Sentiment Analysis and Topic Modeling of BTS' (방탄소년단) English Lyrics

Fatima Lois E. Suarez

NLPIR Project Presentation

Outline

- Introduction
- Objectives
- Datasets
- Sentiment Analysis
  - Methodology
  - Results
  - Challenges
- Topic Modeling
  - Methodology
  - Results
  - Challenges

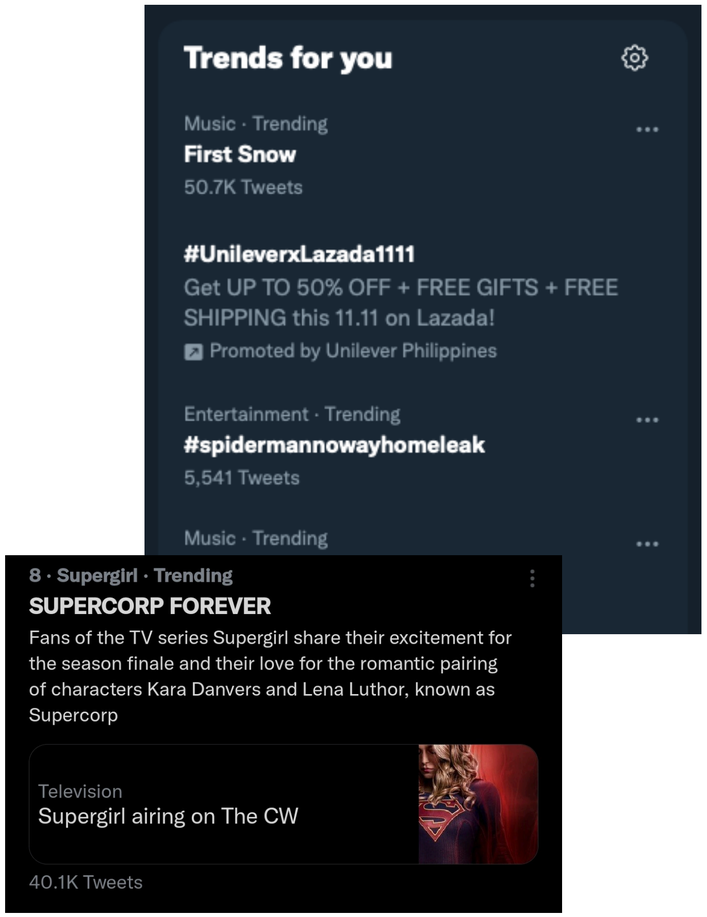
Sentiment Analysis

- determining whether a particular text expresses a **positive, negative, or neutral opinion**
- can be expanded to a range of different **emotions**

Topic Modeling

- assigning topics to unlabelled text documents

  ⇢ determining **underlying/ hidden topics or themes** in a body of text and **how much of a topic exists** in it



BTS Lyrics Data

- Sourced from: https://www.kaggle.com/kailic/bts-lyrics
  - Contains: track ID, album titles, album release date, English and Korean titles, **English translation of lyrics (Genius)**, special track indicators (hidden, remix, featuring, repackaged, version), performers, language
- 19 studio albums and EPs from **2013 to 2020**

Challenges encountered with BTS lyrics data

- 19 albums but only **131 unique songs with lyrics**
  - ***repackaged albums*** - common in K-Pop to release new albums with the same songs + a few new songs
- overlapping songs in dataset ⇢ *cannot analyze by album*
- data is unlabeled (for sentiment analysis)

Objective: Sentiment Analysis

- Train a binary classifier on a set of non-K-Pop lyrics

  ⇢ *How well can non-K-Pop lyrics evaluate the sentiment expressed in BTS' lyrics?*

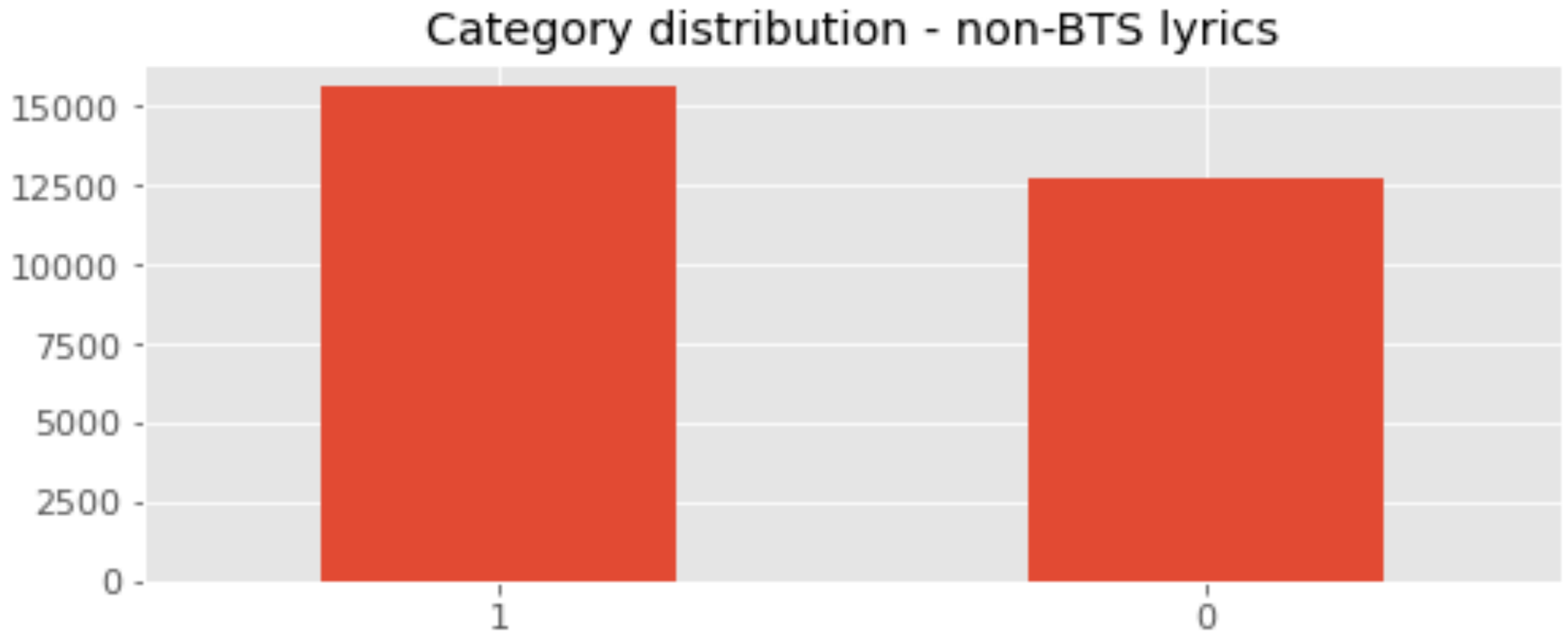
Lyrics Dataset

- Sourced from: https://data.mendeley.com/datasets/3t9vbwxgr5/3
- list of ~20,000 lyrics from 1950 to 2019
- includes *metadata* of all songs

Labeling the lyrics data for sentiment analysis

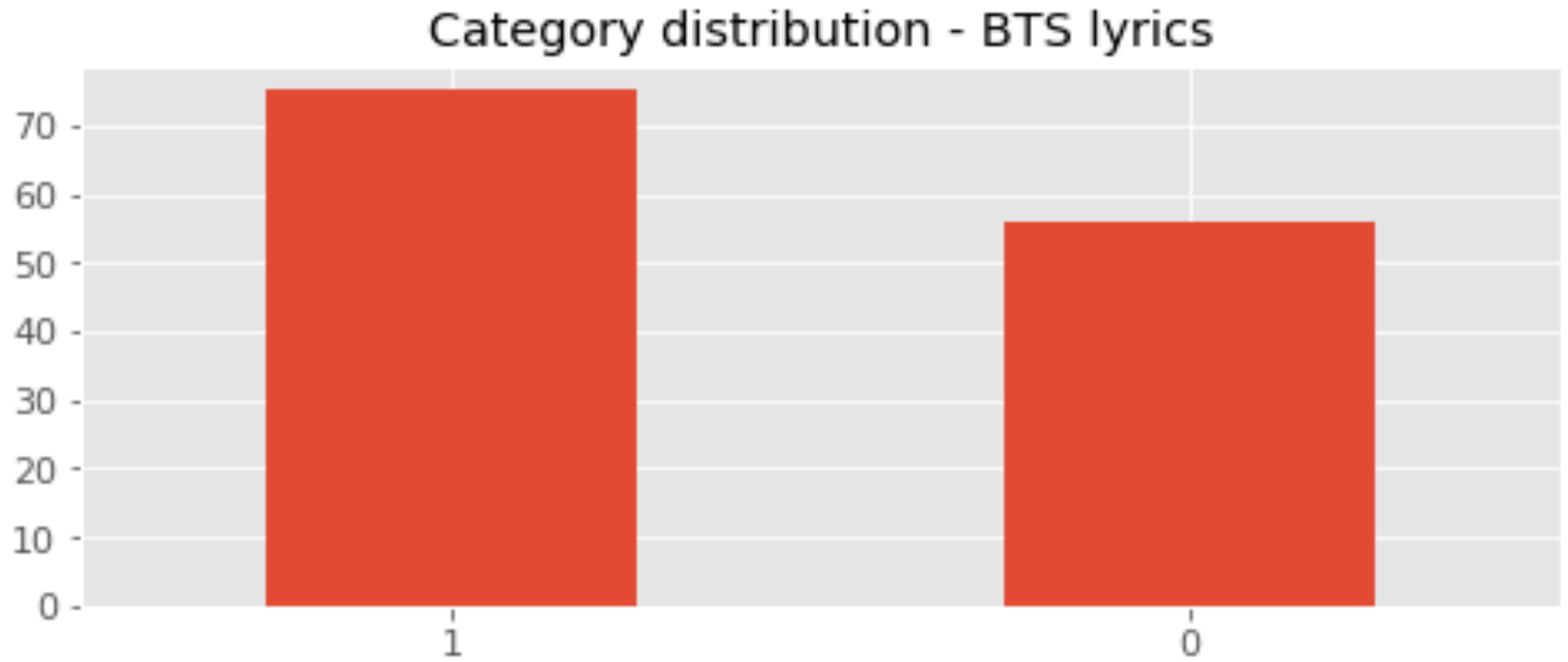
- **valence**: describes musical positiveness conveyed by a song
  - *high valence*: more positive (e.g. happy, cheerful)
  - *low valence*: more negative (e.g. sad, angry)
- non-BTS lyrics: metadata
- BTS lyrics: Spotify API was used to extract metadata for all unique songs
- labels: 0 == valence $<$ 0.5; 1 == valence $\geq$ 0.5

Category Distribution: Non-BTS Lyrics



- 0 == valence < 0.5; 1 == valence ≥ 0.5

Category Distribution: BTS Lyrics



- 0 == valence < 0.5; 1 == valence ≥ 0.5

Sentiment Analysis

1. **Data Preprocessing**
    - remove instrumentals, skits, duplicate songs
    - tokenization, remove punctuation, lemmatization
2. **Analysis**
    - feature extraction
        - word frequency
        - TF-IDF
    - binary classifier: Multinomial Naive Bayes
    - evaluation: classification report + confusion matrix

Source code: Introduction to Social Networks - Lab Homework 2

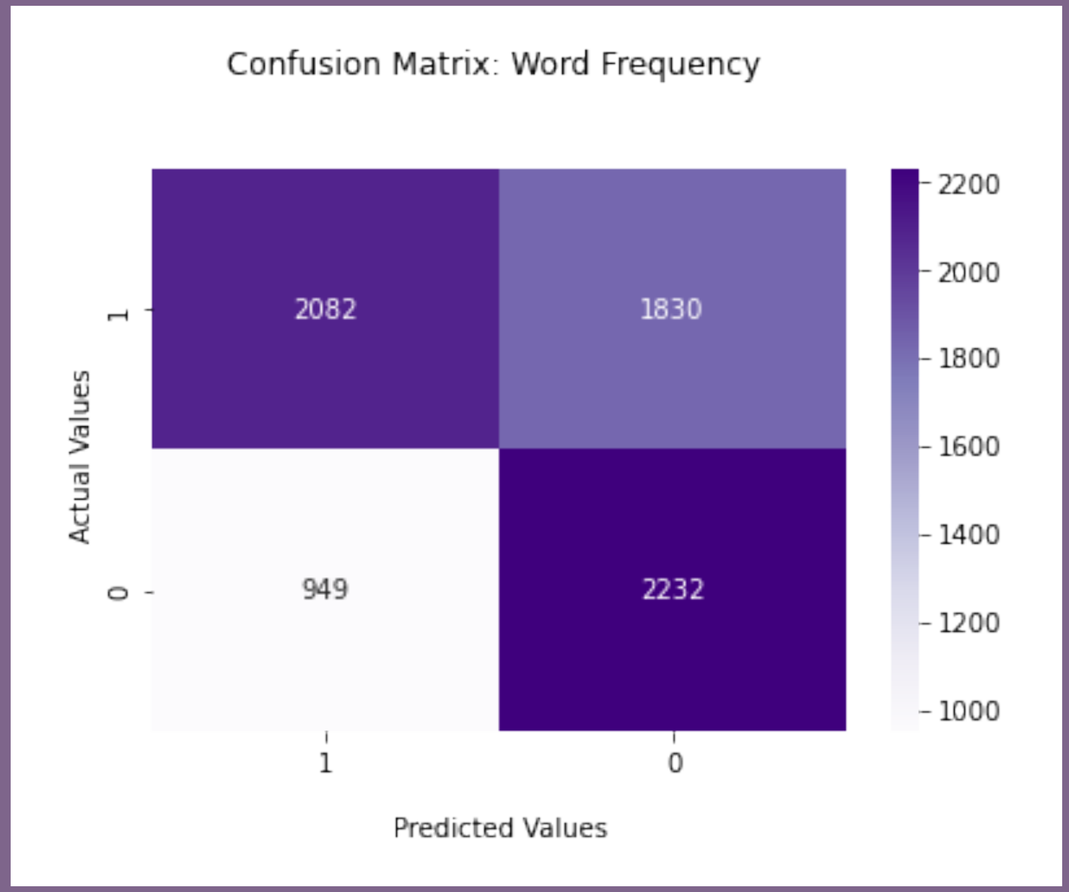
Training + testing binary classifier on non-K-Pop lyrics

Word Frequency:

Accuracy: 0.6082052728041731

Classification Report

	precision	recall	f1-score	support
0	0.55	0.70	0.62	3181
1	0.69	0.53	0.60	3912
accuracy			0.61	7093
macro avg	0.62	0.62	0.61	7093
weighted avg	0.63	0.61	0.61	7093



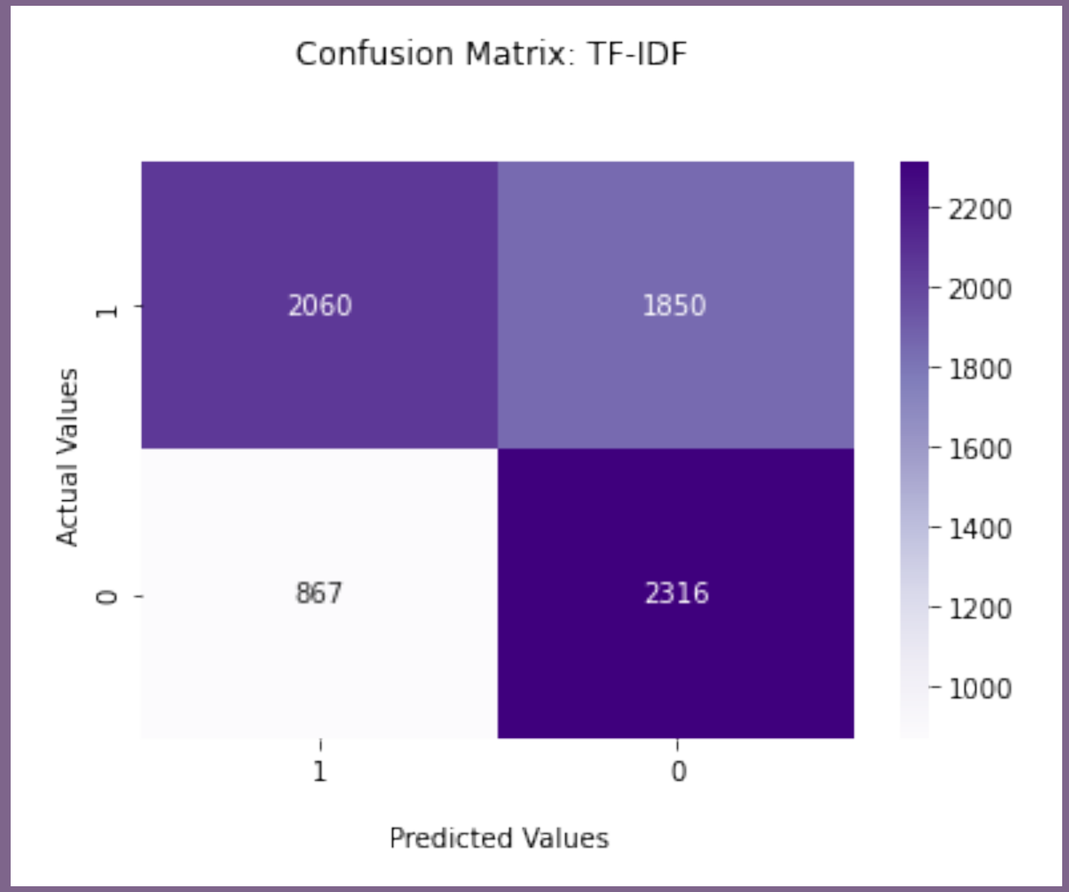
Training + testing binary classifier on non-K-Pop lyrics

TF-IDF:

Accuracy: 0.6169462850697871

Classification Report

	precision	recall	f1-score	support
0	0.56	0.73	0.63	3183
1	0.70	0.53	0.60	3910
accuracy			0.62	7093
macro avg	0.63	0.63	0.62	7093
weighted avg	0.64	0.62	0.62	7093



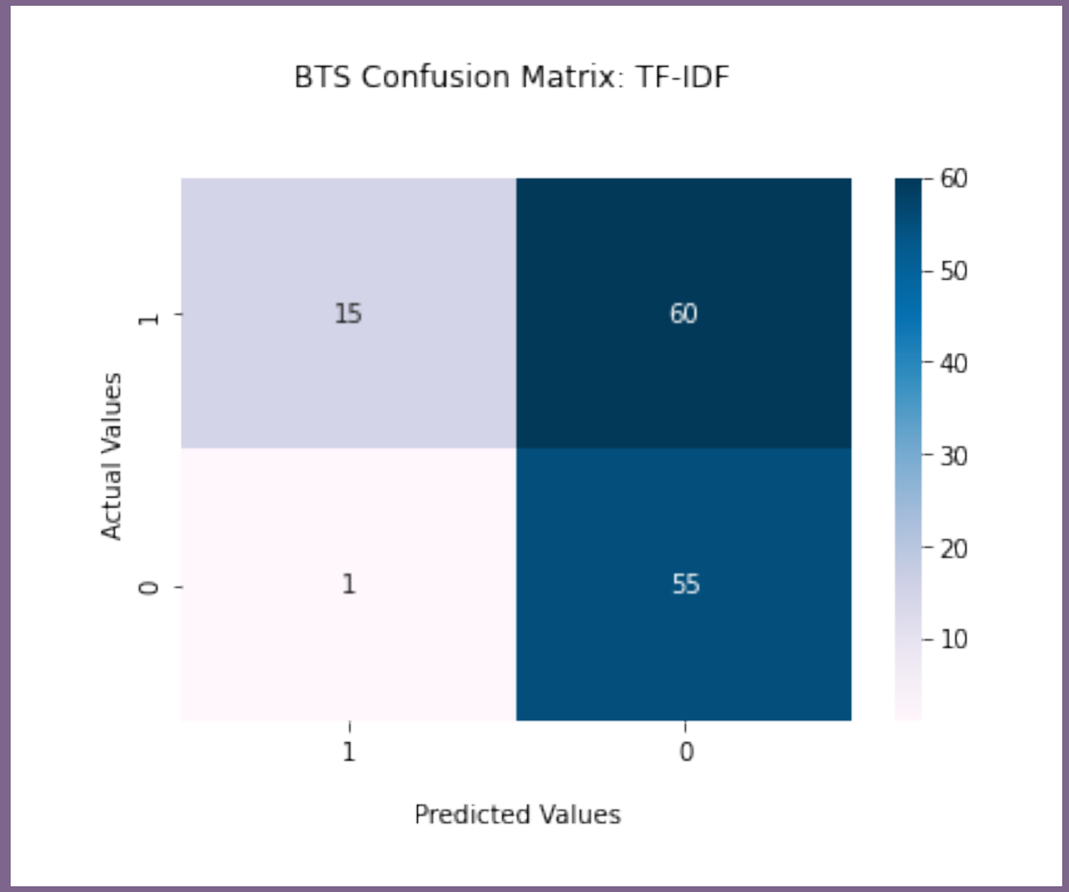
Using classifier on BTS’ lyrics

Word Frequency:

Accuracy: 0.5343511450381679

Classification Report

	precision	recall	f1-score	support
0	0.48	0.98	0.64	56
1	0.94	0.20	0.33	75
accuracy			0.53	131
macro avg	0.71	0.59	0.49	131
weighted avg	0.74	0.53	0.46	131



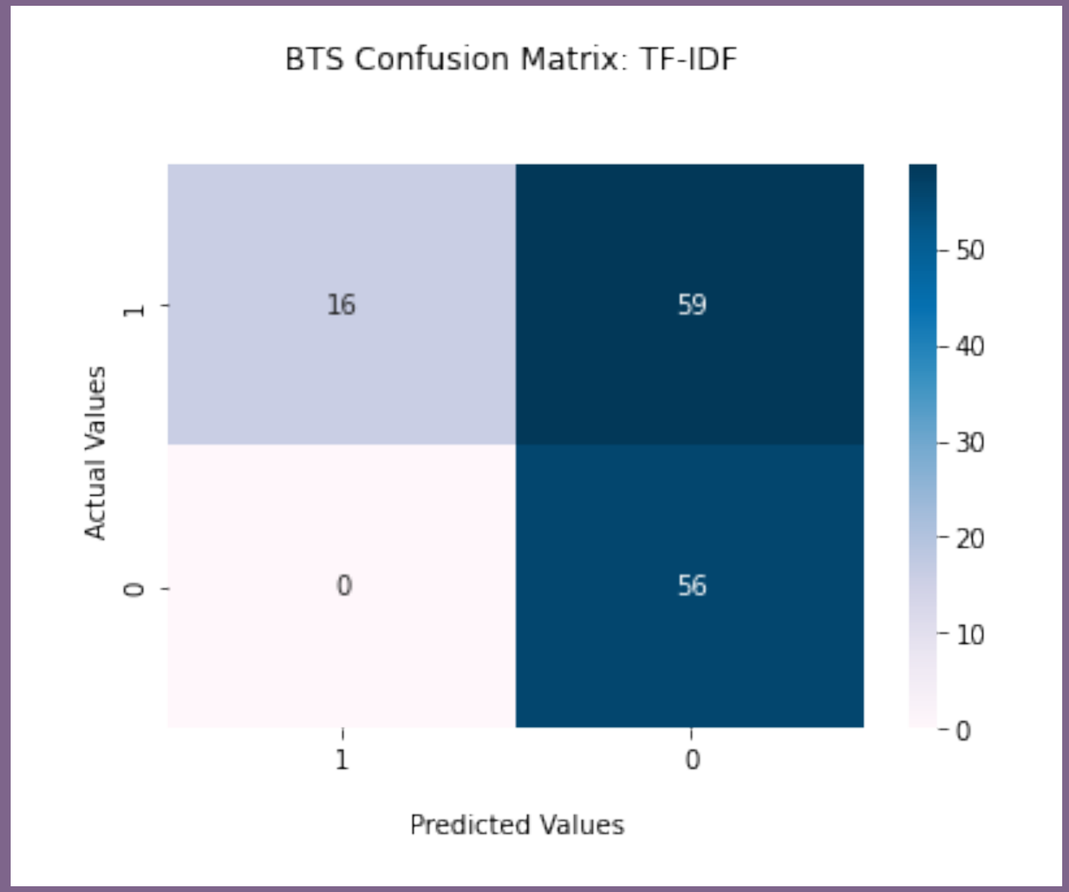
Using classifier on BTS' lyrics

TF-IDF:

Accuracy: 0.549618320610687

Classification Report

	precision	recall	f1-score	support
0	0.49	1.00	0.65	56
1	1.00	0.21	0.35	75
accuracy			0.55	131
macro avg	0.74	0.61	0.50	131
weighted avg	0.78	0.55	0.48	131



Remarks

- valence as a label
  - musical features vs. lyrics only
- computing the feature matrix
  - BTS lyrics included so that classifier can be applied

  ⇢ increased number of features
- pre-processing
  - lemmatization: ex. lemmatize(*“was”*) ⇢ *“wa”*
  - Korean words without direct English translations

Objectives: Topic Modeling

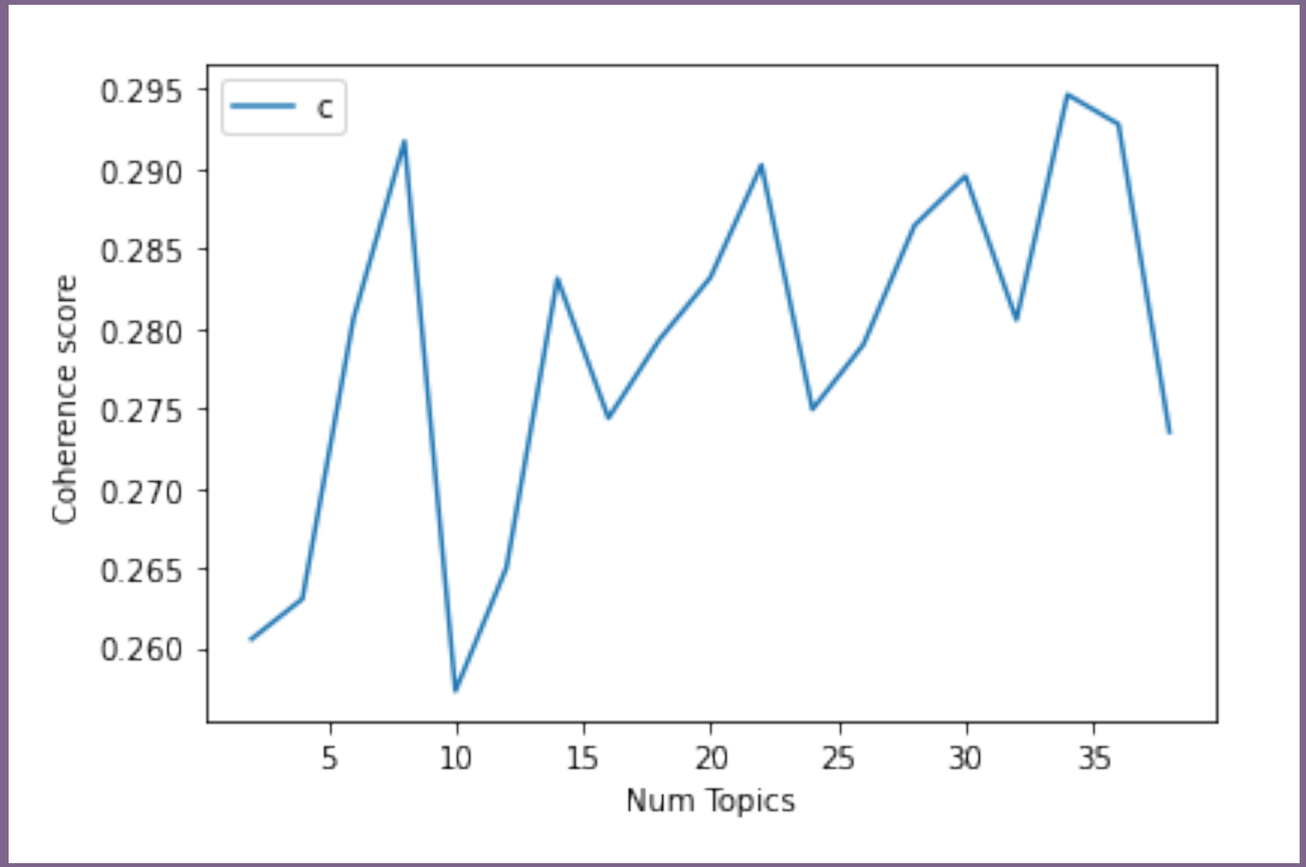
- What are the prevalent topics and themes discussed in their music?
- How do the themes of their songs vary over time?

Topic Modeling (Latent Dirichlet Allocation)

1. **Data Preprocessing**
    - remove instrumentals, skits, duplicate songs
    - tokenization, remove punctuation, lemmatization
2. **Analysis**
    - create document term matrix from processed data
    - choosing number of topics: coherence score
    - train LDA model
    - classify song lyrics into topics
        - plot song topics over time

Source code: https://github.com/anshieh12/Taylor_swift_song_recommender/blob/master/notebook/01_Topic_modeling.ipynb

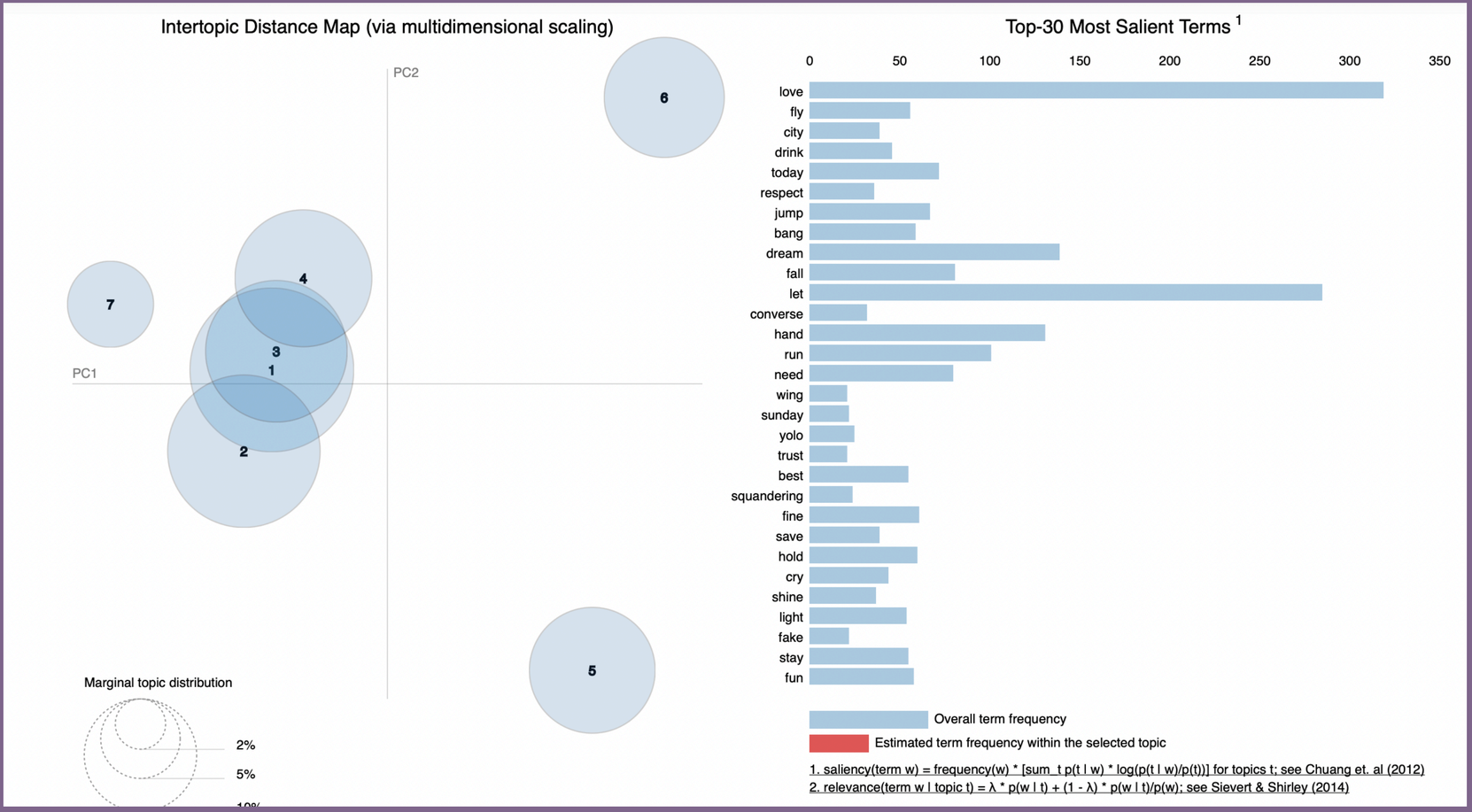
Choosing number of topics



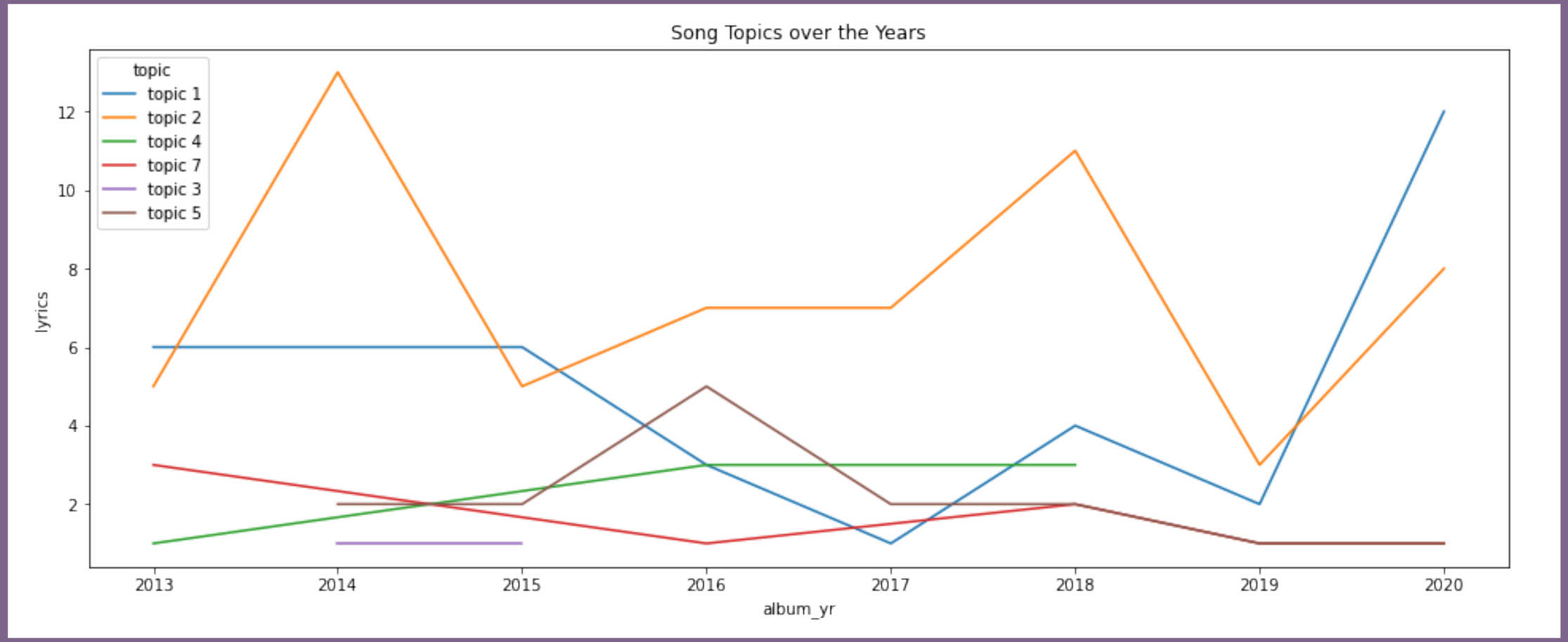
Generated topics

topic	top words
1	know, let, day, bang, fine, want, stay, way, fun, moment
2	know, converse, want, save, high, get, stay, yolo, time, squandering
3	dream, say, make, want, world, love, know, life, got, crazy
4	love, know, drink, fall, respect, say, live, look, fake, shot
5	let, jump, get, time, feel, hand, want, look, say, day
6	city, today, fly, hand, wing, night, want, trust, come, shine
7	love, let, run, need, say, keep, day, stop, sunday, turn

Intertopic Distance Map



Song Topics over Time



Remarks

- topic similarity to improve clustering + topic assignment
- assigning topic names based on top words

References

- *‘The Data Science of “Someone Like You” or Sentiment Analysis of Adele’s Songs’* by Preetish Panda, Prompt Cloud. Retrieved from: *https://www.kdnuggets.com/2018/09/sentiment-analysis-adele-songs.html*
- *TIGP SNHCC* - Introduction to Social Networks - Text Mining Lab Homework
- *‘Using Machine Learning to Analyze Taylor Swift's Lyrics’* by Ian Freed. Retrieved from: *https://www.codecademy.com/resources/blog/taylor-swift-lyrics-machine-learning/*
- *‘Taylor Swift Song Recommendation with NLP and Topic Modeling’* by Annie Shieh. Retrieved from: *https://medium.com/@annieshieh12/taylor-swift-song-recommendation-with-nlp-and-topic-modeling-8594636608fa*
  - Github (source codes): *https://github.com/anshieh12/Taylor_swift_song_recommender*
- *‘What Makes A Song Likeable?’* by Ashrit Shetty. Retrieved from: *https://towardsdatascience.com/what-makes-a-song-likeable-dbfdb7abe404*
- *Text Mining with R: A Tidy Approach* by Julia Silge and David Robinson (2017)
- *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions* by Bing Liu (2015)
- *‘Data Analysis of K-POP: Playing with Spotify API’* by Nancy Yan. Retrieved from: https://nancyyanyu.github.io/posts/63adf3bb/
- *‘The Data Science of K-Pop: Understanding BTS through data and A.I.’* by Haebichan Jung. Retrieved from: https://towardsdatascience.com/the-data-science-of-k-pop-understanding-bts-through-data-and-a-i-part-1-50783b198ac2
- *‘Is K-pop still K-pop without the ‘K’?’* by Haley Yang. Retrieved from: https://koreajoongangdaily.joins.com/2021/10/12/entertainment/kpop/twice-the-feels-english-kpop-songs-boa-eat-you-up/20211012155647679.html

Thank you very much!