Analysis of Cultural Impact on Tech Discussions: Reddit vs YouTube

Aditya Anand Doshi
adoshi3@binghamton.edu
State University of New York at Binghamton
Binghamton, New York, USA

Sharon Yang
syang20@binghamton.edu
State University of New York at Binghamton
Binghamton, New York, USA

Sagar Naresh Sidhwa
ssidhwa@binghamton.edu
State University of New York at Binghamton
Binghamton, New York, USA

ABSTRACT

Why Tech and Culture? The global nature of technology means products launched in one country can become talking points across the world. By studying discussions around tech topics, we can uncover cultural biases, preferences, and unique perspectives. Our project gathers data from two primary sources: Reddit and YouTube. From Reddit, posts, comments, upvotes, and downvotes from tech-related subreddits are collected. YouTube provides data on video metadata, comments, likes, dislikes, and views from tech reviewers across different cultural backgrounds or regions. Both Reddit's and YouTube's API are scheduled to run daily.

1 INTRODUCTION

In the era of globalization, technology is not just a catalyst for innovation but also a medium through which cultural nuances are expressed and shared. Social media platforms, such as Reddit and YouTube, act as virtual melting pots, bringing together diverse perspectives on technology from different regions and cultures. The discourse around technology in these platforms is rich and multi-faceted, providing an opportunity to delve into understanding how culture influences discussions about technological advancements, products, and events.

This research aims to analyze the cultural impact on tech discussions by examining and comparing the dialogues and engagement metrics on Reddit and YouTube. Reddit, known as "the front page of the internet," provides an anonymous space for individuals to discuss and debate a myriad of topics, including technology. On the other hand, YouTube, as a video-sharing platform, offers visual insights into tech topics through product reviews, news, and opinion pieces presented by influencers from various cultural backgrounds. By studying these platforms, we intend to explore how different cultures perceive and interact with technology-related content.

Through this analysis, we aim to uncover patterns, biases, preferences, and unique cultural insights that shape the narrative around technology in the online sphere. This endeavor is significant in understanding the intertwining of tech discourse and cultural context, thereby providing valuable insights for tech developers, marketers, and policymakers.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 DATA SOURCE

2.1 Reddit

Reddit, often described as "the front page of the internet," is a vast platform of user-generated content, discussions, and community engagement. For the purpose of this study, we've focused our attention on a specific selection of subreddits that sit at the intersection of technology and culture.

2.1.1 Selection Criteria for subReddits. The process for selecting the 20 subreddits involved a systematic approach:

(1) **Primary Analysis:** We began by identifying subreddits that directly align with technology discussions, e.g., /r/technology, /r/gadgets. Additionally, we considered subreddits centered around cultural discussions for potential overlaps.
(2) **Keyword Analysis:** Employing keywords such as 'tech', 'culture', and 'innovation', we analyzed which subreddits often hosted relevant discussions.
(3) **Engagement Analysis:** Subreddits were ranked based on engagement metrics to ensure they were not just relevant but also active.
(4) **Overlap Analysis:** A search was conducted to find subreddits that discuss both tech and culture, allowing us to capture nuanced discussions.
(5) **Diversity Check:** Ensuring a global perspective, we included region-specific tech and culture subreddits.

Following this methodology, we selected the following subreddits: 'technology', 'tech', 'ScienceAndTechnology', 'compsci', 'badscienceculture', 'SocietyAndCulture', 'CultureAndGenerations', 'Futurology', 'gadgets', 'Apple', 'Android', 'AskTechnology', 'AskReddit', 'programming', 'pcgaming', 'worldnews', 'science', 'cyberpunk', 'AskScience', 'dataisbeautiful'

2.2 YouTube

YouTube, a globally renowned video-sharing platform, offers a plethora of content categories. For this project, we targeted content that provides insights into technology from diverse cultural perspectives.

2.2.1 Video Category and Attributes Selection. The process for video data collection involved:

(1) **Category Selection:** We primarily zeroed in on 'Science & Technology', 'Education', and 'News & Politics'. These

categories are ripe with discussions about tech's cultural implications.

(2) **Keyword Filtering:** Videos were filtered based on keywords such as 'technology', 'culture', 'innovation', and 'digital culture'.
(3) **Channel Source:** Data was collected from a set of prominent channels including 'TED', 'TEDx Talks', 'TED-Ed', 'Veritasium', 'WIRED', 'Vsauce', 'Marques Brownlee', 'ColdFusion', 'Simplilearn', 'Linus Tech Tips', 'Quantum Tech HD', 'Bloomberg Technology', 'Yahoo Finance', 'The Wall Street Journal', 'CNN', 'BBC News', 'Unbox Therapy', 'Technical Guruji', 'Dave Lee (Dave2D)', 'Austin Evans', 'Arirang News', 'Krystal Lora', 'TechAltar', 'Geekyranjit', 'C4ETech', and 'TechnoBuffalo'.
(4) **Engagement Metrics:** Videos with higher engagement metrics (likes, comments) were prioritized as they signify relevance and impact.

3 DATA COLLECTION

Data collection is a crucial step in our research, allowing us to gather the necessary information for analysis. We focused on extracting data from two major platforms: Reddit and YouTube. Reddit, a popular social discussion site, and YouTube, a video-sharing platform, both provide extensive APIs that facilitate the extraction of vast amounts of data. This data enables us to gain insights into various aspects of user interactions and preferences.

3.1 Reddit Data Collection

To effectively gather and understand the nature of the data derived from Reddit, we interfaced with Reddit's API using Python. The Reddit API requires an access token for authentication. This token was obtained by combining the client ID, client secret, username, and password with the designated authentication endpoint. To uniquely identify our application, we employed a generated user agent. To avoid potential request blocks due to repetitive agent strings, we developed a custom function to create random user agents, ensuring a smoother data collection experience.

Our data collection focused on several key areas:

Subreddit Data Collection: Using the endpoint https://oauth.reddit.com/r/[subreddit_name]/about.json, we aimed to gather detailed information about specific subreddits, such as its title, description, and subscriber count. The gathered data was stored in the 'subreddits' table within our PostgreSQL database, which includes fields like subreddit_title, subreddit_name, and subreddit_id.

Post Data Collection within a Subreddit: We extracted the latest posts from specific subreddits using the endpoint https://oauth.reddit.com/r/[subreddit_name]/new.json. This data was saved in the 'subreddit_posts' table in our database, which maintains details like post_id, subreddit_id, and post_name.

Comments Data Collection for a Post: To collect top-level comments associated with specific posts in a subreddit, we used the endpoint https://oauth.reddit.com/r/[subreddit_name]/comments/[post_id].json. Comments were stored in the 'subreddit_posts _comments' table, with fields such as comment_id, post_id, and post_name.

Nested Comment (Replies) Data Collection: For recursive data extraction to fetch nested comments or replies under a specific comment, we used the endpoint https://oauth.reddit.com/r/[subreddit_name]/comments/[post_id]/[comment_id].json. Similar to the top-level comments, these replies were stored in the 'subreddit_posts_comments' table.

Our data extraction script was designed to cycle through a predefined list of subreddits repeatedly. After extracting data from all subreddits, the script paused for 7200 seconds before initiating another cycle. We utilized the psycopg2 library to establish a connection with our PostgreSQL database named 'socialgood'. Both the extraction from Reddit and insertion into the database were managed within this script.

Error handling was crucial to our process. Exceptions were meticulously managed to ensure that potential API or database connection errors did not disrupt the entire data collection procedure. Given the rate limits associated with Reddit's API, we monitored our total request count diligently. After data extraction from each subreddit, the script paused for 121 seconds before moving on, ensuring compliance with Reddit's API request frequency guidelines.

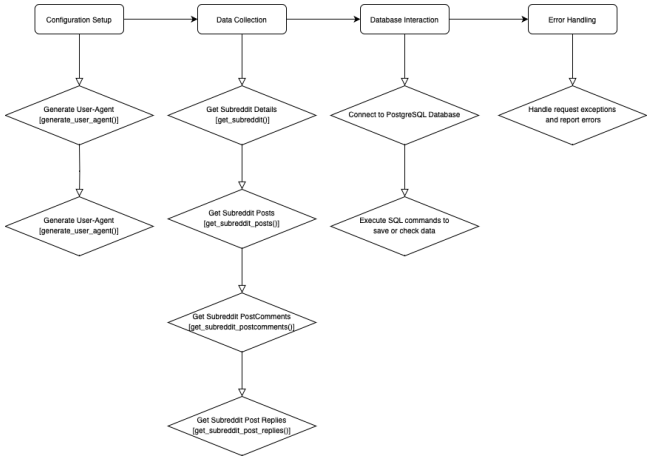

Figure 1: Subreddit Data Collection Impelementation.

3.2 YouTube Data Collection

The objective of the data collection process from YouTube was to comprehensively gather pertinent information about YouTube channels, their associated videos, and user comments. We utilized YouTube's v3 API, facilitated through the `googleapiclient.discovery` library.

3.2.1 Connection with PostgreSQL Database. Our choice for data storage was the PostgreSQL database due to its robustness and versatility. A connection was established to this database, ensuring that all data fetched would be stored systematically for easy retrieval and further analysis.

3.2.2 YouTube API Utilization. Given YouTube API's stringent usage constraints, it was pertinent to strategically use the API keys. To maximize our data collection efficiency, we employed a cyclical

use of multiple API keys. This approach allowed us to almost continuously fetch data while adhering to YouTube's terms of service.

3.2.3 Modularized Data Collection. We adopted a modular approach to data extraction by creating distinct functions tailored to gather specific data types:

- Metadata related to YouTube channels.
- Replies to comments under specific videos.
- Top-level comments from videos.
- Metadata for videos under specific channels.

Each function ensured that data duplication was avoided by checking the new data against the already stored data in our database. If new and unique, the data was then stored systematically in its respective table.

3.2.4 Iterative Approach for Data Collection. An iterative mechanism was implemented for continuous data collection. By cycling through a predefined list of channel IDs and cycling between different API keys, we ensured that the maximum amount of data was collected within the given API constraints. To prevent any potential breaches of rate limits set by YouTube, specific intervals were maintained between each data collection cycle.

3.2.5 Error Management. Considering the unpredictability of data collection processes, we integrated an exception handling mechanism. This ensured that any unforeseen errors, particularly during database connectivity or during data extraction, were appropriately logged. This error-logging allowed for immediate action and troubleshooting, ensuring the system's smooth functioning.

In conclusion, our approach to YouTube data collection was methodical, automated, and efficient, ensuring comprehensive data extraction while strictly adhering to YouTube's API constraints and guidelines.

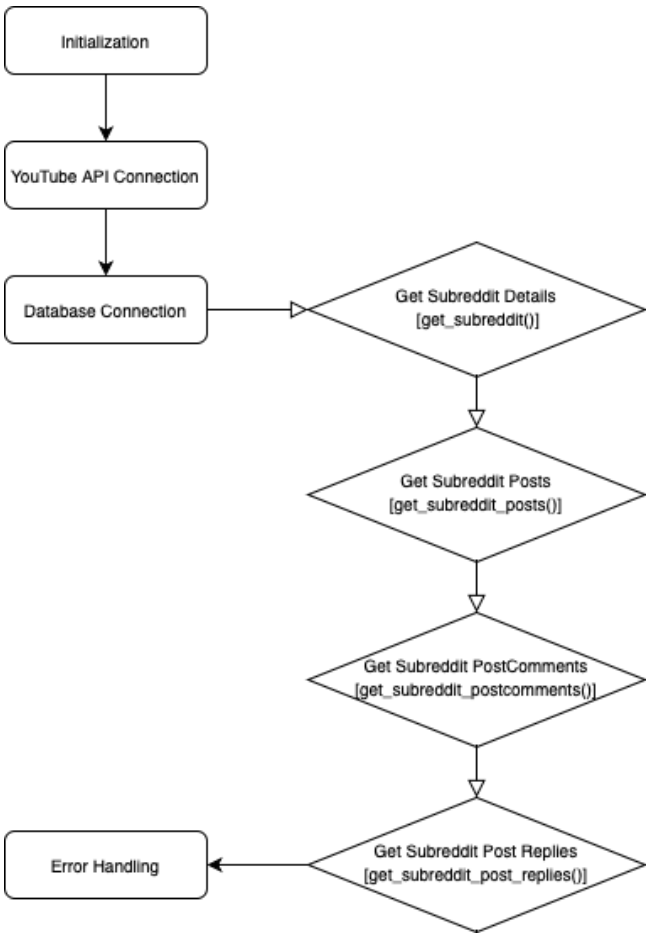

Figure 2: YouTube Data Collection Impelementation.

4 TECHNOLOGY FRAMEWORK

For our project, we have chosen **Python** as the primary programming language due to its versatility and extensive library support. Our data collection is powered by the **Reddit API** and **Google's YouTube Data API v3**, which allow us to access a wealth of information from these platforms.

Our analysis employs a range of Python libraries such as **Pandas** for data manipulation, **NumPy** for numerical computations, and **Matplotlib** for data visualization. We conduct natural language processing using **NLTK** and **TextBlob**, which provide tools for sentiment analysis and keyword extraction. For creating word clouds, we utilize the **WordCloud** library, which helps in visualizing the most frequent terms from our datasets.

To assess the toxicity levels in textual content, we integrate the **ModerateHatespeech API** into our analysis. This specialized tool allows us to quantify and evaluate the nature of discussions, contributing a significant dimension to our research.

Our relational database system of choice is **PostgreSQL**, selected for its robust data storage capabilities. We manage backend operations and API integrations through **Flask**, a lightweight web framework in Python. Task scheduling is efficiently handled by **Flask-APScheduler**, and database connectivity is facilitated by **Psycopg2**, which serves as the PostgreSQL adapter.

For large-scale data processing and analysis, we also make use of **Jupyter Notebooks**, which provide an interactive coding environment. This allows us to document our analysis process and share results in a reproducible manner.

Documentation and drafting are carried out on **Overleaf**, a collaborative cloud-based LaTeX editor. Its LaTeX support and collaborative features are indispensable for maintaining the integrity and quality of our project reports.

5 ANALYSIS APPROACHES

Our study's primary focus is to understand the cultural nuances in tech discussions across various platforms, mainly Reddit and YouTube. To this end, we employ multiple analytical approaches:

Engagement Metrics Analysis: This method identifies patterns of engagement, such as likes, comments, and shares. We investigate whether these engagement metrics correlate with the cultural or regional origin of the content or its audience. Our analysis includes a detailed assessment of engagement trends over time, offering insights into the evolving nature of user interactions on these platforms.

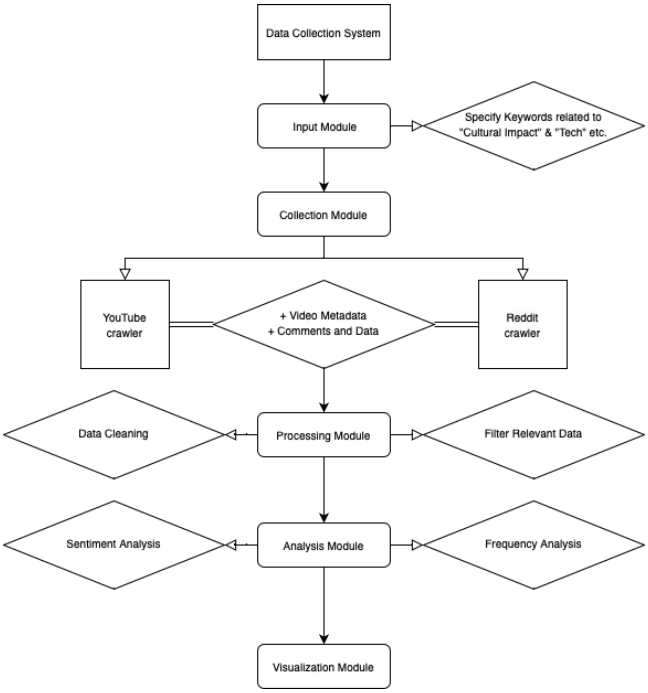

Figure 3: Proposed System Impelementation.

all_tables_description

	subreddit	subreddit_posts	subreddit_posts_comments	youtube_channels	youtube_channel_videos	youtube_videos_comments
0	[{'s_id': 'SERIAL PRIMARY KEY'}]	[{'p_id': 'SERIAL PRIMARY KEY'}]	[{'c_id': 'SERIAL PRIMARY KEY'}]	[{'yc_id': 'SERIAL PRIMARY KEY'}]	[{'ycv_id': 'SERIAL PRIMARY KEY'}]	[{'ycvc_id': 'SERIAL PRIMARY KEY'}]
1	[{'subreddit_date_time': 'timestamp'}]	[{'post_date_time': 'timestamp'}]	[{'comment_date_time': 'timestamp'}]	[{'yc_date_time': 'timestamp'}]	[{'ycv_date_time': 'timestamp'}]	[{'ycvc_date_time': 'timestamp'}]
2	[{'subreddit_title': 'VARCHAR'}]	[{'post_id': 'VARCHAR'}]	[{'comment_id': 'VARCHAR'}]	[{'yc_name_title': 'VARCHAR'}]	[{'ycvideo_id': 'VARCHAR'}]	[{'ycvcomment_id': 'VARCHAR'}]
3	[{'subreddit_name': 'VARCHAR'}]	[{'subreddit_id': 'VARCHAR FOREIGN KEY REFEREN...	[{'post_id': 'VARCHAR FOREIGN KEY REFERENCES s...	[{'youtube_channel_id': 'VARCHAR'}]	[{'youtube_channel_id': 'VARCHAR FOREIGN KEY R...	[{'ycvideo_id': 'VARCHAR FOREIGN KEY REFERENCE...
4	[{'subreddit_id': 'VARCHAR'}]	[{'subreddit_url': 'TEXT'}]	[{'post_name': 'VARCHAR'}]	[{'yc_url': 'TEXT'}]	[{'yc_name_title': 'VARCHAR'}]	[{'ycv_url': 'TEXT'}]
5	[{'subreddit_url': 'TEXT'}]	[{'post_name': 'VARCHAR'}]	[{'post_url': 'TEXT'}]	[{'yc_custom_url': 'TEXT'}]	[{'yc_url': 'TEXT'}]	[{'youtube_channel_id': 'VARCHAR FOREIGN KEY R...
6	[{'subreddit_description': 'TEXT'}]	[{'post_url': 'TEXT'}]	[{'subreddit_id': 'VARCHAR FOREIGN KEY REFEREN...	[{'yc_description': 'TEXT'}]	[{'ycv_name_title': 'TEXT'}]	[{'yc_url': 'TEXT'}]
7	[{'subreddit_subscribers': 'BIGINT'}]	[{'post_title': 'TEXT'}]	[{'subreddit_name': 'VARCHAR'}]	[{'yc_data': 'JSON'}]	[{'ycv_url': 'TEXT'}]	[{'ycvc_body': 'TEXT'}]
8	[{'subreddit_data': 'JSON'}]	[{'post_data': 'JSON'}]	[{'subreddit_url': 'TEXT'}]	[{'Data Count': 16}]	[{'ycv_description': 'TEXT'}]	[{'ycvc_data': 'JSON'}]
9	[{'Data Count': 20}]	[{'post_categorized': 'INTEGER'}]	[{'comment_body': 'TEXT'}]	[-]	[{'ycv_data': 'JSON'}]	[{'ycvc_categorized': 'INTEGER'}]
10	[-]	[{'post_class': 'VARCHAR'}]	[{'comment_data': 'JSON'}]	[-]	[{'ycv_categorized': 'INTEGER'}]	[{'ycvc_class': 'VARCHAR'}]
11	[-]	[{'post_confidence': 'NUMERIC'}]	[{'comment_categorized': 'INTEGER'}]	[-]	[{'ycv_class': 'VARCHAR'}]	[{'ycvc_confidence': 'NUMERIC'}]
12	[-]	[{'Data Count': 15882}]	[{'comment_class': 'VARCHAR'}]	[-]	[{'ycv_confidence': 'NUMERIC'}]	[{'Data Count': 19232}]
13	[-]	[-]	[{'comment_confidence': 'NUMERIC'}]	[-]	[{'Data Count': 1193}]	[-]
14	[-]	[-]	[{'Data Count': 41940}]	[-]	[-]	[-]

Figure 4: Overview of the database schema illustrating the structure and relationships of tables within the social media analytics platform.

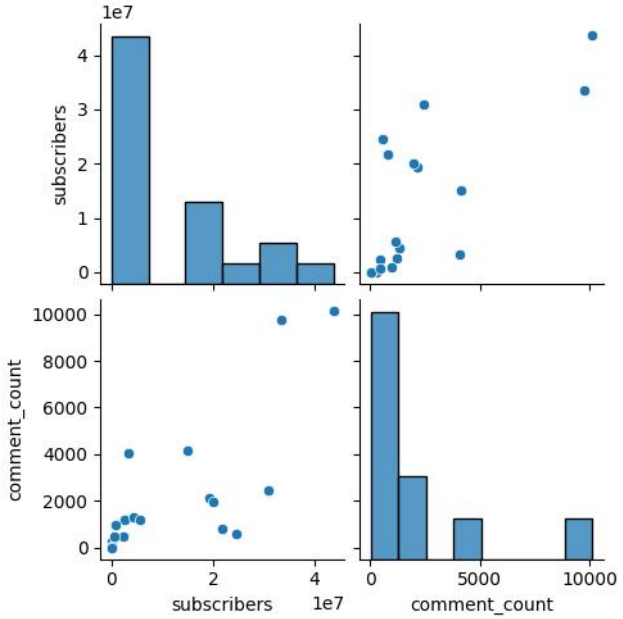

Figure 5: Engagement metrics analysis for Reddit, correlating the number of subscribers to the comment counts in different subreddits.

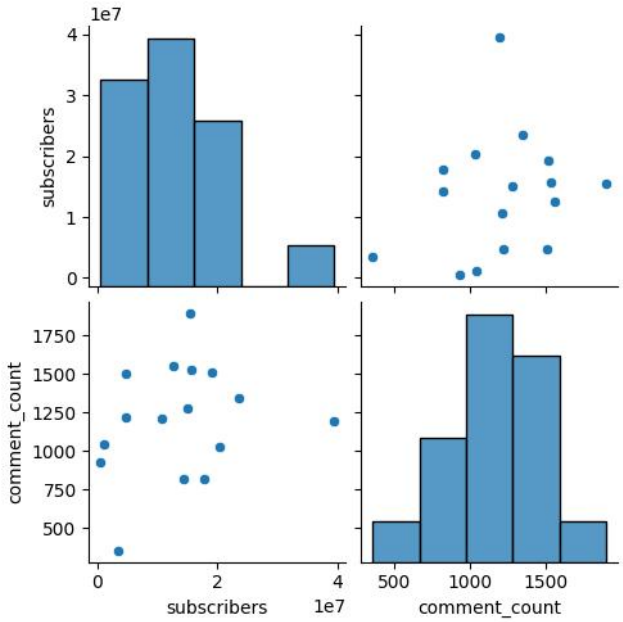

Figure 6: Analysis of engagement patterns on YouTube, exploring the relationship between channel subscribers and comment activity.

Content Popularity Analysis: We analyze the plethora of tech topics discussed daily to understand what's trending across platforms. Our approach involves quantitative analysis of post and video titles on Reddit and YouTube, helping us pinpoint trending tech subjects or products. This analysis is crucial in identifying the most influential and widely discussed tech topics in different cultural contexts.

Cultural Keyword Analysis: Our approach includes an in-depth analysis of the language used in tech discussions. We look for culturally specific terms or phrases that could indicate particular biases or preferences. By identifying and analyzing these recurring

Figure 7: Word cloud representing the most frequently occurring words in Reddit posts, highlighting prevalent themes in tech discussions.

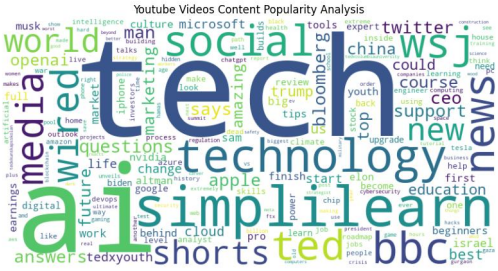

Figure 8: Word cloud visualization for YouTube video titles, emphasizing dominant topics and language in technology-related videos.

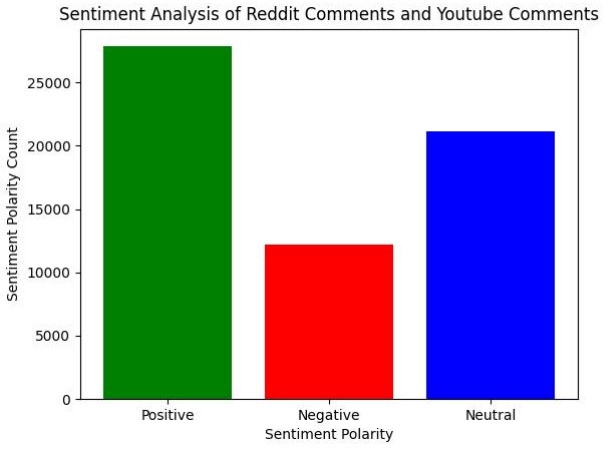

Figure 9: Comparative sentiment analysis of Reddit and YouTube comments, indicating the distribution of positive, negative, and neutral sentiments.

phrases or terms, we gain a clearer picture of cultural biases, values, or perspectives. This analysis is enhanced by sophisticated text analysis tools, enabling us to process and interpret large volumes of textual data efficiently.

Daily Counts and Trend Analysis: A new addition to our analytical toolbox, this approach focuses on the daily counts of posts and comments on Reddit and YouTube. We analyze these counts to identify trends and patterns over time, providing insights into the ebb and flow of tech discussions on these platforms. This analysis is supported by detailed data visualizations, illustrating the dynamics of user engagement and content creation day by day.

6 EXPLORATION AND MEASUREMENT OF DATA

6.1 YouTube

Our data tracking includes 20 channels that focus on technology, culture, and information. The analysis revealed the following insights:

- On average, each channel posts about 1 new video per day.
- The average number of comments per video is found to be 100.
- We collect basic details for each video, such as the title, view count, like/dislike count, and comments.

However, our data indicates fluctuations in daily video posts and comments, as visualized in the attached figures.

Total YouTube data per week: Based on our data, the weekly average equates to approximately 140 videos and 14,000 comments, considering the variances observed in daily activity.

6.2 Reddit

Our monitoring spans 20 technology-oriented subreddits. The data analysis has provided the following observations:

- Each subreddit witnesses an average of 50 new posts per day.
- On average, there are 40 comments per post.

Similar to YouTube, Reddit's daily post and comment counts also show variability, which is depicted in the graphs below.

Total Reddit data per week: The collected data suggests an average of 7,000 posts and 280,000 comments per week, with daily counts subject to change as demonstrated in the plots.

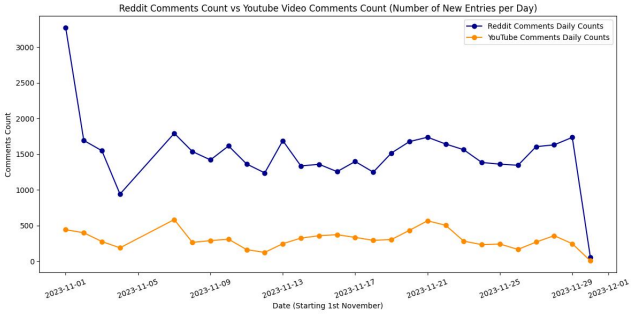

Figure 10: Comparison of daily comment counts on Reddit versus YouTube, illustrating the fluctuations in user engagement over time.

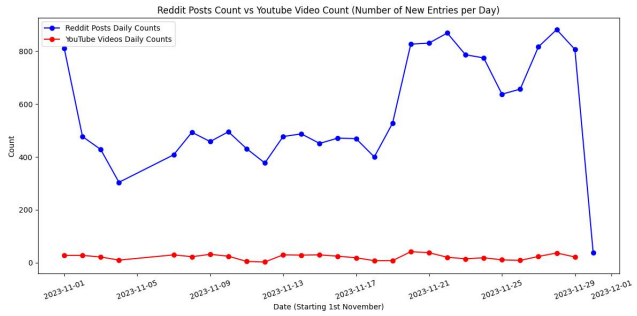

Figure 11: A comparison of daily post counts on Reddit with daily video counts on YouTube, revealing the dynamics of content generation on both platforms.

7 TOXICITY MEASUREMENT

7.1 Registration and Access

Our team has registered on the ModerateHatespeech platform and gained access to their Application Programming Interface (API), which is crucial for our toxicity analysis operations. An API key was obtained for authenticating and executing our requests to their service.

7.2 API Requests

We have thoroughly reviewed the ModerateHatespeech API documentation and utilized the endpoint https://api.moderatehatespeech.com/api/v1/moderate/ for retrieving toxicity scores. Our Python-based scripts have been making HTTP requests to this endpoint, submitting text data for evaluation and processing the responses.

7.2.1 System Robustness. We have designed our system with robustness in mind to handle potential outages from ModerateHatespeech. Sophisticated error handling and debugging mechanisms have been implemented, ensuring our operation's resilience and smooth data processing.

7.3 Receiving Toxicity Scores

The toxicity analysis of our dataset has produced substantial results. We have generated visuals that classify the dataset into 'normal' and 'flagged' categories, providing insights into the prevalence of toxicity within the comments and posts from Reddit and YouTube.

8 BACKGROUND AND RELATED WORK

8.1 Background

The advent of social media has revolutionized the way people discuss and disseminate information about technology. Platforms such as Reddit and YouTube have become primary sources for sharing opinions, experiences, and knowledge related to tech products, services, and innovations. These discussions are not only reflective of technological trends but also offer a window into the cultural and societal influences that shape public perception and discourse.

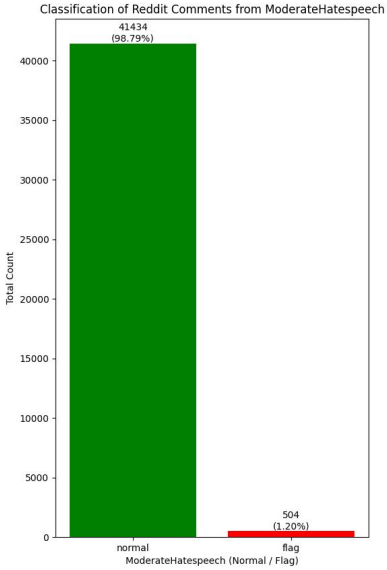

Figure 12: Bar chart showing the classification of Reddit comments into 'normal' and 'flagged' categories based on the analysis from ModerateHatespeech.

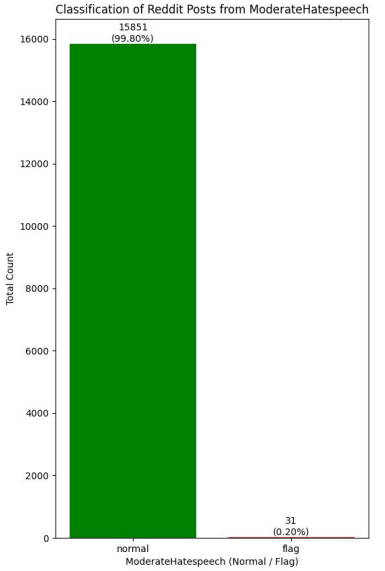

Figure 13: Bar chart indicating the toxicity classification of Reddit posts, as analyzed by the ModerateHatespeech API.

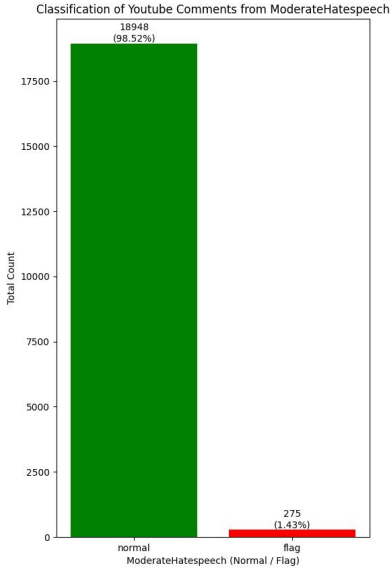

Figure 14: Distribution of YouTube comments categorized by toxicity levels, as determined by ModerateHatespeech API's analysis.

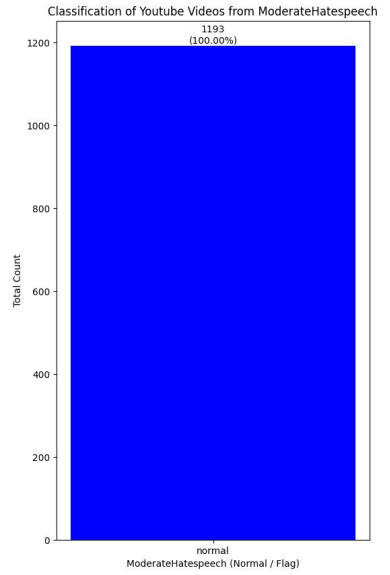

Figure 15: Analysis of YouTube videos classified into 'normal' based on toxicity measurements from ModerateHatespeech.

8.2 Related Work

Several studies have explored the intersection of technology and culture. For instance, Mondal et al. [1] have examined how cultural differences manifest in online discussions and the role that anonymity plays in the nature of discourse. Poecze, Ebster, and Strauss [2] have focused on engagement metrics, analyzing how they correlate with content popularity and user interaction. Further, research by Nakayama and Wan [3] has contributed to understanding sentiment analysis, revealing how sentiments in tech discussions vary across different platforms and cultural contexts.

Our study builds upon these foundations, aiming to provide a more granified analysis of tech-related discussions and how cultural nuances influence them. We employ a multi-faceted approach to dissect various aspects of online interactions, from sentiment and engagement to the popularity and toxicity of content, offering new insights into the cultural underpinnings of technology discourse.

9 DISCUSSION

Our investigation into the cultural nuances of tech discussions on Reddit and YouTube has yielded significant insights. Through the meticulous analysis of engagement metrics, content popularity, and sentiment comparison across these platforms, we have identified distinct patterns that reflect the cultural impact on technology discourse. For instance, the engagement metrics analysis revealed that certain cultural backgrounds are more inclined to interact with technology-related content, which can be leveraged to tailor content more effectively for global audiences.

10 CONCLUSION

In conclusion, this study has affirmed the powerful influence of cultural context on technology discussions online. Our findings suggest that cultural biases and preferences do indeed shape online narratives, which has profound implications for content creators, tech companies, and policy-makers who seek to understand or engage with diverse global audiences.

However, our work is not without limitations. The scope of our data collection was restricted to specific subreddits and YouTube channels, which may not comprehensively represent the global discourse. Additionally, the sentiment analysis, while insightful, relies on the nuances of language that automatic tools may not fully capture.

10.1 Future Work

To address these limitations and further our research, future work could expand the range of sources for data collection, incorporating more diverse platforms and languages to capture a broader cultural spectrum. Additionally, employing more advanced sentiment analysis techniques, such as deep learning models that can understand context better, would refine our understanding of the sentiments expressed in tech discussions.

10.2 Research Questions

(1) How do cultural differences influence the engagement metrics of technology-related content on social media platforms?

(2) What are the emerging trends in technology discussions across different cultures, and how do they vary between text-based and video-based platforms?
(3) Can the sentiment analysis of online discussions predict the market performance of tech products across different cultural regions?

With these questions as our guide, we will delve deeper into the cultural dynamics of technology discourse, aiming to provide more nuanced insights.

REFERENCES

[1] Mainack Mondal, Denzil Correa, and Fabrício Benevenuto. Anonymity effects: A large-scale dataset from an anonymous social media platform. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, page 69–74, New York, NY, USA, 2020. Association for Computing Machinery.
[2] Flora Poecze, Claus Ebster, and Christine Strauss. Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts. *Procedia Computer Science*, 130:660–666, 2018.
[3] Makoto Nakayama and Yun Wan. The cultural impact on social commerce: A sentiment analysis on yelp ethnic restaurant reviews. *Information Management*, 56(2):271–279, 2019.