NLP and Clustering of YouTube Music & Movie Videos

YouTube sentiment analysis Project

- **Course:** Advanced topics in machine learning
- **Lecturer:** Dr. Chen Hajaj
- **Team Members:** Dor Ingber & Itai Bekenshtein

The Problem We're Addressing

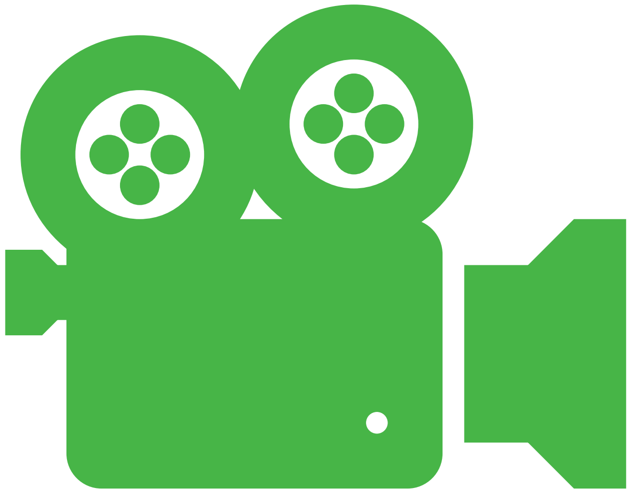
User-generated content on YouTube is vast and reflects diverse sentiments.

Understanding audience sentiment towards music and movies is valuable for content creators and marketers.

Current approaches to sentiment analysis often neglect the unique context of YouTube video descriptions and titles.

Original Project Goals

- Extract sentiment from YouTube video titles and descriptions.
- Cluster videos based on their dominant sentiment.
- Provide insights for content creators and marketers to tailor their offerings and marketing strategies.



Other Methods & Techniques

- **Other Potential Approaches:**
    - **Supervised Machine Learning:** Requires large, labeled datasets to train models. Models learn to predict sentiment based on provided labels (positive, negative, etc.).
    - **Deep Learning:** Advanced techniques using neural networks. Can excel at complex sentiment analysis but often require substantial computational resources and massive datasets.

Other Methods & Techniques

- **Why We Chose Unsupervised Learning:**
  - **No Labeled Data:** Manually labeling sentiment for a large YouTube video dataset can be extremely time-consuming and expensive.
  - **Exploratory Analysis:** We wanted to uncover patterns in the data and identify the dominant sentiments without prior assumptions about specific emotions.

Dataset and Features

- First Data:
- Source: YouTube Data API v3
- Features: ID, Title, Description, Date

	Id	title	date	description
0	v4KXWsMw8Fc	Relaxing Music For Stress Relief, Anxiety and ...	2024-03-18T08:40:05Z	Relaxing Music For Stress Relief, Anxiety and ...
1	NgGJaXDC0wU	Best Praise and Worship Songs 2023 ✝️ Nonstop...	2024-03-18T13:49:38Z	► Music and Video Copyright belongs to @Praise...
2	mLW35YMzELE	Creepy Nuts「Bling-Bang-Bang-Born」 × TV Anime「マ...	2024-03-03T09:00:37Z	「Bling-Bang-Bang-Born」 (2024.1.7.Digital Relea...
3	BxPhT3mVVQw	🔴 Relaxing Music 24/7, Sleep Music, Stress Rel...	2024-03-18T08:51:47Z	Enjoy our latest relaxing music live stream: y...
4	h8Cq1BwdTsg	Ozoda - Ko'k jiguli (Official Music Vide...	2024-02-21T13:39:33Z	Composer: OZODA\nLyrics: OZODA\nArrangement: D...
...	...	...	...	...
1274	ZygQeQ4paLc	3 Movies From My Big Fat Greek Wedding Flim Fr...	2024-03-17T21:12:46Z	Please Subscribe my YouTube channel.\n\n\n\n\n...
1275	qAPsrv5qFwc	2 Movies of The Evil Dead Flim Franchise.	2024-03-17T20:42:16Z	Please Subscribe my YouTube channel.\n\n\n\n\n...
1276	HUrsuH5KRJM	Scene 👀 #hollywoodmovie #englishmovie #moviecl...	2024-03-17T11:11:12Z	یا کونسا ایسا ھےجو تمھیں روزی دے اگر وہ اپنی ...
1277	Wwd1sfHSwcQ	A hilarious and outrageous comedy video #movie...	2024-03-19T18:10:02Z	Click here 🗣️ \nDarkness Lurks Every Corner \nh...
1278	3a49xBNA98Q	Beast From The Haunted Cave (1959) | Michael F...	2024-03-17T21:30:00Z	Beast From The Haunted Cave (1959) | Michael F...

1279 rows × 4 columns

Dataset and Features

- Second Data: After NLP process
- Features: ID, Title, Description, Date, title&description , emoji, processed, emotions, emoji_grade, scaled_emotions

	Id	title	date	description	title&description	emoji	processed	emotions	emoji_grade	scaled_emotions
0	v4KXWsMw8Fc	Relaxing Music For Stress Relief, Anxiety and ...	2024-03-18 08:40:05	Relaxing Music For Stress Relief, Anxiety and ...	relaxing music for stress relief anxiety and d...	['🎹', '🌞', '✔']	relax music stress relief anxiety depressive s...	{'joy': 0.2, 'positive': 0.28421052631578947, ...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	{'joy': 0.6923076923076924, 'positive': 0.9999...
1	NgGJaXDC0wU	Best Praise and Worship Songs 2023 ✝️ Nonstop...	2024-03-18 13:49:38	► Music and Video Copyright belongs to @Praise...	best praise and worship songs 2023 nonstop chr...	['✝️', '🔻', '📌', '📌', '📌', '📌']	best praise worship song nonstop song time pra...	{'joy': 0.32653061224489793, 'positive': 0.346...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	{'joy': 0.9375, 'positive': 1.0000000000000002...
2	mLW35YMzELE	Creepy Nuts「Bling-Bang-Bang-Born」 × TV Anime「マ...	2024-03-03 09:00:37	「Bling-Bang-Bang-Born」 (2024.1.7.Digital Relea...	creepy nutsbling bang bang born tv anime mashl...	['©']	anime mashle collaboration music video bbbb bl...	{'joy': 0.02912621359223301, 'positive': 0.077...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	{'joy': 0.16666666666666669, 'positive': 1.0, ...
3	BxPhT3mVVQw	🔴 Relaxing Music 24/7, Sleep Music, Stress Rel...	2024-03-18 08:51:47	Enjoy our latest relaxing music live stream: y...	relaxing music 24/7 sleep music stress relief ...	['🔴']	relax music sleep music stress relief music sp...	{'joy': 0.3020408163265306, 'positive': 0.4265...	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound...	{'joy': 0.7024390243902439, 'positive': 1.0, '...
4	h8Cq1BwdTsg	Ozoda - Ko'k jiguli (Official Music Vide...	2024-02-21	Composer: OZODA\nLyrics: OZODA\nArrangement: D...	ozoda ko 39 k jiguli official music video	['▪', '▪']	official music video composer ozoda lyric	{'trust': 0.047619047619047616,	{'neg': 0.0, 'neu': 1.0, 'pos': 0.0,	{'trust': 0.0, 'joy': 0.0, 'positive': 1.0, 's...

Dataset and Features

- Third Data: emotions data frame
- Features: Id, joy, positive, sadness, negative, anger, anticipation, fear, trust, disgust, surprise

	Id	joy	positive	sadness	negative	anger	anticipation	fear	trust	disgust	surprise
0	v4KXWsMw8Fc	0.692308	1.000000	0.538462	0.461538	0.115385	0.153846	0.115385	0.192308	0.000000	0.000000
1	NgGJaXDC0wU	0.937500	1.000000	0.187500	0.000000	0.000000	0.437500	0.187500	0.812500	0.000000	0.000000
2	mLW35YMzELE	0.166667	1.000000	0.500000	0.000000	0.833333	0.000000	0.833333	0.166667	0.000000	0.000000
3	BxPhT3mVVQw	0.702439	1.000000	0.521951	0.043902	0.000000	0.146341	0.004878	0.175610	0.000000	0.063415
4	h8Cq1BwdTsg	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Dataset and Features

- Forth Data: Clusters data frame
- Features: Id, joy, positive, sadness, negative, anger, anticipation, fear, trust, disgust, surprise

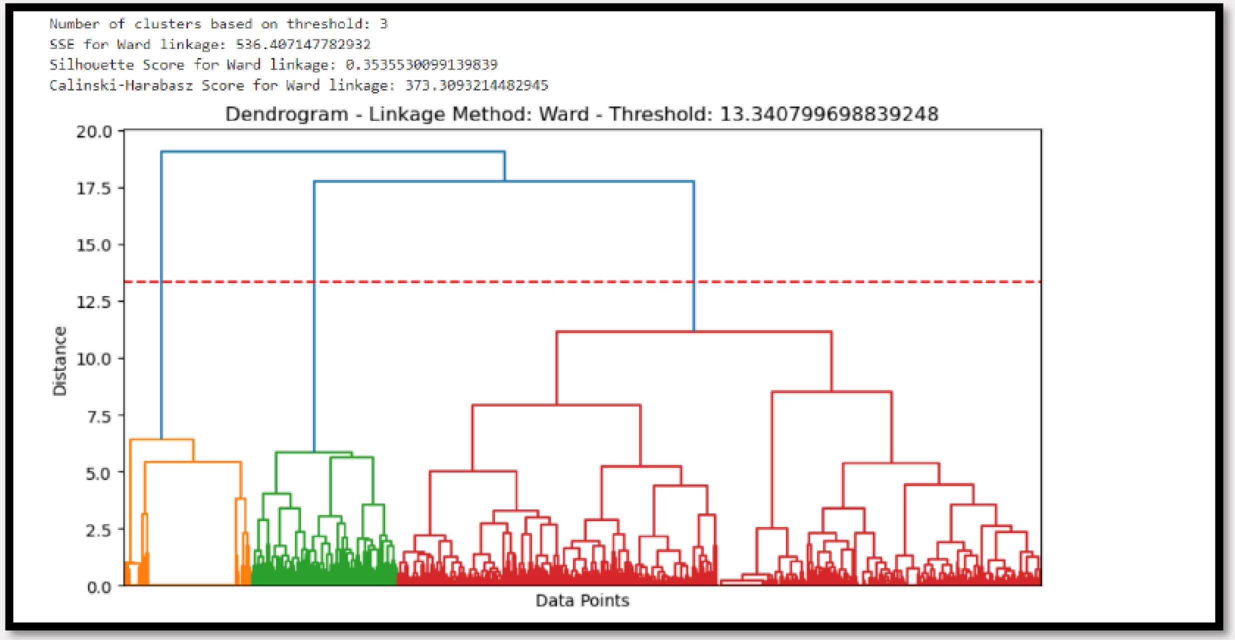
	Id	joy	positive	sadness	negative	anger	anticipation	fear	trust	disgust	surprise	cluster_label
0	mLW35YMzELE	0.166667	1.000000	0.500000	0.000000	0.833333	0.000000	0.833333	0.166667	0.000000	0.000000	2
1	3cbnNwxtUUA	0.137931	0.206897	0.206897	1.000000	0.068966	0.000000	0.862069	0.000000	0.034483	0.758621	2
2	t_4ob8SB2UI	0.609756	1.000000	0.487805	0.804878	0.365854	0.560976	0.243902	0.536585	0.000000	0.048780	2
3	pRpeEdMmmQ0	0.333333	0.666667	0.833333	0.833333	0.166667	1.000000	0.333333	0.666667	0.000000	0.000000	2
4	36vjwGx-Vzc	0.000000	0.222222	0.111111	1.000000	0.111111	0.000000	0.111111	0.000000	0.111111	0.888889	2

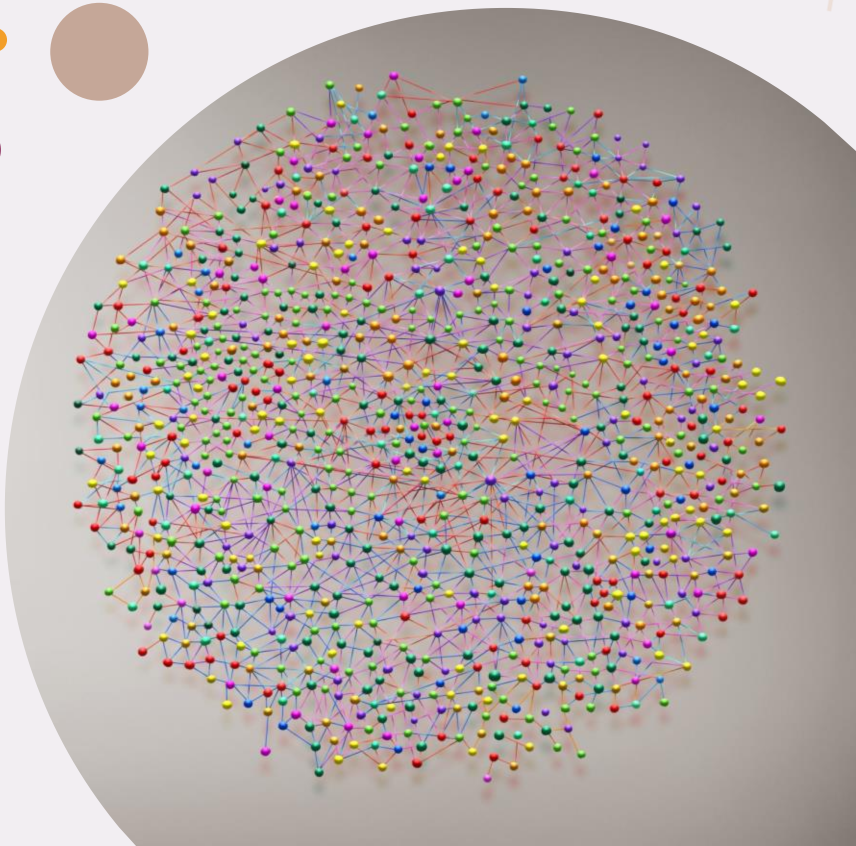
Methodology

- **NLP**
  - Libraries used (VaderSentiment, spaCy, NRCLex).
  - vaderSentiment - for analyzing emotions from emojis.
  - spacy - Stop word removal, Lemmatization, and entity removal.
  - NRCLex - dealing with sentiment analysis from text.
  - We used MIN MAX Scaler to normalize sentiment scores for each video. We wanted to normalize the sentiments to highlight dominant emotions and reduce the scores of minor emotions.

Methodology

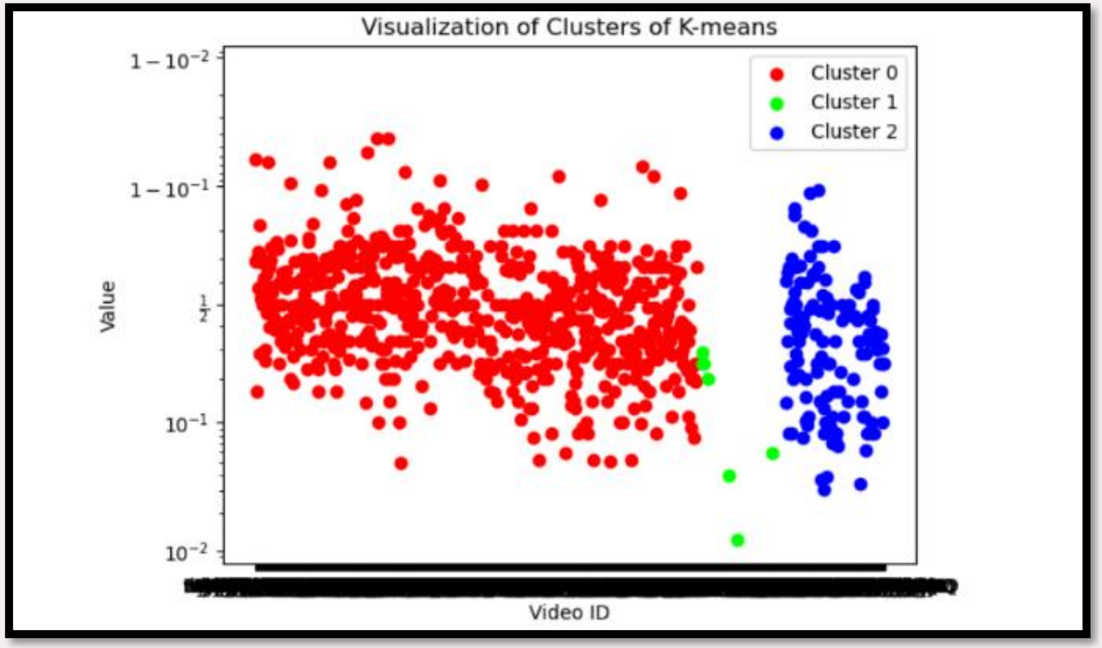
- ## Clustering
  - ### Hierarchical Clustering:
    - Forms a hierarchy of clusters based on distances between data points.
    - Offers a visual representation of the data's hierarchical structure (dendrogram). and identifying natural clusters in the data.
    - Threshold explains 70% of the data

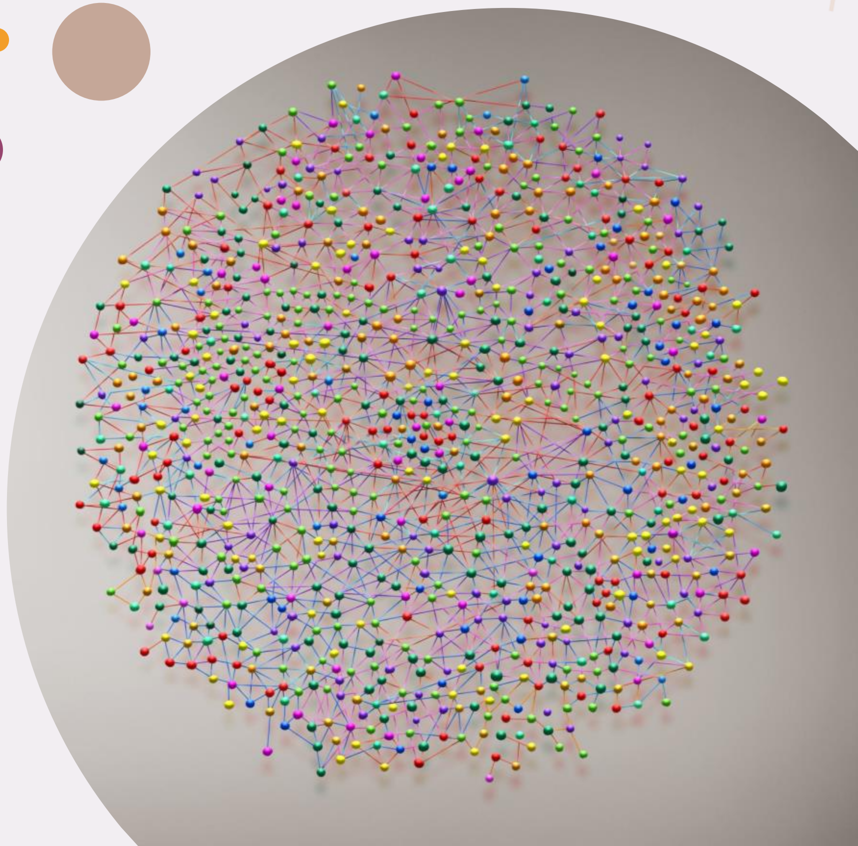




Methodology

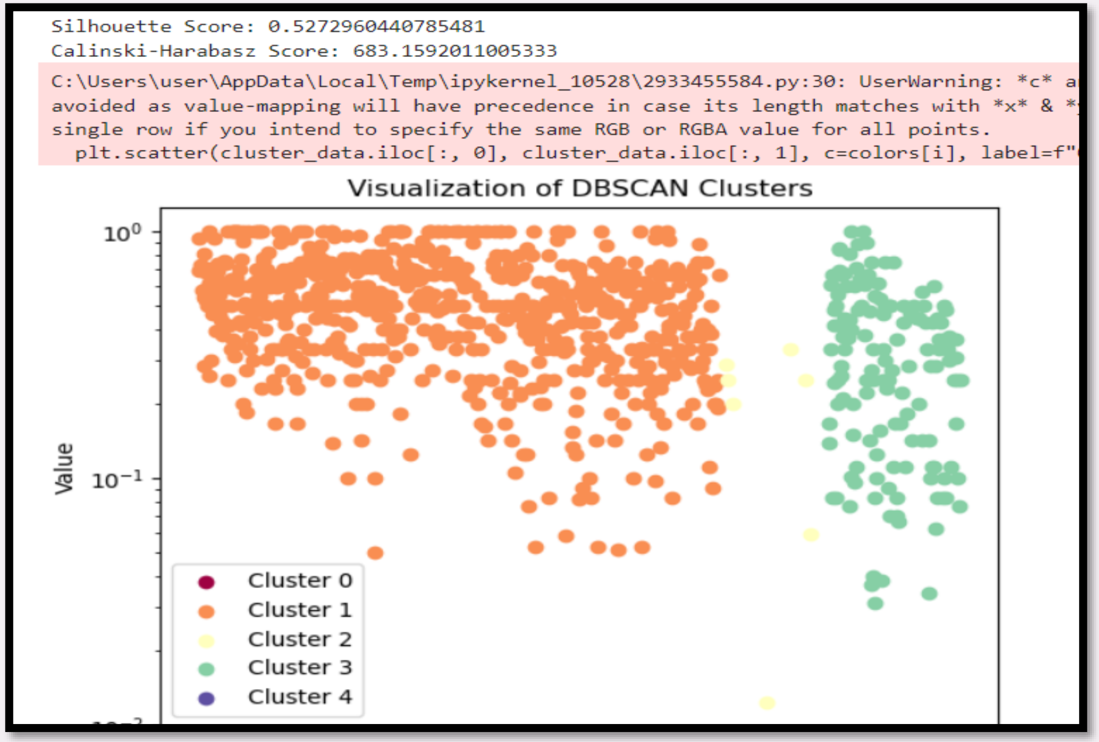
- **Clustering**
  - **K-means Clustering:**
    - Partitions data into a fixed number (k) of clusters.
    - We tried this model because its clear separation of data points into clusters makes it easy to identify and understand the different groups present in the data.

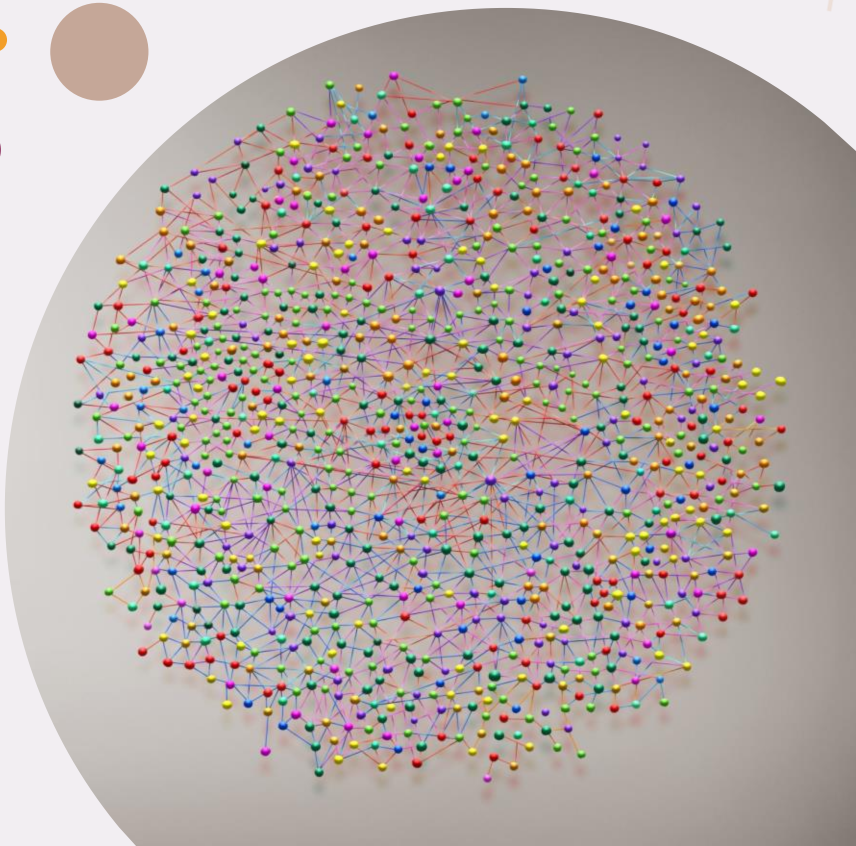




Methodology

- **Clustering**
  - **DBSCAN Clustering:**
    - Groups data points based on their density in a neighborhood.
    - Does not require a pre-defined number of clusters.
    - We tried this model because we wanted to diagnose a model that works on density and not on distance calculation





Experiments

- **Parameter Tuning:**
  - **Hierarchical Clustering:**
    - We measured the results for changes in the hyperparameters of this model. Calculating result for each linkage method.
  - **K-means Clustering:**
    - To find K, we pairs of emotions with the highest correlation between them. We used “networkx “to create a graph to find groups of closely related emotions based on their correlation. We checked the natural clusters from the Hierarchical model as well.
    - We measured the results for changes in the hyperparameters of this model. Calculating distance, number of iterations, and different choices for the initial location of the cluster centers.
  - **DBSCAN Clustering:**
    - We measured the results for changes in the hyperparameters of this model. Calculating different values for epsilon and minimum points parameters.

Experiments

- **Evaluation Metrics:**
  - **SSE, Silhouette Score, Calinski-Harabasz Index**

DBSCAN

	n_clusters	Silhouette	Calinski-Harabasz
0	11	0.463057	587.949247
1	9	0.442910	640.402853
2	5	0.402639	453.131065
3	11	0.519986	629.993677
4	6	0.504436	490.094060
5	4	0.497020	587.239534
6	7	0.535639	569.389848
7	5	0.527806	676.603092
8	5	0.527296	683.159201

Hierarchical Clustering

	n_clusters	SSE	Silhouette	Calinski-Harabasz	linkage_methods
0	2	808.942570	0.494122	1100.374446	ward
1	6	392.219389	0.467451	485.282546	complete
2	5	417.659876	0.494371	579.700808	average
3	21	237.429446	0.466082	135.651756	single

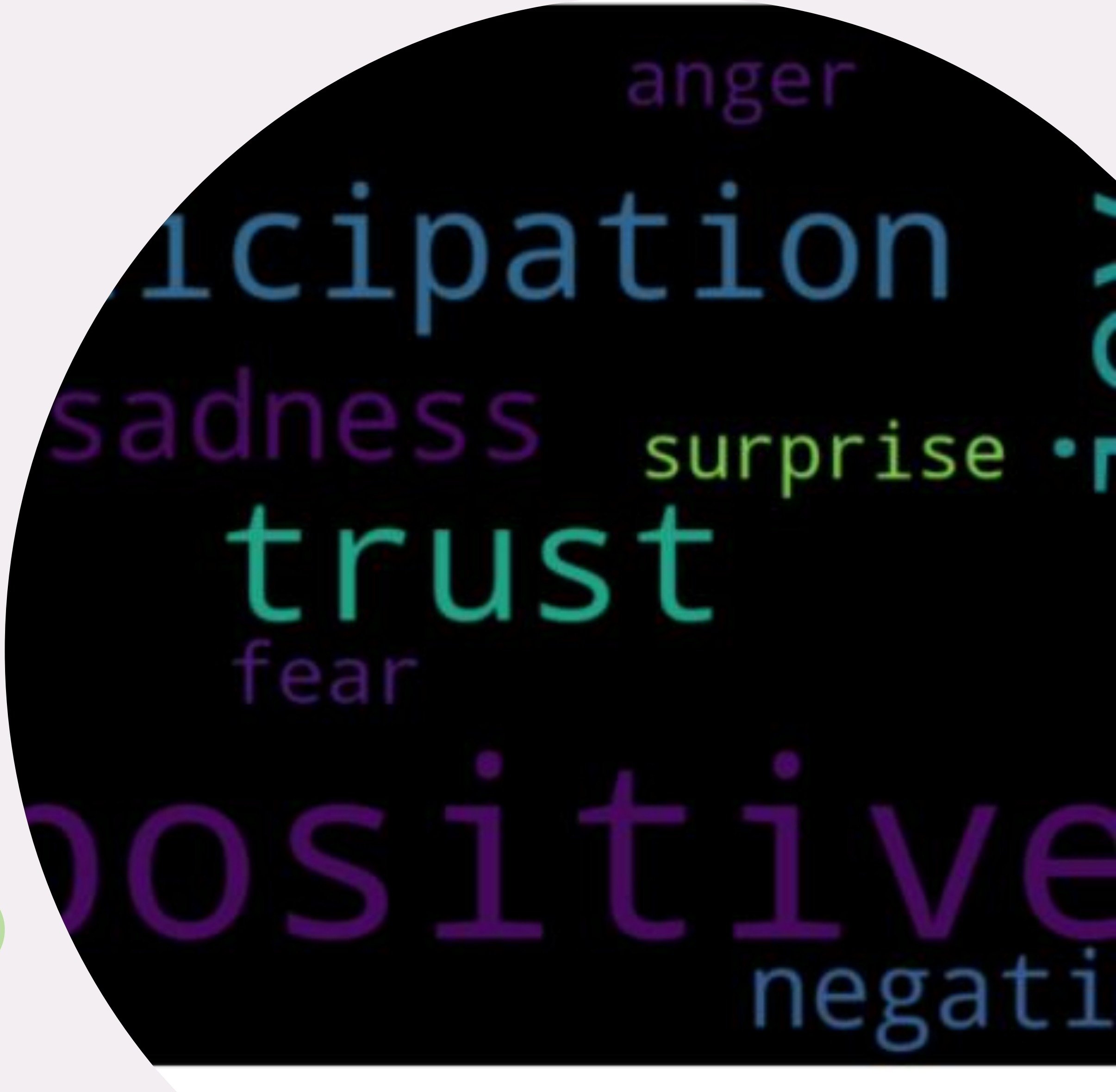
K-means

	n_clusters	SSE	Silhouette	Calinski-Harabasz
0	2	808.942570	0.515813	1214.487828
1	3	536.407148	0.520630	1223.672644
2	5	417.659876	0.317243	870.716273

frequencies_dict is: Counter({'positive': 79360, 'joy': 37622, 'trust': 32405, 'anticipation': 26019, 'sadness': 15333, 'negative': 10815, 'fear': 6537, 'anger': 5896, 'surprise': 5672, 'disgust': 1525})

Results

- Best-performing model was k-means.
- Insights from the clusters based on sentiment patterns:
- **Cluster 0: Predominantly Positive**
  - **Dominant Emotions:** 'positive', 'joy', 'trust', 'anticipation'
  - Videos in this cluster seem to convey strong positive emotions, likely expressing enthusiasm, happiness, a sense of trust, and excitement about the content. This cluster might include music videos with upbeat tunes or videos with messages of optimism and hope.



```
frequencies_dict is: Counter({'anticipation': 2976, 'sadness': 1052, 'trust': 984, 'negative': 794, 'fear': 760, 'positive': 680, 'joy': 415, 'surprise': 235,
'anger': 179, 'disgust': 125})
```

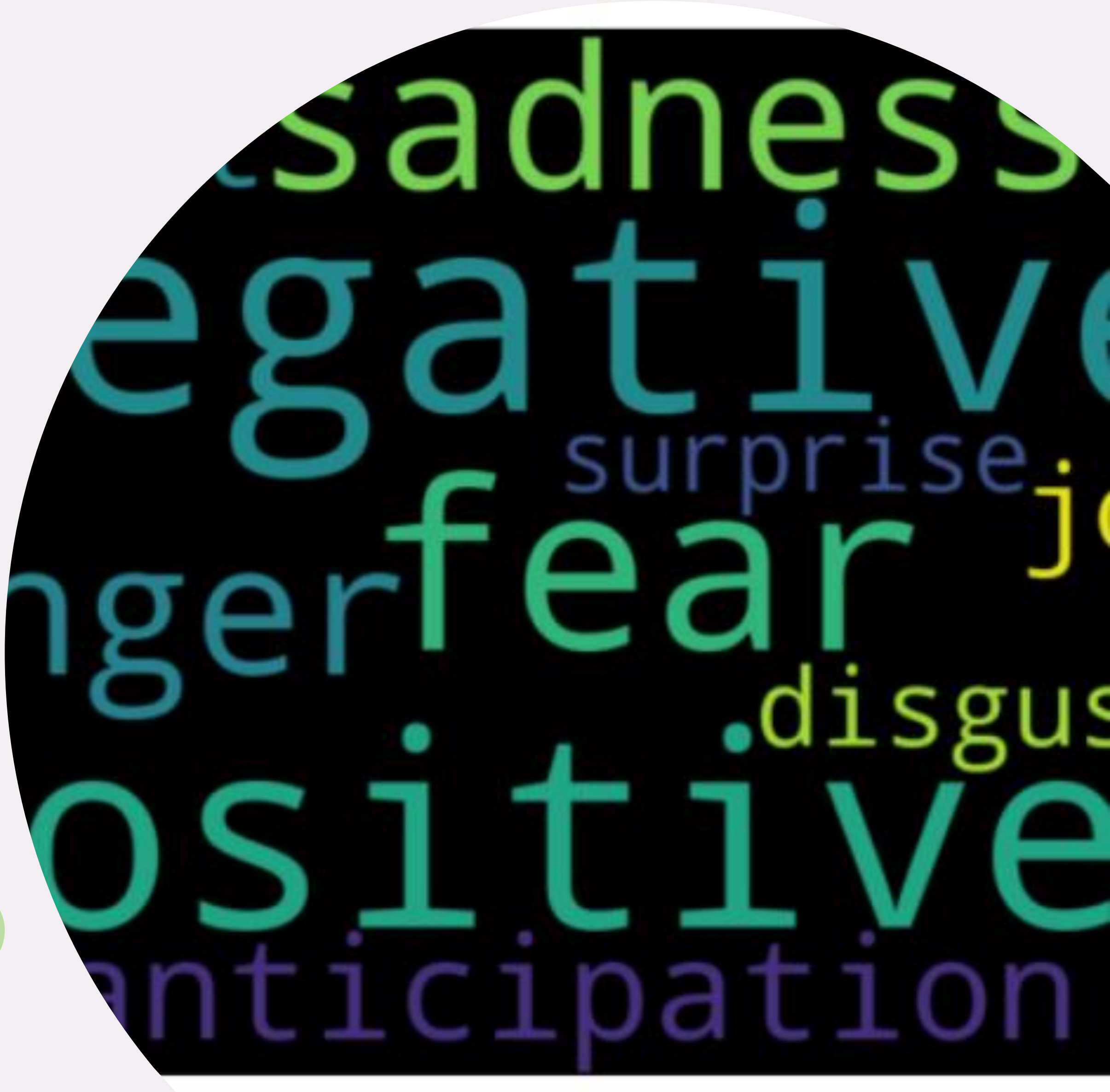
Results

- **Cluster 1: Mixed Emotions with a Lean Towards Sadness**
  - **Dominant Emotions:** 'anticipation', 'sadness', 'trust', 'negative', 'fear'
  - This cluster suggests a mix of emotions, with a noticeable presence of sadness and negativity. Some anticipation and trust might suggest underlying hopefulness, while fear adds a touch of apprehension. This cluster could include dramatic movie scenes or videos dealing with loss or difficult life experiences.

```
frequencies_dict is: Counter({'negative': 17189, 'positive': 15729, 'fear': 11873, 'sadness': 8276, 'anger': 7318, 'anticipation': 7232, 'trust': 6814, 'joy': 4924, 'disgust': 4021, 'surprise': 3472})
```

Results

- **Cluster 2: Complex Mix of Positive and Negative Emotions**
  - **Dominant Emotions:** 'negative', 'positive', 'fear', 'sadness', 'anger'
  - This cluster showcases a complex mixture of emotions. There's a clear presence of negativity, fear, sadness, and anger, but there's also a strong undercurrent of positive sentiment. Videos in this cluster might explore difficult topics, videos that spark debate or controversy, or videos depicting emotionally-charged moments.



Any Questions?