Topic Modeling and Classification Modeling on some IMDB Reviews.

Anna Maria Teodori (matr. 889903)

Michele Gazzola (matr. 825267)

Master's Degree Course in Data Science, Department of Informatics, Systematics and Communication DISCo. University of Milan Bicocca. Viale Sarca, 336 - 20126, Milan

Abstract

This study undertakes a thorough examination of IMDb reviews, addressing the dual objectives of sentiment classification modeling and topic modeling. The dataset undergoes meticulous preprocessing, including tokenization, punctuation removal, and other crucial steps to ensure the quality of textual data. Our exploration of IMDb reviews employs a multifaceted approach to uncover latent themes. We began by identifying common tokens, frequent words, and the top 10 adjectives in positive and negative reviews, gaining valuable insights into prevailing expressions within the dataset.
Following this, we initiated topic modeling, utilizing N-gram analysis to delve into contextual relationships via both bigram and trigram networks. Implementing Latent Dirichlet Allocation (LDA) significantly enhanced our understanding, revealing hidden topics and patterns within the IMDb review corpus. In our sentiment analysis, we employ a diverse range of models tailored to discern IMDb review polarity. Our testing covers Bag of Words (BoW) with both single-word (Unigrams) and two-word (Bigrams) representations, utilizing binary encoding (multi-hot) and TF-IDF. Additionally, we explore a sequence model, incorporating both word embedding without masking and with masking.

1 Dataset and data collection

This study is rooted in the exploration of sentiment analysis within the realm of film critiques, utilizing the extensively curated Large Movie Review Dataset v1.0. Developed by

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, this dataset is specifically designed for binary sentiment classification. It surpasses its predecessors in scale, featuring 25,000 highly polar movie reviews for both training and testing, accompanied by additional unlabeled data for further exploration. The dataset is meticulously organized, with separate folders for training and testing. Each of these folders is subdivided into positive and negative review categories, reflecting the polarized nature of the sentiments expressed in the movie critiques. Each review is annotated with its corresponding sentiment label, indicating whether it originates from the positive or negative category. To facilitate our analysis, we have aggregated the 50,000 reviews into a structured Data Frame. This Data Frame features two crucial columns: one housing the reviews and another labeled 'Sentiment,' indicating whether the sentiment of a given review is positive or negative. This consolidation allows for a holistic examination of sentiments expressed across diverse movie reviews within a unified framework.

2 Pre-Processing

The preprocessing phase is integral for effective text mining and sentiment analysis on IMDb reviews. The systematic approach includes:

2.1 Lowercasing and Text Cleaning:

- *Lowercasing:* Converts all reviews to lowercase for uniformity.
- *Text Cleaning:* Removes textual noise, such as website links, ensuring a refined dataset.

2.2 Tokenization and Removal of Numbers:

- *Tokenization:* Breaks reviews into individual words for detailed linguistic analysis.
- *Removal of Numbers:* Eliminates numerical characters, streamlining text for sentiment analysis.

2.3 Punctuation Removal, Stopword Removal, and Lemmatization:

- *Punctuation Removal:* Eliminates special characters for focus on semantic content.
- *Stopword Removal:* Removes common English stopwords to reduce noise.
- *Lemmatization:* Reduces words to base form, ensuring coherence and standardization.

2.4 Compilation and Storage:

- Aggregates processed reviews into a DataFrame with a 'clean_review' column.
- Saves the refined dataset as a CSV file for subsequent analyses.

This systematic preprocessing ensures subsequent analyses operate on meaningful, noise-free text, facilitating nuanced exploration of sentiment patterns and topic dynamics within movie critiques.

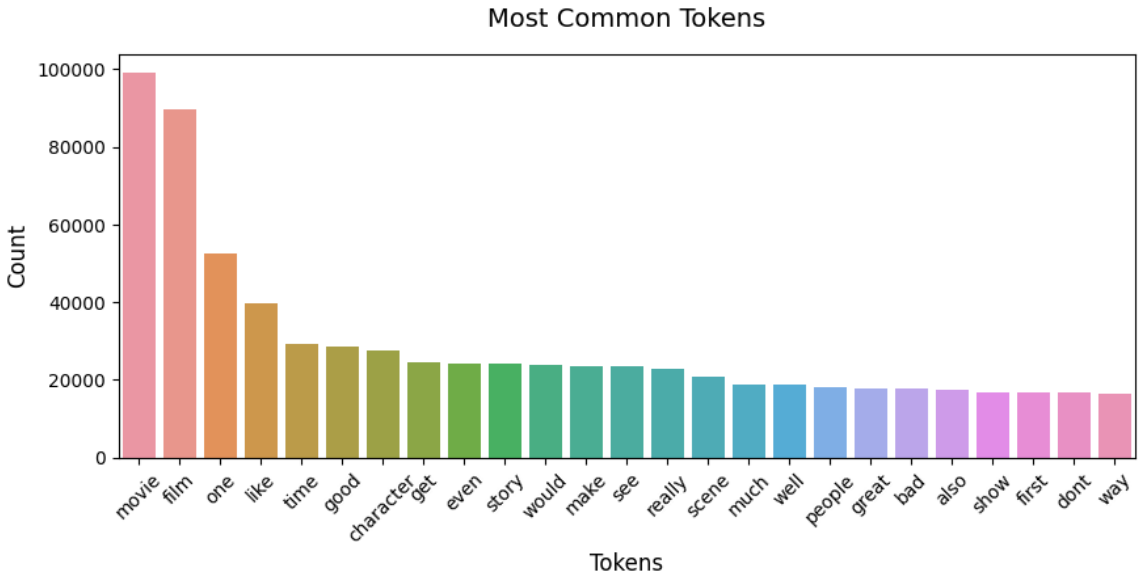
3 Data Exploration and Visualization

In this section, data exploration and visualization techniques are employed to unravel the most common tokens, frequent words in positive and negative reviews, and the prevalence of popular adjectives in general and for the different type of reviews. This journey not only aids in understanding the language dynamics but also sets the stage for a more profound analysis of sentiments within the cinematic narrative.

3.1 Identification of Popular Tokens:

The initial step in our exploration involves uncovering the most frequently occurring tokens within the IMDb reviews. The output presents a snapshot of the top 25 tokens and their corresponding frequencies.

1. *Movie and Film Dominance:*
    - Notably, "movie" and "film" emerge as the dominant tokens, reflecting the central theme of the dataset—movie reviews. Their high occurrence underscores the dataset's focus on cinematic discussions.
2. *Common Descriptors:*
    - Tokens like "one," "like," and "time" indicate common language patterns used in reviews. These versatile terms might play a role in expressing personal preferences or commenting on the temporal aspects of the movie-watching experience.
3. *Positive Sentiment Indicators:*
    - Positive sentiment is hinted at by terms like "good," "great," and "well." These positive descriptors suggest a general inclination toward favorable expressions in the reviews.
4. *Character-Centric Discussions:*
    - The presence of "character" and "story" emphasizes the significance of character development and narrative in the discussions. This aligns with expectations, as these elements often play a crucial role in movie critiques.
5. *Critique and Evaluation:*
    - Words like "bad," "dont," and "even" imply critical evaluations and potentially negative sentiments. The dataset appears to encapsulate a spectrum of opinions, from positive acclaim to critical assessments.
6. *Broad Audience Reference:*
    - The inclusion of "people" and "see" suggests a consideration of the broader audience perspective, indicating that reviewers may contemplate the film's appeal to a wider demographic.



In essence, this initial exploration of popular tokens provides a foundational understanding of the recurring themes and language dynamics within IMDb reviews. Subsequent analyses can delve deeper into these linguistic patterns to extract richer insights into sentiment and thematic elements.

3.2 Frequent Words:

Following the identification of the most popular tokens in IMDb reviews, a nuanced approach was adopted to refine the visualization of common words in both positive and negative sentiments. Recognizing the potential bias introduced by ubiquitous terms like 'film' and 'movie'—inherent to a dataset centered around movie reviews—a strategic decision was made to exclude these terms. This exclusion aims to spotlight authentic sentiments expressed in the reviews, ensuring that the visual representation is free from the influence of generic language. The refined word cloud, devoid of these overarching terms, provides a more accurate and insightful portrayal of the distinctive language used in positive and negative sentiments. This careful curation ensures that the visual narrative aligns with the objective of capturing the true essence of IMDb reviews, transcending the dominance of generic movie-related language.

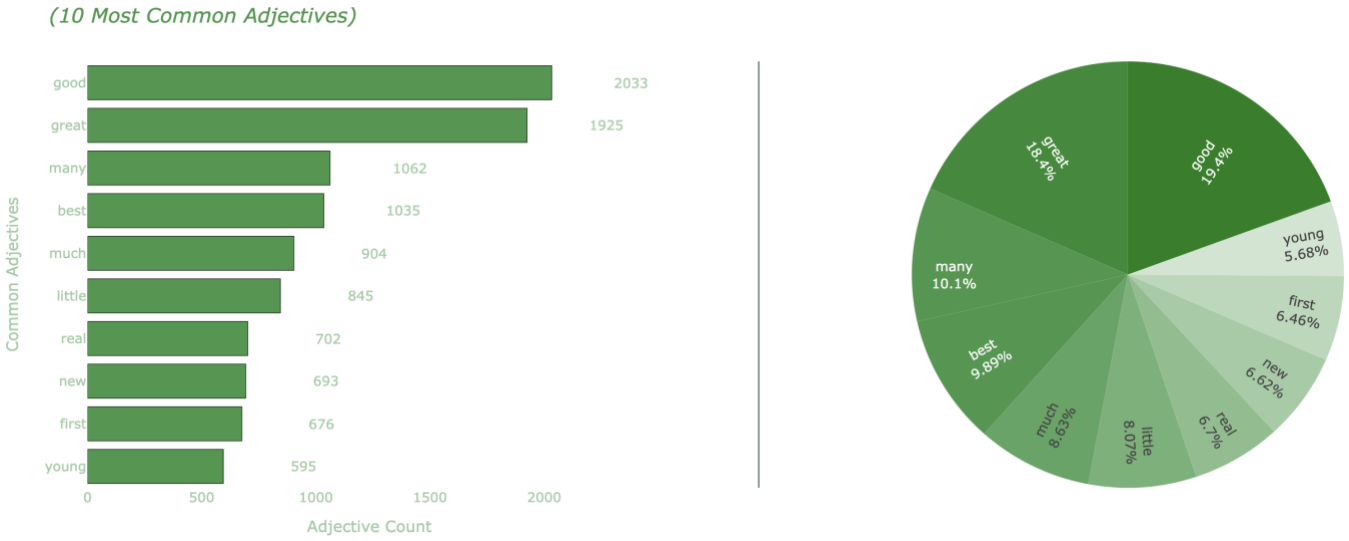
1. *Frequent words from positive reviews:* as possible to see the prevalent terms in positive reviews, such as 'story,' 'good,' 'make,' and 'sense,' underscore the audience's appreciation for well-crafted narratives and a positive viewing experience. The prominence of 'life' and 'show' suggests an acknowledgment of films that resonate with real-life scenarios, emphasizing the impact of compelling storytelling on the audience's sentiments.

2. *Frequent words from negative reviews:* In contrast, the word cloud for negative reviews reveals insights into areas of dissatisfaction. Noteworthy terms like 'bad,' 'even,' and 'nothing' express critical sentiments, while the unexpected inclusion of 'good' emphasizes a potential context-specific usage or a nuanced expression within negative reviews. The presence of 'dont' and 'people thing' highlights instances of negative opinions and varied perspectives, underscoring the diversity of criticism within the dataset.

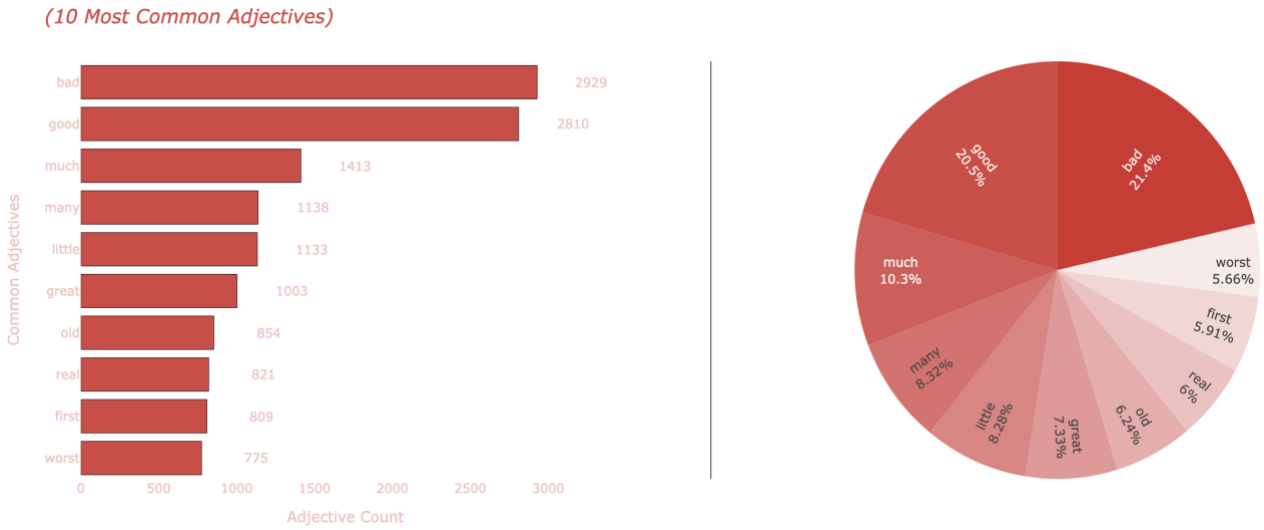
3.3 Most Popular Adjectives:

Recognizing the nuanced role that adjectives play in shaping sentiment, a deliberate exploration into the ten most frequently used adjectives in both positive and negative reviews was conducted. Adjectives, being descriptors that carry inherent evaluative nuances, can offer a more nuanced insight into the emotional tone of IMDb reviews.

1. *Frequent adjectives from positive reviews:* In the realm of positive sentiments, adjectives such as 'good,' 'great,' and 'best' dominate the landscape, reflecting the audience's appreciation for high-quality, exceptional content. The prevalence of 'many' and 'much' suggests a positive acknowledgment of abundance, while descriptors like 'real' and 'young' point to an appreciation for authenticity and youthfulness in the narrative.



2. *Frequent adjectives from negative reviews:* Contrastingly, negative sentiments are characterized by adjectives such as 'bad,' 'worst,' and 'old,' signifying critical evaluations of the filmic elements. The recurring use of 'much' and 'many' in negative reviews may indicate disappointment or excessiveness, contributing to unfavorable sentiments. The presence of 'little' suggests dissatisfaction with the quantity or impact of certain elements, emphasizing the diverse aspects of discontent within negative sentiments.



4 TOPIC MODELING

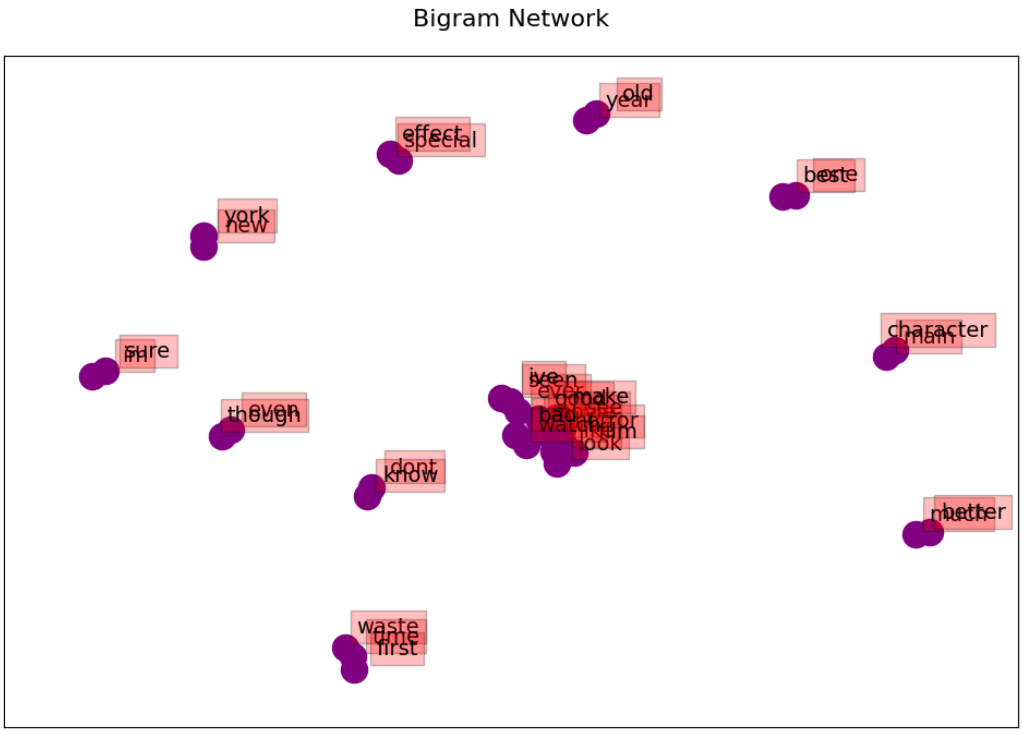
To address the topic modeling task, we initially conducted an N-gram analysis, followed by the implementation of Latent Dirichlet Allocation (LDA).

4.1 N-gram Analysis

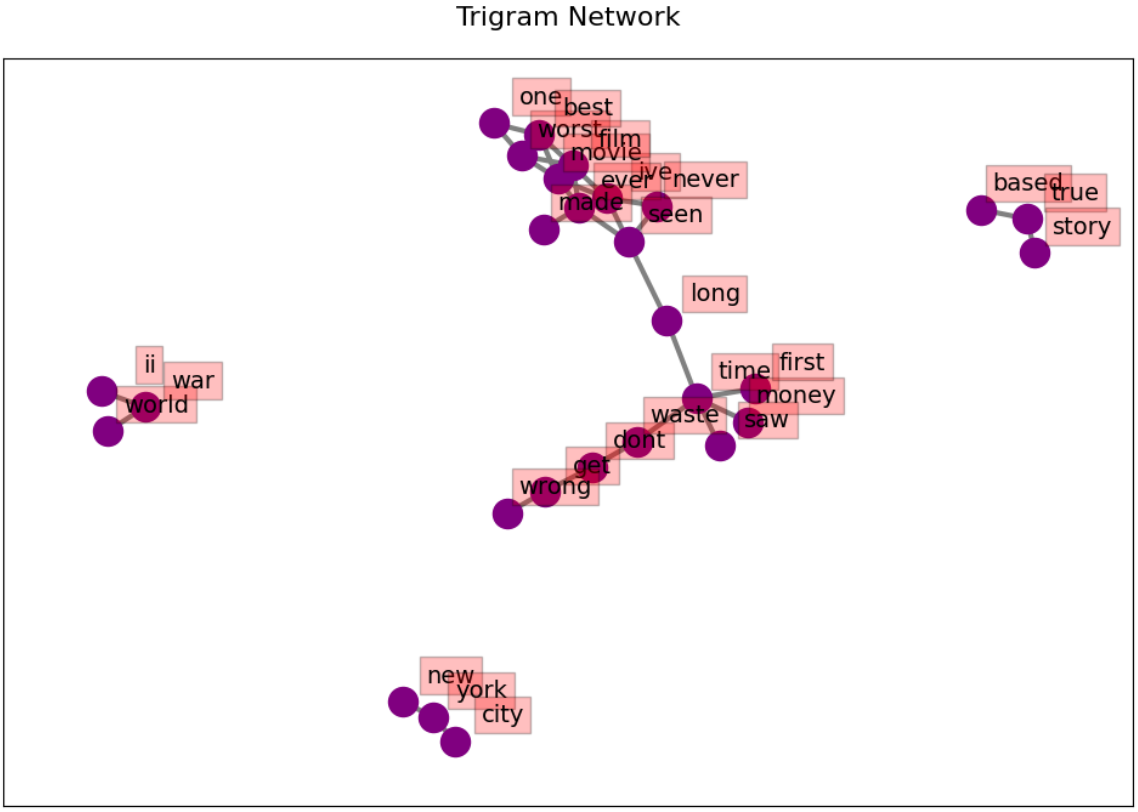
The main objective of this analysis was to explore linguistic patterns and the occurrence of words and phrases in movie reviews. We followed several steps to achieve this goal:
We began by tokenizing the reviews, breaking down the text of the reviews into individual words or tokens using the 'word_tokenize' function from the Natural Language Toolkit (NLTK) library.
Next, we focused on identifying bigrams, which are frequent pairs of consecutive words in the reviews. Using the 'bigrams' function from NLTK, we generated these pairs for each review, creating a list for further analysis.
We calculated the frequency of each bigram in the dataset using Python's 'Counter' class, resulting in a structure called 'bigrams_freq' that linked each bigram to its frequency.
We selected and visualized the top 25 most common bigrams from 'bigrams_freq,' displaying them in descending order of frequency. This highlighted the most frequently used word pairs in the reviews.

```
[(('look', 'like'), 2879),
 (('ever', 'seen'), 2532),
 (('special', 'effect'), 2200),
 (('ive', 'seen'), 2148),
 (('dont', 'know'), 2054),
 (('main', 'character'), 1878),
 (('even', 'though'), 1867),
 (('one', 'best'), 1843),
 (('year', 'old'), 1736),
 (('movie', 'like'), 1714),
 (('movie', 'ever'), 1546),
 (('see', 'movie'), 1495),
 (('horror', 'movie'), 1493),
 (('good', 'movie'), 1461),
 (('waste', 'time'), 1441),
 (('much', 'better'), 1426),
 (('horror', 'film'), 1374),
 (('watch', 'movie'), 1363),
 (('first', 'time'), 1327),
 (('im', 'sure'), 1324),
 (('bad', 'movie'), 1319),
 (('ive', 'ever'), 1301),
 (('make', 'movie'), 1280),
 (('new', 'york'), 1279),
 (('film', 'like'), 1250)]
```

We constructed a network graph using the NetworkX library to visually represent the relationships between these common bigrams, with individual words of the bigrams serving as nodes and connections between nodes indicating word co-occurrence.



We applied a similar process to trigrams, which are sequences of three consecutive words. After calculating their frequency, we identified the most common trigrams in the dataset.



Just like with bigrams, we selected and visualized the top 25 most common trigrams, providing further insight into frequent word sequences in the reviews.

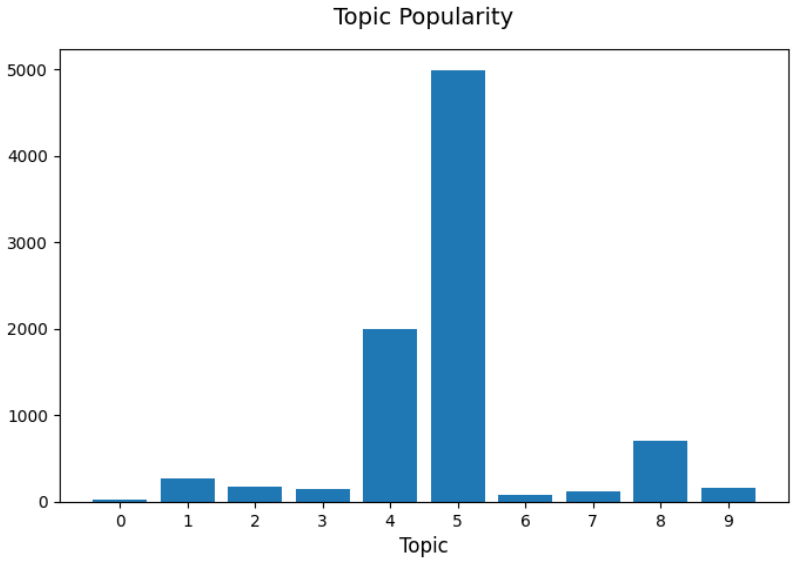
For trigrams as well, we created a network graph to illustrate their co-occurrence patterns.Our analysis revealed several common bigrams in IMDb reviews. For example, the bigram "('look', 'like')" emerged as the most frequent, indicating a recurring expression in the reviews. The network visualizations of both bigrams and trigrams offered an intuitive perspective on the connections between these word pairs and sequences, highlighting co-occurrence relationships and providing insights into contextual associations between words in the reviews.

4.2 LDA

The critical step is text vectorization, where the script employs the `CountVectorizer` technique to convert the text data into a numerical format suitable for machine learning algorithms. Each movie review is transformed into a unique row, and every word in the entire corpus of reviews becomes a column, with each cell indicating the word's frequency within a particular review. This transformation is achieved using the `fit_transform` method, allowing the script to learn the vocabulary and generate a Document-Term Matrix (DTM). The resulting feature names represent the entire lexicon shared across all reviews. A vital element of this script is the incorporation of the Latent Dirichlet Allocation (LDA) model. LDA is a statistical model designed to categorize observations, in this case, words in movie reviews, into distinct topics or categories. The script chooses to create ten topics, a crucial parameter in LDA. It employs the `fit` method to identify these topics within the DTM and utilizes the `transform` method to assign topic probabilities to each review. Moving forward, the script proceeds to reveal the discovered topics. It provides a list of the top ten words most strongly associated with each specific topic. This is achieved by accessing the `components_` attribute of the LDA model, which holds the word-topic matrix, indicating the importance of each word within its respective topic.

```
0 ['german', 'prot', 'victor', 'world', 'man', 'would', 'film', 'new', 'gram', 'team']
1 ['film', 'time', 'one', 'story', 'scene', 'life', 'great', 'take', 'well', 'play']
2 ['movie', 'one', 'scene', 'like', 'love', 'story', 'end', 'get', 'make', 'show']
3 ['movie', 'even', 'film', 'story', 'great', 'well', 'good', 'character', 'one', 'find']
4 ['film', 'movie', 'one', 'story', 'character', 'like', 'time', 'life', 'would', 'even']
5 ['movie', 'film', 'one', 'like', 'good', 'really', 'bad', 'time', 'get', 'see']
6 ['film', 'war', 'man', 'good', 'one', 'version', 'character', 'cast', 'would', 'performance']
7 ['film', 'like', 'get', 'time', 'bad', 'one', 'movie', 'good', 'match', 'monster']
8 ['film', 'one', 'scene', 'story', 'good', 'character', 'well', 'time', 'also', 'like']
9 ['film', 'series', 'performance', 'one', 'great', 'well', 'best', 'time', 'story', 'also']
```

The script goes beyond topic extraction; it also intelligently assigns topics to individual movie reviews based on the topic with the highest probability. This results in the creation of a new column named `topic` within the dataframe, signifying the topic index for each review. To offer a more comprehensive view, the script displays a segment of the dataframe, showcasing the original review text, sentiment, cleaned review text, and the assigned topic for the first few rows. Additionally, the script presents a visual representation of topic popularity through a bar plot. By counting the number of documents associated with each topic and arranging them by index, it provides a clear depiction of how prevalent each topic is among the reviews. This bar plot is generated using matplotlib's visualization capabilities.



Interpreting the script's output is a crucial aspect of its utility. The "Topic Popularity" bar plot reveals insights into the distribution of topics across the reviews. For example, Topic 5 emerges as the most prevalent, closely followed by Topic 9 and Topic 4, while Topics 0 and 8 are less frequent. Simultaneously, the list of words associated with each topic highlights the central themes or subjects of reviews linked to that topic. For instance, Topic 0's words like "german," "prot," and "victor" suggest a theme related to war or conflict, while Topic 1's words like "film," "time," and "story" indicate a focus on narrative and storytelling elements. The dataframe snippet further illustrates how topics are distributed among individual reviews, mapping each `clean_review` to a `topic` number, revealing the dominant theme for each review.

5 CLASSIFICATION: SENTIMENT ANALYSIS

To tackle the sentiment classification task on IMDb reviews, we pursued two distinct approaches: the Bag of Words (BoW) and sequence models. The Bag of Words approach involves treating each document as an unordered set of words, disregarding the sequence in which they appear. In contrast, sequence models consider the sequential structure of the text, capturing dependencies and contextual nuances. These diverse strategies offer complementary insights into sentiment analysis, with BoW emphasizing word occurrences and sequence models delving into the contextual flow of language.

Before constructing the models derived from the two different approaches, we initially partitioned the dataset into three subsets: training, validation, and test sets. Training set comprised 60% of the data, while the validation and test sets each comprised 20%. This division allowed for independent training and evaluation of the sentiment classification models on distinct subsets of the IMDb reviews dataset. Subsequently, we built three models using a Bag of Words approach and two models using the sequence model approach. Each model underwent training for 10 epochs, with checkpoints to preserve the

best-performing version. To assess and compare the models, we printed the test accuracy, recall, f1 score, and precision results.

5.1 Bag of Words (BoW):

- Single Words (Unigrams) with binary Encoding (Multi-hot):

In this model, we utilized a TextVectorization layer to convert IMDb reviews into a binary-encoded representation, focusing on individual words or unigrams. The layer, configured with a maximum of 20,000 tokens and set to "multi-hot" output mode, adapted to the training data. We then trained a neural network with a hidden layer and dropout to classify sentiments based on the binary-encoded unigrams. The model was compiled with binary cross-entropy loss and RMSprop optimizer, and checkpoints were saved to monitor and preserve the best-performing configuration during training. This approach aimed to effectively capture review patterns and enable accurate sentiment predictions.

- Two Words (Bigrams) with binary Encoding (Multi-hot):

In this phase, the focus shifted to binary encoding (multi-hot) for bigrams, two-word combinations, in sentiment analysis. A dedicated TextVectorization layer was employed, configured with ngrams=2 to consider bigrams, a maximum of 20,000 tokens, and "multi-hot" output mode. The training process and neural network architecture were similar to the unigram model, with the key distinction being the utilization of bigrams. This approach captures nuanced relationships between adjacent words, potentially enhancing the model's ability to discern sentiment patterns based on two-word combinations.

- Two Words (Bigrams) with TF-IDF Encoding:

In this model variant, we adopt TF-IDF encoding for bigrams, a departure from the previous binary encoding methods. The TextVectorization layer is specifically configured for bigrams with the "tf-idf" output mode, emphasizing not just the presence of bigrams but their relative importance. The subsequent steps, including neural network architecture and training, share similarities with earlier models, featuring a dense hidden layer with ReLU activation and dropout for regularization. This refined approach aims to elevate sentiment analysis by evaluating the significance of bigrams in conveying sentiment, building upon the foundation laid by our prior models.

5.2 Sequence Models:

- Word Embedding without Masking:

  In building this model, we employed a TextVectorization layer with a maximum token limit set at 20,000 to convert IMDb reviews into integer sequences.

These sequences, constrained to a maximum length of 600 tokens, were then embedded into a high-dimensional space using an Embedding layer. To capture contextual information comprehensively, we integrated Bidirectional LSTM layers, enabling the model to consider both forward and backward directions. For regularization purposes and to prevent overfitting, we introduced a dropout layer with a dropout rate of 0.5. The final touch involved incorporating a dense layer with a sigmoid activation function, empowering the model to perform sentiment classification by making binary predictions based on the learned representations of the input sequences.

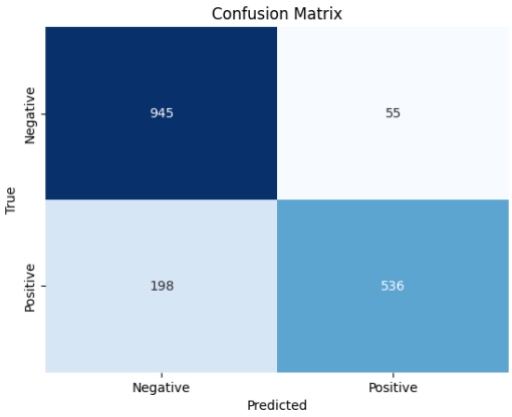
- Word Embedding with Masking:

  In contrast to the previous model, our Word Embedding with Masking model incorporates a subtle yet impactful modification by setting mask_zero=True in the Embedding layer. This adjustment enables the model to explicitly recognize and disregard padded sequences during training. By doing so, the model becomes more adept at handling varying sequence lengths, enhancing its flexibility and potentially improving performance on sequences of different sizes.

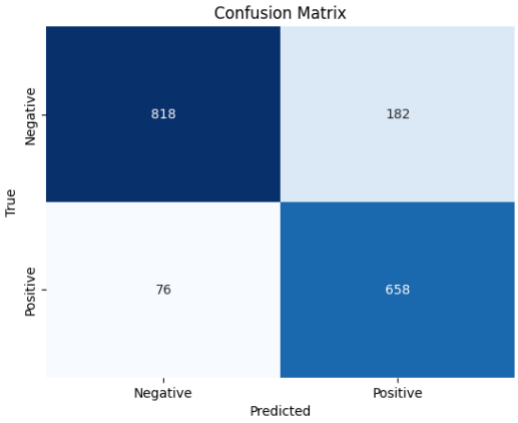
5.1 Results:

	Accuracy	Precision	Recall	F1-Score
Word Embedding with Masking	85.41%	0.91	0.73	0.81
Word Embedding without Masking	85.12%	0.78	0.90	0.84
Two Words (Bigrams) with TF-IDF Encoding	89.04%	0.85	0.90	0.87
Two Words (Bigrams) with binary Encoding (Multi-hot)	88.29%	0.86	0.87	0.86
Single Words (Unigrams) with binary Encoding (Multi-hot)	88.12%	0.86	0.86	0.86

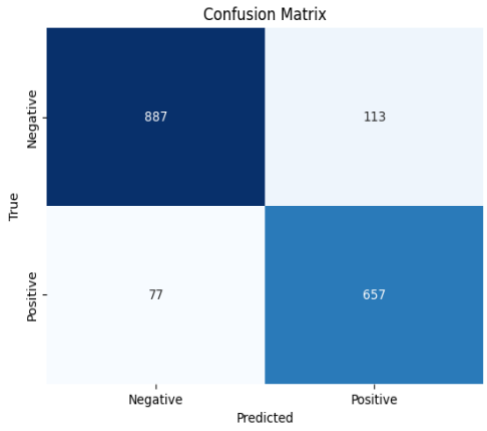
Word Embedding with Masking:



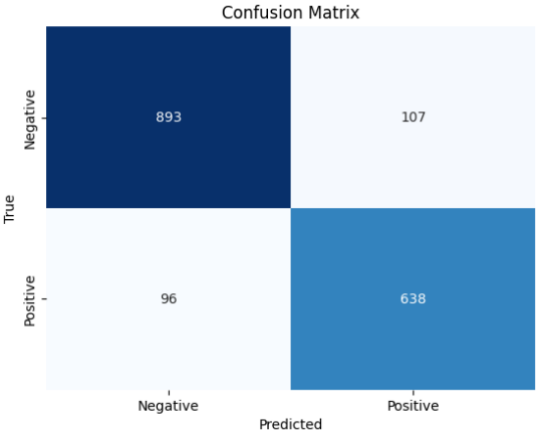
Word Embedding without Masking:



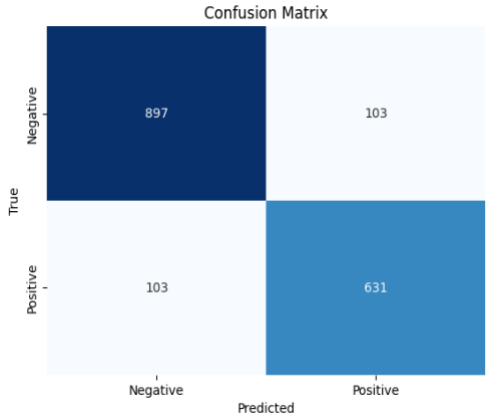
Two Words (Bigrams) with TF-IDF:



Two Words (Bigrams) with binary(Multi-hot):



Single Words (Unigrams) with binary (Multi-hot):



Conclusions

As regards the topic modeling task, the N-gram Analysis, encompassing bigrams and trigrams, provided profound insights into linguistic patterns within IMDb reviews. Visualizations, including network graphs, unveiled recurrent expressions like "('look', 'like')" and depicted co-occurrence relationships, enhancing our understanding of prevalent sentiments. LDA-based topic modeling efficiently uncovered distinct themes in movie reviews. By intelligently assigning topics and visually representing their prevalence, it facilitated a nuanced exploration of IMDb user sentiments, adding depth to our understanding.

As regards the classification task, the Two Words (Bigrams) with TF-IDF Encoding model emerges as the most effective, boasting an accuracy of 89.04% and well-balanced precision, recall, and F1-score. This outcome aligns with expectations, showcasing the model's aptitude in discerning sentiment nuances through TF-IDF-weighted bigrams. This result suggests that, for sentiment analysis in the context of IMDb reviews, the BoW approach with TF-IDF encoding for bigrams provided a more accurate and nuanced representation compared to the sequence model. The BoW approach, which treats each document as an unordered set of words or bigrams, proved to be well-suited for this task, emphasizing the importance of considering specific word combinations and their weighted impact on sentiment.

Bibliography

- ✓ Torin Retting, “NLP Sentiment Analysis & Topic Modeling for Movie Reviews’” (https://torinrettig.net/TLJ-NLP-Sentiment-Topics)
- ✓ Ashok Chilakapati, "Word Bags vs Word Sequences for Text Classification” (https://xplordat.com/2019/01/13/word-bags-vs-word-sequences-for-text-classification)
- ✓ “A Guide to Text Classification and Sentiment Analysis”, Abhijit Roy, Jul 11, 2020 (https://towardsdatascience.com/a-guide-to-text-classification-and-sentiment-analysis-2ab021796317)