Classifying Music Genres from Lyrics

DS4002 - F25 Project 1

Data Devils (Group 7): Michael Hijduk, Anthony Jiang, Laura Espuna Santana

UVA | SCHOOL of DATA SCIENCE

1. Motivation & Context
2. Hypothesis & Research Question
3. Data Acquisition & Ethics
4. Analysis Plan
5. Tricky Analysis Decision
6. Bias & Uncertainty
7. Results & Conclusions
8. Next Steps

Motivation & Context

Why classify music genres by lyrics?

- **The importance of music:** Genres shape culture, mood, and how people connect
- **Most genre classification is audio-based**
- **Lyrics = overlooked signal:** Contain semantics, themes, and stylistic patterns
- **Research Gap:** Limited exploration of lyrics-only classification despite availability of text data
- **Impact:** A successful lyrics-only model would expand tools for music recommendation, cultural analysis, and NLP applications in art

UVA | SCHOOL of DATA SCIENCE

Research Question

Hypothesis:

Song lyrics exhibit genre-specific word frequency, semantics, and syntax, then we can train a logistic regression model to classify songs with ~80% accuracy

Research Question:

To what extent can lyrics alone be used to classify music genre?

Modeling Approach:

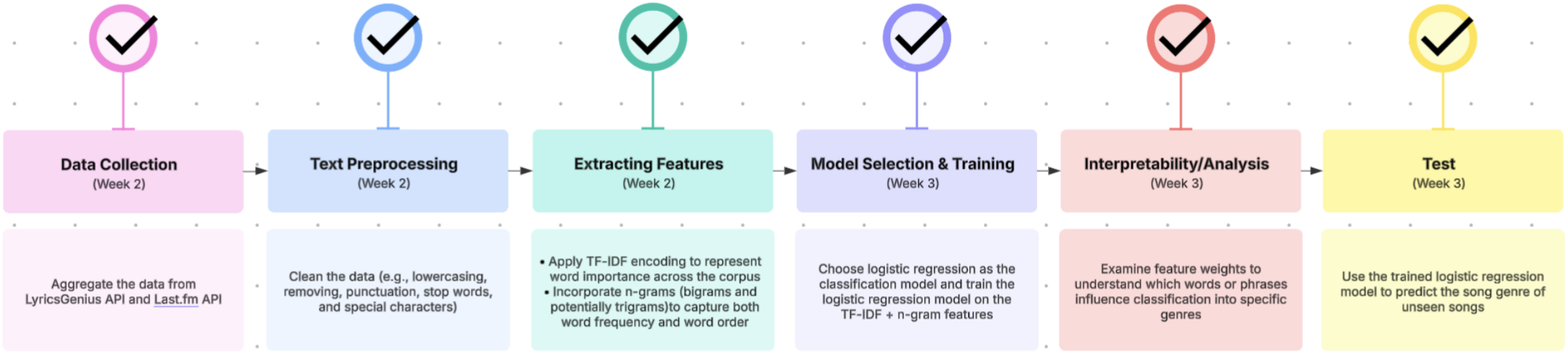
Dataset with 2,000 songs across 8 genres. Features include TF-IDF with unigrams & bigrams (and possible trigrams), Model chosen is logistic regression for classification

UVA | SCHOOL of DATA SCIENCE

Data Acquisition

- **Key Features:**
  - **Format:** Text-based (lyrics + metadata) stored in JSON format
  - **Size:** ~2,000 tracks and 250 per genre (pop, rock, electronic, country, metal, hip-hop, jazz, R&B)
  - **Song Fields:** name, artist, genre, lyrics
- **Acquisition:**
  - Pulled from **LastFM API** (metadata) + **Genius API** (lyrics)
  - Automated collection with cleaning pipeline (removing headers, retries for failed pulls)
- **Challenges & Considerations:**
  - **Missing lyrics**: some songs not retrievable
  - **Genre labeling**: relies on LastFM definitions (potential bias)
- **Licensing:**
  - Code & metadata which is sharable under MIT license
  - Lyrics which are copyright protected, use limited to academic/research purposes

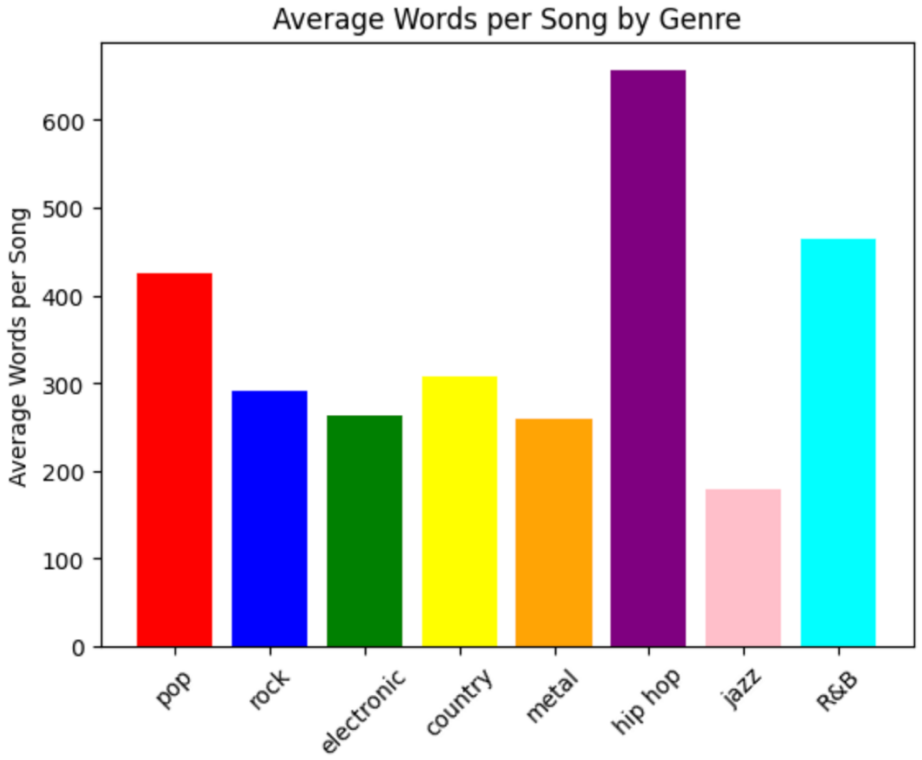
Analysis Plan

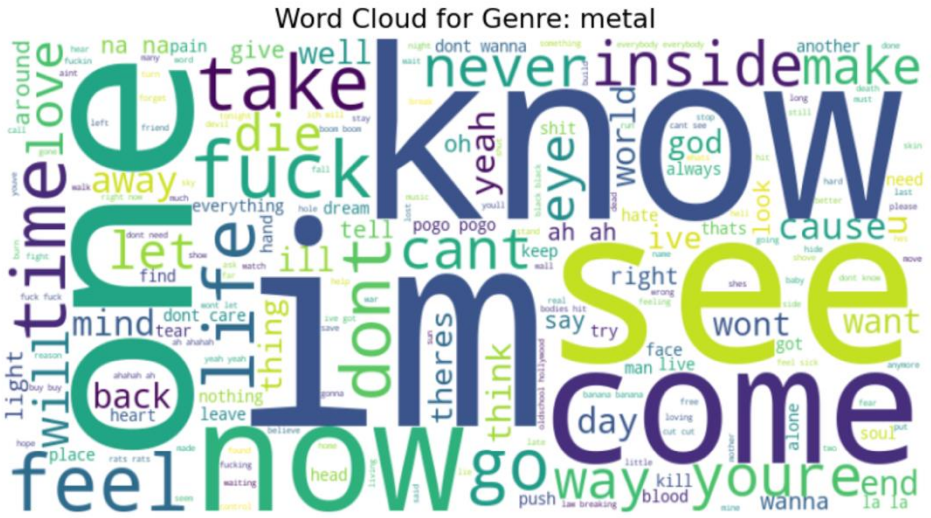


Interpretability & Analysis was a challenging step

UVA | SCHOOL of DATA SCIENCE

Exploratory Data Analysis





UVA | SCHOOL of DATA SCIENCE

Tricky Analysis Decision

Decision: Whether to include **trigrams** in TF-IDF, beyond unigrams and bigrams

- Could capture **short phrases** (ex. “love of my,” “out of time”) which could result in potentially better genre signals
- But adds **huge feature growth**, risking overfitting on a ~2k-song corpus
- **Why it was tricky:**
  - **Signal vs. noise trade-off**: more context vs. many rare phrases
  - **Compute & time**: feature space explodes (memory, training time), limiting iterations
  - **Data regime**: with 250 songs/genre, many trigrams occur **once** = unstable weights
- **Impact of the decision:**
  - **Faster** training and more hyperparameter sweeps/better diagnostics
  - **Lower variance** & improved **generalization** (reduced sparsity/overfit risk)
  - **Deviation from MI2:** Initially planned to test trigrams; we **pivoted** after pilot results showed no meaningful performance gains.

Bias & Uncertainty

Types of Bias

- **Selection Bias:** Only *top 250 songs per genre (2025)* may not represent full genre diversity
- **Genre Labeling Bias:** Genres depend on **LastFM's taxonomy**
- **Lyrics Availability Bias:** Not all Genius API queries succeeded and missing lyrics reduced representation for certain genres
- **Word Frequency Bias:** Some genres have longer songs so word count imbalance affects feature density

Corrections Applied

- Cleaning pipeline: Dropped entries missing lyrics or genre

Uncertainty in Reported Values

- Cross-validation (5-fold) used to estimate model stability across splits
- Majority-class baseline recorded for sanity check
- Macro-F1 & per-genre metrics reported to capture variability, not just accuracy

Results & Conclusion

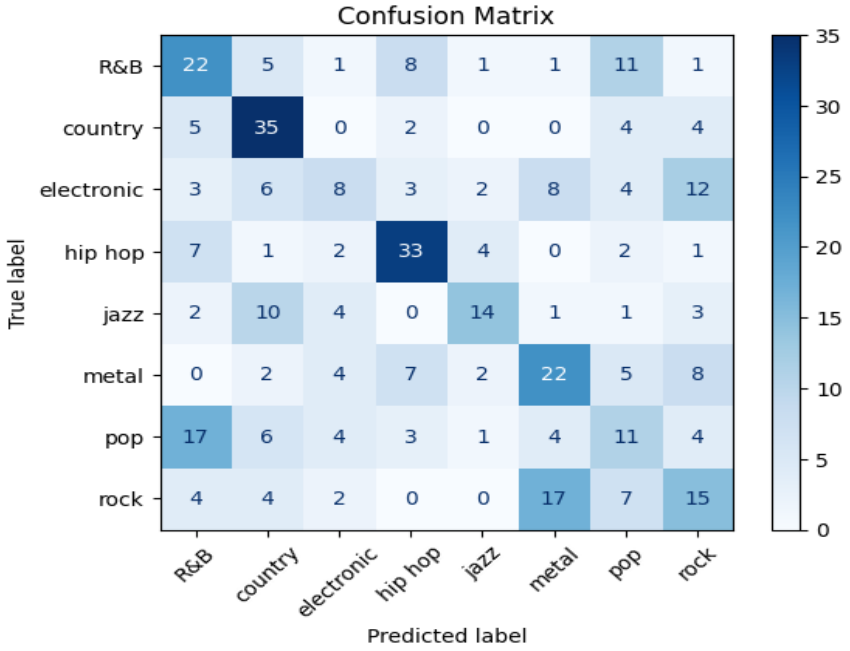
The results show that the model had the best performance against hip-hip and country songs and the worst performance against electronic and pop songs

- The model had major trouble differentiating between R&B and Pop
- The model had major trouble differentiating between Metal and Rock

The model only had a **42.1%** test accuracy

- This is better than the baseline **12.5%** (random guessing)
- Failed to prove the hypothesis (building a model that can predict song genre solely on lyrics with 80% accuracy)

Conclusion: A model that is only trained on lyrics is not sufficient to reliably to predict song genre




```
Cross-validation scores: [0.43092105 0.4620462  0.44884488 0.3960396  0.37623762]
Mean CV accuracy: 0.4228178738926524
Test accuracy: 0.42105263157894735

Classification report:
              precision    recall  f1-score   support

         R&B       0.37      0.44      0.40        50
     country       0.51      0.70      0.59        50
  electronic       0.32      0.17      0.23        46
     hip hop       0.59      0.66      0.62        50
        jazz       0.58      0.40      0.47        35
       metal       0.42      0.44      0.43        50
         pop       0.24      0.22      0.23        50
        rock       0.31      0.31      0.31        49

    accuracy                           0.42       380
   macro avg       0.42      0.42      0.41       380
weighted avg       0.41      0.42      0.41       380
```

UVA | SCHOOL of DATA SCIENCE

Next Steps

New Lines of Exploration:

- Model Selection: Experiment with deeper neural networks, ensemble methods, or transformer models.
- Multi-label Classification: Many songs fit into more than one genre. Exploring multi-label classification could make results more realistic.

Improvements:

- Data Quality: Expand the dataset for better balance across genres and avoid bias toward popular genres.
- Cross-modal Models: Combining text analysis (lyrics) with acoustic features for better performance.

New Questions:

- Multi-language songs: How will the model handle songs in languages other than English, or code-switching within lyrics?
- Temporal Drift: Will a model trained in the past need to be replaced/retrained to take into account song genre evolution?

References

[1] R. Mayer, R. Neumayer, and A. Rauber, "Rhyme and style features for musical genre classification by song lyrics," in *Proc. 9th Int. Conf. Music Information Retrieval (ISMIR)*, Philadelphia, USA, 2008, pp. 337–342. [Online]. Available: https://www.researchgate.net/publication/220723682_Rhyme_and_Style_Features_for_Musical_Genre_Classification_by_Song_Lyrics [Accessed Sept. 11, 2025].

[2] M. Fell, C. Sporleder, “Lyrics-based Analysis and Classification of Music,” in Proc. 25th Int. Conf. Computational Linguistics, Aug. 2014, pp. 620-631. [Online]. Available: https://aclanthology.org/C14-1059.pdf. [Accessed Sept. 11, 2025].

[3] M. Schedl, H. Zamani, C. Chen, Y. Deldjoo, M. Elahi, “Current challenges and visions in music recommender systems research,” International Journal of Multimedia Information Retrieval, vol. 7, pp. 95-116, [Online]. Available: https://link.springer.com/article/10.1007/s13735-018-0154-2. [Accessed Sept. 11, 2025].

Github: https://github.com/michaelhijduk/DS-4002

UVA | SCHOOL of DATA SCIENCE

Any Questions?