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.

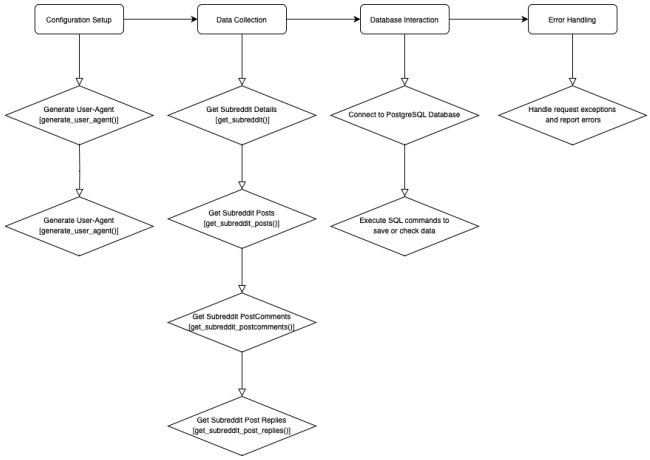

Figure 1: Subreddit Data Collection Impelementation.

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.

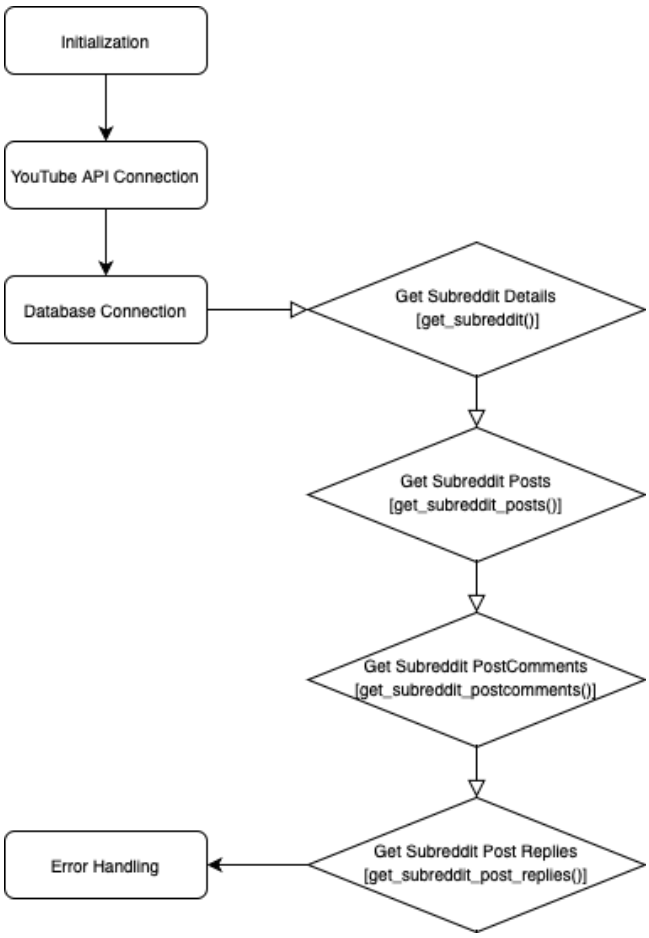

Figure 2: YouTube Data Collection Impelementation.

4 TECHNOLOGY FRAMEWORK

The primary language used for the development of our project is **Python**. To collect data from specific subreddits, we utilize the **Reddit API**. On the other hand, Google's **YouTube Data API v3** is employed to access public content on YouTube, encompassing video details and comments. The **google-api-python-client** serves as our Python client library, particularly for interfacing with Google's discovery-based APIs, inclusive of the YouTube API. For our data storage, we've adopted **PostgresQL**, a relational database system. The project also incorporates **Flask**, a lightweight Python web framework. Task scheduling within Flask is managed using **Flask-APScheduler**, while **Psycopg2** functions as our PostgreSQL adapter, streamlining database connections. Additionally, for drafting and documentation purposes, we've chosen the cloud-based LaTeX editor, **Overleaf**.

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:

Cultural Sentiment Analysis or Sentiment Analysis by Region/Country: This involves evaluating sentiments surrounding tech products or events from specific regions or countries. The aim is to discern potential cultural biases or preferences that may be inherent to a region.

Engagement Metrics Analysis: This method is tailored to identify patterns of engagement, such as likes, comments, and shares. We investigate whether these engagement metrics have any correlation to the cultural or regional origin of the content or its audience.

Content Popularity Analysis: With the myriad of tech topics being discussed daily, it becomes vital to understand what's trending across platforms. This analysis helps in pinpointing those trending tech subjects or products.

Cross-Platform Sentiment Comparison: A side-by-side comparison of sentiment metrics between Reddit and YouTube offers a richer perspective. By contextualizing sentiments within cultural backgrounds, we can derive a more comprehensive insight into possible cultural biases or preferences.

Cultural Keyword Analysis: This involves an in-depth dive into the language used in tech discussions. We look for culturally specific terms or phrases that could indicate a particular bias or preference. By identifying and analyzing these recurring phrases

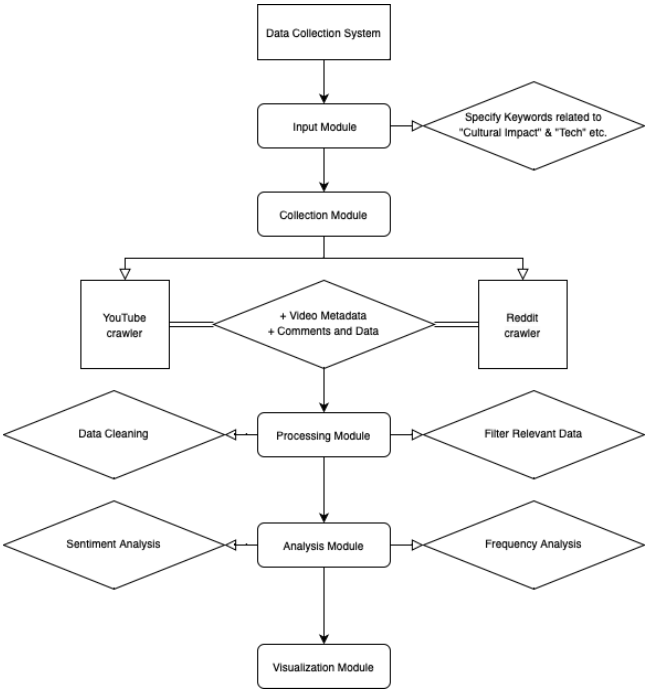

Figure 3: Proposed System Impelementation.

or terms, we can gain a clearer picture of cultural biases, values, or perspectives.

6 EXPLORATION AND MEASUREMENT OF DATA

6.1 YouTube

We track 20 tech, culture and information-related channels Each channel has an average 1 new video posted per day. Each video has an average of 100 comments. For each video, we collect basic details (e.g., title, view count, like/dislike count) and the aforementioned comments. Total YouTube data per week = 20 channels * 1 video/channel/day * 7 days = 140 videos. Total comments from YouTube per week = 100 comments/video * 140 videos = 14,000 comments.

6.2 Reddit

We track 20 tech-related subreddits. Each subreddit has an average of 50 new posts per day. Each post has an average of 40 comments. Total Reddit posts per week = 20 subreddits * 50 posts/subreddit/day * 7 days = 7,000 posts. Total comments from Reddit per week = 40 comments/post * 7,000 posts = 280,000 comments.

7 CHANGES AND CHALLENGES IN DATA COLLECTION

In our initial proposal, our assumptions were formulated based on preliminary research and trends observed in the domain. However, as we transitioned into the actual data collection phase, adjustments were made, and challenges were encountered.

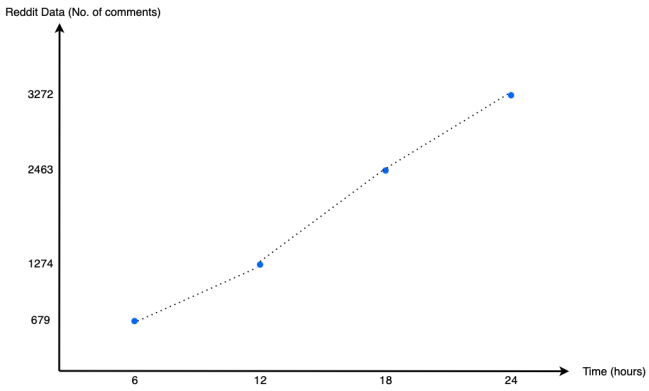

Figure 4: Reddit Comment Data collection over time plot.

7.1 YouTube Data Collection Adjustments and Challenges

:

(1) **Channel Selection and Diversity**: Initially, we aimed to track 100 tech-related channels. However, the final data collection covered only 20 channels which now included a mix of tech, culture, and information-related channels. This adjustment was done to capture a broader spectrum of discussions and get insights beyond the strict tech domain.
(2) **Comment Volume**: Our initial assumption was that each video would receive around 500 comments. In practice, the average number of comments per video was found to be 100. This can be attributed to several factors, including the nature of content, the engagement level of the channel's audience, and more.
(3) **Technical Challenges**: During the YouTube data collection process, we encountered situations where comments were disabled on some videos. These instances posed challenges as our tools were expecting comment data for every video. We implemented error-handling mechanisms to address this issue and ensure the continuity of our data collection process.

7.2 Reddit Data Collection Adjustments and Challenges

:

(1) **Post Volume**: In our proposal, we assumed each subreddit would have around 100 new posts per day. However, the actual number was closer to 50 new posts per day. This difference could be due to seasonal trends, the nature of the subreddits tracked, or other external factors.
(2) **Comment Volume**: Our initial prediction was 50 comments per post. In reality, the average came down to 40 comments per post, indicating slightly less engagement or discussion than anticipated.
(3) **Content Limitations**: We observed that tech-related topics had relatively limited content compared to more mainstream topics, such as entertainment. This meant that while our data collection was niche, the volume was not as expansive as broader categories.