# Text Mining and Sentiment Analysis

1. **Title**

This task aims to demonstrate the practical application of data science techniques in the analysis of text data, to provide a more robust understanding of the sentiments of the said data. The title of this task is “Analyzing Netflix’s Trends: Streaming sentiments and text mining”.

1. **Introduction**

In the current time, digital media is producing millions of terabytes of data every second (*How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read*, n.d.)-, making the analysis of these data types very crucial.

Taking advantage of the robustness of python libraries, this analytical research will cover tetual data preprocessing as well as sentiment intensity using NLTK. Emotional content across various dimensions such as genre would be quantified.

This approach is not only methodological and data-driven, it is narrative, telling a story of the complex horizon that is the streaming content industry.

1. **Dataset Description**

Data from notable Media giant “Netflix” is the data set of interest –Netflix Movie and TV Show dataset from [Kaggle](https://www.kaggle.com/code/sainischala/netflix-shows-sentiment-analysis-using-python/input).

The dataset is comprised of 1,910 instances (excluding column names) and 12 features

This particular dataset is of interest as it provides the material for this kind of textual analysis. The variables are versatile enough to offer a perspective of how context interacts within the available time frame in the dataset.

|  |  |
| --- | --- |
| **Variable** | **Variable Type** |
| Show\_id | Identifier variable |
| Type | Binary/Boolean |
| Title  Director  Cast  Country  Rating  Listed\_in | Nominal |
| Description | Textual |
| Date\_added  Release\_year | Date/Time |
| duration | Ordinal |

*Table 1: Variable Information of the Netflix dataset from Kaggle* (SHIVAM BANSAL, 2021)

1. **Text Exploration and Preprocessing**

Given the nature of the dataset and the objectives within this analysis, the preprocessing steps for this dataset are essential for data quality and analysis readiness

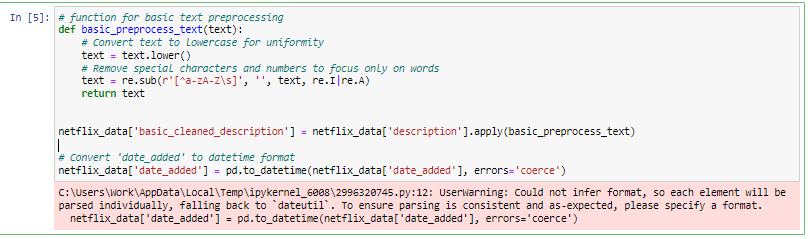
1. Import necessary libraries, Load Data and Inspect the Data



1. Converting text to lower case, Removing special characters,Converting ‘date\_added’ to datetime format

All text is converted to lower case to ensure uniformity of text.

Special characters and numbers are removed to allow the algorithms focus on “words” without skewing the data.



1. Tokenization, Stop words removal and Lemmatization

Tokenization involves breaking down full text into “token” words. Eliminating sentence fillers like “the”, “a”, etc. is performed by removing stop words(from nltk.corpus import stopwords).

Lemmatization functions as a word base reducer, reducing words to their base form o consolidate different forms of the same word.

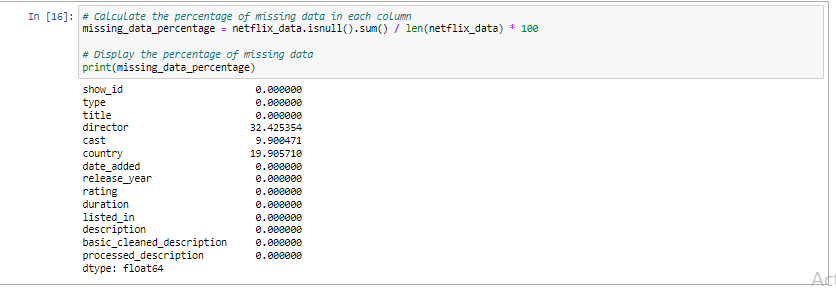
These processes ensure that the textual data is free from unnecessary additions that may impact the results of the analysis negatively, hence painting a wrong picture.

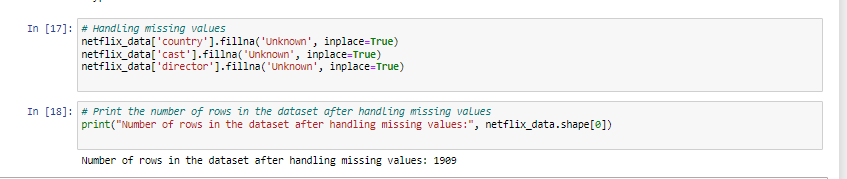


1. Handling Missing Data

There are missing data values within the dataset. Missing values summary:

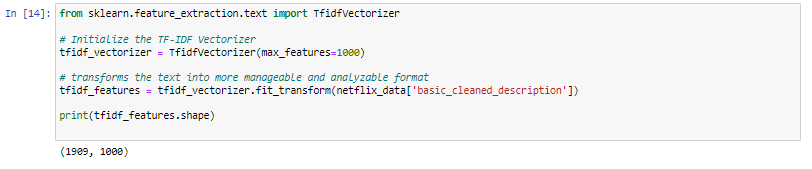
* director --- 32.4% missing data
* cast --- 9.9% missing data
* country --- 19.9% missing data



To handle the missing values, these can be replaced with a placeholder like 'Unknown'. This ensures that the data for analysis is preprocessed without introducing bias.

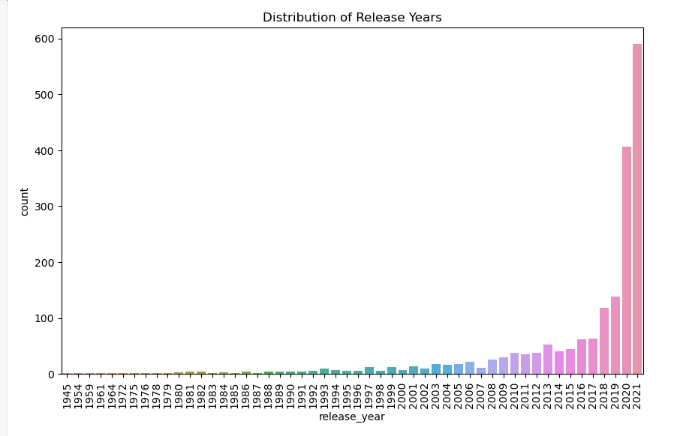
1. Data Transformation using TF-IDF Vectorization

Due to the nature of the ‘description’ column, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization will be applied, to simply transform the text data into a format suitable for machine learning models.



1. Visual Exploration

This plot visually explores the steady rise of the movies released year on year, giving a clear view on the progression of the industry across the years.

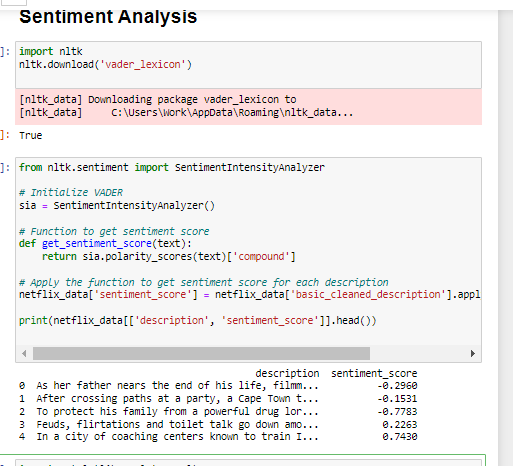


1. Analysis Implementation
2. Sentiment Analysis

Sentiment analysis is a systemic technique used in determining the emotional tone in text(s). On a general basis, it is often used to differentiate information that are subjective in content like reviews, comments, posts, etc. The emotion can either be negative, positive or neutral.

1. Set sentiment score

The vader\_lexicon is used here to determine the sentiment of any piece of text. The VADER (Valence Aware Dictionary and sEntiment Reasoner) is a tool for sentiment analysis.



A "good" sentiment score refers to a positive sentiment. Sentiment scores range from \_-1 to +1 in the context of the NLTK library.

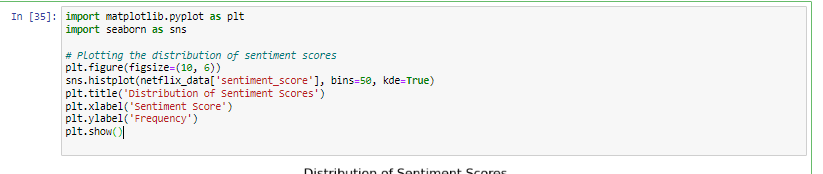
-Negative Sentiment: Scores below 0 indicate negative sentiment. The closer the score is to -1, the more negative the sentiment.

-Neutral Sentiment: A score around 0 suggests a neutral sentiment, meaning the text is neither positive nor negative.

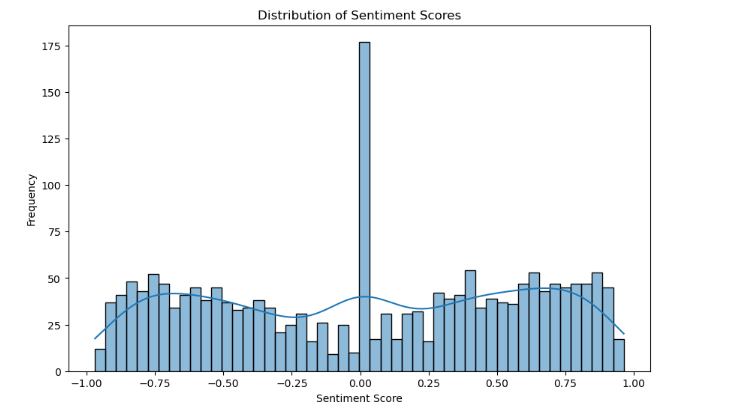
-Positive Sentiment: Scores above 0 represent positive sentiment. The closer the score is to 1, the stronger the positive sentiment.

1. Sentiment Score Distribution

The histogram plot is used to visualize the sentiments across the entire dataset.



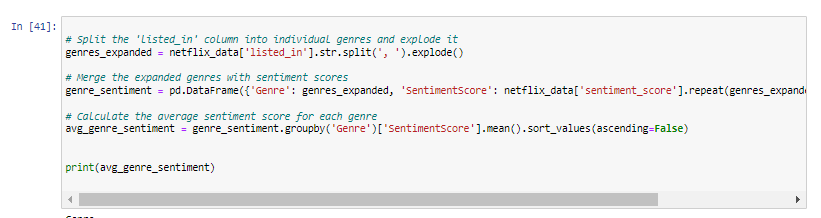
On inspection, the histogram reveals a distribution of sentiment score with a significant peaking around the neutral zone. Even though the histogram reveals multiple sentiment intensity across board, it could indicated that the highest peaking are just from more factual emotional expression- especially due to the fact that the analysis is not in singularity.

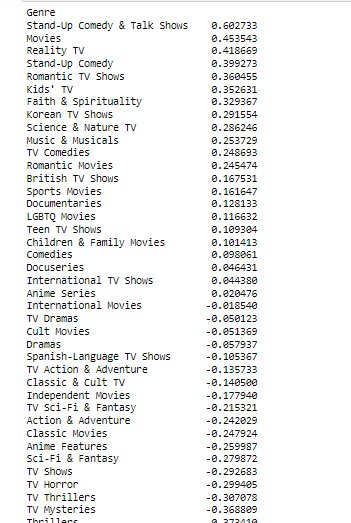


1. Genre Sentiment: Splitting and Expanding Genres in the “Listed\_in” Column

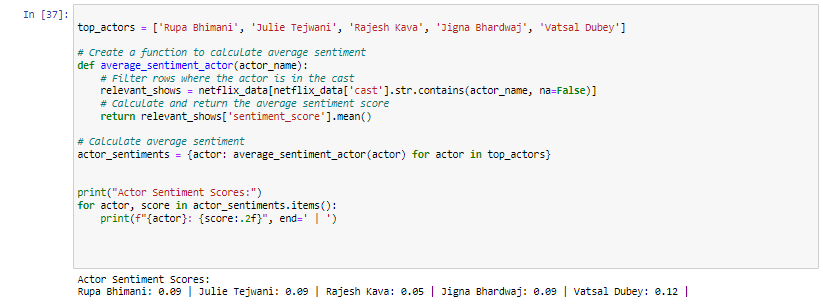
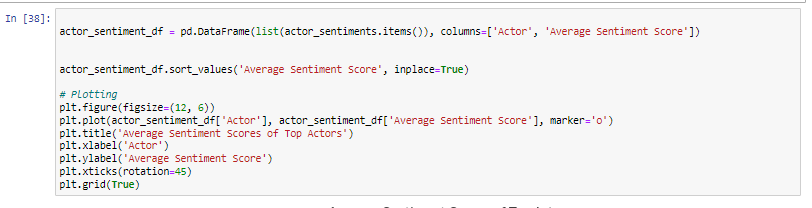
In this step, the average sentiment genre is computed. It provides insights into interesting trends in the context of genre.

-Splitting Genre: The genre is residing in the ‘listed\_in’, but listed as multiple genres per line. The str.split() uses the common delimiter, comma, to separate the genres.

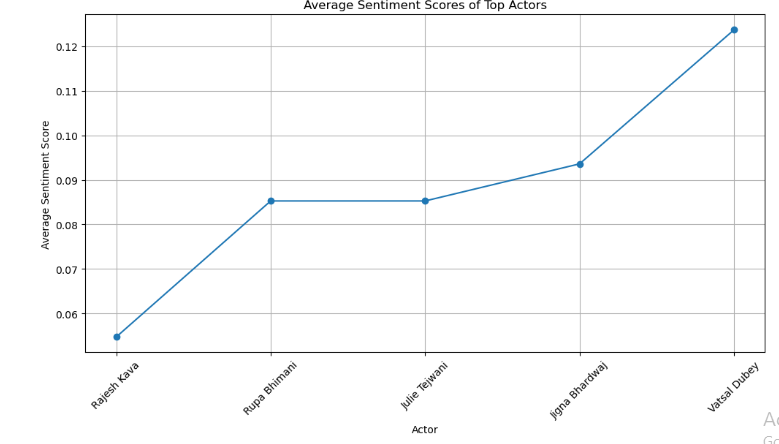
-Expanding Genre: After splitting, the explode() is used to transform each element to a list-like row; creating readability for the frame.



1. Average Sentiment Scores of Top Actors:



The plot shows the average sentiment scores for top actors inferring a positive correlation between specific actors and the sentiment of their associated genre. Actors with higher sentiment scores may indicate more positive descriptions or have more appearances in uplifting and engaging content (standup comedy & talk shows are the top genres0.



1. Sentiment of Content on Netflix by Release Year

The stacked column chart visualizes the gradual trends of sentiments for total contents vis a vis the year they were released. The columns are distinguishable by colors, each color representing the mode of the sentiment.



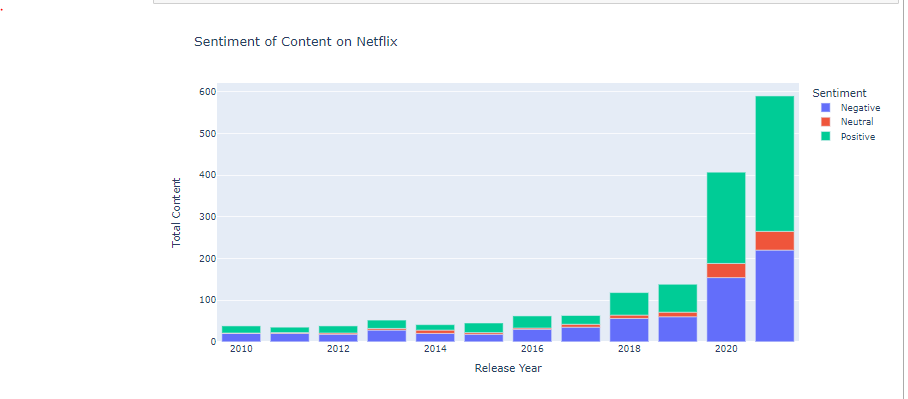
On observation of the plot:

-There is a significant growth in the number of contents produced as the years’ progresses

-Neutral comments were not significant until more content options were available.

-As the years’ progresses, there are more positive sentiments towards Netflix movies

-A significant jumping-rise of the total content released from 2020-2021.



1. Text Mining Analysis

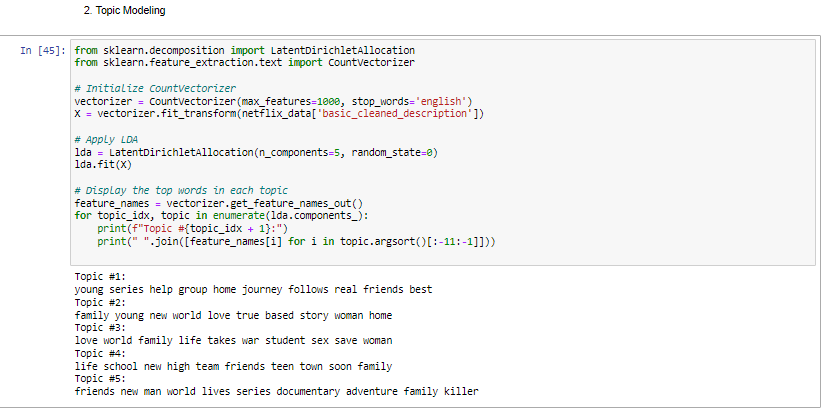
“The text mining analysis is the process of extracting important insights and trends from unstructured text. It involves analyzing large volumes of text to uncover relationships, trend and other meaningful information that may not be obvious at initial inspection”-@@@@-

1. Frequency Analysis

This analysis identifies, then counts the most frequently occurring singular words in the movie or descriptions, revealing the most common terms used in the Netflix dataset.

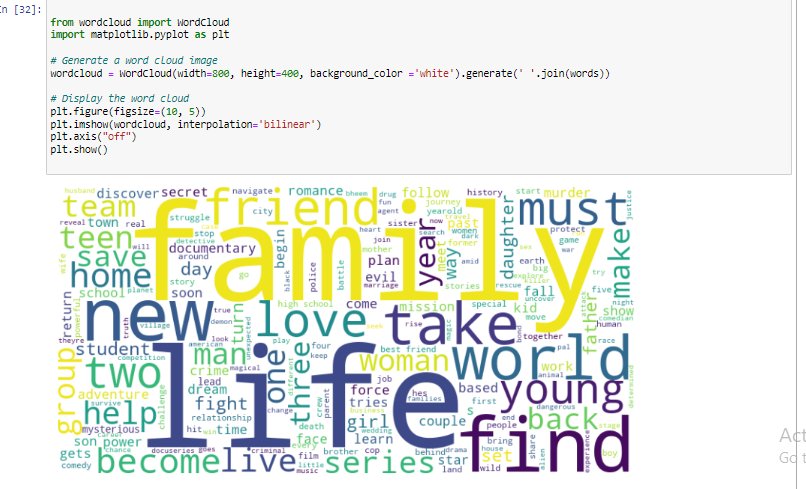
1. Topic Modelling

Topic modelling uses Latent Dirichlet Allocation (LDA) to uncover unapparent topics within the movie descriptions, grouping them into different themes based on the pattern of the words within the dataset (or any media text).



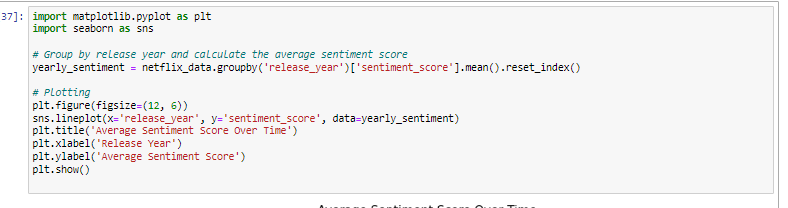
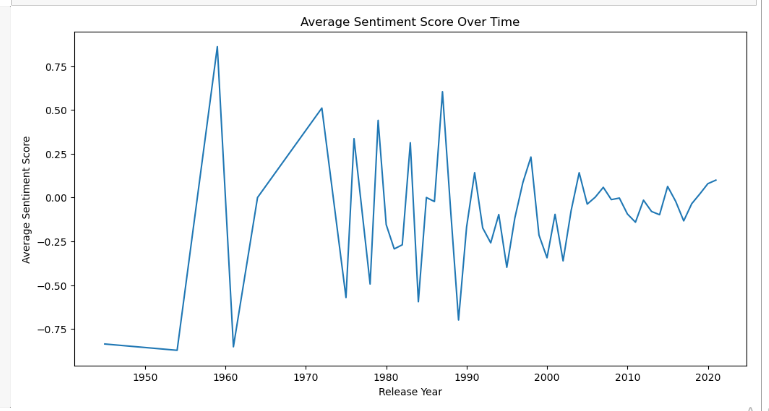
1. Word Cloud

The function is in the name. Word cloud creates a cloud of the most common words.



1. Sentiment of Time

This plot show how sentiment trends have evolved over time, (by calculating the average sentiment score for each release year in the Netflix dataset)



1. Named Entity Recognition (NER)

The NER extracts and identifies named entities (like people, places, etc. ) from the movie descriptions. This aids in providing insights into common themes in the content.



1. N-gram Analysis

N-gram identifies a particular number of word combination based on specifications, bigram in this case. It identifies and counts the frequency of bi-grams (two-word combinations) descriptions, spotlighting the most prevalent phrases or themes in the dataset



1. Result Evaluation
2. Sentiment Analysis
3. Trends of sentiments: Across the Netflix data itself, there is a significant peaking towards the neutral sentiment Though, the peaks at both ends are a mix of highly positive and negative content. The more contents produced, the more diverse the sentiments.
4. More content have been put out as the years progresses, the sentiment score is observed to be responding by presenting diverse emotional tones. This is progressive.

* Actor Sentiment Correlation

1. Top ranking actors like Rupa Bhimani and Vatsal Dubey are associated with content with higher positive scores which influences viewer/user satisfaction and interaction.

However, it is not mathematically confirmed if the top performing genre/content are top performing because of the actors or the actors are top performing because of genre they feature in.

* Genre Sentiment Dynamics

1. Genres such as 'Stand-Up Comedy & Talk Shows', 'Movies', and 'Reality TV’ are top performing in the context of their sentiment scores
2. Crime TV Shows' and 'Horror Movies' are low performing in the context of their sentiment scores, which perhaps align with the typically intense nature of these genres.

* Genre Popularity

1. 'Dramas' and 'International Movies' are the highest occurring genre.Though, sentiment score-wise they are not top performers.

* Country Specific Content

1. Top content production originates from US,India and the UK respectively. The mix of these nationality would ensure diversity of the Netflix content.
2. Text Mining

* Common Theme and Topics:

1. Commonly occurring words such as "life," "family," "new," and "love", indicates that family systems, romantic scenarios, etc. are central in Netflix content.
2. Topic modelling reveals patterns related to youth, family- i.e. indicating content with better demographic preference.

* Sentiment Fluctuations over Time

1. The fluctuations of sentiments vary and may either peak or fall dependent on then content theme, time and the external societal inference.

* Named Entity Recognition

"Johnson" and "New Orleans" are entites that stood out. This may imply content with specific geographical and character focus.

* N-gram Analysis

1. Bi-grams like "high school," "new york," and "best friend" are the most frequent bi-grams and it spotlights common settings and relationships in the data.
2. Conclusion and Recommendation

* Content Strategy:

- Marketing and content development focus should be on life, family, and friendship themes. This recommendation is adviced in the context of word frequency analysis showing high occurrence of words like "life" (226 mentions) and "family" (171 mentions).

-Marketing/content strategy should be dependent on sentiment score trends, with an emphasis on genres with positive sentiment like 'Stand-Up Comedy & Talk Shows' (average sentiment score of 0.60).

* Marketing and Promotion:

-Harness themes of "new beginnings" and "love" in promotional content, aligning with top word frequencies in text mining results. Ads, all promotional content would need to align with the themes of "love", "family". This is advised in the context of top word frequencies in text mining results

-Implement location-based NER insights, such as "New Orleans," to build marketing campaigns

* Personalization and Recommendation:
* Integration of actor sentiment data (e.g., Vatsal Dubey's positive score of 0.12) to personalize viewer recommendations. Allow viewers select preference based on individual actors instead of a Cast.
* Align recommendations with sentiment progression; progression of sentiment scores for top actors and genres over time.
* Ethical Considerations:
* Since these data are personal data, precautions should be taken to ensure sentiment analysis tools are applied ethically, respecting user privacy.
* All data collection or analysis processes must be free from bias and cultural appropriation.