Trending Now: Leveraging Social Media to Identify and Analyze Movie and Genre Trends

Dimple Singh
Computer Science
Binghamton University
Binghamton, NY, USA
dsingh27@binghamton.edu

Jay Balaram Sankhe
Computer Science
Binghamton University
Binghamton, NY, USA
jsankhe1@binghamton.edu

Debangana Ghosh
Computer Science
Binghamton University
Binghamton, NY, USA
dghosh2@binghamton.edu

Jeremy Anton
Computer Science
Binghamton University
Binghamton, NY, USA
ajeremy1@binghamton.edu

Ritika Kale
Computer Science
Binghamton University
Binghamton, NY, USA
rkale2@binghamton.edu

Abstract

Our project, "Trending Now: Leveraging Social Media to Identify and Analyze Movie and Genre Trends," has progressed through two distinct phases. Earlier, we focused on developing a system for extracting relevant data from Reddit and The Movie Database (TMDb). This phase involved scraping data about movies, TV shows, related Reddit posts and comments, and their respective ratings. We organized this data in a database, grouping movies and shows by genre and correlating them with social media discussions.

This project advances our work by moving into the analysis phase. Here, we introduce a crucial tool for sentiment analysis: the ModerateHatespeech API. This third-party service evaluates the text of comments, assigning a rating of 'normal' or 'flag' with a corresponding confidence score. The application of this tool is pivotal in assessing the tone and sentiment of discussions around movies and TV shows, particularly in identifying the toxic or non-toxic nature of these conversations. The insights gained from this sentiment analysis, when juxtaposed with the ratings data, aim to reveal deeper correlations between public sentiment and the popularity or perception of various genres in entertainment.

This abstract encapsulates our journey from data collection to sophisticated analysis, highlighting our efforts to understand and interpret the complex dynamics of movie and genre trends as reflected in social media discussions.

Keywords: Reddit API, TMDb API, ModerateHatespeech API, PostgreSQL, Movies, Data Collection, Analysis, Plotting, Visualization

ACM Reference Format:
Dimple Singh, Jay Balaram Sankhe, Debangana Ghosh, Jeremy Anton, and Ritika Kale. 2023. Trending Now: Leveraging Social Media to Identify and Analyze Movie and Genre Trends. In *Proceedings of ACM Binghamton conference (Binghamton'23).* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/1234567890

1 INTRODUCTION

In an era where social media significantly influences public opinion and trends, understanding the dynamics of online discussions about movies and TV shows has become increasingly important. Our project, "Trending Now: Leveraging Social Media to Identify and Analyze Movie and Genre Trends," aims to bridge the gap between social media chatter and measurable insights into entertainment trends.

The first phase of our project aid the groundwork by establishing a robust data collection framework. We utilized the Reddit API and The Movie Database (TMDb) API to gather a wide range of data, including posts, comments, and ratings related to movies and TV shows. This data was meticulously categorized by genre and stored in a database, setting the stage for a comprehensive analysis.

In our current project, we have shifted our focus towards the analytical aspect of the data. A key development in this phase is the integration of the ModerateHatespeech API, a tool vital for conducting sentiment analysis. By analyzing the text of social media comments, this API helps us classify discussions as either 'normal' or 'flagged' for toxicity, with an accompanying confidence score. This feature is pivotal

Binghamton'23, Binghamton, NY, USA,

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/1234567890

in assessing the nature of online conversations around different movies and shows, particularly in identifying toxic or non-toxic dialogues.

The objective of this phase is to uncover correlations between the sentiments expressed in these discussions and the ratings and popularity of the movies and TV shows. By doing so, we aim to provide a nuanced understanding of how online discourse influences and reflects the success and perception of various genres in the entertainment industry.

This report outlines our approach, methodologies, challenges encountered, and the insights garnered from our analysis. It represents a significant step in using social media data to derive meaningful conclusions about current trends and audience preferences in the entertainment sector.

2 DATASET DESCRIPTION

2.1 DATA SOURCE

Here, in this project, we worked with collected information about top-rated movies and tv shows. We used the TMDb rating data that we obtained from the TMDb API and the posts and comments data obtained from Reddit from r/movies and r/tv through Reddit Stream API and analyzed it by using ModerateHatespeech API for measurement of toxicity of the discussions on the collected data based on genre. For implementation, we used Python, and some of its libraries to fetch and handle the Requests and store the data collected in PostgreSQL.

2.2 API METHODS

The data collection process leverages various API methods to interact with the Reddit and TMDb APIs, as well as the ModerateHatespeech API for sentiment analysis. We use GET requests for the API's.

Reddit API: We use get_reddit_token for authentication and get_reddit_posts_paginated_per_media to fetch and paginate posts from relevant subreddits.

TMDb API: The tmdb_result function is utilized to retrieve and store information about movie and TV titles, including their ratings and popularity.

ModerateHatespeech API: The API plays a crucial role in our sentiment analysis by assessing the toxicity of text in Reddit discussions. It classifies posts and comments as 'normal' or 'flagged,' accompanied by a confidence score. Our function, get_toxicity_score, enables us to quantitatively evaluate sentiment. All data is efficiently processed and stored in PostgreSQL for accessibility and analysis.

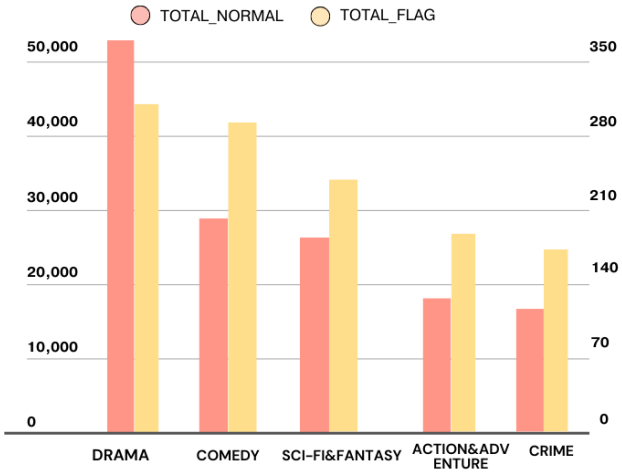

Figure 1. TV Genre Normal and Flagged

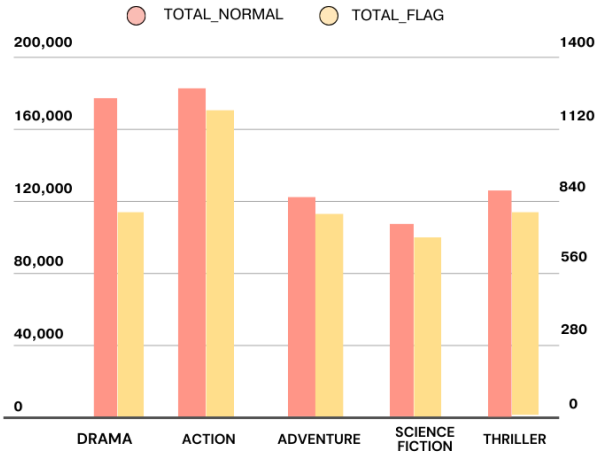

Figure 2. Movies Genre Normal and Flagged

3 DATA EXPLORATION

In delving into the sentiment of social media discussions, we paid particular attention to the genre-specific distribution of toxicity in movie and TV show discussions based on the Reddit data.

Television Genre Toxicity Analysis: As collaborated in Figure 1, our analysis revealed the Drama genre to be a focal point for engagement in television discussions, with 51,713 comments, and 313 of these flagged as toxic. In contrast, the Comedy genre in television discussions displayed a higher toxicity proportion with 305 flagged comments out of 28,836, suggesting that this genre may elicit more polarizing and contentious discussions compared to others.

Movie Genre Toxicity Analysis: Within the domain of movies, in the figure 2, the Action genre exhibited a substantial volume of discussions, totaling 183,502 comments with 1,275 flagged as toxic. Meanwhile, the Comedy genre for movies showed a similar trend with a higher number of toxic comments (651 out of 106,632), indicating that the Comedy genre tends to generate more toxic discussions across both movies and TV.

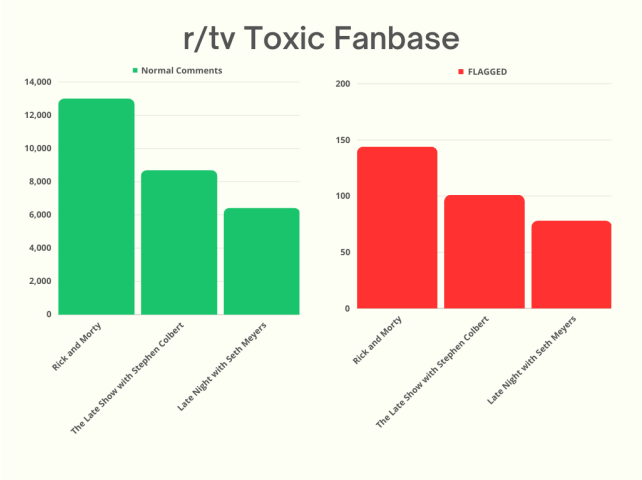

Figure 3. Top 3 Toxic Engagements for TV

Figure 3 shows that *The Walking Dead* has the highest number of toxic comments, followed by *Rick and Morty* and *The Late Show with Stephen Colbert* as collected in our data.

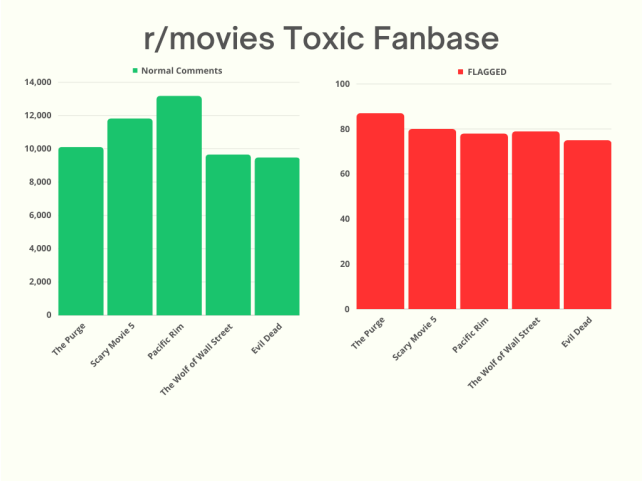

Figure 4. Top 3 Toxic Engagements for Movies

Figure 4 shows that *The Purge* has highest number of toxic engagement followed by *Scary Movie 5*, *The Wolf of Wall Street*, *Pacific Rim*, *Evil Dead*

Comparative Genre Analysis: The data from both reports indicates that certain genres, specifically Drama and Comedy, are more prone to toxic discussions. For example, in the Comedy genre, the total number of comments on television and movies combined is 135,468, with a sum of 956 toxic comments, which is higher compared to other genres when combined. This analysis provides a foundational understanding of how different genres attract varying levels of discussion engagement and sentiment. It also raises questions about the impact of content format on user interaction and sentiment, which we will explore further in subsequent analyses.

4 Politics Collection

We have collected around 2984 posts and 486458 comments from November 1st to November 29th. This averages around 130 posts and 16215 comments per day. We can see the trend of daily submissions to r/politics in the graphs below (Figure 3 and Figure 4).

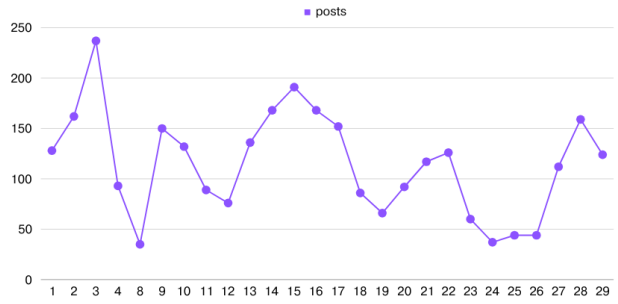

Figure 5. r/politics post collection

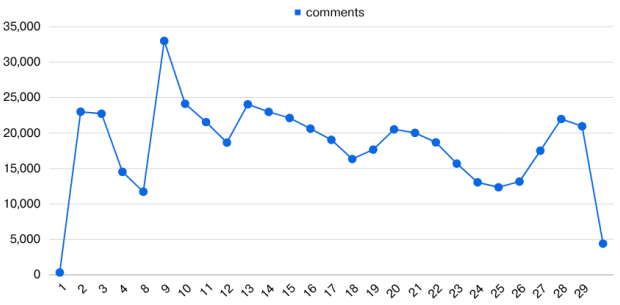

Figure 6. r/politics comments collection

5 BACKGROUND RELATED WORK

With the increasing dominance of social media in shaping public discourse and trends, there has been a growing interest in leveraging platforms like Reddit to understand discussions surrounding popular movies and TV shows. Social media

platforms serve as rich sources of real-time data, providing insights into audience sentiments and preferences.

Data Collection and Analysis of Movie and TV Discussions: The focus is on comprehensively collecting and analyzing discussions related to currently popular movies and TV shows. We recognize that the dynamics of online discussions significantly impact the perception and success of entertainment content. Therefore, our research aims to bridge the gap between social media chatter and quantifiable insights into entertainment trends.

Integration of ModerateHatespeech API: A pivotal development in our project is the integration of the ModerateHatespeech API. This API, specializing in real-time toxicity measurement, allows us to conduct sentiment analysis on discussions about movies and TV shows. By classifying discussions as either 'normal' or 'flagged' for toxicity with a confidence score, we gain a nuanced understanding of the nature of online conversations.

Challenges and Opportunities: The project encountered several challenges, particularly in obtaining timely data. Initially, we faced issues with receiving limited data from Reddit, which could have skewed our analysis. However, we overcame this challenge by delving into Reddit's pagination API documentation, which provided a solution for accessing older discussions. This experience not only highlighted the dynamic nature of social media data but also demonstrated the project's capacity for problem-solving and innovation.

The ability to adapt to unforeseen hurdles underscores the project's resilience and presents an opportunity for developing data collection methodologies. It also emphasizes the importance of thorough documentation and proactive problem-solving in the field of social media analytics.

While integrating the ModerateHateSpeech API, we encountered instances where it provided empty responses for some input. To address this issue, we refined our script to discard these responses and handle potential exceptions. Additionally, due to time constraints, we were unable to process the entire body of data, resulting in many posts and comments that did not receive a sentiment score.

With additional time, we can gather more data on comments and posts, leading to a more comprehensive understanding of toxicity across genres and the titles that contribute to this toxicity and report more accurate findings.

Ethical Considerations: Given the influence of social media, ethical considerations in handling user-generated content and sentiment analysis are paramount. We acknowledge the importance of respecting user privacy and ensuring responsible use of the collected data.

In summary, our work builds upon the foundation laid by previous research in social media analytics and sentiment analysis. By combining data from Reddit discussions, TMDb ratings, and toxicity measurements, we aim to contribute valuable insights into the ever-evolving landscape of entertainment trends and audience preferences.

6 DATA FLOW

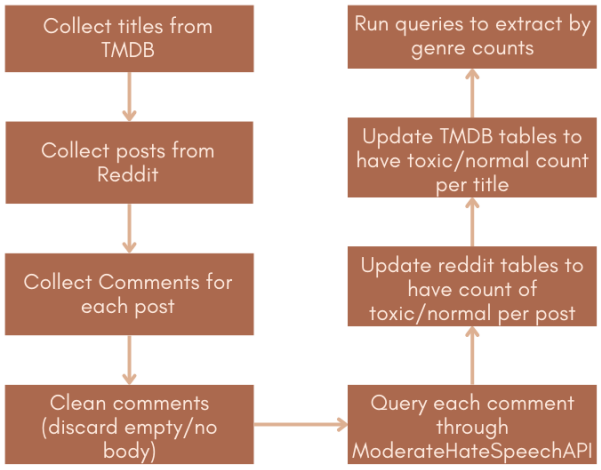

Figure 7. Flow Diagram

From Figure 5, the data flow diagram, we can see the Reddit API is used to collect posts and comments from Reddit.The ModerateHatespeech API is used to extract hate speech from the collected posts and comments.The comments are cleaned to remove irrelevant characters and symbols, and to ensure the data format is aligned with the ModerateHatespeech API's requirements.The Reddit and TMDb tables are updated to have the number of toxic and normal comments per post and per title respectively.

7 OBJECTIVES

The central objectives of our project are to analyze and understand the social media discourse surrounding movies and TV shows and to investigate the sentiment and toxicity of these discussions as they relate to various genres. These objectives are aimed at contributing to both academic research and practical applications in media analysis and trend forecasting. Moreover, we are focusing on the following research questions to extend our work further:

- Is there a correlation between the popularity trend and ratings of movies and television? In other words, do highly-rated titles attract more discussion?
- Does the popularity of movies or television correlate with the amount of toxic content on online forums? Additionally, which genres generate the highest levels of toxic content?
- How does the toxicity of the most popular titles change over time?

8 OBSERVATION

API Integration Impact: The inclusion of the ModerateHatespeech API enhances the project's analysis by providing real-time toxicity measurement, offering insights into sentiment dynamics.

Clear Data Flow and Objectives: The report effectively outlines data flow and project objectives, providing a clear roadmap for analyzing social media discourse and measuring toxicity.

Future Analytical Directions: The report sets the stage for future analyses, indicating a shift towards more in-depth genre-specific and audience engagement patterns exploration.

9 CONCLUSION

Our data collection system for Reddit and TMDB has come a long way since the project started. We've overcome initial challenges, successfully gathered data, and handled a lot of information efficiently. Our improved projections make sure we collect data within storage limits.
This progress positions us well for further analysis, allowing us to dig into the details of the collected data. The strong infrastructure we now have in place lets us explore and answer the research questions we've raised, forming a solid foundation for a comprehensive understanding of social media discussions about movies and TV shows. Moving forward, our refined data collection system paves the way for a clearer insight into trends and patterns on Reddit and TMDB, aligning with the main goals of our project.

10 ACKNOWLEDGEMENT

We would like to acknowledge the Reddit and TMDb APIs for providing access to valuable data. We also appreciate the guidance and support received during the project's development.

References

[1] Meet Singh, Dhrubasish Sarkar. *Content-Based Movie Recommendation System with Sentiment Evaluation of Viewer's Reviews.* https://link.springer.com/chapter/10.1007/978-981-16-6893-7_15

[2] Sandipan Sahu, Raghvendra Kumar, Mohd Shafi Pathan, Jana Shafi, Yogesh Kumar, Muhammad Fazal Ijaz. *Movie Popularity and Target Audience Prediction Using the Content-Based Recommender System.* https://ieeexplore.ieee.org/abstract/document/9758691