Minecraft : Player Insights Matter

PURDUE UNIVERSITY | The Data Mine | MOJANG STUDIOS

The Data Mine Corporate Partners Symposium 2024

Students:-Aadit Rohit Bennur, Cailey Jen, Chi- Hao Law, David Christopher Price, David Martin Caedo Calalang, Nathan Wu, Paraj Goyal, Shubhaang Agarwal, Shubham Kose, Soumi Mukherjee, Srushti Mahesh Tavare, Sukrati Gautam

Goals and Objectives

Understanding players interaction with Minecraft on Social Media platforms

- Correlate social media trends with Minecraft engagement –Explore other social media platforms for data collection
- Competitor Analysis – Identify competitors and predict their impact on Minecraft gaming time
- Develop an end to end working pipeline from data collection to prediction of sentiment scores
- Develop a dashboard for timeseries analysis of the data depicting sentiment scores and topics across competitors.

Competitor Analysis

Fortnite
- Competitive, Creative modes
- Consistent social interaction, User generated content
- Monthly Users: 231 Million

Minecraft
- Creativity, Social Interaction, Competition, Nostalgia
- Monthly Users: 163.67 Million

Roblox
- User-generated content, Create Games, Social Platform, Social Interactions
- Monthly Users: 202 Million

Terraria
- 2D sandbox style, Resource gathering, Exploration and combat
- Structured progression
- Monthly Users: 600 Thousand

Listeners

YouTube
- Uses a selenium web scraper to collect comments from relevant YouTube videos
- Automated the process of opening the YouTube webpage and interacting with it using the HTML tags.
- Script collects URLs of videos from the last week and their top comments.
- Refined comment relevance by targeting specific game review and update videos.
- Databricks pipelines established for weekly scraping operations.
- Compiled comments and collection dates into a tsv for sentiment analysis processing.

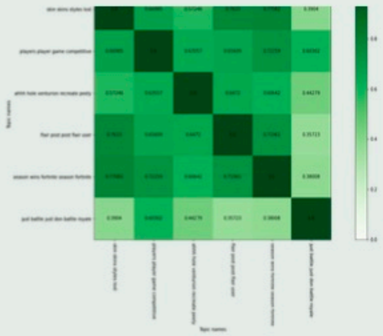
Reddit
- Initially used a Selenium scraper, like YouTube process, but faced reliability issues.
- Shifted strategy to secure Reddit API access due to enhanced web scraping protections.
- Successfully obtained API access through Minecraft to retrieve data from Reddit.
- Identified and focused on specific subreddit "flairs" pertinent to our research on four games.
- Concentrated on flairs relating to game updates, bugs, and user experience for relevance.
- Established a weekly data collection pipeline in Databricks, outputting to a timestamped tsv file for subsequent sentiment analysis by the data science team.

Data Cleaning

The image to the left depicts how data is processed to convert emojis to text format using NLP and ML models. The gaps and null values are removed from the data for analysis.

BERT Topic Modeling

- BERT leverages language context and semantics.
- Moves beyond word frequency, understanding word nuances in context.
- Results in more accurate topic distributions in texts.



The image on the left depicts a similarity matrix where similarity scores above a user-set threshold of 0.6 are displayed. Each cell in the matrix represents the similarity score between two topics. The similarity score indicates how closely related two topics are, based on their content and semantic meaning

BERT Sentiment Analysis

- Multilingual model used (English, Dutch, German, French, Spanish, and Italian)
- Analyzes the sentiments through a function that takes in a string and outputs a score (1 to 5)
- Lambda function to run the sentiment analysis function and place results in new data frame column

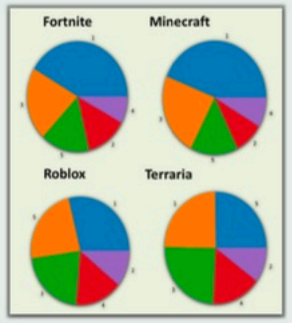
Linking Market Research with Sentiment analysis and Topic Modelling

- Selected Fortnite update data for a specific week
- Aimed to extract meaningful insights about this update.
- Analyzed 19 topics featuring keywords such as "glitch," "emote," "skin," and "Disney," pertinent to the update, and conducted sentiment analysis.

	topics	Sentiment Score
11	disney pickaxes midas lightsaber	5
16	emote slide shuffle emotes	3
18	shop item item shop days	5
26	happened glitch happens second	1

The image on the leftshows the sentiment scores for the relevant topics, indicating which aspects of the update were viewed positively or negatively. For instance, Keyword 'glitch' had negative sentiment, indicating update problems, while 'Disney' was positively received, showing players liked the collaboration.

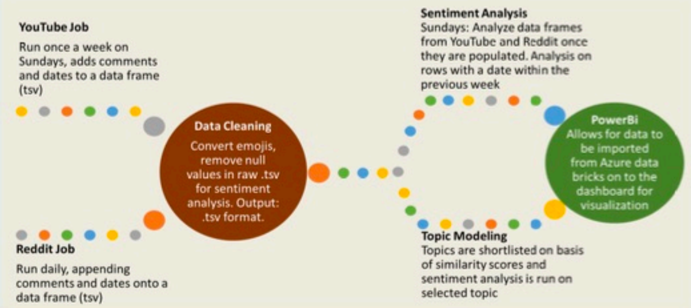
Data Visualization



- The image above (left) depicts an interactive PowerBI time series visualization that compares the average sentiment scores for the games on a specific week. This dashboard features date filtering, which allows for displaying comparisons on specific days. It connects to data on Azure Databricks that was garnered by our data engineering team and utilizes Direct Query to enable a smooth and automated dataflow to the dashboard.
- The image above (right) represents the distribution of sentiment scores for various competitors during the February 23-27, 2024 week from Reddit. In this case, the scores span from 1 to 5, with 1 being negative and 5 being positive.

Future Goals

- Currently developing pipeline. (Refer to image below)
- Need to integrate topic modeling data with dashboard and market research keywords.
- Automate storage of data frames post-sentiment analysis from social media.
- Consolidate sentiment scores and topics into one accessible data frame.
- In the process of creating Instagram API listener for game-related data.



References

- https://maartengr.github.io/BERTopic/
- https://umap-learn.readthedocs.io/en/latest/
- https://hdbscan.readthedocs.io/en/latest/
- https://blog.turtlebeach.com/when-does-fortnite-update-28-30-release/

Acknowledgements

Gratitude to our TA, Mihir Mutyampeta, and Corporate mentor, Timothy Olojede, for heir constant support and guidance. Thanks to Nick Rosenorn, Cai Chen and Data Mine staff for all their support and assistance.