**Sentiment Analysis Report**

**Youtube Comments**

**Introduction**

The goal of this project is to analyze the sentiments expressed in YouTube comments using Natural Language Processing (NLP) techniques. Sentiment analysis, a subfield of NLP, involves identifying and categorizing opinions expressed in textual data as positive, negative, or neutral. This project leverages R for data preprocessing, sentiment scoring, and visualization.

The dataset comprises comments extracted from the YouTube video ["I Want It That Way | Brooklyn Nine-Nine"](https://www.youtube.com/watch?v=HlBYdiXdUa8). This video features a memorable scene from the popular television show *Brooklyn Nine-Nine*, which has garnered significant engagement and feedback from viewers. These comments serve as the foundation for exploring patterns and sentiments.

Several techniques, including Term-Document Matrices (TDM), Document-Term Matrices (DTM), TF-IDF scoring, Latent Semantic Analysis (LSA), and Latent Dirichlet Allocation (LDA), are utilized to uncover insights into user opinions. Visualizations such as word clouds and sentiment distributions are used to illustrate the findings.

This report presents the methods employed, the visualizations generated, and the insights derived from analyzing YouTube comments. The analysis aims to highlight trends in viewer feedback, identify recurring themes, and uncover variations in sentiment across the dataset.

A graph with a blue line

Description automatically generated

The Net Sentiment Distribution plot shows a peak at 0, meaning a lot of comments are neutral, probably because people are just talking about facts from the video without much emotion. The plot is positively skewed, suggesting that most viewers have good opinions, likely due to the quality or appeal of the content. There are fewer negative scores, which might mean that not many viewers found major faults to mention.

A graph with a green line

Description automatically generated

Sentiment Score Distribution by Category shows that neutral comments peak at 0, reflecting factual or balanced discussions without strong feelings. Positive comments group around higher scores, showing that viewers were generally satisfied and enjoyed the content. Negative comments are fewer and spread across lower scores, indicating that fewer viewers had strong negative reactions.

* Positive Term-Document Matrix (TDM)

A screenshot of a computer code

Description automatically generated

The Positive Term-Document Matrix (TDM) shows how often each term appears in the first three documents with positive sentiment. Words like "backstreet" and "boys" show up only in the first document, pointing to specific mentions. In the second document, terms like "buena" and "original" appear, showing different content focuses in the comments.

* Negative Term-Document Matrix (TDM)

A screenshot of a computer code

Description automatically generated

The Negative Term-Document Matrix (TDM) lists how often terms appear in the first three documents with negative sentiment. Words like "going" and "one" appear in the first document, reflecting common negative phrases. Terms like "prison" and "lonely" suggest specific negative topics being discussed.

* Neutral Term-Document Matrix (TDM)

A screenshot of a computer code

Description automatically generated

The Neutral Term-Document Matrix (TDM) shows term frequencies in the first three neutral sentiment documents. Words like "cena" and "brilhante" are frequent, indicating discussions about specific scenes or elements. The variety of terms across documents shows a range of neutral topics without strong sentiment.

* All Term-Document Matrix (TDM)

A screenshot of a computer code

Description automatically generated

The All Term-Document Matrix (TDM) lists term frequencies in the first three documents across the whole set. Like the neutral TDM, terms like "cena" and "brilhante" are common, pointing to recurring themes in neutral comments. This spread shows the overall discussion topics in all comments.

* Positive Document-Term Matrix (DTM)

A screenshot of a computer code

Description automatically generated

In the Positive Document-Term Matrix (DTM), words like "backstreet," "boys," and "guys" only show up in the first document, indicating specific references. Other documents don't have these terms, showing a varied focus in positive comments and reflecting the diversity of topics discussed positively.

* Negative Document-Term Matrix (DTM)

A screenshot of a computer program

Description automatically generated

In the Negative Document-Term Matrix (DTM), terms like "going," "one," and "prison" appear in the first document, suggesting specific negative contexts. A few other documents also mention "one" and "prison," indicating recurring themes in negative comments. The sparse distribution shows varied negative experiences.

* Neutral Document-Term Matrix (DTM)

A screenshot of a computer code

Description automatically generated

In the Neutral Document-Term Matrix (DTM), words like "brilhante," "causa," and "cena" mostly appear in the first document, indicating specific neutral discussions. These terms aren't in other documents, showing varied neutral content, which means neutral comments cover diverse topics.

* All Document-Term Matrix (DTM)

A screenshot of a computer code

Description automatically generated

In the All Document-Term Matrix (DTM), terms like "brilhante," "causa," and "cena" also show up in the first document, similar to the neutral DTM. This distribution matches the neutral comments as they are part of the overall dataset, indicating that overall comments share common themes with neutral sentiments.

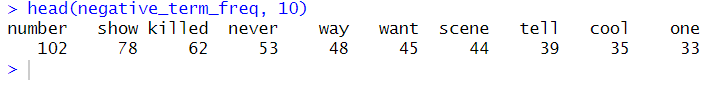
* Positive Term Frequency

A black and white text

Description automatically generated

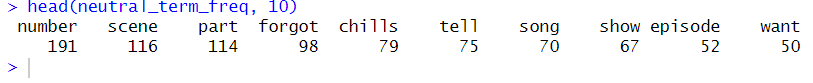
In Positive Term Frequency, words like "number," "scene," "song," "like," and "best" show up a lot, indicating positive discussions about specific parts of the video. The frequent use of "love" and "best" reflects strong positive reactions, highlighting what viewers enjoyed the most.

* Negative Term Frequency



In Negative Term Frequency, common terms like "number," "show," and "killed" suggest critical discussions about specific elements. The term "killed" might relate to a specific line or scene in the video. Frequent negative terms like "never," "want," and "tell" indicate dissatisfaction or strong opinions.

* Neural Term Frequency



In Neutral Term Frequency, frequent terms like "number," "scene," and "part" indicate factual or descriptive comments about the video. Words like "forgot," "chills," and "tell" suggest neutral recounting of experiences or observations, showing viewers' engagement without strong positive or negative emotions.

* All Term Frequency

A black text on a white background

Description automatically generated

Overall Term Frequency shows that high frequencies of "number," "scene," and "song" indicate these are central topics across all sentiments. Words like "show," "part," and "best" are also common, reflecting what viewers discussed frequently, showing what aspects of the video were most engaging to the audience.

* Top 10 Terms in Positive Comments

A graph of blue bars

Description automatically generated

The Top 10 Terms in Positive Comments include words like "number," "scene," "song," and "like," showing viewers' positive reactions to specific parts. The frequent use of "best" and "love" reflects strong positive feelings towards the content, highlighting what aspects viewers appreciated the most.

* Top 10 Terms in Negative Comments

A graph of a number of blue bars

Description automatically generated with medium confidence

For the Top 10 Terms in Negative Comments, frequent terms like "number," "show," and "killed" indicate critical discussions about specific parts of the video. Words like "never," "way," and "want" suggest dissatisfaction or negative reactions, highlighting the aspects that viewers found problematic.

* Top 10 Terms in Neutral Comments

A graph of blue rectangular bars

Description automatically generated with medium confidence

In the Top 10 Terms in Neutral Comments, common terms like "number," "scene," and "part" show that neutral comments often discuss specific video elements factually. The term "forgot" suggests a neutral recounting of events or experiences. Frequent use of "chills," "tell," and "song" indicates engagement without strong positive or negative sentiment.

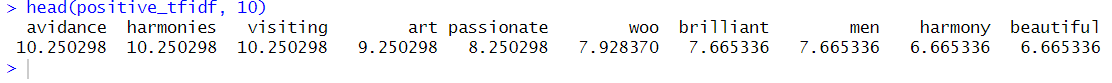
* Top 10 Terms in All Comments

A graph of blue bars

Description automatically generated

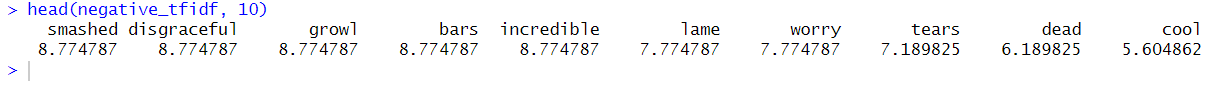
In the Top 10 Terms in All Comments, "number," "scene," and "song" are common, indicating these are central topics in the discussion. The frequent use of "show" and "part" reflects overall engagement with specific parts of the video. Words like "like" and "best" suggest that the general sentiment is mostly positive.

* Positive TF-IDF



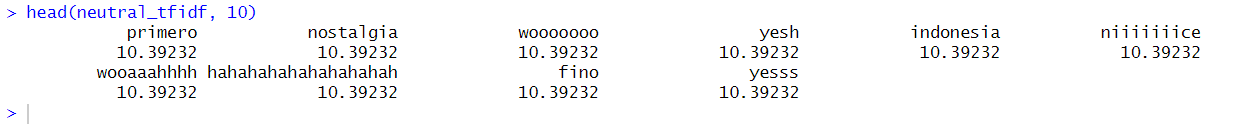
In Positive TF-IDF, terms like "avidance," "harmonies," and "visiting" have high scores, showing they are unique and important in positive comments. High scores for "art," "passionate," and "brilliant" reflect strong positive reactions to specific elements, highlighting what viewers found uniquely positive about the video.

* Negative TF-IDF



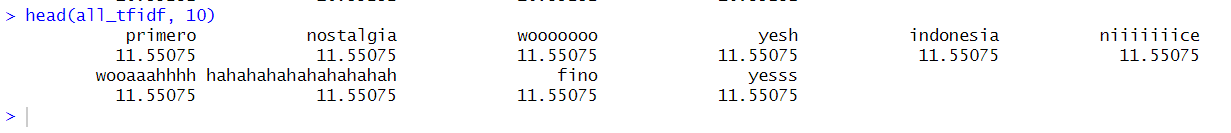
In Negative TF-IDF, terms like "smashed," "disgraceful," and "growl" have high scores, indicating specific negative reactions. High scores for "lame," "worry," and "dead" suggest critical sentiments toward certain aspects, showing viewers' strong negative responses to particular elements of the video.

* Neutral TF-IDF



In Neutral TF-IDF, terms like "primero," "nostalgia," and "wooooo" have high scores, indicating unique, factual, or descriptive comments. High scores for "indonesia" and "niiice" suggest neutral discussions about geographical or subjective observations, reflecting the diverse, non-emotional nature of neutral comments.

* All TF-IDF



In All TF-IDF, terms like "primero," "nostalgia," and "wooooo" appear frequently with high scores across all sentiments. High scores for "yesh" and "indonesia" indicate common, distinctive terms in the entire comment set, highlighting the general themes and unique elements discussed in the video.

Latent Semantic Analysis

**A blue background with black text

Description automatically generated**

**Clustering of Terms**:

* **say, tell, want, way**: These terms have higher values in the second dimension, suggesting they are closely related in the context of the documents.
* **love, song, sing, voice**: These terms have similar coordinates in both dimensions, indicating they might represent a common theme, perhaps related to music or emotional expression.
* **episode, show, series**: These terms cluster together, likely representing content related to TV shows or serialized media.

**Trying lsa with dim = 5**

A blue rectangle with black text

Description automatically generated

For the 2-dimensional case, the singular values are: [12.01326085, 8.81255803]

*  For the 5-dimensional case, there is a warning about zero singular values, which means some dimensions have zero variance.
* Conclusion : dims = 2 is better

A blue and black text

Description automatically generated

* The rows represent individual terms or words in the vocabulary, while the columns represent each document in the corpus.
* For example, if you look at the row labeled love, the values in each column represent the frequency or weight of the term "love" in each document.

A blue and black text

Description automatically generated

Each row corresponds to a term or feature, while each column represents a different dimension or component of the term-space. These column indicate the strength or importance of each term/feature in the given dimensions of the term-space. For example, term 'love.' In the first dimension (column 1), it has a weight of a -0.0201, and in (column 2), it has a weight of approximately -0.0631. meaning that it was used more in column 1 .

A blue rectangle with black text

Description automatically generated

We are trying this with dims 5 so we can compare how documents are related .

Latent Dirichlet Allocation (LDA)

In natural language processing, Latent Dirichlet Allocation (LDA) is a popular topic modelling approach that reveals latent theme patterns in massive text corpus. It functions according to the idea that documents are made up of several subjects, each of which is made up of a group of words with different probabilities. The first step of the LDA method is to randomly allocate each word in a text to one of the k predefined themes. The word's likelihood inside the subject and the topic's percentage throughout the text are the two essential probabilities for each word that are then repeatedly refined by the system. Finding the top terms for each subject, looking at the topic distribution inside documents, and figuring out the document distribution across topics are all part of analysing the output of an LDA model. This aids in comprehending the prevailing themes and how frequently they occur across the corpus. More efficient exploration and interpretation of the subjects may be achieved with the use of visualisation tools like word clouds and interactive plots, such those offered by pyLDAvis (Kulshrestha, 2019).

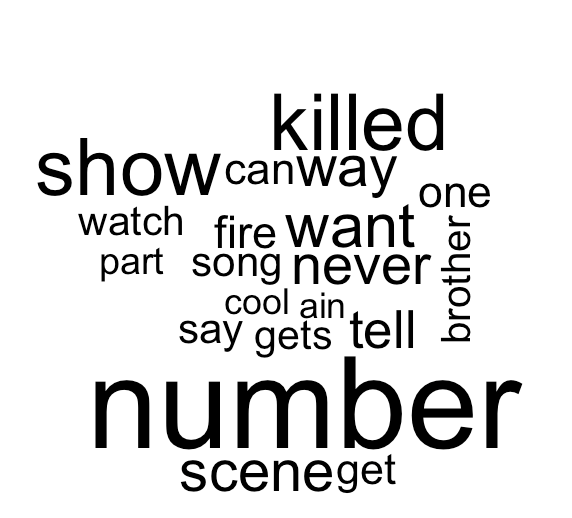
Word Cloud

A word cloud is a type of data visualisation that shows the terms that appear most frequently in a sizable body of text. Words are shown graphically as graphical components, and the size of each word indicates how frequently it appears in the text. Word clouds are useful for rapidly locating important subjects, themes, or patterns in text data because the most important words are displayed in the visualisation as bigger and more conspicuous terms. With its simple and clear depiction of textual data, this visualisation tool is helpful for understanding the organisation and content of sizable text collections. In text analysis, content summarising, and topic modelling, word clouds are frequently utilised to present a visual overview of the most significant phrases in the data (Kibe, 2024).

A close-up of a number

Description automatically generated A close-up of words

Description automatically generated

Figure 1 Word Cloud for Entire Corpus Figure 2 Word Cloud for positive sentiment A close-up of words

Description automatically generated

Figure 3 Word Cloud for Negative sentiment Figure 4 Word Cloud for Neutral sentiment

In these figures we see that the most frequent used term in the entire corpus and all sentiments is “number” as it is having largest font size and a bold font. In the word cloud of the entire corpus the word “scene” is also a frequent term as well in other sentiment but more frequent in the entire corpus and positive sentiment. In the positive sentiment, words such as “best”, “like” and “love” are used frequently which indicates these terms are used for positive comments. There are many words such as “song” which used across many sentiments but that frequently used.

LDA for each sentiment and for the entire corpus

A group of black text

Description automatically generated

In this topic 1 we can say that the words “love” and “best” indicate the positive reaction of the viewer, topic 2 it is talking about a “part” or “section” of the video, topic 3 is generally about the song in the video “I want it that way” and topic 4 talk about the tv show in general with words “show”, “watching” and “time”.

A screenshot of a computer

Description automatically generated

The topic 1 can indicate the positive reaction to the song, topic 2 a strong positive reaction to the scene, topic 3 could be about the impact the song has on the viewers with words like “can” and “ever” and topic 4 is also about the positive reaction to the song with maybe with the “know” indicating familiarity and recognition. Overall, it helps in understanding the different aspects of the scene that resonate with viewers who had positive comments.

A screenshot of a computer screen

Description automatically generated

Topic 1 negative reaction on the scene and show likely due the word “killed”, same can said for topic 2 and topic 3 with topic 2 more on the song and topic 3 could more on the tv show as a whole. Topic 4 terms like “want”, “way”, “tel”l is likely from the song, “never” can be said used to emphasis the negative reaction. But in the video, there a line which say, “Number 5 killed my brother”. The term killed likely used in that context and it is likely ended up in the frequently used term in negative sentiment due to its meaning.

A screenshot of a computer program

Description automatically generated

In here topic one could be about watching the scene maybe once or multiple times, it does not directly indicate any negative and positive reaction. Topic 2 talks in the context of lines in the scene as words like “number”, “chills” and “sing” are words said in the video. Topic 3 talks in context of lines in the video. Topic 4 could be talking about reaction of viewers with the word “can”, it could be used in context of “I can not believe”.

Topic Word Distribution

A key concept in topic modelling is topic word distribution, which is the probability that a word will appear in each subject that has been determined. This distribution is essential for comprehending and evaluating the subjects taken from a corpus of text documents because it captures the likelihood of seeing particular terms given a given topic. Through an analysis of word distributions across topics, analysts are able to identify the key themes and subjects that are commonly found in the text data. Higher probability words are seen to be distinctive of a topic and are often used to identify or characterise it. Researchers can compare subjects, find similarities and differences, and learn more about the fundamental organisation and substance of the text corpus by using this approach. In the end, topic word distribution interpretation allows for a more profound comprehension of the hidden themes found in the data, which facilitates insightful textual analysis and interpretation. (ChatGPT, n.d.)

A screenshot of a graph

Description automatically generated A screenshot of a graph

Description automatically generated

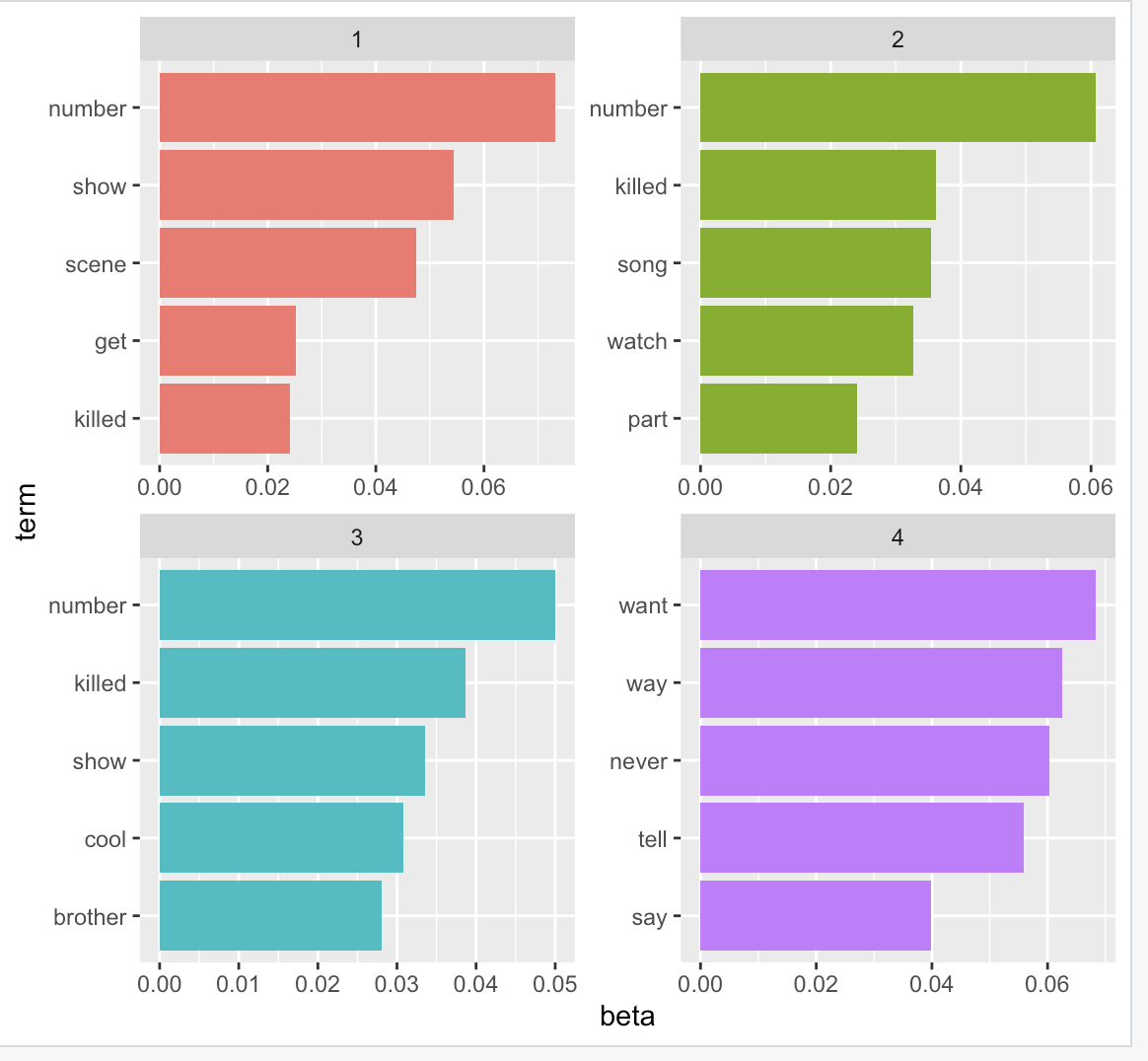
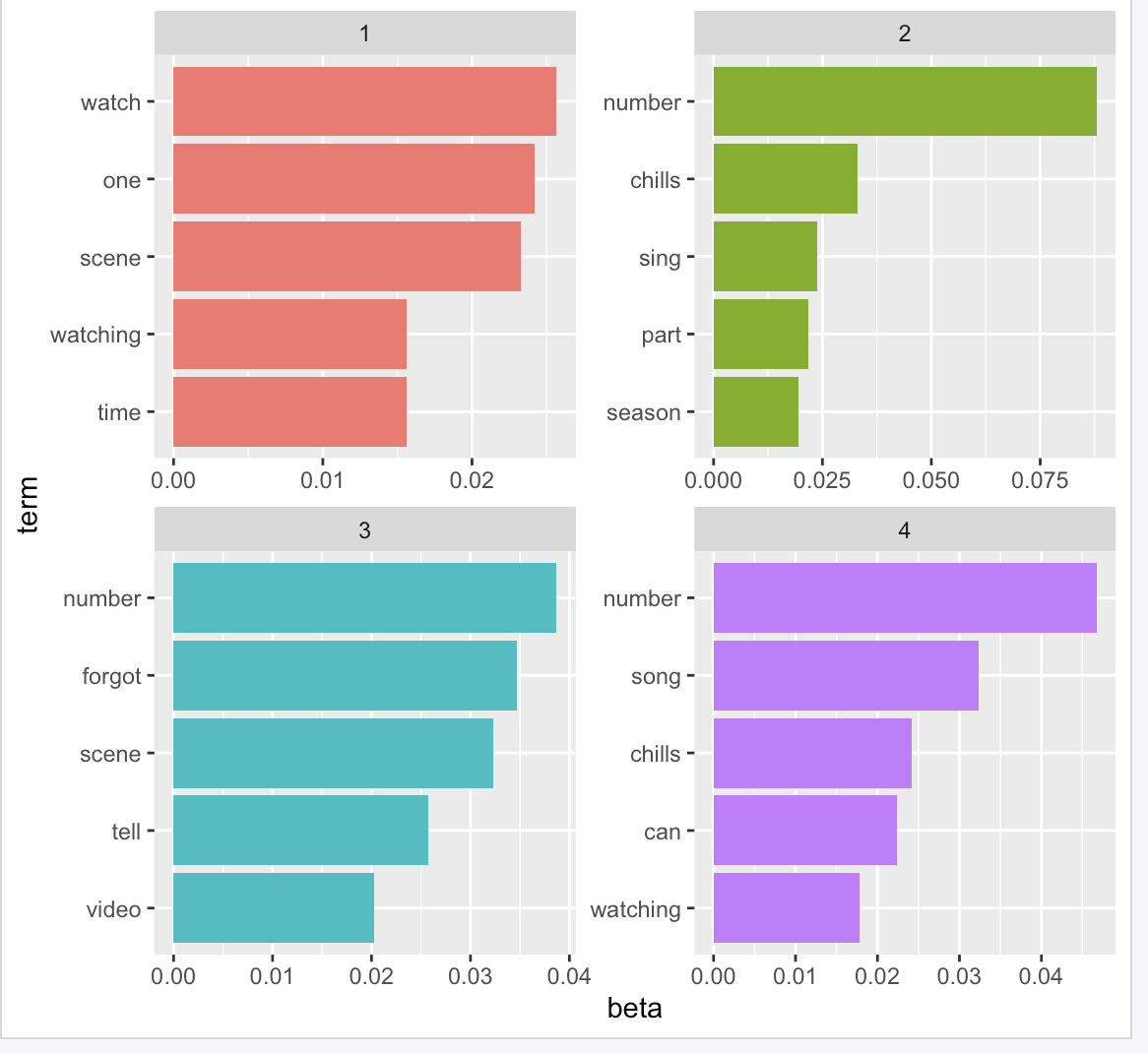
Figure 5 Topic-Word Distribution: Entire Corpus Figure 6 Topic-Word Distribution: Positive Sentiment 

Figure 7 Topic-Word Distribution: Negative Sentiment Figure 8 Topic-Word Distribution: Neutral Sentiment

This is term by probability distribution. In this we can say the word “number “is the frequently used term across the entire corpus and all sentiments. It is term with highest probability in most of the topics. In the topic-word distribution of the entire corpus in topic 1 and 2 the probability of most of the terms being used is less than 0.02. In topic 3 and 4 the probability of most of the terms being used is between 0.02 and 0.03. In the positive sentiment, words “like”, “love” and “best” are which can describe positive reactions. In this we can see that the probability of the word “love” in topic 1 is just slightly higher 0.03 and topic 2 it is the approximately 0.018 and the ranks 5th in top 5 frequently used terms in topic 2. The word “best” in topic has highest probability and third highest in topic 4. The word “like” seems to have lower probability of being used except in topic 4 where it is the second highest. In the negative sentiment, the most frequent word in each topic seems to have a higher probability than both the entire corpus and positive sentiment, with term “number” in topic 1 with probability around 0.07, term “number” in topic 2 with 0.06 probability, term “number” in topic 3 with 0.05 and word “want” in topic with more than 0.06 probability. It can be also said that most terms in topic 2 have probability less 0.04 being used and topic 4 most terms have probability of 0.04 or higher. In neutral sentiment, also have the term “number” has probability in 3 out 4 topics but with varying probabilities with topic 2 with highest with over 0.075, followed by topic 4 then 3. Overall, this shows the difference in significance and usage of terms across different topics and sentiments and showing how certain words dominate in one context while others change according to sentiment.

LDA Interactive Visualization

Entire corpus:

A graph of a number of people

Description automatically generated with medium confidenceA graph with numbers and a red circle

Description automatically generated

A graph of a bar graph

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

In these figures we can see that topic and 1 and 3 are distinct from the others due to the distance of the circles. The terms in these topics are more distinct and are less likely to overlap with terms in other topics, indicated by this the separation.While topic two and four have similaries as there is an overlap between the two circles . This overlap shows that there are more terms or patterns in common between these two topics.Even the chart of topics two and four has similar structure even if the terms are not all the same .

Positive Sentiment :

A graph of a number of words

Description automatically generated with medium confidence A graph with a bar chart and a diagram

Description automatically generated with medium confidence  A graph with numbers and circles

Description automatically generated with medium confidence

A graph with numbers and circles

Description automatically generated with medium confidence

There is no overlap between any of the topics in the positive sentiment which can be identified as all four circles are far away from each other. The separation suggests that there are few common terms or themes among the topics in the positive sentiment , making each one distinct. Even the graphs for each is very different from each other.

Negative Sentiment:

A graph with a circle and a number

Description automatically generated with medium confidence A graph with a red circle

Description automatically generated

A graph with red and blue bars

Description automatically generated A graph with a number of text

Description automatically generated with medium confidence

A graph with a number of different colored bars

Description automatically generated with medium confidence

In here there is high overlap between topics 2 and 4 which indicates a significant similarity between them. This overlap inticates that topics 2 and 4 likely share common terms or themes, making them more related to each other compared to the other topics.

Neutral Sentiment:

A graph with circles and numbers

Description automatically generated with medium confidence A graph with red circles and blue dots

Description automatically generated

A graph of a graph with a number of circles and a number of text

Description automatically generated with medium confidence A graph with numbers and a red circle

Description automatically generated

A graph of a graph with a bar chart

Description automatically generated with medium confidence

There is no overlap between topics in the  the neutral sentiment, suggesting that each topic is unique and represents a separate aspect of the comments. The idea that topics 2 and 4 are distinctively different from one another and that each one represents a distinct  aspect within the neutral sentiment comments can be said due to the approximate equal distance between topics 2 and

# References

Kulshrestha, R., 2019. *A Beginner’s Guide to Latent Dirichlet Allocation(LDA).* [Online]   
Available at: https://towardsdatascience.com/latent-dirichlet-allocation-lda-9d1cd064ffa2

Kibe, K., 2024. *Topic Modeling Using Latent Dirichlet Allocation (LDA).* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2023/02/topic-modeling-using-latent-dirichlet-allocation-lda/#:~:text=Word%20cloud%20is%20a%20data,and%20content%20of%20the%20data.

ChatGPT, n.d. *ChatGPT.* [Online]   
Available at: https://chatgpt.com/c/b0e57878-926b-4911-916f-472415a262fa

GeeksforGeeks. "Understanding TF-IDF (Term Frequency-Inverse Document Frequency)." GeeksforGeeks, 13 April 2023, <https://www.geeksforgeeks.org/understanding-tf-idf-term-frequency-inverse-document-frequency/>. Accessed 26 May 2024.