Analyzing the Success Factors for New Songs and Artists on YouTube

A Comprehensive Multisource Data Study

Utsav Chaudhary
uchaudh3@binghamton.edu
SUNY Binghamton
Binghamton, New York, USA

ABSTRACT

Music holds a profound place in the human experience, serving as a universal language that conveys emotions and connects individuals. It has therapeutic effects, fostering well-being and alleviating stress. In the era of abundant data, the confluence of music and data science presents intriguing possibilities. This study delves into the world of music by analyzing the factors that drive the success of new songs and artists on YouTube, a platform that shapes modern music consumption. Leveraging a multi-source data approach, including YouTube, Reddit, and NewsAPI, this research comprehensively examines engagement metrics, demographics, Reddit sentiment, trending hashtags, and upcoming news to unravel the intricate dynamics of musical success. By exploring these diverse dimensions, we aim to shed light on the art and science behind the resonance of music in the digital age.

INTRODUCTION

In the digital age, the Internet has revolutionized the music landscape, fundamentally altering how music is created, shared, and consumed. It has democratized the industry, offering a global stage for emerging independent artists to showcase their talent. Today, music enthusiasts can easily express their passion for music online, instantly access the latest releases, and engage in vibrant discussions about their favorite songs and artists.

The proliferation of social media platforms, music streaming services, and online channels has generated a vast reservoir of data, capturing the pulse of an ever-evolving music industry. This abundant data presents an unprecedented opportunity to delve deep into the realm of music analysis, gaining insights into people's reactions and preferences as they navigate this dynamic landscape.

In this era of data-driven exploration, we embark on a comprehensive and multisource data study. By harnessing the power of three diverse data sources, including YouTube, Reddit, and NewsAPI, we aim to unravel the intricate factors that underpin the success of new songs and artists on YouTube, one of the primary drivers of contemporary music discovery and enjoyment.

Our journey will take us through the rich tapestry of engagement metrics, demographic nuances, sentiment on Reddit, trending hashtags, and upcoming news. By leveraging these multifaceted dimensions of data, we seek to illuminate the art and science that orchestrates the resonant impact of music in the digital age. However, in our pursuit of knowledge, we acknowledge the evolving landscape of data privacy and the challenges it poses. Yet, we remain steadfast in our commitment to unveil the remarkable story behind the melodies that shape our world.

DATA COLLECTION

For this project we will be collecting data from three primary sources: YouTube, Reddit, and NewsAPI. Each source provides unique insights into the factors that contribute to the success of new songs and artists on YouTube.

1 YOUTUBE

To extract data from YouTube using Python, we will develop a custom script to interact with the YouTube Data API. First, we must obtain API credentials, including an API key, from the Google Developers Console. Once obtained, we can use the API key to access YouTube's vast repository of video data.

Our Python script will be responsible for making API requests to retrieve specific data from YouTube. This data can include video metadata (titles, descriptions, upload dates), engagement metrics (likes, dislikes, views, comments), and even information about the channel and demographics of the video's audience. To maximize efficiency and relevance, we can pass various parameters to our API requests, such as video IDs or search queries.

Upon receiving the data from YouTube, we will process and store it in a MySQL database, ensuring that the information is organized and accessible for further analysis.

Additionally, we can implement error handling and pagination in our Python script to handle potential issues with API requests and to retrieve large volumes of data effectively. By creating a well-structured Python script and utilizing MySQL databases, we can efficiently extract, store, and manage the YouTube data needed for our comprehensive analysis of the success factors for new songs and artists on the platform.

This approach will enable us to gather the necessary data from YouTube for in-depth examination, facilitating the exploration of engagement metrics, demographics, and other critical factors influencing the success of new songs and artists.

1.1 URL AND ENDPOINTS

Fetching Channel Data:

- Endpoints: ` /channels `
- Parameters: ` part=snippet,contentDetails,statistics ` and ` id=<comma-separated channelIDs> `

Collecting Video IDs from Playlists:

- Endpoint: ` /playlistItems `
- Parameters: ` part= snippet,contentDetails, playlistId=<playlistId> `, and ` maxResults=50 `

Retrieving Video Statistics:

- Endpoint: ` /videos `
- Parameters: ` part= snippet,contentDetails,statistics ` and ` id=<comma-separated video_ids> `

1.2 TABLE STRUCTURE

All Channel Details: channelName, channelID, subscriberCount, totalViews, totalVideos, playlistID
All Video IDs: channelID, videoID, videoTitle, channelTitle, uploadDate, tags, duration
First Data Collection: dataCollectionDate, videoID, viewCount, likeCount, commentCount
Daily Data: dataCollectionDate, videoID, viewCount, likeCOunt, commentCount
Comment Data: commentID, videoID, comment, publishedAt

2 REDDIT

To extract data from Reddit using Python, we will create a custom Python script specifically designed to interact with the Reddit API. Initially, we will obtain a temporary OAuth token from Reddit to authenticate our API requests, ensuring secure access to the Reddit data we need.

Subsequently, our Python script will be engineered to collect data from Reddit, focusing on the most recent posts and user-generated content within specified subreddits. We will employ a range of parameters to fine-tune the data retrieval process, including setting limits to control the number of items fetched and using before/after parameters to filter posts by specific dates.

Once we've retrieved the data, we will meticulously structure it into a well-organized dataframe. This dataframe will include essential attributes such as post titles, timestamps, content, and user interactions, encompassing valuable information like comments, upvotes, and downvotes. We will also consider integrating additional optimization parameters to further refine and enhance our data extraction process.

By implementing these strategies, we will effectively collect, structure, and store Reddit data, enabling us to conduct a thorough analysis of the factors contributing to the success of new songs and artists across the two platforms.

2.1 TABLE STRUCTURE

Daily Data: dataCollectionDate, publishedAt, postID, userID, title, upvoteRatio, ups, totalComments, url

3 NEWSAPI

For data extraction from NewsAPI using Python, we will develop a specialized script to interface with the NewsAPI. First, we will acquire the necessary API key from the NewsAPI platform to enable access to their extensive repository of news articles and content.

Our Python script will be designed to make API requests to fetch relevant data. We will utilize two main endpoints: the "Everything Endpoint" to search for articles from various sources related to artists and their songs published within the last five years, and the "Top Headlines Endpoint" to retrieve breaking news headlines pertinent to the music industry, artists, and songs. By making these API requests, we will collect a wealth of information, including news articles, headlines, publication dates, and content related to artists and songs.

In addition to this, we will implement a sentiment analysis component, wherein we analyze the news and blogs sourced from reputed NEWS providers towards the songs and artists. This sentiment analysis will provide valuable insights into the perceptions and reactions of journalists and public court, which will be a crucial element in our comprehensive study of the success factors for new songs and artists on YouTube.

This approach will enable us to systematically store and access news articles, ensuring that we remain well-informed about current events and news that may influence the reception and success of new songs and artists on YouTube.

3.1 TABLE STRUCTURE

Daily News Data: dataCollectionDate, celebName, sourceName, title, description, publishedAt

METHODOLOGY

The proposed Methodologies comprises the following stages:

Data Extraction: Begin by crafting Python scripts tailored to each data source. For YouTube, authenticate using the YouTube Data API and retrieve video-related metrics, descriptions, and demographic information. For Reddit, use the Reddit API to fetch data on posts, comments, upvotes, downvotes, and trending keywords. Utilize NewsAPI to collect news articles and headlines related to artists and songs.

Data Splitting: Organize the extracted data into relevant categories, such as YouTube engagement metrics, Reddit discussions, and news articles. Create distinct dataframes or data structures for each data type to facilitate efficient handling.

Data Storage in MySQL Database: Employ a MySQL database, to store the data. Design appropriate database schemas to accommodate the structured nature of the data. Store YouTube metrics, Reddit discussions, and news articles in separate tables within the database.

Data Cleaning and Transformation: Initially, we'll preprocess the gathered data, addressing missing values, resolving format inconsistencies, and handling outliers. We'll also standardize data types and structures across the dataset as needed to ensure consistency.

Post Data Storage in a MySQL Database: Following data cleaning and transformation, we'll proceed to store the finalized dataset in a MySQL database. With the data securely stored, we'll embark on an in-depth analysis to uncover the key factors influencing the reach and success rate of artists on YouTube.

CONCLUSION

The main goal of this project is to understand what makes new songs and artists successful on YouTube. We're looking at things like views, who's watching, what people are saying, and even trending topics in the news. By using data from different sources, we want to uncover the secrets behind music's impact in the digital age. We hope our findings will be useful to creators, marketers, and music lovers, helping them replicate success patterns on online platforms. In the future, we plan to expand our research to other social media and streaming services to get a complete picture of artists' online presence and how different platforms influence each other.