FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Spotify Sentiment Analysis

Lauren Jachimczyk

Machine Learning Fall 2023

Background

- Spotify's API allows you to access your personal song data, including song titles, artist names and valence scores
  - Valence Score (according to Spotify) = a measure of a song's positiveness
- I was interested in the concept of songs that are upbeat but actually about sad topics
  - A lot of my favorite songs fall into this category and I was curious about 2 questions:
    - **How would Spotify's valence score classify them vs. a ML sentiment analysis based on the lyrics vs. my own classification?**
    - **How to find the right labels in a sentiment analysis, since they can be quite subjective?**

Pre Processing

Size of the datasets and cleaning

- Due to limitations of Spotify's API, they only allow you to retrieve your top 50 songs and they do not allow you to retrieve the lyrics of those songs.
- I used a different API to retrieve song lyrics, however some of them came in completely wrong, so I needed to remove 5 of them.
- Lastly, 1 song is instrumental/has no lyrics so it also needed to be removed
- This left us with 44 songs
- Due to the small size of the dataset, I did not introduce a 3rd class of Neutral sentiment. The songs were labeled as either Positive (1) or Negative (0)

Labels

- Valence score from Spotify is between 0 and 1. Used 0.5 as threshold to create an updated label of either 0 or 1
- Self Assigned labels were determined by me. I exported the top songs to excel, added my labels (determined by pre existing knowledge of the songs, listening to the songs again, watching their music videos and searching Google to see if the artist has made any comments about the meaning of the song) and imported the updated excel sheet back in

Tokenization and NLP

- Used Ntlk's word_tokenize and sklearn's TfidfVectorizer to convert the song lyrics into features
- Removed stop words and song lyric-related words (i.e. chorus, verse)

Train Test Split

Datasets before Tokenization

1

	Name	Artist	Lyrics	Valence	Binary_Valence
0	Oceans of Darkness	The War On Drugs	8 ContributorsOceans of Darkness Lyrics[Verse ...	0.782	1
1	Cleopatra	The Lumineers	53 ContributorsTranslationsPortuguêsCleopatra ...	0.485	0
2	It's Not Living (If It's Not With You)	The 1975	134 ContributorsTranslationsPortuguêsIt's Not ...	0.526	1
3	Friend of the Devil	Mumford & Sons	5 ContributorsFriend of the Devil Lyrics[Verse...	0.179	0
4	Ends of the Earth	Lord Huron	26 ContributorsEnds of the Earth Lyrics[Intro]...	0.407	0

2

	Name	Artist	Lyrics	Self Assigned Label
0	Oceans of Darkness	The War On Drugs	8 ContributorsOceans of Darkness Lyrics[Verse ...	0
1	Cleopatra	The Lumineers	53 ContributorsTranslationsPortuguêsCleopatra ...	0
2	It's Not Living (If It's Not With You)	The 1975	134 ContributorsTranslationsPortuguêsIt's Not ...	0
3	Friend of the Devil	Mumford & Sons	5 ContributorsFriend of the Devil Lyrics[Verse...	0
4	Ends of the Earth	Lord Huron	26 ContributorsEnds of the Earth Lyrics[Intro]...	0

Datasets after Tokenization

1 X_Valence.head()

	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.292591	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.080108	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.048308	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.033998	0.049780	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0

2 X_Self_Assigned.head()

	0	1	2	3	4	5	6	7	8	9	...	990	991	992	993	994	995	996	997	998	999
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.290658	...	0.0	0.0	0.0	0.0	0.0	0.066169	0.0	0.0	0.079579	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.048136	0.000000	0.000000	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.033946	0.049704	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	0.0

Naive Bayes Model

Naive Bayes ML algorithm was used because it is a commonly used algorithm for Sentiment Analysis, it was particularly useful for this project because Naive Bayes Models:

- ✓ Works well with discrete values
- ✓ Does not require large amounts of training data

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Probability of A happening, given that B has occurred
B is the evidence and A is the hypothesis

Results

Accuracy for Valence dataset:
78%

Classification Report for Valence Dataset:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	7
1	0.00	0.00	0.00	2
accuracy			0.78	9
macro avg	0.39	0.50	0.44	9
weighted avg	0.60	0.78	0.68	9

Accuracy for Self Assigned Label dataset:
67%

Classification Report for Self Assigned Label Dataset:

	precision	recall	f1-score	support
0	1.00	0.40	0.57	5
1	0.57	1.00	0.73	4
accuracy			0.67	9
macro avg	0.79	0.70	0.65	9
weighted avg	0.81	0.67	0.64	9

Results continued

	Name	Artist	Valence	Binary_Valence	Self Assigned Label	Predicted_Label
0	Oceans of Darkness	The War On Drugs	0.7820	1	0	0
1	Cleopatra	The Lumineers	0.4850	0	0	0
2	It's Not Living (If It's Not With You)	The 1975	0.5260	1	0	1
3	Friend of the Devil	Mumford & Sons	0.1790	0	0	1
4	Ends of the Earth	Lord Huron	0.4070	0	0	0
5	Landslide	Dagny	0.4130	0	0	1
6	The Night We Met	Lord Huron	0.1000	0	0	0
7	It's Called: Freefall	Rainbow Kitten Surprise	0.2810	0	0	1
8	Sleep On The Floor	The Lumineers	0.2750	0	1	0
9	Red Eyes	The War On Drugs	0.4670	0	1	1
	Mirror	Porter Robinson	0.3500	0	1	0
11	Agnes	Glass Animals	0.2330	0	0	1
12	Babel	Mumford & Sons	0.6480	1	1	1
13	Graveyard girl	M83	0.1420	0	0	0
	Equal	ODESZA	0.1800	0	0	0
15	Gloria	The Lumineers	0.6590	1	0	1
	At the Bottom of Everything	Bright Eyes	0.7840	1	0	0
	Wake Me	Bleachers	0.4570	0	1	0

This song is about suicide

This song is about addiction

This song is about a plane crash

Conclusion

- After running both datasets through the Naive Bayes model, the Valence dataset had a higher accuracy rate
- Nuances of music and sentiment potentially not captured
- Issues with limitations of Spotify's API as well as preprocessing being more time consuming than anticipated