Analyzing The Influence of Celebrities on Today’s Generation

Shruti Iyengar
Computer Science
SUNY at Binghamton
Binghamton, NY, USA
siyenga1@binghamton.edu

Rushabh Kothari
Computer Science
SUNY at Binghamton
Binghamton, NY, USA
rkothar1@binghamton.edu

Bhavit Yogesh Shah
Computer Science
SUNY at Binghamton
Binghamton, NY, USA
bshah5@binghamton.edu

Mukul Dev Chhangani
Computer Science
SUNY at Binghamton
Binghamton, NY, USA
mchhang1@binghamton.edu

ABSTRACT

In the digital era, the impact of celebrities on the younger generation is increasingly mediated through social media platforms. This study aims to analyze the extent and nature of this influence quantitatively and qualitatively. Utilizing the APIs of two major social media platforms, YouTube, and Reddit, we systematically gather a substantial dataset, which is then meticulously stored in a MongoDB database for robust analysis. Our methodology encompasses a multi-faceted analytical approach, including an Influence Matrix, Trends Analysis, Content Analysis, and Sentiment Analysis, to provide a comprehensive understanding of celebrity impact.

KEYWORDS

Reddit API, YouTube API, MongoDB, Data Collection, Analysis, Plotting, Visualization, Sentiment Analysis, Influence Matrices, Content Analysis, and Trends Analysis

1. INTRODUCTION

In the modern digital era, the impact of celebrities on public opinion and the younger generation has gained significant attention. Social media platforms like YouTube and Reddit have transformed how celebrities connect with their audience, fostering more direct and personal interactions. This shift is crucial for understanding how celebrity influence shapes cultural trends, consumer behavior, and social norms. This study aims to investigate the influence of celebrities on today's generation through data from these social media platforms.

Social media has empowered celebrities to amplify their influence and reach. Platforms like YouTube and Reddit not only allow for wider distribution of content created by celebrities but also facilitate interactive and engaging relationships with their audience. Understanding these new dynamics of celebrity influence in the digital age is essential, yet it presents complex challenges that require advanced analytical methods and effective data management.

To address this challenge, our study employs a methodical approach by harnessing the capabilities of the YouTube and Reddit APIs. These platforms are chosen for their widespread popularity and diverse user demographics, providing a rich source of data on celebrity interactions and audience responses. The data collected through these APIs are systematically stored in a MongoDB database, ensuring efficient management and retrieval for analysis.

Our analytical framework comprises four key components: an Influence Matrix, Trends Analysis, Content Analysis, and Sentiment Analysis. The Influence Matrix aims to quantify the reach and engagement of celebrities, providing a metric for their influence. Trends Analysis examines the evolution of discussions and topics over time, shedding light on shifting interests and the ebb and flow of celebrity popularity. Content Analysis delves into the substance of the interactions, exploring the themes and narratives that resonate with today's generation. Finally, Sentiment Analysis offers insights into the emotional tone of the discourse, reflecting public perceptions and attitudes towards

2. BACKGROUND AND RELATED WORK

Social media platforms have revolutionized the landscape of celebrity influence, as discussed in "The Culture of Connectivity: A Critical History of Social Media" by José van Dijck. These platforms, especially YouTube and Reddit, have become pivotal in shaping how celebrities engage with and influence their audiences. YouTube, with its vast user base and diverse content, offers a unique window into the ways celebrities interact with fans and disseminate content, as explored in "The YouTube Reader" edited by Pelle Snickars and Patrick Vonderau. Reddit, known for its community-driven discussions and diverse subcultures, provides a different yet equally important perspective on celebrity influence, particularly in how public opinion is formed and expressed in online communities.

The use of YouTube and Reddit APIs is a significant methodological consideration in our study. These APIs allow for the systematic collection of large-scale data on user interactions, content dissemination, and audience engagement, as outlined in "Analyzing Social Media Networks with NodeXL" by Derek Hansen, Ben Shneiderman, and Marc A. Smith. By accessing these APIs, our research can harness a wealth of data that is critical for understanding the nuances of celebrity influence in the digital age.

Furthermore, integrating these platforms into our study aligns with the current trends in social media usage and the evolving nature of celebrity-fan interactions. As discussed in "Social Media and Public Relations: Fake Friends and Powerful Publics" by Judy Motion et al., celebrities now have unprecedented tools at their disposal to craft their public image and engage with their audience, making the analysis of these interactions more relevant than ever.

In summary, the integration of the importance of YouTube and Reddit, along with the use of their APIs, into the Background and Related Works of our study, provides a solid foundation for understanding the multifaceted nature of celebrity influence today. This approach not only situates your research within the current technological and cultural context but also leverages the latest methodologies in social media analytics to offer novel insights into the dynamics of celebrity influence on today's generation.

3. DATASET

We have collected the data using YouTube API and reddit API and stored them into the mongoDB database. For both the sources, while collecting the data, we fetch the data based on the array of celebrities' names as the keywords.

3.1 Data Sources

First Data Source : We used Reddit API to get real-time related posts to the celebrities
Second Data Source : We used YouTube API to collect the YouTube data such as comments and title of the published videos related to the celebrities

3.1.1 Reddit API

In our project on analyzing celebrity influence, Reddit, accessed via its REST API, serves as a crucial data source. This platform is a rich repository of user opinions and discussions about celebrities, offering insights into public sentiment. The Reddit API allows for the extraction of user-submitted content and engagement metrics (like upvotes and comments), which are key to understanding how celebrities are perceived and discussed by the public.
Specifically, our project utilizes the Reddit API to gather data on the top 10 celebrities from the Forbes list, ensuring a focus on influential figures recognized for their economic and social impact. This approach enables a detailed analysis of how these celebrities are viewed and talked about on Reddit, providing a window into the dynamics of celebrity influence among today's generation. Reddit's diverse and active user base makes it an invaluable resource for capturing a wide range of perspectives and reactions in the digital landscape.

3.1.2 YouTube API

In our project, the YouTube Data API is a key data source for analyzing celebrity influence. It allows you to access a wealth of video content and viewer comments related to the top 10 celebrities identified from your web search. This API enables the collection of both video data and audience reactions, providing insights into how these celebrities are portrayed and perceived on one of the world's largest video-sharing platforms. The comments section is a rich resource for sentiment analysis, offering direct insights into public opinions and attitudes towards these celebrities. This makes the YouTube Data API an invaluable tool for understanding the multifaceted influence of celebrities in today's digital age.

3.1.3 Modern Hate Speech API

We used the modern hate speech API to find the toxicity score of the gathered data from reddit and YouTube API. We fetch all the gathered data (comments, subreddits, video titles) from the database and used the Modern Hate Speech API to find the toxicity score between 0 to 1 whereas any score lying between 0 to 1 shows the confidence level of the prediction, and class type which shows the toxicity type as “toxic” or “normal”.

3.1.3 Data Collection

In our project, preliminary testing revealed that not all comments on YouTube and Reddit posts directly mention the targeted keywords related to the top 10 celebrities from Forbes. Despite this, we were able to collect the data using the reddit API about 27.5k and using YouTube API about 3.5k.
Our analysis focuses on these celebrities: Cristiano Ronaldo, Lionel Messi, Selena Gomez, Kylie Jenner, Dwayne Johnson, Ariana Grande, Kim Kardashian, Beyoncé, Khloé Kardashian, and Kendall Jenner. You're monitoring specific subreddits dedicated to each celebrity and using the YouTube API to fetch their top 10 trending videos. The relevant comments and posts are then stored in a MongoDB database for further analysis. This approach aims to comprehensively assess the influence of these celebrities based on their online presence and audience engagement on these platforms.

MongoDB Collection: database name: team_caffeine

```
collection_1: YouTube_data
{
        "id": ...,
        "text": ...,
        "comments": ...
}

collection_2: Reddit_data
```

```
{
        "id": ...,
        "title": ...,
        "comments": ...
}

collection_3: modernHateSpeech
{
        comment_text : …,
        class : …,
        confidence : …,
        response : …,
        timestamp : …
}
```

4. METHODOLOGY

We aim to implement various analyses of influential celebrities. We will be collecting the data using the YouTube and Reddit APIs and updating every new entry into the database in real-time. To analyze the collected data, we will retrieve data from the database periodically and perform the various visualization graphs using Python. We are performing sentiment analysis, trend analysis, and comparative analysis and utilization of modern hate speech API to analysis of toxicity to answer the insightful questions.

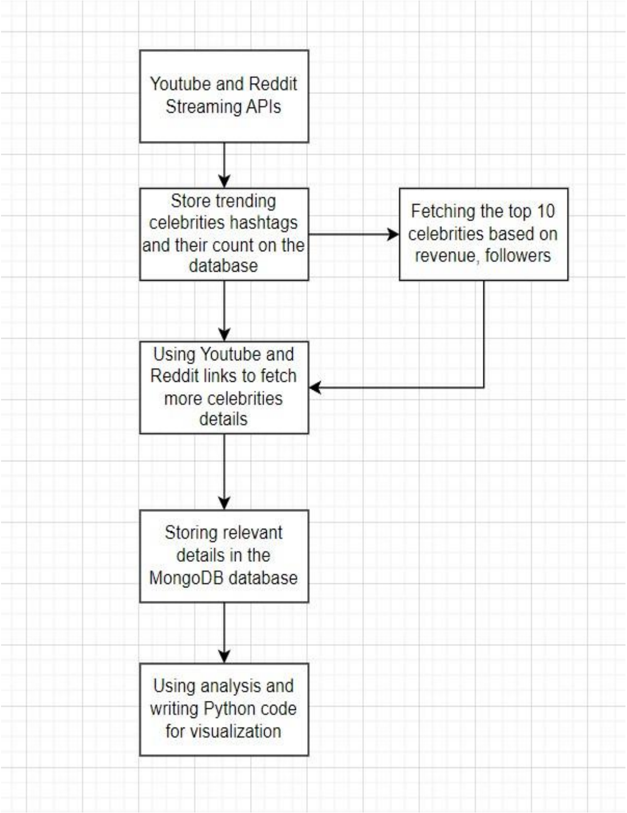

Youtube and Reddit Streaming APIs
Store trending celebrities hashtags and their count on the database
Fetching the top 10 celebrities based on revenue, followers
Using Youtube and Reddit links to fetch more celebrities details
Storing relevant details in the MongoDB database
Using analysis and writing Python code for visualization

Figure 1- Working model of the Project

4.1 Comparative analysis:

To perform a comparative analysis in our project on celebrity influence using YouTube and Reddit data, we followed below steps:
Define Metrics: We chose user engagement as our parameter.
Collect and Process Data: We Used APIs to gather data, then clean and process it for analysis.
Analyze Data: We Extract key metrics and plot the two graphs to understand the popularity of the celebrities on the two social media platforms
The two graphs provided show user engagement with content related to various celebrities on Reddit and YouTube.

4.1.2 Reddit Engagement:

In Figure 2, the bar chart for Reddit shows user engagement for different celebrities. Selena Gomez and Ariana Grande have the highest engagement, followed closely by Kendall Jenner. The engagement levels decline for other celebrities, with Dwayne Johnson and Cristiano Ronaldo receiving the least engagement among those listed.

4.1.2 YouTube Engagement:

In Figure 3, the YouTube graph also measures user engagement but shows a different pattern. Here, Ariana Grande leads significantly, followed by Selena Gomez. Other celebrities like Dwayne Johnson and Cristiano Ronaldo have more engagement on YouTube compared to Reddit.

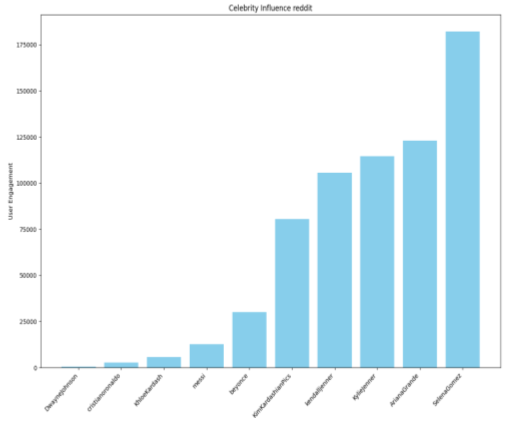


Figure 2- Reddit Celebrity Influence

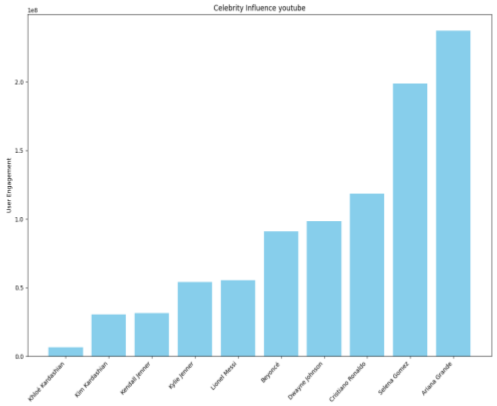


Figure 3- YouTube Celebrity Influence

4.2 Influence Metrics:

To perform an Influence Matrix analysis in your project:
Define Metrics: Select key metrics for influence, like we selected the “upvotes” as our parameter
Collect Data: Gather data from YouTube and Reddit using APIs, focusing on these metrics for your chosen celebrities.
Process Data: Clean and normalize the data for consistency.
Analyze: Compare and analyze the influence of each celebrity based on the matrix.
Visualize: We used libraries like matplotlib for easier interpretation of the matrix .

4.2.1 Engagement Data Analysis on YouTube:

In Figure 4, the bar chart presents the number of likes (which is a proxy for user engagement) on YouTube content related to various celebrities. The data shows that Dwayne Johnson and Ariana Grande have the highest engagement levels, with likes reaching into the tens of millions. This indicates a strong positive response from the platform's users to content associated with these celebrities. Notably, the engagement for other celebrities like Cristiano Ronaldo, Lionel Messi, and the Kardashians varies significantly, suggesting differing levels of influence or interest among YouTube users.

4.2.2 Engagement Data Analysis on Reddit:

In Figure 5, the bar chart details user engagement on Reddit, again measured by likes. Here, Kylie Jenner and Kendall Jenner lead in engagement, followed by Kim Kardashian and Dwayne Johnson. The distribution is different from YouTube, indicating platform-specific patterns of celebrity influence. This might reflect the content preferences of Reddit users, or the nature of the discussions sparked by posts related to these celebrities.

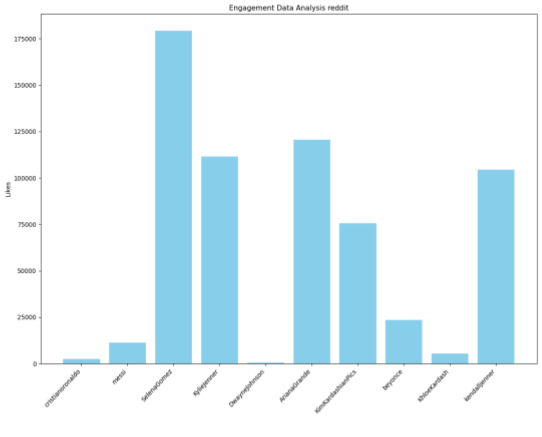


Figure 4- Engagement data- Reddit

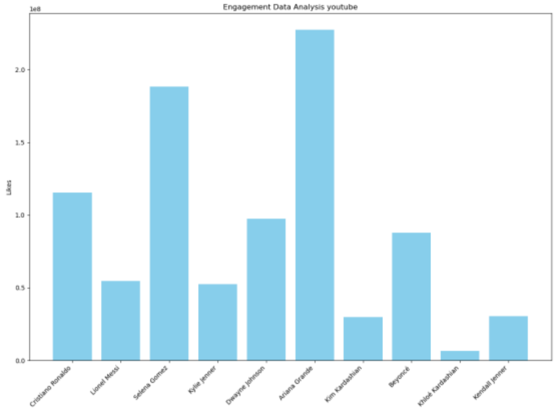


Figure 5- Engagement data- YouTube

4.3 Sentiment analysis:

To perform sentiment analysis in our project on celebrity influence using YouTube and Reddit data:
Collect Text Data: We Use APIs to gather relevant text data (comments, posts) from Reddit related to the selected celebrities
Preprocess Data: Clean the text data (remove special characters, non-relevant information, etc.) and standardize it for analysis.
Choose a Sentiment Analysis Tool: We used Sentiment Intensity Analyzer from nltk library
Analyze Sentiment: Apply the tool to your text data to classify sentiments (positive, negative, neutral).
Aggregate Results: Compile the sentiment scores for an overall sentiment picture for each celebrity.
Interpret and Integrate: Interpret the sentiment data in the context of celebrity influence and integrate these findings into your overall analysis.

4.3.1 Toxicity Levels in Reddit (Entertainment Subreddits):

In Figure 6, the graph shows sentiment levels across various entertainment-related subreddits, such as movies, Netflix, Bollywood, videos, and fantasy football. The 'movies' subreddit has a high level of negative sentiment compared to the others. This might indicate a more critical or contentious environment, which could be influenced by divisive opinions on films or celebrity actions within the movie industry.

4.3.2 Toxicity Levels in Reddit (Sports Subreddits):

In Figure 7, the Sentiment levels across different sports-related subreddits are depicted here, including NFL, CFB (College Football), Cricket, Baseball, and Formula1. NFL and CFB exhibit a higher proportion of negative sentiment, which might reflect intense fan rivalries or controversies surrounding the sports or athletes.

4.3.3 Toxicity Levels in Reddit (Music Subreddits):

In Figure 8, the graph displays sentiment analysis for subreddits dedicated to music and individual artists like Selena Gomez, Ariana Grande, Taylor Swift, and Travis Scott. The subreddit for Taylor Swift shows a notably high level of negative sentiment, which could be due to various factors such as fandom dynamics or discussions surrounding her public image and musical works.

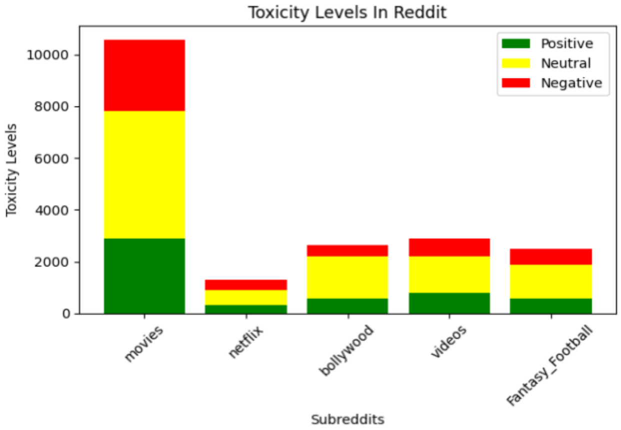


Figure 6- Toxicity Levels in Reddit- Entertainment

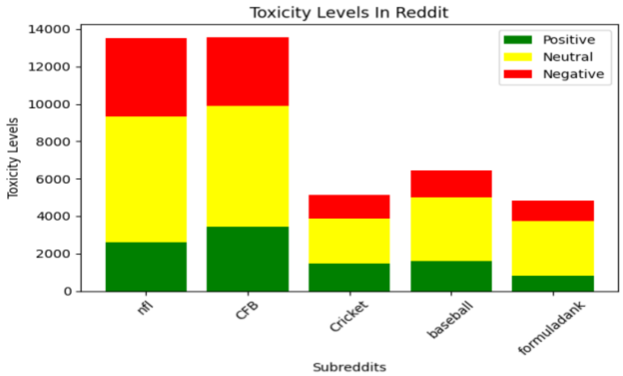


Figure 7 – Toxicity Levels in Reddit- Sports

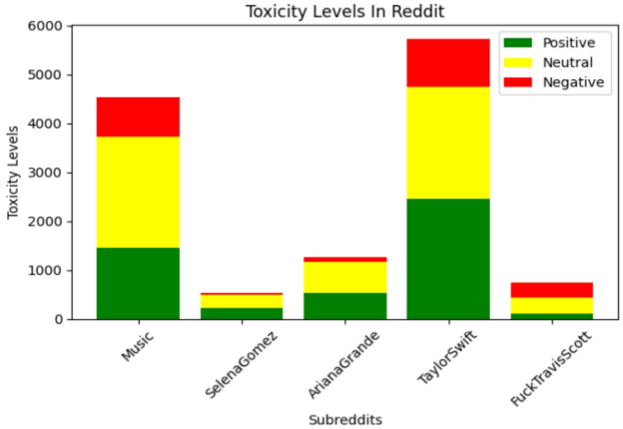


Figure 8- Toxicity Levels in Reddit- Music

4.3.4 Quantitative Comparison of Toxicity Levels in Reddit

In Table 1, the comparison reveals that within the entertainment category, the 'movies' subreddit exhibits the highest negative sentiment, whereas 'fantasy football' shows the lowest. In sports, the 'NFL' subreddit stands out for high negative sentiment, while 'Formula1' has comparatively lower negativity. For music, 'Taylor Swift' garners the most negative sentiment, contrasting with 'Selena Gomez', which has the least. This quantitative comparison provides insight into the varied reception of different content types on Reddit, useful for understanding the impact of celebrities and their associated discussions on today's generation.

Subreddit Category	High Negative Sentiment Subreddit	High Negative Sentiment Level	Low Negative Sentiment Subreddit	Low Negative Sentiment Level
Entertainment	Movies	10,000	Fantasy, Football	2,000
Sports	NFL	12,000	Formula1	2,000
Music	Taylor Swift	5,000	Selena Gomez	1,000

Table 1- Toxicity Score Comparison of Subreddits

4.4 Trends Analysis :

To perform a trends analysis in your project on celebrity influence using YouTube and Reddit data:
Collect Data: Gather data from YouTube and Reddit APIs focusing on selected celebrities.
Clean and Organize Data: Remove irrelevant entries and arrange data chronologically.
Identify Key Metrics: Choose metrics like mention frequency, sentiment scores, and engagement rates.
Time Series Analysis: Plot these metrics over time to identify patterns and changes.
Compare Trends: If analyzing multiple celebrities, compare their data to identify differences or similarities.

4.4.1 Cumulative Distribution Function of Daily View Counts :

Figure 9 - likely represents the probability distribution of daily views for celebrity-related content. The X-axis shows the view counts on a logarithmic scale, indicating a wide range of values. The Y-axis is the cumulative probability (CDF), which increases from 0 to 1. The steep curve in the middle suggests that a small number of posts receive a high number of views, which is a common characteristic of online content distribution where a few items are extremely popular.

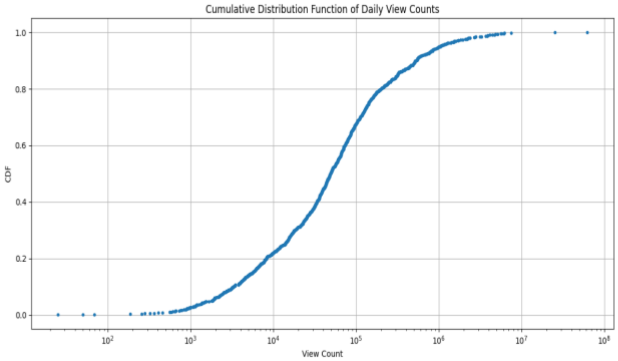


Figure 9- CDF of Daily View Count

4.4.2 Trends analysis of celebrities influence on reddit platform:

4.4.2.1 Post Count by Day:

The Figure 10, shows the number of posts made on Reddit over a series of dates. There is a noticeable peak at the beginning of the time series, on 2023-11-17, followed by a decline and then a plateau. This suggests that there might have been a particular event or a series of posts related to a celebrity that sparked a lot of initial interest, leading to a surge in posts. The following days show variability but maintain a relatively steady average, indicating a sustained interest over time but not at the initial surge level.

4.4.2.2 Comment Count by Day:

The second graph displays the number of comments made on Reddit posts over the same set of dates. Similar to the post count, there is a significant peak in comments on 2023-11-17, which implies a high level of user engagement on that date. The subsequent days show a decrease but with intermittent spikes, which may correlate with ongoing discussions or additional events related to celebrities that revived interest and interaction among users.

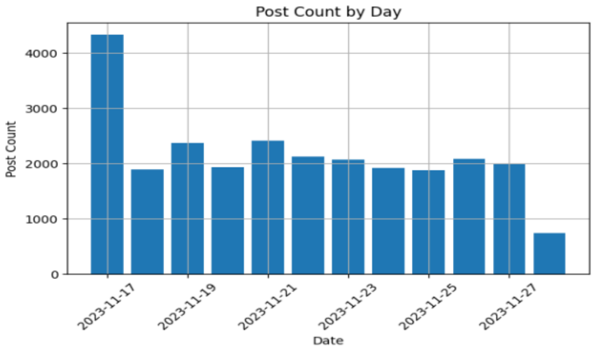


Figure 10- Posts Count/Day

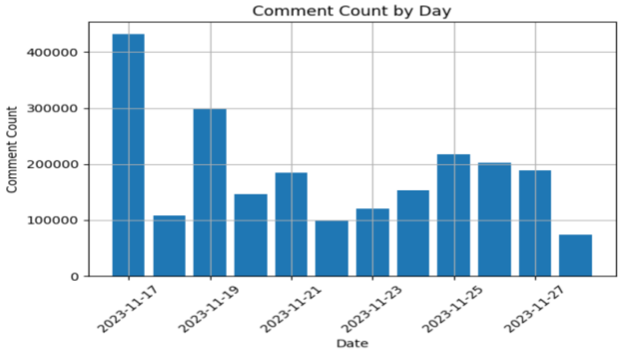


Figure 11- Comment Count/Day

4.5 Trend Analysis of /politics subreddits :

4.5.1 Post Count by Day:

In Figure 12, the bar chart tracks the number of posts made each day within the /politics subreddit. There's a high number of posts on the first day shown (2023-11-17), with a noticeable decrease afterward. The post frequency fluctuates over the subsequent days, suggesting varying levels of user activity and possibly reflecting the influence of current events or popular discussions.

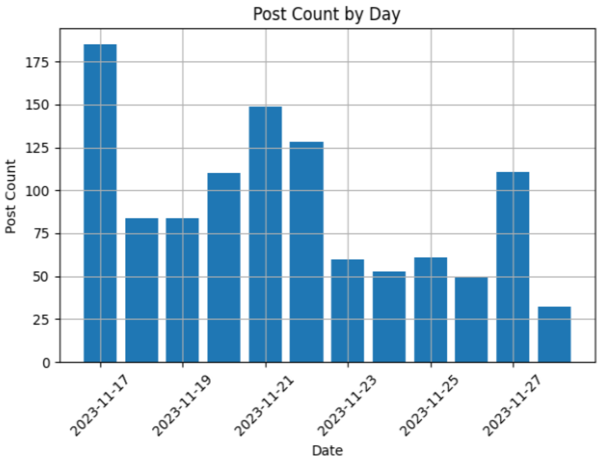


Figure 12- Post Count/ Day

4.5.2 Comment Count by Day:

In Figure 13, the bar chart illustrates the daily comment count within the /politics subreddit. Similar to the post count graph, the highest number of comments is on 2023-11-17, decreasing afterward but with less pronounced fluctuation than the post count. This implies that while fewer posts may be made after the initial peak, the discussions they generate remain robust, indicating sustained user engagement with the content.

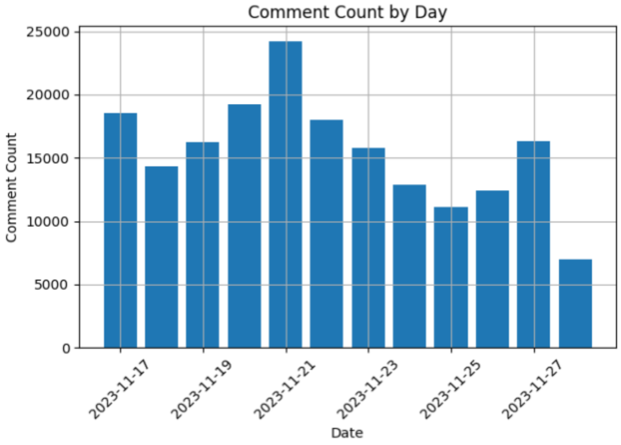


Figure 13- Comment Count/ Day

4.5.3 Hourly Comment Count on /politics:

In Figure 14, The line graph illustrates the hourly comment counts on the /politics subreddit, showing significant fluctuations in user activity. Two prominent spikes suggest events or discussions that drove heightened interaction. The pattern of comments suggests active user engagement, with periods of intense discussion likely tied to specific political events or news. This graph is indicative of the dynamic nature of political discourse on Reddit.

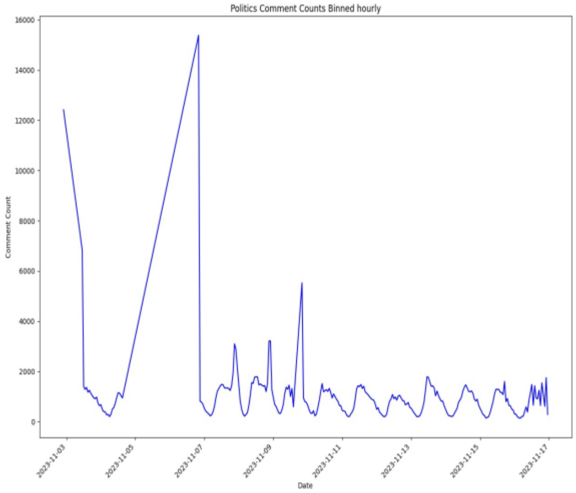


Figure 14-Politics Comments Count Binned Hourly

4.5 Modern Hate Speech API:

We have used modernHateSpeech API to retrieve a predicted class and confidence score on the gathered data from YouTube and reddit platform. In Figure 15, the graph displays the cumulative distribution function (CDF) of toxicity scores from YouTube (orange line) and Reddit (blue line) over time. Both lines closely track each other, indicating similar distributions of toxicity scores on both platforms. The sharp rise at the high end of the score spectrum suggests a smaller proportion of comments are classified with high toxicity confidence. The closeness of the two lines throughout suggests that the levels of toxicity, as well as the confidence in those measurements, are quite similar between YouTube and Reddit.

-

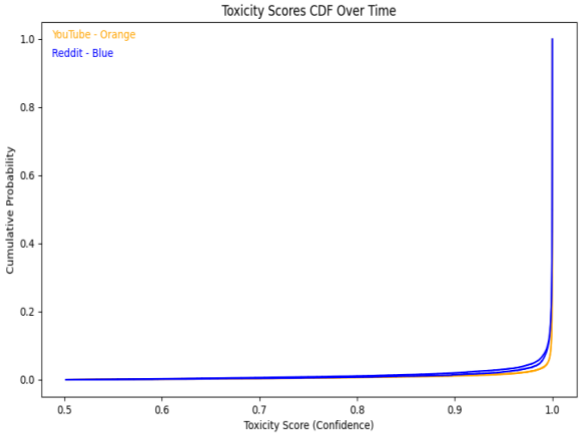


Figure 15- YouTube, Reddit Toxicity Score

The below graph Figure 16 compares the number of comments we received from all the data source such as r/politics subreddit, YouTube and the various celebrities from the various subreddits as discussed in the above section. The number of comments we received in reddit were the highest based on the given time frame from the 17th of November to 28th November.

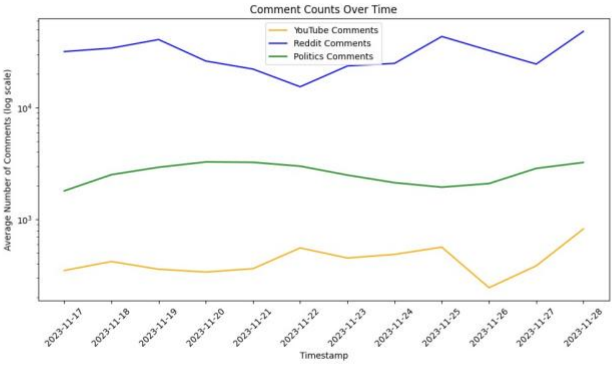


Figure 16- Reddit, YouTube, Politics Subreddit Comment Counts

5. DISCUSSION

This project is focused on the influence of celebrities in the group of people based on the videos, comments, likes, view count and toxicity score of the comments of the two data source YouTube and Reddit. We took subreddits such as the one given below:

Subreddit:

"r/Entertainment","r/netflix","r/Music","r/videos","r/pics","r/nfl","r/CFB","r/memes","r/bollywood","r/Fitness","r/TikTok","r/Cricket","r/baseball","r/formuladank","r/SelenaGomez","r/KylieJenner", "r/hiphopheads","r/ArianaGrande","r/KimKardashianPics","r/Fantasy_Football","r/TaylorSwift","r/KendallJenner","r/bangtan","r/movies","r/FuckTravisScott","r/celebrities"

The reason of taking these subreddits because they were highly active in terms of discussion, public opinion and also, we focused on the top celebrities in the various sectors such as sports, music, acting, politics, entertainment because we wanted to study the top celebrities impact on the huge crowd of people.

We did the same for YouTube as well, the channels we focused on were entertainment, sports, music, fitness. We created separate collections in the MongoDB where we stored comment counts, likes, views, number of posts for detailed study.

We made log errors that we got in the collection to do a focused study on the API response and how a huge peak impacted the functioning of the API. We developed jobs in the way of batches of timeframe for periodic fetching using multiple API Keys.

Our three research questions are that we are performing sentiment analysis, trend analysis, and comparative analysis to answer the insightful questions such as: 1) Sentiment Analysis: Implement sentiment analysis on both YouTube comments and Reddit discussions to gauge public sentiment towards celebrities. Are comments generally positive, negative, or neutral? 2)Trend Analysis: Analyze trends in celebrity-related content on YouTube. Which celebrities are currently trending, and why? Are certain types of content (e.g., interviews, behind-the-scenes footage) more popular? Explore the most discussed celebrities on Reddit. Are there particular events or controversies that lead to spikes in discussions? 3)Influence Metrics: Develop influence metrics to measure the impact of celebrities on viewers and readers. Consider factors like engagement, subscriber growth, and content popularity.

Sentiment: We observed that posts and comments in the subreddits were increased significantly during a peak event, the comments that we received were majorly neutral in the movies subreddits because majority of the discussion was regarding the opinions on new movies and releases. Also, we observed that negative comments regarding one fashion was the most toxic In this subreddit. It has the highest number of posts and comments received. Also, in the sports section, the subreddit that matched same as movies was nfl. In the music section, most toxic and neutral posts and comments were received in the Taylor Swift subreddit.

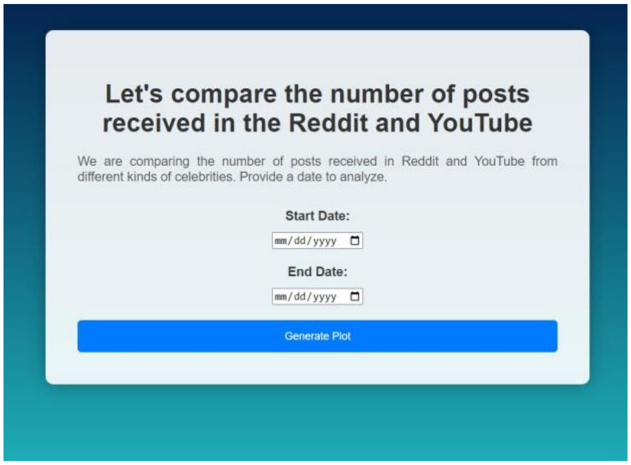
Trend: We observe that the number of posts and comments we received was on 17th November because there was an event of festival/political during the time. In reddit and YouTube we observed number of posts and videos to be very high with active engagement of the user. The count was 175 posts per half an hour. Also, there was a controversial discussion on the recent election which led to a spike in the comments on 21st November. This is also based on the /politics subreddit.

Comparative: In this analysis we compared the top celebrities based on the above analysis on the various data set such as Reddit and YouTube, the aim was to find the toxicity score, engagement level of the same celebrities on different platforms. We observed that Selena Gomez and Ariana Grande were highly trending on Reddit and YouTube based on the posts and the comments that we received.

6. PROJECT 3 IMPLEMENTATION

For the dashboard, we ended up making a web-app using Python flask and HTML, which takes number of parameters as an input based on the analysis:

1-Comparative: In this analysis we are taking range of two dates as an input in YYYY-MM-DD format which will then call HTML file which is mapped to the python script comparative.py, the webapp is hosted locally on the following pair of IP address and the local-host: port 5000, please check the attached URL here which is given below (http://127.0.0.1:5000/comparativeAnalysis), which will query the MongoDB collection based on the given name and generate a plot which will show comparison between the number of posts that was received on both Reddit and the YouTube as shown in the figure.



Let's compare the number of posts received in the Reddit and YouTube

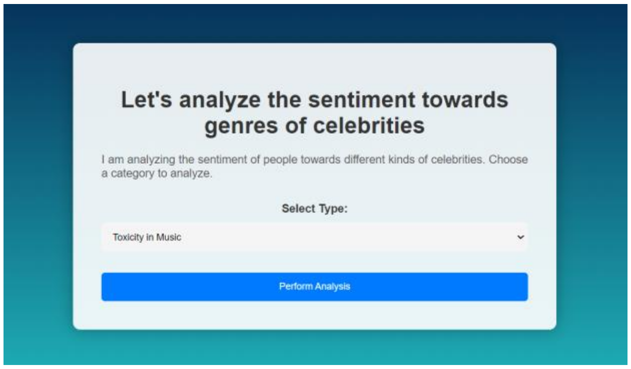
We are comparing the number of posts received in Reddit and YouTube from different kinds of celebrities. Provide a date to analyze.

Start Date: mm/dd/yyyy

End Date: mm/dd/yyyy

Figure 17- Comparative Dashboard

2-Sentiment Analysis: In this analysis we are focusing on three genres of celebrities that is music, entertainment, and the sports. We will find toxicity in the analysis based on the date. This analysis worked by calling the HTML file of display and then the script sentiment.py which will query the MongoDB collection and generate a plot which will show a bar chart, with three sections each representing the data source. The website link which is given as follow: - (http://127.0.0.1:5000/sentiment_analysis)



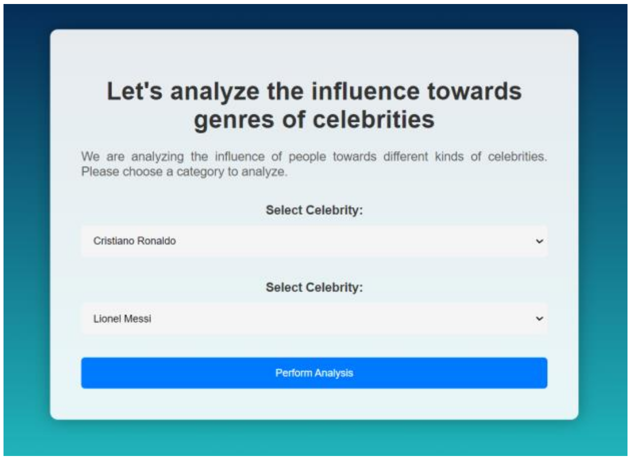
Let's analyze the sentiment towards genres of celebrities

I am analyzing the sentiment of people towards different kinds of celebrities. Choose a category to analyze.

Select Type: Toxicity in Music

Figure 18- Sentiment Dashboard

3-Influence Analysis: In this analysis we will be focusing on the influence of celebrities in the group of people. For this, we take two celebrities as an input and based on the input, our HTML files will call script named sentiment.py which will query the MongoDB collection and generate a plot which will show comparison based on the likes, posts, comments of the celebrities. All the plots are executed by the matplotlib. We used SSH to access the VM and make the dashboard which has host IP 127.0.0.1. The URL is given below :- (http://127.0.0.1:5000/influence).



Let's analyze the influence towards genres of celebrities

We are analyzing the influence of people towards different kinds of celebrities. Please choose a category to analyze.

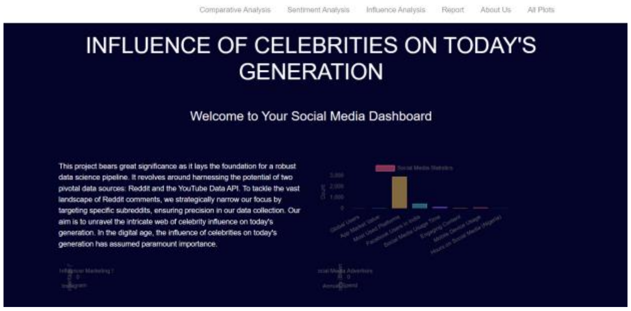
Select Celebrity: Cristiano Ronaldo

Select Celebrity: Lionel Messi

Figure 19- Influence Dashboard

7. DASHBOARD

Given below Figure 20 shows the main dashboard which is hosted on the URL :- (http://127.0.0.1:5000/) it represents the home page of the flask based web-app where we have the flexibility to monitor the influence of celebrities by three analysis as discussed above in Section 6. We have also given a button for Plots gallery which will show all the plots that we generated so far in the project. Also, we have given the option of About Us: where our teammates efforts and guidance are declared. In the analysis, we have the data from 9th November to 28th November. We have used plotting mechanisms such as matlabplotlib that plots the live data fetched from the MongoDB database and our code is written in HTML for front end display and in backend we have used Python Flask. We have also implemented basic error handling mechanisms for the date such as if two dates are same, then error will be displayed, and both the dates cannot be null. Also, we have focused on the logic that the end date can never be less than the start date, which is a better way of representing the date. In case a user tries to give a beyond future based date, we are throwing an error. Our aim is to gather a learning of top influential celebrities so that we can see which celebrity is most popular and how their behavior impacts the youth of our country. Toxicity in sports and music has also been a big concern since in sports, most toxic players can be studies and their patterns can be recognized.



INFLUENCE OF CELEBRITIES ON TODAY'S GENERATION

Welcome to Your Social Media Dashboard

This project bears great significance as it lays the foundation for a robust data science pipeline. It revolves around harnessing the potential of two pivotal data sources: Reddit and the YouTube Data API. To tackle the vast landscape of Reddit comments, we strategically narrow our focus by targeting specific subreddits, ensuring precision in our data collection. Our aim is to unravel the intricate web of celebrity influence on today's generation. In the digital age, the influence of celebrities on today's generation has assumed paramount importance.

Figure 20- Home Dashboard

8. BUG REPORT

We observed that the plotting of the graph takes some time to produce since they are using Reddit and YouTube based API to send requests, retrieve data in real time from the MongoDB database, parse the answer and then provide a data frame. Despite this, we did not face a major bug.

9. RESULTS

We observed that the Taylor Swift, Selena Gomez, and Ariana Grande were the major influential and most active engagement celebrities of the music field on both Reddit and YouTube. Out of this, the most toxic comments we observed were about Selena Gomez due to her recent controversy. In the field of movies and entertainment, we observed subreddit name r/movies have been the most active, controversial, and toxic from the last 1 month. In the field of sports, Fantasy Football, NFL was the most engaged with maximum posts and comments.

10. CONCLUSION

In conclusion, as per our work, we observed Reddit has the highest user engagement as compared to YouTube. Our work was confined to the monitoring of the top 10 celebrities of various fields including music, entertainment, and sports. More celebrities’ lists can be added on subreddit and the YouTube channels. The future scope lies in the fact that we can use other NLP and ML based advanced learning methodologies rather than our Sentiment Intensity Analyzer strategy.

7. REFERENCES

[1] The link for the retrieval of the top 10 celebrities based on the Instagram https://www.forbesindia.com/article/explainers/most-followed-instagram-accounts-world/85649/1
[2] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10165405/.
[3] Impact of the celebrity endorsement on purchase decisions : A study among the youth of Barely, India by Tarun Gupta - ISSN: 2249-7196
[4] "Analyzing Social Media Networks with NodeXL" by Derek Hansen, Ben Shneiderman, and Marc A. Smith
[5] "Social Media and Public Relations: Fake Friends and Powerful Publics" by Judy Motion et al
[6] https://github.com/cjhutto/vaderSentiment
[7] https://github.com/pallets/flask
[8] https://www.forbes.com/advisor/business/social-media-statistics/
[9] https://sproutsocial.com/insights/social-media-statistics/
[10] https://datareportal.com/social-media-users