YouTube sentiment analysis Project by unsupervised machine learning

Course: **Advanced topics in machine learning**

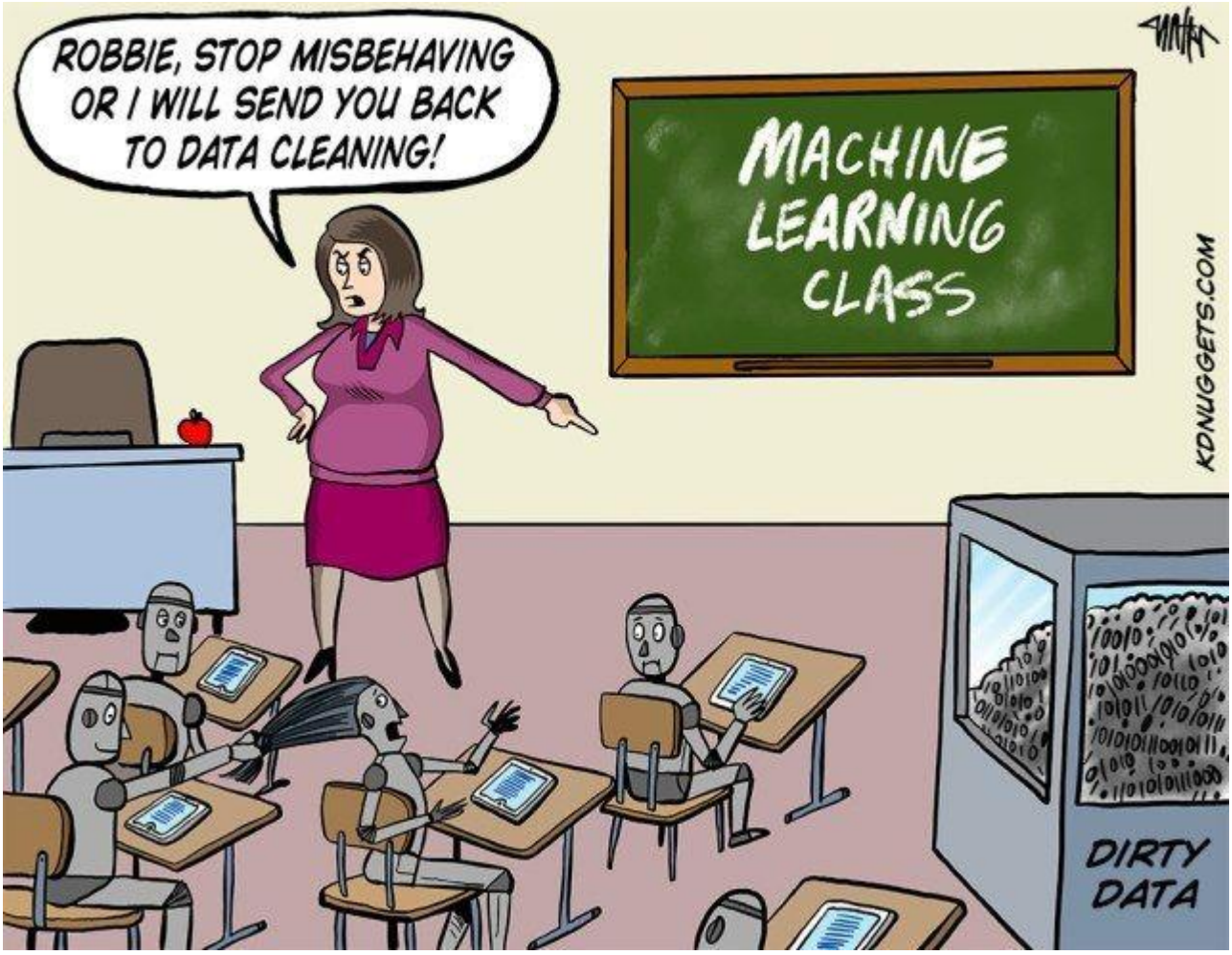
Lecturer: **Dr. Chen Hajaj**

Team Members:

Dor Ingber 316080159

Itai Bekenshtein 315338582

Table of Contents



Abstract

In today's digital world, understanding audience sentiment is essential for creators and marketers. This project aims to analyze emotions in YouTube videos related to music and movies. Our motivation is to understand audience sentiment towards music and movie content on YouTube offers valuable insights for content creators and marketers. We start from the premise that content creators choose to give a description and a name to the video they created and let them choose to describe from their point of view what the video is about. The methodology in This project employed natural language processing (NLP) techniques to extract and analyze sentiment from YouTube video titles and descriptions. The videos were then clustered based on their dominant sentiment. The project successfully identified predominant sentiments within music and movies-related YouTube content. Additionally, it created clusters that organizes videos based on their emotional tone. The Value of this work provides a tool for understanding reactions to media content, allowing content creators to refine their work, and helping marketers target their campaigns more effectively.

Introduction

In a world saturated with online content, understanding audience reactions is extremely important for content creators, marketers, and anyone seeking to gauge public opinion. YouTube, a dominant video-sharing platform, offers a vast pool of user-generated content and reflects a wide spectrum of sentiments toward various topics. This project delves specifically into the field of music and movies on YouTube, with the aim of analyzing the descriptions and titles that the content creators chose to give to the content they created and to understand from them what sentiments their content deals with. Sentiment analysis, powered by natural language processing (NLP), has the potential to reveal important trends in audience preferences. The goal of the project is to develop a system that extracts sentiment from YouTube video metadata, ultimately cluster videos according to their emotional impact. By mapping the sentiment landscape of popular music and movie content, this work can empower creators to tailor their offerings, optimize marketing strategies, and provide insights into the changing tides of audience preferences.

Dataset and Features

The API we utilized is the YouTube API. YouTube offers several APIs, each serving a different purpose. We chose the API called YouTube Data API v3. Through this API, it's possible to retrieve information about videos from the platform. Initially, we wanted to ensure that we're fetching videos related to music and movies, because we assumed that videos dealing with these areas express strong sentiments. We pulled data based on the categories. At this stage, we retrieved the video's ID, its title, and the date it was published. We noticed that if we fetch the video's description at this stage, we only get partial information. Therefore, we performed another process to obtain detailed data for each video.
In the second stage, we took data from the API using the video IDs obtained in the first stage. This process allowed us to create a single data frame containing ID, title, date, and description for each video. To maximize the amount of data without payment, we also added code to search for the "nextPageToken" in the response we received, indicating the key to the next page. We paid attention to this and proceeded to the next pages to fetch more data as long as the "nextPageToken" variable was present in the response.
Then we moved to the preliminary stage before data analysis. Firstly, we checked our data frame and noticed NULL values in the description column. We examined these videos and found out they only have titles, and they are “Shorts” videos, meaning short videos lasting no more than 60 seconds. These videos are characterized by short titles and no descriptions. We decided to create a new column that combines the video's title and description. Therefore, at this stage, we decided to continue the process as usual and try to analyze sentiments from these videos through the newly created column. We also converted the date column to datetime type.

We created a new column containing emojis for each video. Later, we also analyzed emotions from these emojis. We created a function to remove internet links, unnecessary symbols (@, |, etc.), emojis, and reduce excessive spaces to a single space. We used REGEX for this purpose. Text cleaning function is essential at this stage as it removes unnecessary information, simplifies, and improves the results we will get from NLP processing.
It is important to note that we made sure to save the data frame in a CSV file and then we read it into a variable. This is an important step at the end of the process of extracting data from an API. This allowed us to work on the data without retrieving the data from the API every time.

The data frame at this stage:

	Id	title	date	description	title&description	emoji
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...	[🎹, 😊, ✔]
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...	[✝, 🔻, 📌, 📌, 📌, 📌]
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...	[©]
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 ...	[🔴]
4	h8Cq1BwdTsg	Ozoda - Ko'k jiguli (Official Music Vide...	2024-02-21 13:39:33	Composer: OZODA\nLyrics: OZODA\nArrangement: D...	ozoda ko 39 k jiguli official music video 2024...	[▪, ▪]

Methodology

The two main processes we conducted in the project are NLP and cluster.
Since we want to understand which sentiments are present in each video based on its description and title, we aim to perform NLP. After obtaining sentiments for each video, we want to cluster videos with similar sentiments.

NLP Process:
The NLP process involved the use of libraries:
vaderSentiment - for analyzing emotions from emojis.
spacy - Stop word removal, Lemmatization, and entity removal. We gave up a tokenization step because the library through which we analyzed sentiment does it in a built-in way.
NRCLex - a library we discovered from an article dealing with sentiment analysis from text. This comprehensive library contains 27,000 words and the emotions they convey, including joy, positive, trust, anticipation, negative, sadness, fear, anger, disgust, and surprise. We preferred using this library over others as it provides more emotions.
We used **MIN MAX Scaler** to normalize sentiment scores for each video.
We chose this normalization because it does not assume a specific distribution for the data. We wanted to normalize the sentiments to highlight dominant emotions and reduce the scores of minor emotions.
At the end of the NLP process, we created a new data frame containing columns for the described sentiments and an ID for each video. We created this data frame to perform clustering based on sentiment scores.

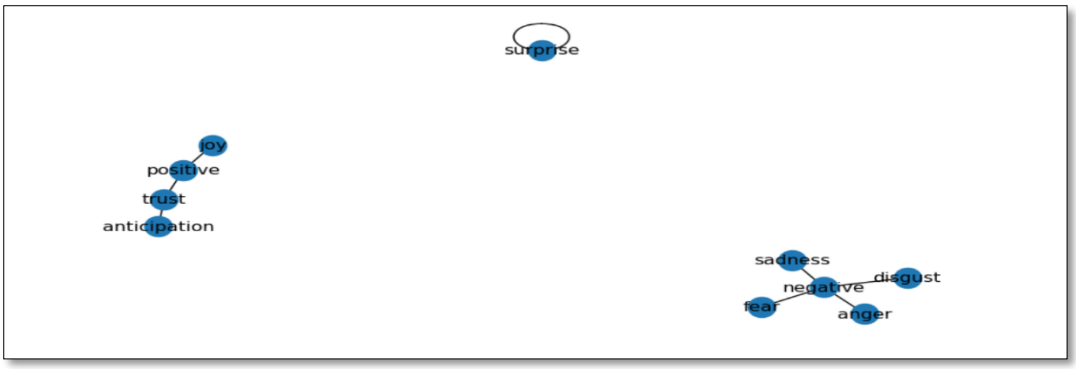
	Id	joy	positive	sadness	negative	anger	anticipation	fear	trust	disgust	surprise
0	v4KXWsMw8Fc	0.692308	1.000000	0.538462	0.461538	0.115385	0.153846	0.115385	0.192308	0.000000	0.000000
1	NgGJaXDC0wU	0.937500	1.000000	0.187500	0.000000	0.000000	0.437500	0.187500	0.812500	0.000000	0.000000
2	mLW35YMzELE	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

Cluster Process:
We explored three models we learned in the course:
Hierarchical clustering
k-means

DBSCAN

The first model we explored was **Hierarchical clustering**. We started with this model because it can provide an optional number of clusters, which is essential for the next model we explored. We evaluated the hierarchical model based on different linkage methods and assessed it using SSE, Silhouette, and Calinski-Harabasz metrics. We defined a threshold of 70% and drew a vertical line around it to determine the number of clusters. We also visualized the results on graphs and saved them to a table to identify the best-performing model. According to the hierarchical model, division by ward and average linkage methods were most effective.

The second model we explored was **K-means**. This model is sensitive to its hyperparameters. Therefore, before running the model, we wanted to find relevant values for K, which signifies the number of clusters we want to divide our data into. We created a heatmap with correlation to check closeness between emotions. Then we extracted 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 settled for a fixed division into 3 clusters.



The best option for the number of clusters was 3 based on correlation and the hierarchical model as well. After examining other hyperparameters such as distance type, initial cluster centroids, and the number of iterations, there were no significant differences between the results. The difference in results arose from the final number of clusters. To evaluate the model's performance, we used SSE, Silhouette, and Calinski-Harabasz metrics and created a graph to identify the inflection point (elbow point). The best result was clearly for 3 clusters according to the k-means model, which was also preferable over the results of the hierarchical model.

The third model we explored was **DBSCAN**. In this model, we also considered its hyperparameters: epsilon and Minimum Points. We ran 9 models of this type for different values of hyperparameters. We evaluated this model based on Silhouette and Calinski-Harabasz metrics. The best result we obtained was for epsilon=0.75, Minimum Points = 15, resulting in 5 clusters. Still, its result was less optimal than the k-means model.
After comparing the different models and finding the optimal models, we ran it and saved its results to CSV files, which we explored to understand the focus of each cluster. In the investigation process, we utilized word clouds and printed the sum of scores for each emotion in the cluster after we scaled them.

Experiments/Results/Discussion

Parameter Choices:

Hierarchical clustering:
Linkage methods: We explored different linkage methods (ward, average, complete, single) to assess their impact on clustering performance. The choice of linkage method influences the shape and structure of the dendrogram, affecting the final clustering result.

Threshold: We defined a threshold to determine the number of clusters in hierarchical clustering. Adjusting the threshold allows us to control the granularity of clustering. We chose 70% granularity.

K-means:
Number of clusters (K): We used various values of K to identify the optimal number of clusters that best represent the underlying structure of the data. The choice of K significantly impacts the compactness and separation of clusters.
Distance type: We experimented with different distance metrics (Euclidean, Manhattan) to measure the dissimilarity between data points. The choice of distance metric affects the clustering result by determining how similarities are calculated.
Init methods: We checked both 'random', 'k-means++' to check the starting point for each cluster center. We didn't try to find it manually because we didn't have any information about which value it should be.

DBSCAN:
Epsilon: We varied the epsilon parameter to control the radius of the neighborhood for density estimation. Adjusting epsilon influences the density of clusters and the number of points considered as core points.
Minimum Points: We experimented with different values of minimum points to specify the minimum number of data points required to form a dense region. This parameter determines the sensitivity of the algorithm to noise and affects the size and shape of clusters.

Evaluation Metrics:

Silhouette Score: We used the silhouette score to evaluate the cohesion and separation of clusters. A higher silhouette score indicates better-defined clusters with instances close to their own cluster and far from other clusters.
Calinski-Harabasz Index: This index measures the ratio of between-cluster dispersion to within-cluster dispersion. A higher Calinski-Harabasz index suggests better cluster separation.
SSE (Sum of Squared Errors): SSE measures the sum of squared distances between each data point and its assigned cluster centroid. Lower SSE indicates tighter clusters.
Visual Inspection: Besides quantitative metrics, we visually inspected dendrograms, scatter plots, and cluster assignments to understand the structure and quality of clusters.

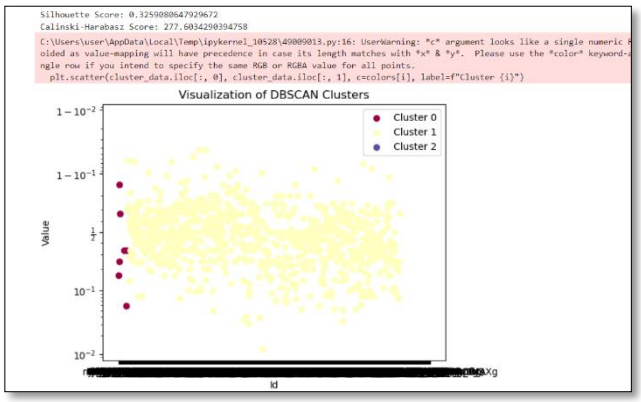
Quantitative and Qualitative Results:

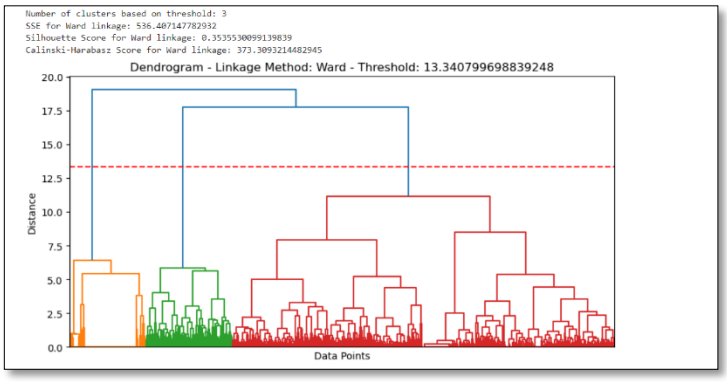
Quantitative Results: We computed the evaluation metrics (Silhouette Score, Calinski-Harabasz Index, SSE) for each clustering algorithm and parameter setting. These metrics provided numerical insights into the performance of different algorithms and parameter choices.

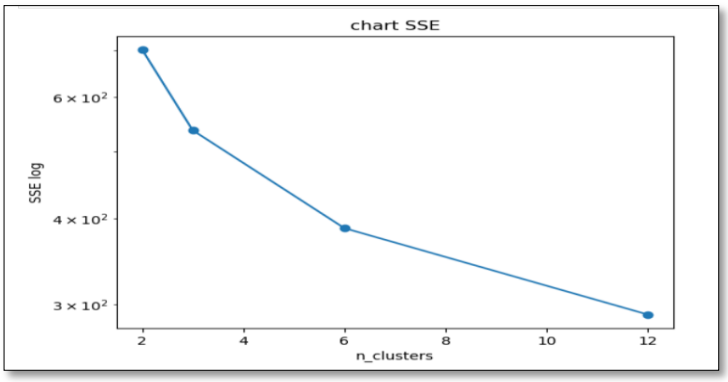
Qualitative Results: We visually inspected the cluster assignments, dendrograms, K-means elbow, and scatter plots to gain qualitative insights into the clustering results. This allowed us to interpret the structure of clusters and understand any patterns. We also used a word cloud to clearly show the dominant emotions in each cluster.

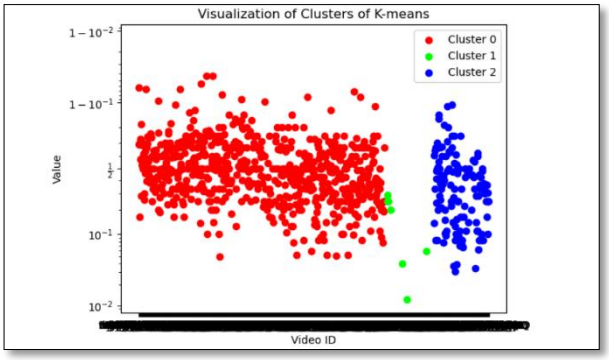
[1562]:

	n_clusters	SSE	Silhouette	Calinski-Harabasz
0	2	702.157475	0.298785	326.203608
1	3	536.416216	0.361904	401.025881
2	6	386.900439	0.271638	315.525302
3	12	290.002838	0.249905	227.026364









Algorithm Performance:

k-means: k-means performed well when the clusters were spherical and evenly distributed. The choice of K significantly influenced its performance. It gave the best result, and we chose it to cluster the data.

Hierarchical Clustering: Hierarchical clustering performed well in capturing hierarchical relationships and identifying natural clusters in the data. We also used the natural clusters size in k-means. It gave similar result as the best model, so we used it to cluster the data as well.

DBSCAN: DBSCAN performed well in identifying clusters of arbitrary shapes and handling noise in the data. It was particularly effective when the epsilon and the minimum points had high value.

This model gives us the lowest results, so we didn't used it.

Conclusion and Future Work

Conclusion

Cluster 0: Mixed Emotions
Dominant emotions: Negative, positive, fear, sadness, anger.
Description: This category may represent YouTube video content that is controversial or touches on disturbing, sensitive, and difficult topics.
Possible genres: Discussions, harsh reviews, suspense movies, horror, depressing clips from dramas, or coverage of traumatic events.
Cluster 1: Very Positive
Dominant emotions: Positive, joy, trust, anticipation.
Description: This cluster represents optimistic, joyful, light-hearted, and supportive content.
Possible genres: Happy music videos, good reviews, humorous clips, inspiring content, movies with a good ending and positive messages.
Cluster 2: Mixed Emotions, Leaning Towards Sadness
Dominant emotions: Anticipation, sadness, trust, negativity, fear.
Description: This cluster reflects content that may be emotional and melancholic. The anticipation and trust indicate an element of hope, but the sadness and negative emotions indicate challenging topics.
Possible genres: Dramas, tragedies, films about dealing with challenges, mixed reviews with negative and positive criticism, documentation of sad events.

Additional Findings from the Project:
Emotional complexity: YouTube videos demonstrate emotional complexity – few videos evoke a single distinct emotion, and many content touches on multiple emotions, sometimes even contradictory.
Importance of context: Sentiment analysis is data-driven, but human interpretation is important to understand the results, considering the genre and purpose of the specific video.

Future Work

The project's contribution lies in creating clusters of videos with similar sentiments, which can assist content creators in understanding the sentiments conveyed in their video titles and descriptions. These clusters can also benefit users on the YouTube platform if a recommendation system is built based on the identified clusters. Furthermore, there is potential to further investigate the topic on a larger dataset to create sentiment-based clusters for other types of videos, not just music and movies.
Throughout the project, we observed that there are "Shorts" videos not tagged by category. We believe there are more types of videos that are not classified into categories, and in such cases, assigning videos to clusters based on the sentiments they convey can expose users to similar videos they may be interested in watching.
We learned from the project that emotions are a complex thing, and a specific video can convey several sentiments. From work we learned that there are sentiments with a strong affinity. Our

findings and the model we built can help people involved in sentiment analysis in social networks. This model, with minor changes, can also be adapted to other social platforms.

Contributions

The two team members worked together on the project and each of us is knowledgeable about the processes we performed. If so, we put emphasis on the division of labor between us. Dor was responsible for getting data from the API and the Cluster process. Itai was responsible for the NLP process and the Clusters investigation.

Appendices

https://www.saifmohammad.com/WebDocs/NRCemotionlexicon.pdf
Article about the NCRLEX sentiment.