Sonic Sentiments

Sentiment Analyzer for Music

Group 5: Lynn Nguyen, Justin Wright, Oliver von Mizener
Data Analytics Boot Camp - Project 4 Presentation

Analysis Goal

Examine music sentiment and genre trends using a dataset of modern music up to 2019.

Project Goal

- Demonstrate a proof of concept for sentiment and genre analysis of music lyrics.
- Build an interactive dashboard that visualizes these trends over time.
- Incorporate real-time user-submitted data via a Flask application to enhance the analysis of music trends.

Our Approach

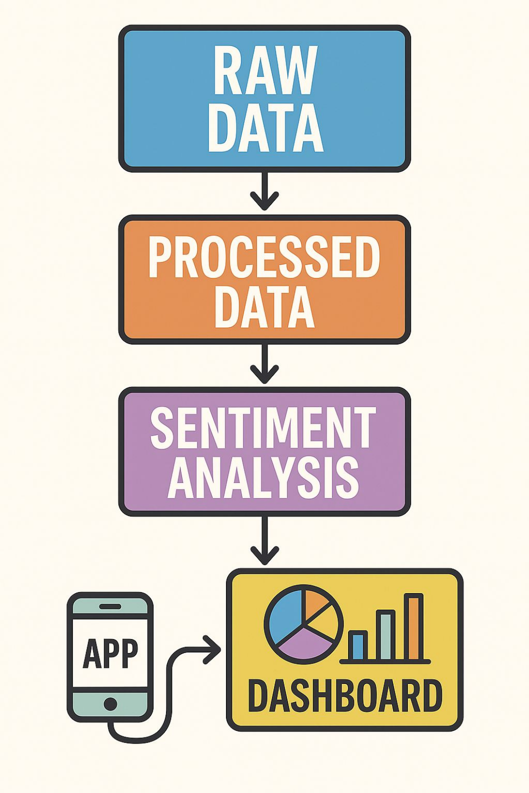
1. Raw data collection and performed cleaning and preprocessing (e.g., extracting release years, normalizing text).
2. Applied sentiment analysis and genre prediction using machine learning models and TextBlob.
3. Developed a **Dash dashboard** for visualizing the processed data alongside the sentiment and genre analysis.
4. Created a separate Flask app to gather user submissions, which are integrated into the dashboard for continuous analysis.

Rationale & Tool Functions

- Understand music trends and evolving sentiments in lyrics
- Support music marketing, playlist curation, and industry insights
- Enable deeper insights into audience mood and genre classification
- Potential Impact:
  - Improved decision-making for music producers/marketers
  - Basis for more advanced research integrating real-time user data

Data Pipeline

- Our team envisioned a pipeline where we could take the raw CSV data to train a model.
- We wanted a dashboard that could accept inputs from both the trained data and processed information, as well as user input data that compares against our original model.
- This let us answer “What if” questions, and fill in the gaps from our dataset.

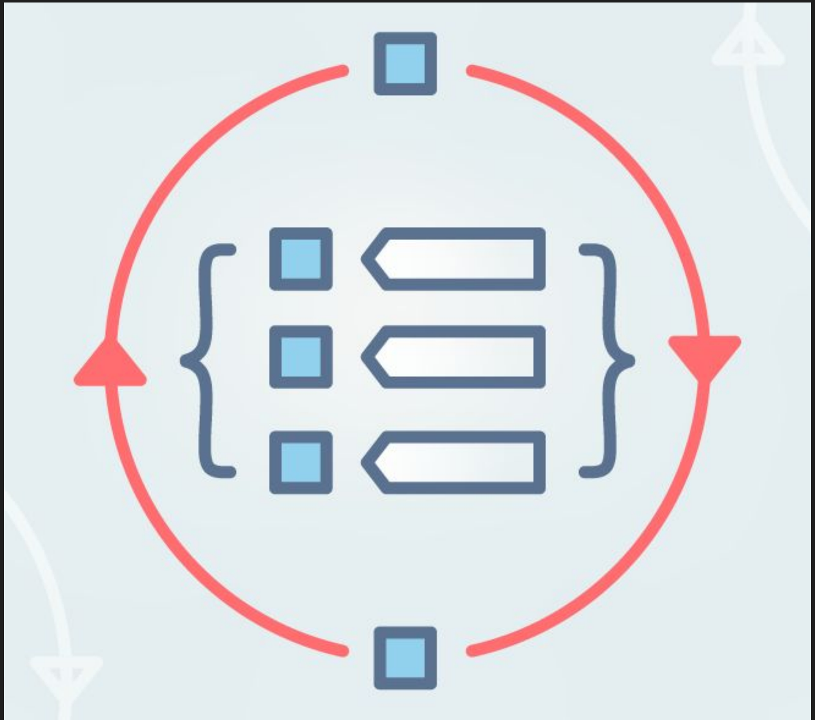


Data Source

The base CSV, processed_tcc_ceds_music.csv, serves as the foundation for our analysis pipeline. This dataset includes key metadata and quantitative measures essential for both sentiment analysis and genre classification. Some of the main headers in the CSV are:

- release_date: The original release dates of the tracks (used to extract the release year).
- artist_name & track_name: Identifiers for each song.
- genre: The genre classification for each track.
- Audio Features: Including columns such as danceability, loudness, acousticness, instrumentalness, valence, and energy, which capture key sonic characteristics.
- Thematic Descriptors: Columns like dating, violence, world/life, night/time, shake the audience, family/gospel, romantic, communication, obscene, music, movement/places, light/visual perceptions, family/spiritual, like/girls, sadness, and feelings that provide deeper context on lyrical themes.

This well-structured dataset, despite its limitations (such as covering data only up to 2019), allows us to cleanly transform raw music data into actionable insights for our proof-of-concept dashboard.



Data Filtering

- **Standardization:**
  - Converted all lyrics to lowercase to ensure uniformity.

- **Removal of Unwanted Characters:**
  - Utilized regular expressions to strip punctuation and special characters from the lyrics.

- **Whitespace Trimming:**
  - Removed extra spaces to facilitate accurate tokenization.

- **Outcome:**
  - These steps produced a clean, standardized dataset that improved the accuracy of our sentiment and genre analysis.

Machine Learning - Sentiment Analysis Model

- We leveraged classical machine learning techniques using TF-IDF to vectorize lyrics for our baseline sentiment and genre analysis. This was leveraged for simplicity of execution on local hardware as well as due to dataset limitations to ensure a functional model for project submission.
- We leveraged deep learning with Random Forest Regression for musical features in testing as our expanded wish-list testing.
  - Pre-trained models (stored as pickle files) were used for sentiment and genre classification.

- **Why This Approach:**
  - The dataset's limited size and scope made conventional methods more practical and interpretable for a proof-of-concept.
  - Our predictive models were pre-trained and stored as pickle files, which allowed us to quickly deploy sentiment and genre analysis without the overhead of training complex deep learning models.
  - In future iterations—with access to larger, more comprehensive datasets—we might explore the benefits of ensemble methods like Deep Learning with Random Forest for all data points to improve accuracy and robustness. But due to limitations of the data set, we opted to ensure accuracy for sentiment and only experiment with other options for extra data points.

DEMO

Limitations/Concessions

Dataset Limitations:

- Extends only to 2019 with limited genre coverage and fewer data points on specific categories than ideal.

Impact on Analysis:

- May produce inaccuracies (e.g., Midwest Emo misclassified as Reggae) due to insufficient data variety. The accuracy rating on this reflects the problems with having such a narrow dataset for this.

Concessions & Future Steps:

- Used the dataset as a proof-of-concept under time constraints.
- Future work would involve integrating professional APIs and merging additional datasets for more robust analysis.
- Running training models on cloud servers would allow for rapid testing and adjustment of deep learning models for better data insights.

Who can use this?

- **Data Analysts working at music streaming services or labels**
  - **Use Case:** find trends and create actionable insights that will help inform future business decisions, either for the service or label.

- **Upcoming music producers**
  - **Use Case:** help them find their audience, sound, and brand as they analyze their own music lyrics to better understand audience reception and current emotional trends.

- **Music Marketers**
  - Use Case: help marketers better understand current and past trends and create actionable insights for future music marketing campaigns.

Thank you!