🎬 Emoji Movie Review & Analysis 🍿

Team

1. **Max Moshe Altman**, moshe.altman@mail.huji.ac.il, 337620959, maxaltman
2. **Noa Greenfield**, noa.greenfield@mail.huji.ac.il, 211601786, Noa.greenfield
3. **Michael Gur**, michael.gur@mail.huji.ac.il, 207555178, michael.gur

Problem Description

People increasingly rely on emojis to convey reactions and emotions when discussing movies and TV shows online. This project set out to uncover how these symbols relate to movie ratings and to each other. We built and cleaned a large dataset of emoji-containing reviews and tweets, then explored three key questions:

1. **Predictive Power** - Can a movie's rating be estimated from the emojis in a review?
2. **Semantic Meaning** - How do individual emojis align with the emotional tone of accompanying text?
3. **Emoji Relationships** - Which emojis frequently appear together, and what thematic clusters do they form?

Through data cleaning, exploratory analysis, and a custom recommendation-style model, we showed that emojis indeed encode sentiment strong enough to predict ratings and reveal broader cultural patterns of use.

Data

We found a dataset (on HuggingFace) that combines posts about movies scraped from Reddit, with data from that movie's listing on Amazon.

- **Size:** the dataset contains 5.3M reviews (1.65GB).
- **Fields:** for every review, the dataset includes
    - The post's title and text content
    - The discussed movie name, as appears in Amazon
    - The movie's rating on a 1-5 scale
    - Other structured features of the movie, including duration, genre, cast, etc.
- **Cleaning:** we applied the following steps:
    - Extracted emoji counts per emoji and review
    - Removed reviews without emojis
    - Manually merged similar genres (e.g. "Adventure" and "Adventures")
    - Deduplicated

  Which left us with 15K reviews.

In addition to the first dataset, in some parts of our project we implemented another dataset (from Kaggle) of tweets containing emojis.

- **Size:** the dataset contains 43 CSV files (102MB) each corrosponding to a different emoji. Each CSV file contains 20,000 tweets that have used the corrosponding emoji.
- **Fields:** for each emoji, there are 20K rows of text. Each row is a separate tweet

Our Solution

Throughout our project we implemented a series of different methods and algorithms learned in class including the PageRank algorithm, cosine distance, etc. In the evaluation section we'll dive deeper into the specific methods used for each of our visualizations.

Evaluation

First, we made sure that there is some observable correlation between certain emojis and movie scores. We grouped reviews by the emojis that appear in them, and took the average ratings.

The dataset presented some challenges. First, the dataset is very sparse - more than 800 unique emojis appear in the reviews, and most reviews contain only 1-2 emojis. For each *n*, we computed the percentage of reviews (out of those with at least one emoji) that had more than *n* emojis. Almost 60% of reviews only contain 1 emoji.

Second, as appears in the first chart, the ratings are skewed positively, with 85% reviews with 4-5 stars, and only 12% reviews with 1-2 stars.

Setup

Our goal is to analyze the relationships between emojis and movie ratings. The skewness of the data meant models like decision trees are prone to overfit and just predict 5-star ratings, without capturing any real insight. The sparsity of the data meant distance measures such as cosine similarity are likely to be inefficient.

That's why we decided to adapt the recommendation system setup, as follows: each emoji is treated as a "user", and its profile is a 5-element array that represents the learned distribution of associated ratings. Then, given a list of emojis, we add up the distributions and return a weighted average as the predicted rating.

Results

To evaluate our system, we used the RMSE metric. We split the dataset into training and evaluation sets, and compared two approaches:

- A "best guess" model that consistently predicts the global average of the training set.
- Our system, trained on the training dataset.

The results of computing the RMSE of the predictions against the true ratings are as follows:

Model	Our System	Baseline "Best Guess"
RMSE	0.8537	1.2268

These confirm that our system is effective (40% better compared to the best guess), and that the emojis contain a signal related to the rating.

Impediments

To understand which emojis were most challenging for the model, we calculated the total squared error contribution for each unique emoji. Below are the top 20

Observing this list, we see emojis such as "😂" and "😭", which we guessed can signify different sentiments based on the context. For example, a "😂" emoji might signify genuine enjoyment

(and a positive rating) of a comedy film, and a sarcastic laughter at a horror film. Our current system predicts a fixed distribution per emoji by design, and can potentially be improved by taking the movie's genre into account.
To test this, we repeated the initial computation of each emojis average rating, but this time we also averaged combinations of emoji × genre. We then plotted the emojis that had the largest difference from a specific genre to their overall average.
Below are plots showing how much an emoji's average rating changed for the "Drama" and "Thriller" genres, compared to that emoji's overall average rating.
Drama movies - we can see emojis like "😭" becoming more positive, possibly indicating genuine reactions to sad movies.
Thriller movies - we can observe generally positive emojis like "🤣" become more negative, possibly indicating sarcasm.

Re-evaluation

Due to the sparsity of the movie review dataset, we decided to pivot to a more in-depth analysis of general emoji use. We mostly used our dataset of tweets containing emojis for this part. These are some visualizations and observations from our analysis:
First, here are two different approaches at creating a co-occurrence network of emojis:

1. Results

The analyses revealed clear clusters of co-occurring emojis, each reflecting semantic or sentiment-based themes. Strongly positive symbols such as 😃, 👍, ❤️, and 😍 emerged as highly central, often appearing together and dominating support-based metrics. Negative or sarcastic emojis (😡, 😭) formed their own smaller clusters, appearing less frequently but still structurally significant. Humor-driven emojis like 😂 and 🤣 acted as global connectors across communities.
Both Apriori itemset mining and Louvain-based network analysis confirmed the presence of meaningful groups:

- Affection clusters (❤️, 😍, 🥰, 😊)
- Celebration/joy (🎉, 🎂, 🥳)
- Humor/laughter (😂, 🤣, 😅)
- Sadness/support (😭, 💔, 🤗)
- Context-specific associations such as 🍿&📽️ in movie reviews

The Louvain method identified 7 major communities with modularity Q ≈ 0.42, indicating moderately strong division. Apriori rules further revealed that some pairs occur far more frequently than chance would predict, highlighting sentiment-driven and context-specific emoji combinations.

2. Visualization choice

For the first graph, we used a co-occurrence network to convey multiple association metrics at once:

- Node size encodes single-emoji support (prevalence across reviews).
- Edge thickness encodes joint support (how often two emojis appear together).
- Edge darkness encodes lift (association strength beyond chance).
- Edge opacity encodes confidence (directional predictiveness).

In the second, nodes were colored by community and displayed with fixed size for clarity. Edge thickness corresponded to co-occurrence frequency, and alpha blending was applied to highlight dense versus sparse links.

3. Issues encountered

Both approaches faced sparsity and imbalance: most emojis occurred infrequently, creating dense graphs dominated by weak edges. To address this, thresholds were applied (minimum support for A-priori; degree and co-occurrence cutoffs for Louvain). Visual clutter was another issue; fixed node sizing, edge weighting, and top-N filtering was required to maintain readability and interpretability.

4. Method and parameters

- **A-priori approach**: Applied with minimum support = 0.001, maximum length = 4. Association rules generated with confidence threshold = 0, then filtered and ranked by lift and confidence. Restricted visualization to top 30 emojis by support for interpretability.
- **Network analysis**: Constructed weighted undirected graphs where nodes are emojis and edges represent co-occurrence frequency. Applied PageRank, HITS, and Louvain community detection. Edge weights normalized with scaling exponent 0.35 and alpha 0.6. Restricted to top 250 emojis by degree, with minimum co-occurrence threshold = 80.

Together, these methods balance statistical association discovery with structural network insights, producing complementary perspectives on how emojis cluster and interact across contexts.

1. Results

The co-occurrence heatmap shows which emojis are most often found together in the same tweets. Strong blocks emerge, for example between ❤️, 😍, 🥰, and 😊, reflecting their shared role in expressing affection or positivity. Another cluster is built around 😂 and 🤣, which may be more humorous tweets. But the strongest connections come from the “easter tweets”, aka 🐣, ⚪, 🐰, 🐰, 🐇. Overall, the results confirm that emoji use is not always random, clear sentiment clusters appear when tweets are aggregated.

2. Visualization choice

A heatmap was selected because it compactly represents pairwise co-occurrence frequencies. Each axis lists emojis, and color intensity encodes how frequently the pair co-appears in tweets. This allows quick detection of strong relationships while still showing weaker ties.

3. Issues encountered

Some emojis appear in vastly more tweets than others, so raw co-occurrence counts were dominated by high-frequency emojis. This was handled by normalizing values for the colormap while still displaying raw counts for transparency.

4. Method, algorithm, scaling

Each CSV contained tweets with a single emoji type specified in the filename. Tweets were scanned, and whenever an additional emoji appeared in the text, a counter for that pair was incremented. This produced a symmetric co-occurrence structure across emojis. Then the top

20 emojis were selected to avoid sparsity and overcrowding. The final co-occurrence matrix was plotted with normalized shading using seaborn's heatmap feature.

https://emoji-data.streamlit.app/

A fun interactive web app that receives text from the user and suggests the most appropriate emoji based on training using the emoji tweets dataset, sentiment analysis, and cosine similarity. Every CSV file in the dataset belongs to one emoji. For each row of text in each file, the code uses NRCLex to score ten basic emotions (anger, joy, trust, etc.). It then averages those scores across the rows of a file. The result is a compact emotional profile for that emoji (all profiles are saved so the model doesn't need to relearn every time). When you type text, the same ten emotion scores are computed for your input. The code then compares your input's profile to each emoji's profile and picks the closest match measured with cosine similarity.

Future Work

There are several exciting directions to extend this project:

- **Context-Aware Modeling** - Incorporate the movie's genre, review text sentiment, or release year into the rating-prediction model so that an emoji like 😂 can adapt its meaning based on context (e.g., comedy vs. horror).
- **Temporal Trends** - Analyze how emoji usage and sentiment change over time, such as the rise of new emojis or shifts in meaning.
- **Interactive Visualization** - Build a web dashboard where users can explore co-occurrence networks and sentiment clusters dynamically, filtering by genre, year, or emoji category.
- **Cross-Language Analysis** - Investigate whether emojis convey similar sentiments across different languages and cultural contexts.

Conclusion

This project demonstrates that emojis carry measurable signals about audience sentiment toward movies. By treating each emoji as a miniature "user profile" and modeling its rating distribution, we achieved significantly lower prediction error than a simple baseline. Our analysis of co-occurrence networks further revealed that emojis cluster naturally into meaningful groups—positive, humorous, or sad—which reinforces their role as a universal visual language. Together, these findings highlight both the expressive power of emojis and their practical value for understanding and predicting audience opinions.