NYU

Steam Review Analysis

PRESENTED BY Rohan Gore, Pooja Thakur

01/23/2025

Problem + Vision

- Steam has 100M+ reviews across thousands of games
- Reviews are noisy, unstructured, and difficult to explore at scale
- Vision: build a dashboard to summarize trends, sentiment, and game popularity from raw reviews

NYU

Why This Is a Big Data Problem

- Volume: 100M+ rows, 23 columns
- Variety: numeric, boolean, text, timestamps, and multilingual content
- Velocity: scraped and preprocessed data
- Computation: Sentiment, filtering, daily trends require distributed analysis (PySpark)

Tech Stack

- Preprocessing: PySpark (Spark 3.x), PyArrow
- NLP: HuggingFace Transformers (DistilBERT SST-2)
- Frontend: Streamlit
- Visualizations: Altair, Plotly
- Storage Format: Apache Parquet
- Languages: Python, SQL

Workflow + Implementation

End-to-End Workflow

- Ingest

  → Clean (Jupyter, Python script, PySpark)

  → Analyze (PySpark + PyArrow + Pandas + NLP)

  → Visualize (Altair + Plotly)

  → Render on Frontend (Streamlit)

Load & Clean Data

- Converted raw csv to Apache parquet + Snappy compression
- Identified damaged rows and dropped them
- Casted 15+ fields to proper types
- Processed all 113M rows, leading to remove 3+ million rows

Features

Engagement Tab

- Inputs:
  - Time slicer
  - Count selector for Top K games
- Output:
  - Top “k” most reviewed game in the selected time frame
  - Author analysis and Influence Score (derived)

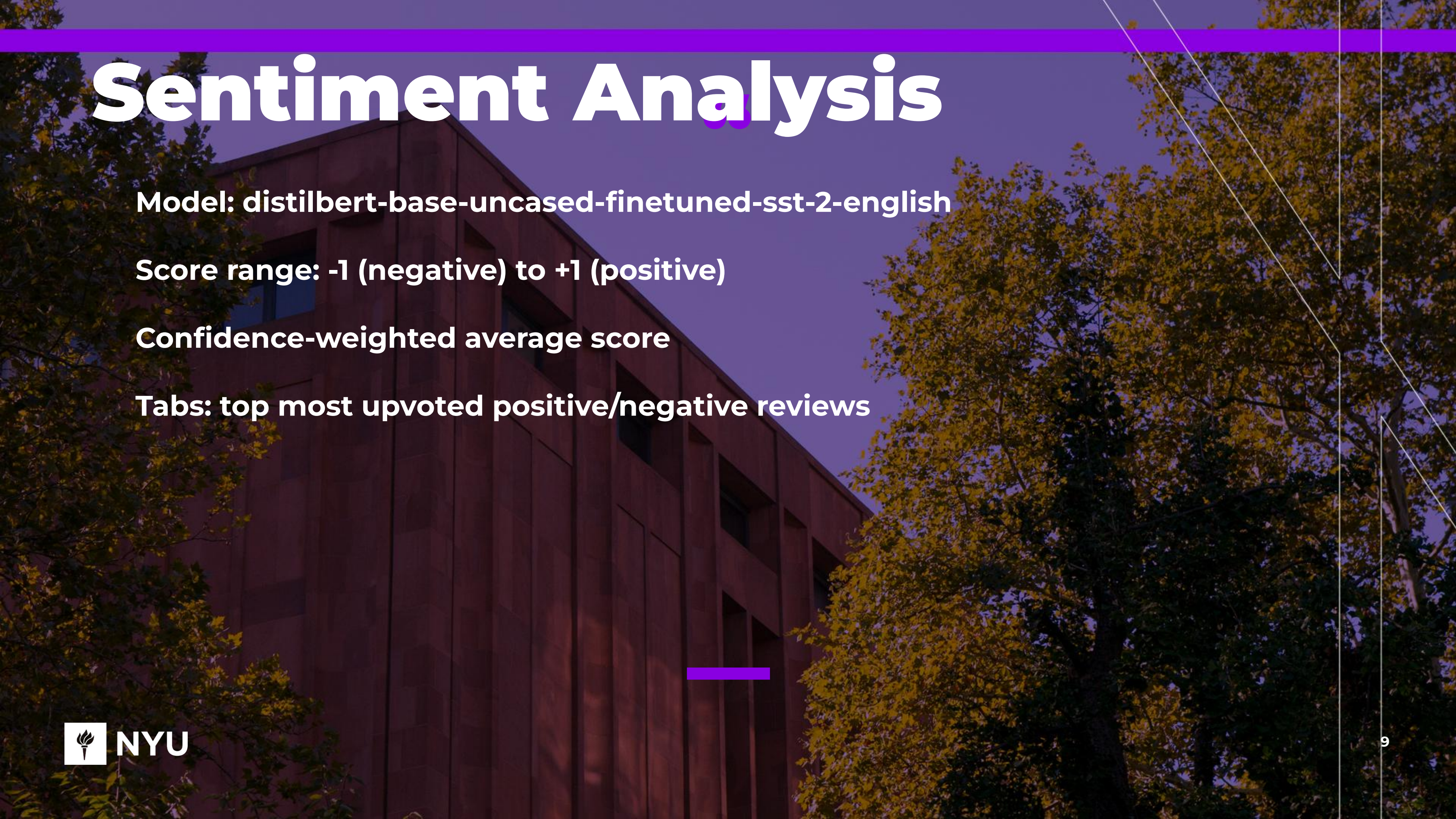
Game Analysis Tab

- Inputs:
  - Game Name
  - TIme Frame
- Output:
  - Review Analysis (multiple views)
  - Sentiment Analysis and top polarizing reviews

Review Cards (Filterable)

1. Filter reviews by language
2. Filter reviews by category:
   a. Upvoted reviews
   b. Funny reviews
   c. Most commented reviews

Sentiment Analysis



Model: distilbert-base-uncased-finetuned-sst-2-english

Score range: -1 (negative) to +1 (positive)

Confidence-weighted average score

Tabs: top most upvoted positive/negative reviews

Results and Observation

- Dashboard processed reviews for 100k+ games
- Sentiment analysis done on 1K top reviews per query
- PySpark aggregations < 1 minute
- In cache filtering for dropdowns

- Free games get more critical reviews
- Review spikes around major updates
- Positive sentiment correlates with high playtime

Challenges

Data Cleaning for complete directory of parquet files

- **Multilingual reviews (non-English model limitations)**
- **Sentiment model accuracy on sarcasm**
- **Handling outliers and false positives**
- **Attempted Kafka integration, but the consumer failed to render data on the UI.**

Future Scope

- Add Kafka stream for live review analysis and dashboarding
- Use XLM-RoBERTa for multilingual sentiment
- Integrate game metadata (price, genre)
- Deploy publicly with Docker + CI/CD

Wrap-Up

GitHub: github.com/rohan-g0re/bigdata_project

Questions?