 topics for news on AMD

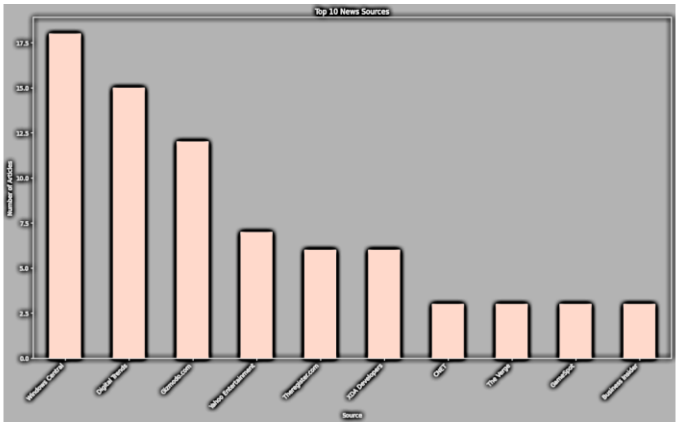


Fig. 7. Top-10 topics for news on AMD

- Windows Central, Digital Trends, Yahoo Entertainment, XDA Developers, and CNET are a few topics that had the most articles for AMD.
- However, The Verge, GameSpot, and Business Insider remains at the bottom with almost 2.5 articles on average.

VI. Conclusion

This project demonstrated how sentiment analysis and topic modeling using LDA [1] could be useful to analyze the video content data and live news, describing trendy insights about user engagements. While sentiment analysis points to the severity of user actions and reactions to determine the video popularity, with the help of LDA [1], a deeper understanding of the themes that capture user attention.

TABLE I. TABLE TYPE STYLES

Table Head	**Table Column Head**		
	Table column subhead	***Subhead***	***Subhead***
copy	More table copy[a]		

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latest Dirichlet Allocation. *Journel of Machine Learning Research*, 3, 993-1022.

[2] Plotly Express [Online]. Available: https://plotly.com/python-api-reference/plotly.express.html

[3] Matplotlib [Online]. Available: https://matplotlib.org/

[4] StreamLit [Online]. Available: https://streamlit.io

[5] Google API https://console.cloud.google.com/apis/credentials

[6] Yahoo Finance API [Online]. Available: https://python-yahoofinance.readthedocs.io/en/latest/api.html

[7] Raksha (2023) [Online]. Available:. *Top 6 Dataset Version Control Tools for your Machine Learning workflow*. https://www.twine.net/blog/top-6-data-version-control-tools/

[8] Vader for Sentiment Analysis [Online]. Available: Welcome to VaderSentiment's documentation! — VaderSentiment 3.3.1 documentation