



Introduction

Methodology

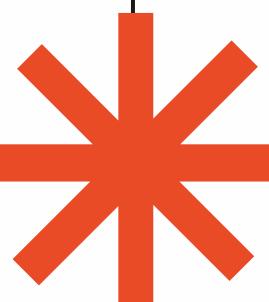
Results

Conclusion

CLASSIFYING MUSIC GENRES USING TOPIC MODELING OF LYRICS

Group 5 Members:

- Philbert - 2702227454
- Izhar Octafirlian Susilo - 2702222144
- Frederick Nicholas Su - 2702220826



```
state={  
  products: storeProducts  
}  
render() {  
  return (  
    <React.Fragment>  
      <div className="py-5">  
        <div className="container">  
          <Title name="our" title="Our Products" />  
          <div className="row">  
            <ProductConsumer>  
              {(value) => {  
                console.log(value);  
              }}  
            </ProductConsumer>  
          </div>  
        </div>  
      </React.Fragment>  
    )  
}
```





INTRODUCTION

This study explores how song lyrics can be used to identify and differentiate music genres using topic modeling.

- Traditional genre classification focuses on musical features (tempo, rhythm, chords)
- Lyrics are an underexplored but informative signal for genre analysis
- NLP and topic modeling have revealed recurring themes in large lyric datasets
- Prior work focuses mainly on sentiment or cultural analysis
- This work applies topic modeling (NMF) to uncover genre-specific lyrical themes

METHODOLOGY



Data Acquisition & Processing

- Dataset: Genius Song Lyrics (Kaggle), 5M+ entries
- Chunk processing: 20k songs per batch, capped at 500k
- Filtering: English lyrics only, removed missing lyrics/genre
- Text cleaning: lowercase, remove non-letters, custom lyric stop words
- Balancing: 10,000 songs per genre → 60,000 total samples

Topic Modeling

- TF-IDF vectorization: converts lyrics into weighted word features
- Topic modeling: Non-negative Matrix Factorization (NMF)
- Output: document-topic vectors representing each song
- Interpretability: each dimension corresponds to a latent lyrical theme
- Topics: number of topics (e.g., 50) tuned as a hyperparameter

Classification & Evaluation

- Dataset is split 80:20 and is stratified by genre.
- Three classifier models: Logistic Regression, Linear SVM, and Random Forest.
- Models are evaluated using Accuracy, Precision, Recall, and F1-Score alongside a confusion matrix.

training Logistic Regression

--> Logistic Regression Results:

Accuracy: 0.4983

Precision: 0.4776

Recall: 0.4983

F1-Score: 0.4827

training SVM (Linear)

--> SVM (Linear) Results:

Accuracy: 0.5129

Precision: 0.4906

Recall: 0.5129

F1-Score: 0.4924

training Random Forest

--> Random Forest Results:

Accuracy: 0.5613

Precision: 0.5484

Recall: 0.5613

F1-Score: 0.5511

WINNER: Random Forest

Accuracy: 56.13%

F1-Score: 0.5511



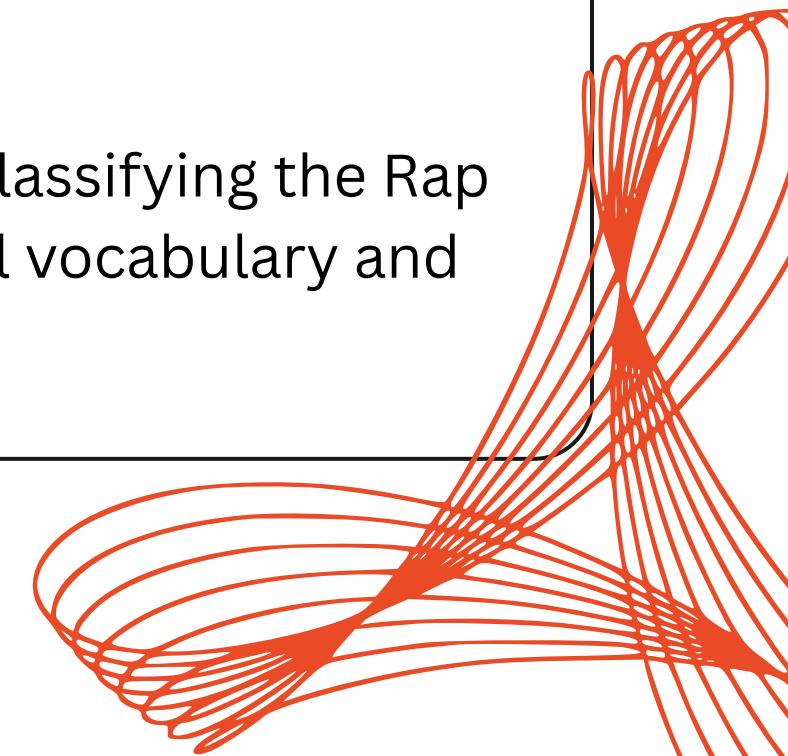
RESULTS & ANALYSIS

Results show similar performance for Logistic Regression and Linear SVM with Random Forest outperforming both slightly on all metrics. From this it can be inferred that lyric-genre relationships are non-linear.

Per-genre results (Random Forest):

- Strong: Rap ($F1 \approx 0.78$), Misc ($F1 \approx 0.77$)
- Weak: Pop and Rock ($F1 < 0.40$)

The random forest model did much better in classifying the Rap genre, most likely because of its distinct lyrical vocabulary and thematic words.



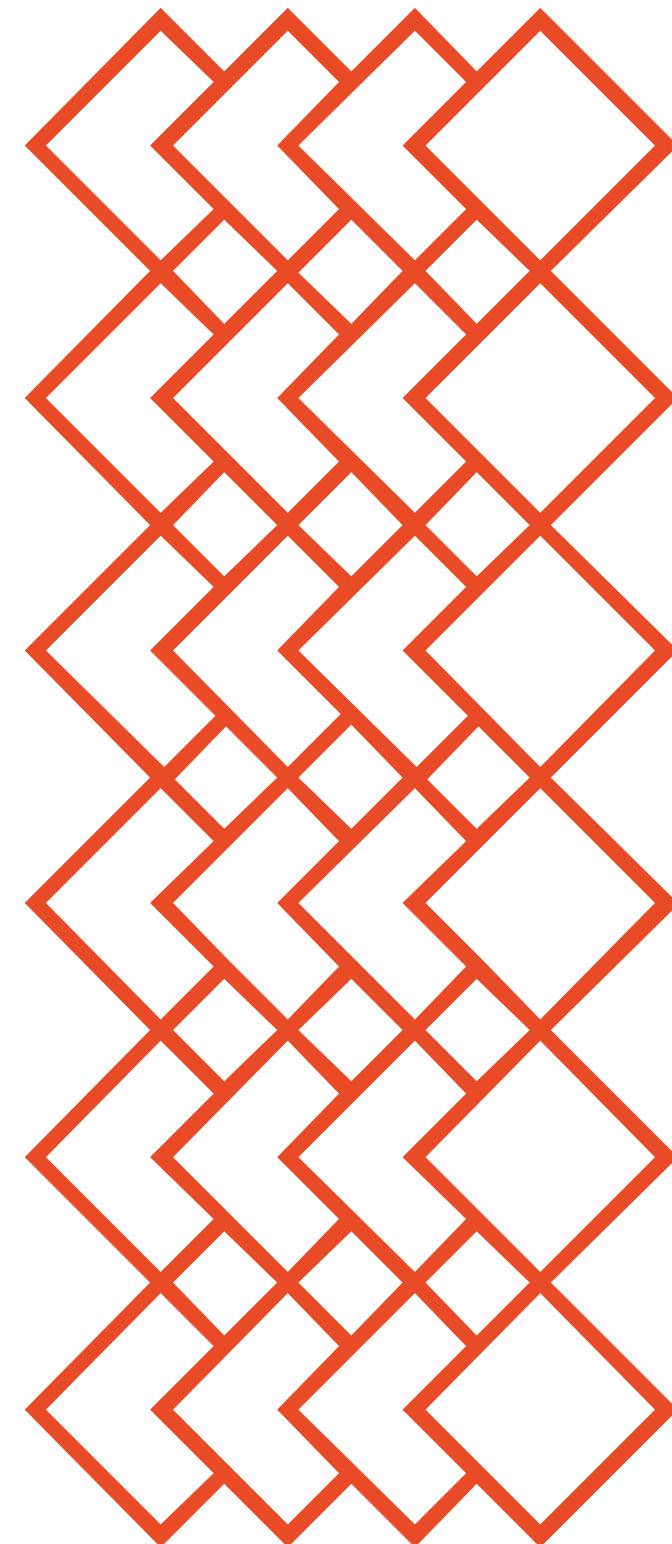
CONCLUSION & FUTURE WORK

This study has given valuable insights into the relationship of the lyrics of a song and its genre.

- Works best for linguistically distinctive genres like Rap.
- Struggles with overlapping genres like Pop, and Rock.
- Linear models have degraded performance trying to classify genres through lyrics.

Future works could higher-level models that work better with non-linear data such as:

- Transformer-based embeddings like BERT.
- Hierarchical or graph-based topic models.
- Combining lyrics with stylistic or audio features.



THANK YOU!

Presented by Group 5.

