University of Washington, Bothell

CSS 581: Machine Learning

Final Report: Screening Gender - A Closer Look at Representation in Movies

December 8, 2023

Poulami Das Ghosh	*2375493*
Sumit Hotchandani	*2328782*
Pratik Jadhav	*2373884*
Manpreet Kaur	*2377653*

Contents

1 Abstract

This project delves into gender representation and bias in the film industry, employing Machine Learning and Natural Language Processing techniques. It addresses issues like one-dimensional character portrayals, biased film ratings, and gender-related disparities in movie attributes. The project's significance lies in the film industry's impact on societal perceptions. Through experiments, including gender classification models and IMDb rating correlation analysis, it aims to contribute to a more inclusive film landscape. However, caution is advised due to data limitations, and the broader societal transformation beyond the film industry is envisioned in the conclusion.

2 Introduction

The film industry, as a medium of storytelling and cultural influence, has long played a significant role in shaping societal perceptions and reflecting prevailing values. However, this influential platform has also been marred by enduring issues related to gender representation and bias. This project delves into various critical facets of the film industry to address these pressing challenges with the help of Machine Learning and Natural Language Processing.

1. **Gender Representation in Movies:** A fundamental issue in the world of cinema is the portrayal of gender. Male and female characters are frequently confined to rigid stereotypes, resulting in a limited and often one-dimensional spectrum of roles. To address this issue, a gender classification model has been developed. This model distinguishes the gender of movie characters based on their dialogues, offering insights into the extent of gender representation in films. Nevertheless, it is important to recognize the limitations of this model, including potential issues with training data quality and its binary approach to gender categorization.
2. **Gender Bias in Film Ratings:** Movies with female-centric themes or strong female lead characters encounter distinctive challenges within the industry, notably the potential for bias in IMDb ratings. Such bias can impact the perception and recognition of these films. This project delves into the correlation between the degree of female-centrism in films and their IMDb ratings. It is vital to remember that IMDb ratings are subjective and can be influenced by personal preferences and biases, and the project does not imply causation but rather explores associations.
3. **Gender Bias in Movie Character Attributes:** The issue of gender bias, unequal opportunities, and harassment within the realm of movie character attributes and dialogues perpetuates stereotypes and social inequalities. An analysis of movie scripts and transcripts employs natural language processing techniques to uncover gender-related disparities in socio-economic status, equal opportunities, and instances of harassment among movie characters. This endeavor aims to raise awareness and advocate for a more equitable portrayal of gender roles in the cinematic landscape.

In the pursuit of addressing these interconnected problems, this project strives to contribute to a more inclusive, diverse, and equitable film industry. The objective is to shed light on these issues, relying on data-driven insights to foster positive change and promote an industry that accurately reflects the diversity and richness of the real world.

In achieving our stretched goals of analysing gender dynamics in movie dialogues we utilized [7]Empath, an advanced natural language processing tool. Empath dynamically generates and validates new lexical categories using deep learning on a vast dataset of modern fiction. This allowed us to gain nuanced insights into gender dynamics by exploring a broad range of topics in movie dialogues.

The data used for this project is derived from the Cornell University curated movie corpus on Kaggle. However, it is imperative to acknowledge the project's limitations:

- The availability and quality of labelled data for the gender classification model and the potential biases present in the data collected for analyzing gender bias in movies.
- Subjectivity and potential biases in IMDb ratings.

- The generalizability of findings, which may be confined to specific datasets, timeframes, genres, and geographic regions.

These limitations underline the necessity of a thoughtful and cautious interpretation of the project's results while conducting a comprehensive analysis of gender-related issues within the film industry.

In this endeavor, we have also trained models like MultinomialNB, Logistic Regression, Random Forest, and Support Vector Classification (SVC) to predict the gender of the speaker of the dialogues in the movie data set.

3 Related Work

3.1 Identifying Gender Bias in Blockbuster Movies

In their 2023 study, Haris et al.[2] explored the portrayal of gender roles in English movies, recognizing the significant influence of this medium on societal beliefs and opinions. The researchers collected scripts from various film genres and applied natural language processing techniques to derive sentiments and emotions.

The scripts were then converted into embeddings, a method of representing text in vector form. Through a detailed investigation, the team identified specific patterns in the personality traits of male and female characters in movies that align with societal stereotypes.

Their analysis revealed biases wherein men are often depicted as more dominant and envious, while women are portrayed in more joyful roles. To achieve this, they introduced a novel technique that converts dialogues into an array of emotions by combining it with Plutchik's wheel of emotions.

This study aims to encourage reflections on gender equality in the film domain and facilitate other researchers in analyzing movies automatically instead of using manual approaches. Their work represents a significant contribution to machine learning and its application in identifying gender bias in movies.

3.2 Gender Bias, Social Bias, and Representation in Bollywood and Hollywood

In a 2022 study, Khadilkar, KhudaBukhsh, and Mitchell performed an extensive analysis of gender and social biases in the English subtitles of popular Bollywood films over 70 years[3]. They applied innovative natural language processing techniques to analyze these subtitles.

The researchers also considered popular Hollywood movies and movies nominated for the Academy Awards to compare their findings. Their study uncovered social biases like colorism, son preference, and geographic and religious representation.

In addition to Bollywood and Hollywood, they analyzed critically acclaimed international films. Their work sparked a dialogue within the film community, gaining widespread media attention.

This study represents a significant contribution to machine learning and its application in identifying gender bias in movies. It provides a valuable reference for other researchers in this field.

3.3 BERT: Bidirectional Transformers for Language Understanding

In 2019, Devlin, Chang, Lee, and Toutanova introduced BERT (Bidirectional Encoder Representations from Transformers)[4], a new language representation model. Unlike previous models, BERT pre-trains deep bidirectional representations from the unlabeled text by conditioning on both the left and right context in all layers.

The pre-trained BERT model can be fine-tuned with a single additional output layer to create state-of-the-art models for various tasks, such as question answering and language inference, without significant task-specific modifications. BERT is conceptually straightforward and empirically robust.

This study represents a significant contribution to machine learning and its application in language understanding. It provides a valuable reference for other researchers in this field.

3.4 Unpacking Gender Stereotypes in Film Dialogue

In 2022, Yulin Yu, Yucong Hao, and Paramveer Dhillon[5], delved into the representation of gender stereotypes in films. They argue that these portrayals can significantly influence societal values and beliefs, reflecting and potentially reinforcing prevailing social norms. The paper decomposes the gender differences portrayed in movies along several socio- and psycho-linguistic dimensions, including the degree of assertion, the degree of confirmation, the valence of emotions, and the topic.

The empirical analyses reveal that the valence of emotions expressed in the dialogue explains the most variation in gender disparity. For certain kinds of dialogue, such as those occurring between different gender actors, the topic of discussion is also a strong predictor of gender differences. The authors find that women tend to use more tentative words, give positive responses to their interlocutors, express more positive emotions, and discuss their families, friends, and acquaintances. In contrast, men are more assertive in expressing their opinions, more likely to interrupt their interlocutors, show more negativity in their emotions, and are more preoccupied with showing off their work and achievements.

4 Data

4.1 Data Overview

4.1.1 Cornell Movie Dialog Corpus

The dataset used for this project is a rich collection of fictional conversations extracted from movie scripts curated by Cornell University[1], which is available on Kaggle.

It contains:

- **220,579** conversational exchanges between **10,292** pairs of movie characters.
- Involves **9,035** characters from **617** movies.
- In total, **304,713** utterances i.e., is a line of dialogue spoken by a character in a movie scene.

The dataset also includes metadata for both movies and characters:

- **Movie metadata:** genres, release year, IMDB rating, number of IMDB votes.
- **Character metadata:** gender (for 3,774 characters), position on movie credits (3,321 characters).

The dataset is divided into several files:

1. **movie_titles_metadata.txt:** Contains information about each movie title, including movieID, movie title, movie year, IMDB rating, number of IMDB votes, and genres.
2. **movie_characters_metadata.txt:** Contains information about each movie character, including characterID, character name, movieID, movie title, gender, and position in credits.
3. **movie_lines.txt:** Contains the actual text of each utterance, including lineID, characterID (who uttered this phrase), movieID, character name, and text of the utterance.
4. **movie_conversations.txt:** Contains the structure of the conversations, including characterID of the first character involved in the conversation, characterID of the second character involved in the conversation, movieID of the movie in which the conversation occurred, and list of the utterances that make the conversation, in chronological order.
5. **raw_script_urls.txt:** Contains the URLs from which the raw sources were retrieved.

4.1.2 IMDB Non-Commercial Data

Based on the feedback received on our mid-term progress, we could also explore how the director's gender could impact the bias against women depicted in the movies. To achieve this goal, we obtained a free

non-commercial version of IMDB's movie details dataset[6]. The dataset includes many files, out of which 3 which served the purpose for us were used:

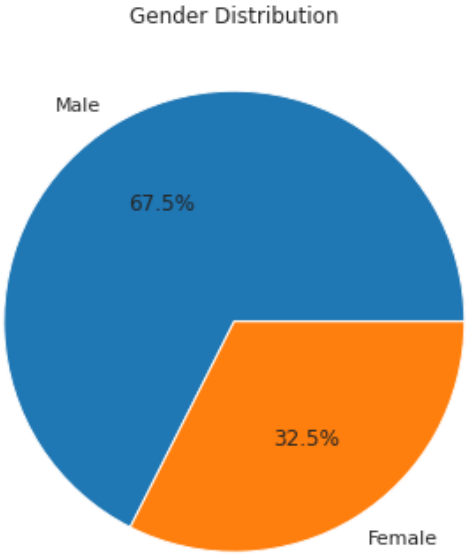
1. **title.basics.tsv:** Contains details about movies/tv show titles e.g. release year, runtime, genres
2. **title.crew.tsv:** Contains metadata of the crew of a movie (connecting data set with movie and crew IDs)
3. **name.basics.tsv:** Contains details about the crew e,g, name, year, primary profession

4.2 Exploratory Data Analysis

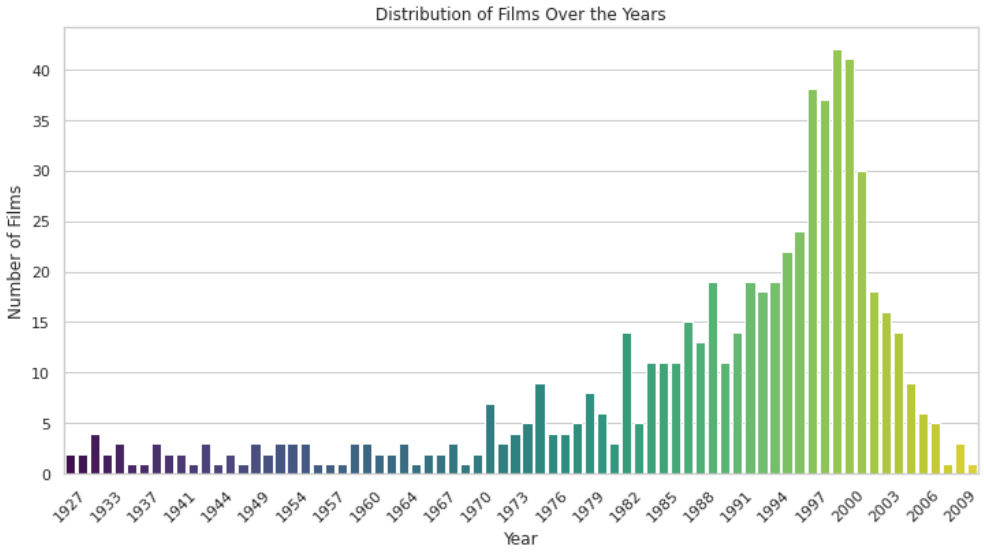
In the following section we have analyzed and visualized datasets with the primary goal of summarizing their main characteristics, often employing statistical and graphical methods.

Based on the initial profiling of the data, below are some key observations:

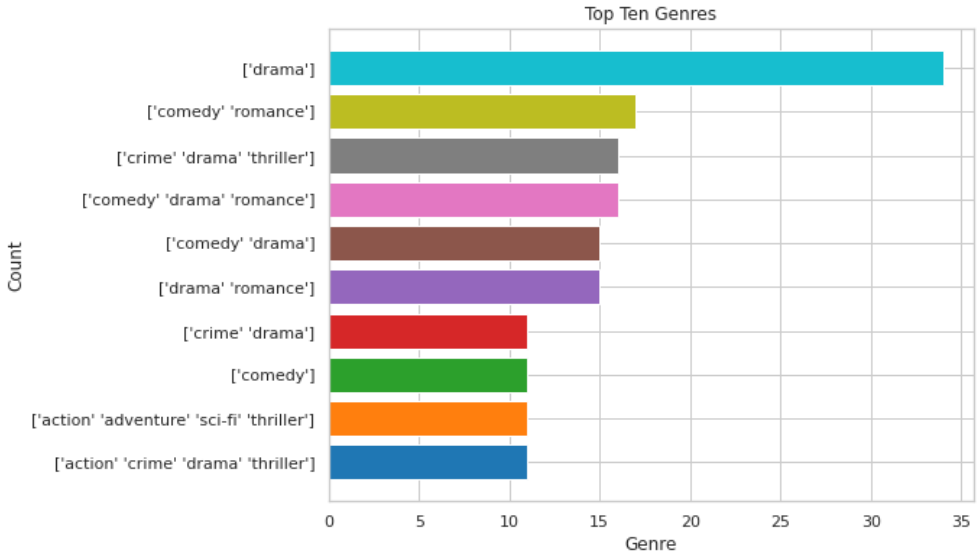
1. **Gender Distribution:** The graph shows % of 'Male' characters to 'Female' characters. Out of 9,035 characters, we have gender labels for 3,026 characters - with a **67.5% (2,044)** to **32.5% (982)** split for 'Male' to 'Female'. This data indicates a disparity in the representation of genders, with 'Male' characters being more prevalent. This aligns well with what we want to analyze from the data and build our model accordingly



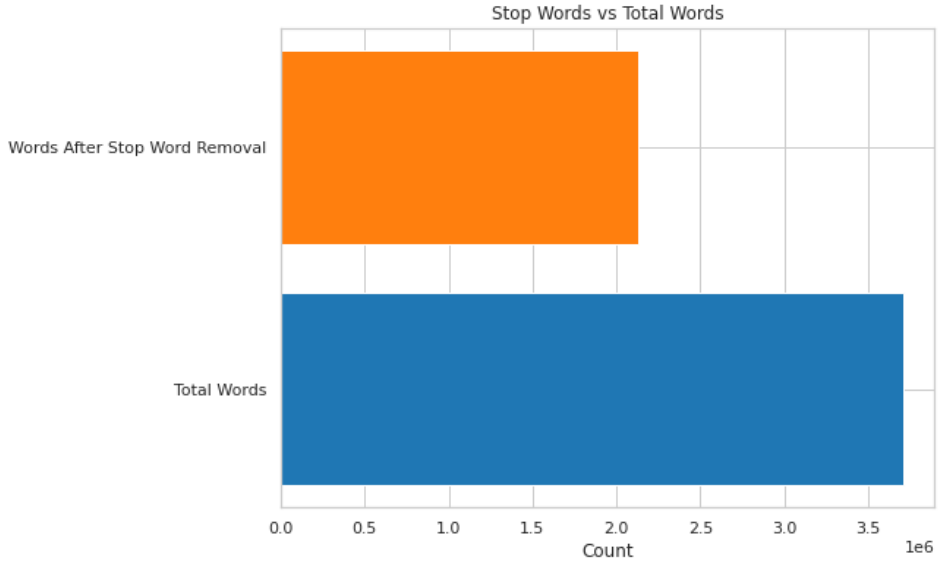
2. **Frequency of movies by year:** The graph depicts the availability of movies by year in the dataset. The data ranges from the year 1926 all the way upto 2010. While we have wide-ranging data, the majority of it lies in the 3 decades of the 1980s, 1990s and 2000s, with **490 of the 617 i.e., ~79%** of the movies in that period. This sample space has enough of a time-spread for us to analyze how representation of women and the general attitude towards them in movies has changed over time.



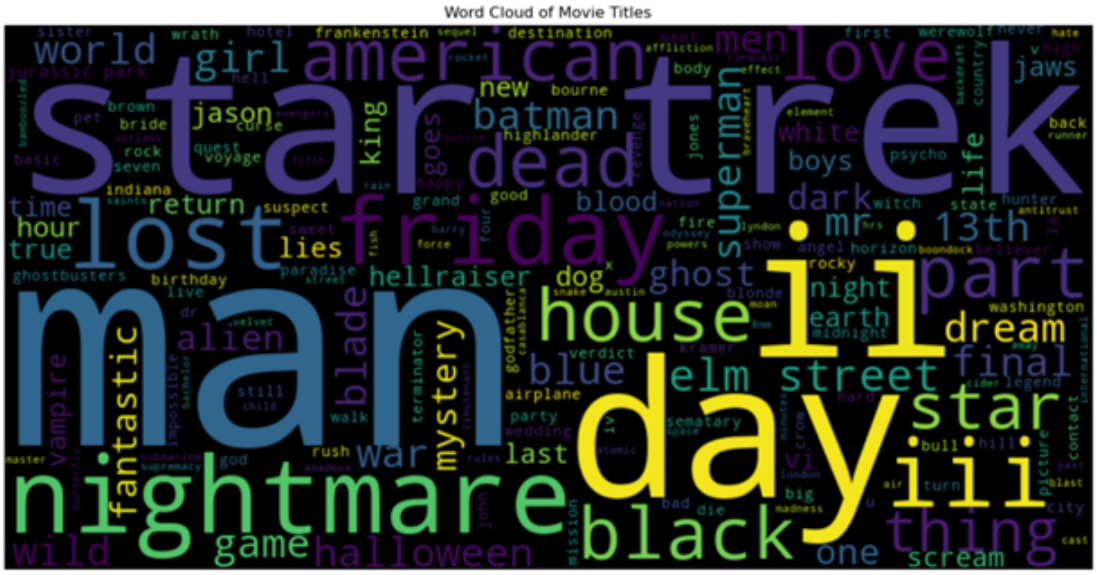
3. **Distribution of Genres:** Distribution of genres can help us understand the types of movies included in the dataset. and how they might affect the representation and portrayal of different genders. For instance, we can see that the most common genre is drama, which could indicate that the dataset contains more movies that focus on the emotional and interpersonal aspects of the characters, rather than the action or adventure, which would give us a better foundation for our analysis.



4. **Stop words vs Total words:** A substantial difference was observed between the overall word count in all dialogues and the count after eliminating common words using the English stopword removal dictionary from the NLTK library. By excluding frequent words such as ”and,” ”the,” ”is,” and so on, we aimed to streamline the word vector for the training of the machine learning model. This process helps in creating a more focused and efficient representation of the language for improved model learning.



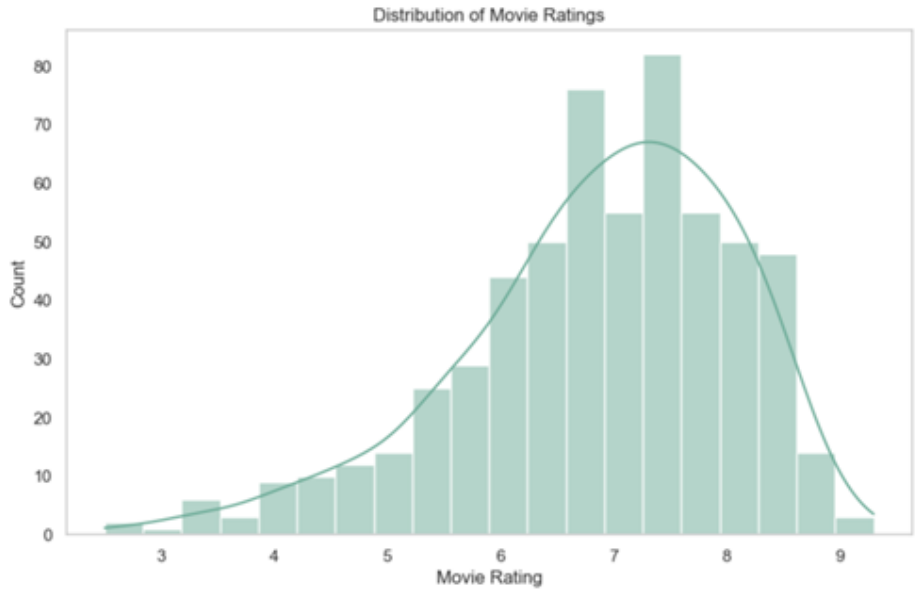
5. **Movie Titles Wordcloud:** A word cloud was created based on movie titles to identify the most frequent words, revealing prevalent themes in the dataset, such as science fiction, horror, action, and adventure genres.



The most frequent words in the movie titles are *“star”, “trek”, “black”, “men”, “day”, “night”, “american”, “dead”, “thing”, “house”, “nightmare”, “elm”, “street”, “batman”, “jurassic”, “park”, “wild”, “lost”, “world”, “brown”, “jaws”, “mission”, “impossible”, “voyage”, “earth”, “game”, “happy”, “hotel”, “meet”, “hard”, “scream”, “basic”, “rush”, “13th”, “part”, “final”, “blue”, “hellraiser”, “dark”, “friday”, “thing”, “scream” *.

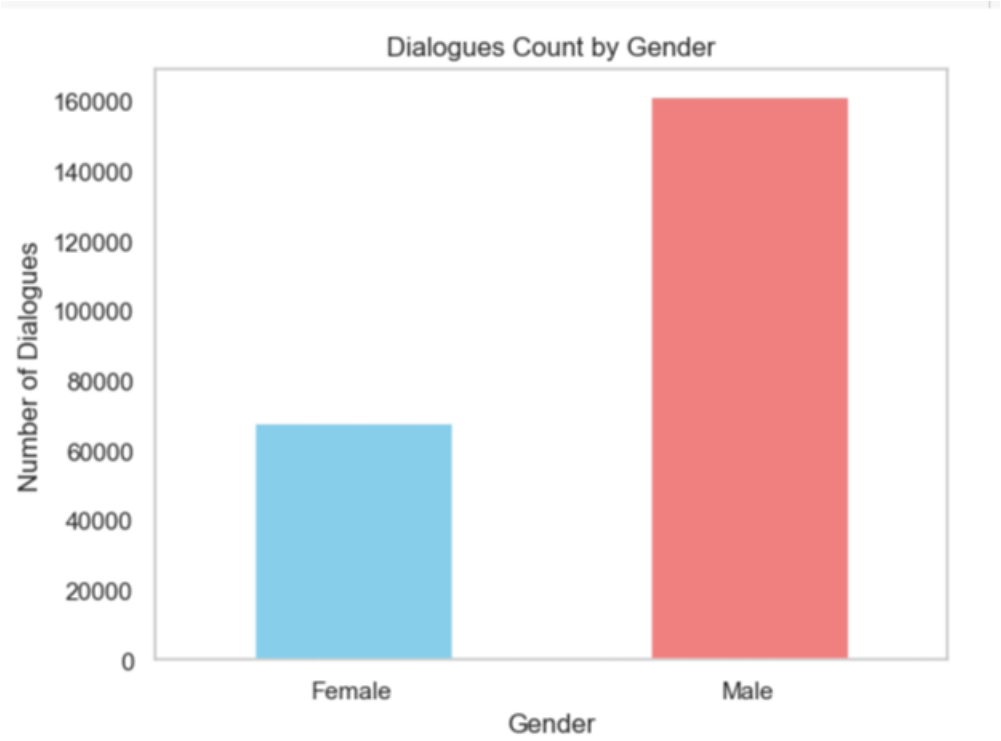
Based on this word cloud, we can conclude that the most common themes in the movie titles are related to science fiction, horror, action, and adventure genres. The words “star”, “trek”, “black”, “men”, “day”, “night”, “american”, “dead”, “thing”, “house”, “nightmare”, “elm”, “street”, “batman”, “jurassic”, “park”, “wild”, “lost”, “world”, “brown”, “jaws”, “mission”, “impossible”, “voyage”, “earth”, “game”, “happy”, “hotel”, “meet”, “hard”, “scream”, “basic”, “rush”, “13th”, “part”, “final”, “blue”, “hellraiser”, “dark”, “friday”, “thing”, and “scream” are indicative of the most popular movie franchises and titles in these genres.

6. **Movie Ratings Distribution:** A histogram illustrated the distribution of movie ratings, indicating the highest-rated movie with a rating of 9.5.



Based on the graph, we can conclude that the majority of the movies have a rating between 6 and 8. The highest number of ratings is around 7.5. This suggests that most movies are rated between 6 and 8, with a few outliers on either end. 9.5 is the highest rating received by any movie in our current dataset.

7. **Dialogue Count per Gender :** The bar graph illustrates the distribution of dialogues between female and male characters.The data indicates that male characters have significantly more dialogues (161244) compared to female characters (67862). This suggests a potential gender imbalance in the dataset or narrative focus.



The substantial difference in dialogue counts raises questions about the representation of female characters in the context of the dataset. The analysis may prompt further investigation into the storyline, genre, or specific characters contributing to these numbers.

4.3 Analysis of Movie Ratings by Gender of Directors

The analysis focuses on exploring trends in movie ratings based on the gender of directors. By comparing the performance of female and male directors across different genres, we aim to uncover potential patterns and disparities in audience reception.

1. **Unique Movies by Gender:** The analysis began by examining the number of unique movies made by female and male directors. The results indicated a significant class imbalance, with a higher number of unique movies made by male directors.

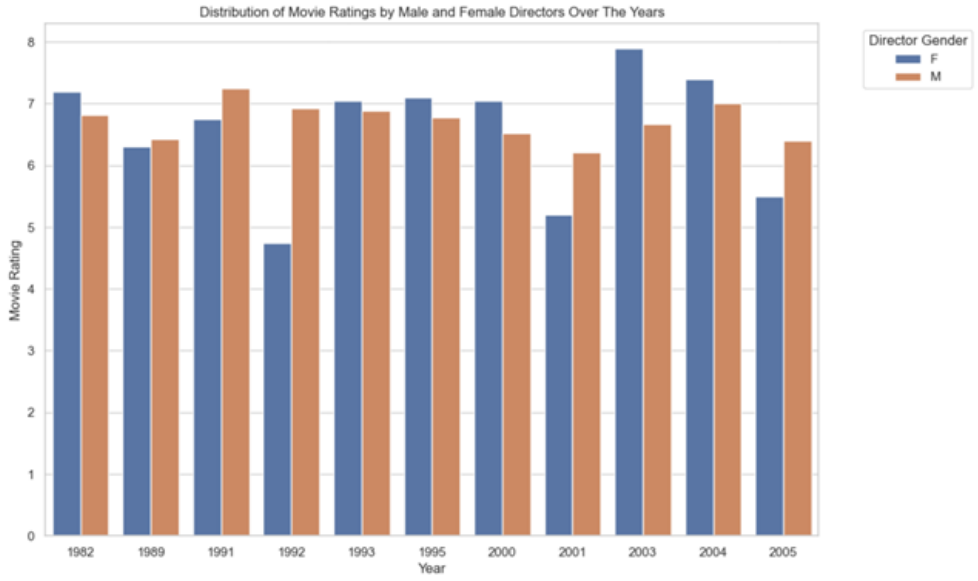
   - Number of unique movies made by female directors: 98
   - Number of unique movies made by male directors: 458

   Additionally, there were no common movies made by both male and female directors in the same year and genre.
   Genre Performance:
   Drama and Drama-Thriller
   • Female directors exhibited superior performance in genres such as Drama and Drama-Thriller, showcasing their proficiency in nuanced storytelling.

   Action Genre
   • Notably, there was a striking absence of female directors in the action genre. This aligns with societal gender norms and prevalent audience preferences.

   Other Genres
   • In other common genres, male directors tended to outperform their female counterparts. This outcome could be influenced by perceived differences in work quality or existing biases, particularly evident in genres like sports and crime.

2. **Rating Trends Over Years:** Despite a detailed analysis, no discernible trend in movie ratings emerged over the years based on the gender of the director. The inconclusiveness is attributed to the insufficient data available for films directed by females.

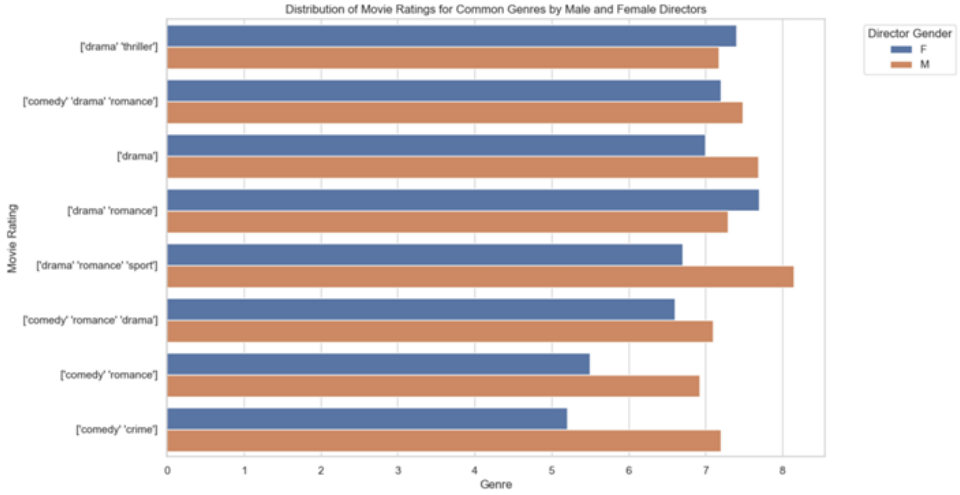


The bar plot showcases the distribution of movie ratings for common years, providing insights into potential trends and variations.

The visual comparison allows for the identification of patterns or shifts in audience reception based on the gender of directors.

Understanding how movie ratings vary over the years based on the gender of directors is crucial for assessing industry dynamics and audience preferences. The visual representation facilitates a comprehensive analysis, guiding future research and initiatives aimed at promoting diversity and inclusivity in filmmaking.

3. **Distribution of Movie Ratings for Common Genres:** To further explore the relationship between movie ratings and the gender of directors, a bar plot was created for the distribution of movie ratings in common genres. The plot visually compares the movie ratings for common genres by male and female directors.



Key Insights

Genre Performance
1. Drama and Drama-Thriller: Female directors demonstrated strength in genres requiring nuanced storytelling, such as Drama and Drama-Thriller.

2. Action Genre: The absence of female directors in the action genre suggests a gender gap in film genres aligned with traditional expectations.

3. Other Genres: Male directors tended to outperform in common genres, potentially influenced by perceived differences in work quality or existing biases.

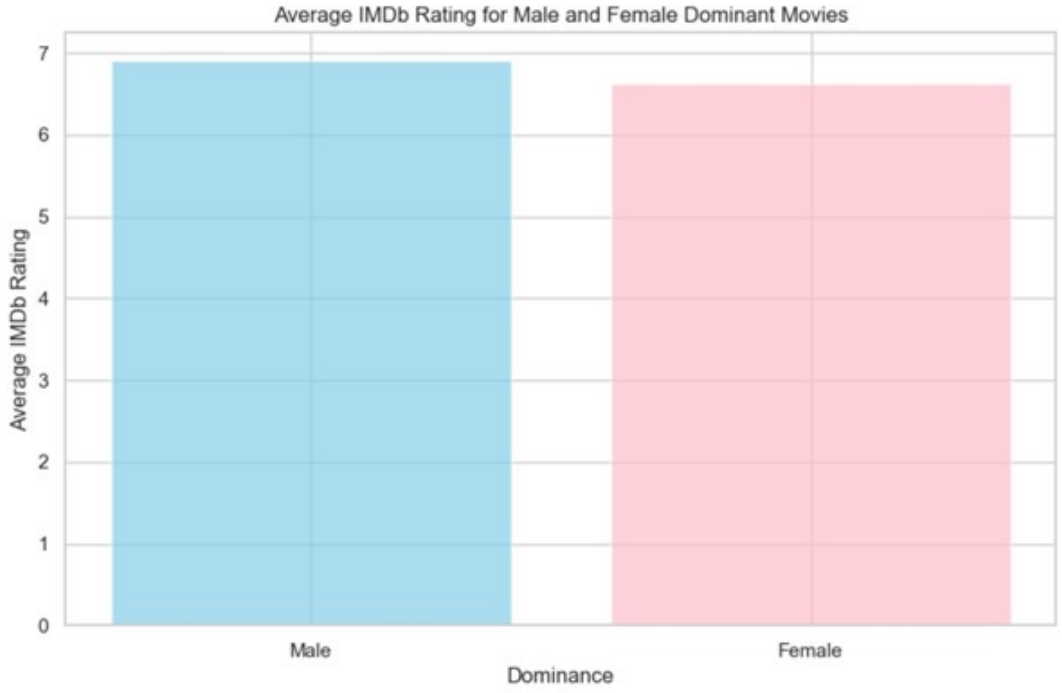
Rating Trends Over Years
• The inconclusive trend in movie ratings over the years highlights the need for more comprehensive data on films directed by females.

4.4 Analysis of Average IMDB Movie Ratings by Gender-Dominant Characters in Movies

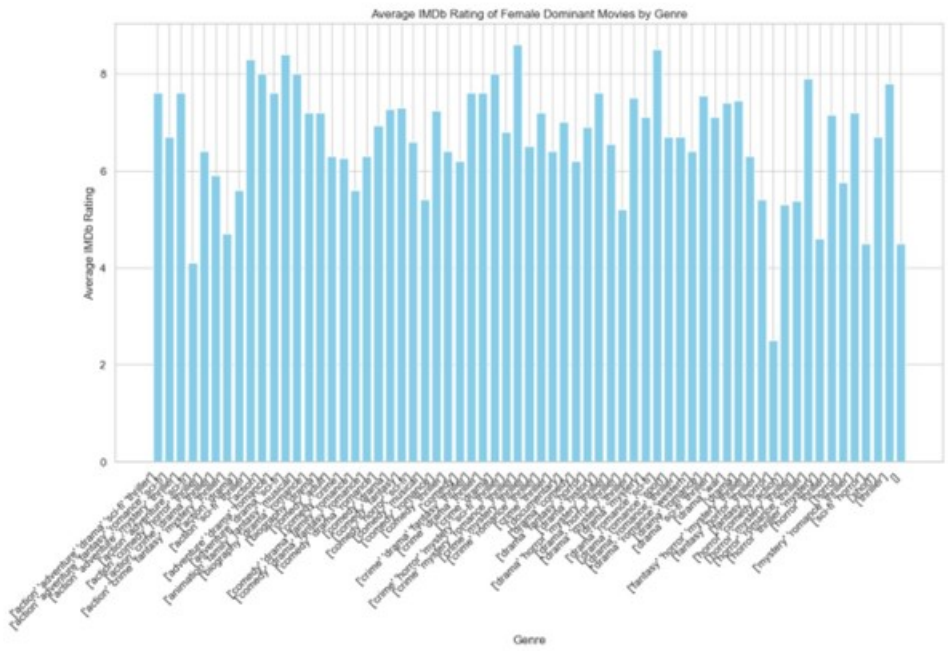
In this analysis, we delve into the intriguing relationship between gender-dominant characters in movies and their impact on average IMDB movie ratings. By exploring the intersection of cinematic representation and audience reception, we aim to uncover patterns and insights that shed light on the influence of gender dynamics on audience perceptions and the overall cinematic experience.

1. **Average IMDB Movie Ratings for Male and Female Dominant Movies:** Male Dominant Movies: The average IMDb rating for male-dominant movies is represented by the sky-blue bar. Male-dominant movies, where male characters have a higher dialogue word count, tend to have an average IMDb rating around 6.9

   Female Dominant Movies: The pink bar depicts the average IMDb rating for female-dominant movies. Female-dominant movies, characterized by a higher dialogue word count for female characters, show an average IMDb rating around 6.62

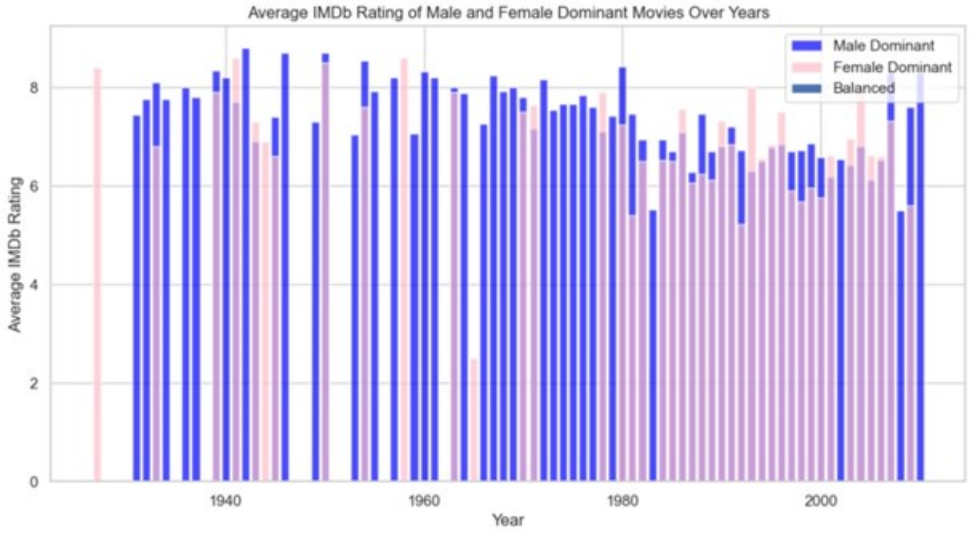


2. **Average IMDb Rating of Female Dominant Movies by Genre:** The analysis of the bar graphs reveals interesting insights into the average IMDb ratings of movies based on dominance (Male Dominant, Female Dominant, or Balanced) within different genres:



The analysis of the bar graphs reveals interesting insights into the average IMDb ratings of movies based on dominance (Male Dominant, Female Dominant, or Balanced) within different genres: Male Dominant Movies: Genres such as Action, Adventure, and Sci-Fi tend to have higher average IMDb ratings for male-dominant movies. These genres, characterized by intense and action-packed narratives, seem to resonate well with audiences when males are prominently featured in dialogue. Female Dominant Movies: Genres like Romance, Comedy, and Drama exhibit higher average IMDb ratings for female-dominant movies. This suggests that movies dominated by female characters, often associated with emotionally engaging storylines, receive positive audience feedback in these genres. Balanced Movies: In genres such as Mystery, Thriller, and Fantasy, where the dialogue distribution is more balanced between male and female characters, movies tend to have competitive IMDb ratings.

3. **Average IMDb Rating of Male and Female Dominant Movies Over Years:** The bar graph illustrates the average IMDb ratings of movies categorized as Male Dominant, Female Dominant, and Balanced (with equal representation of male and female dialogue) over the years.



Male Dominant Movies: Male-dominant movies consistently maintain a relatively stable average IMDb

rating across the years. This suggests that the dominance of male characters in dialogue does not strongly correlate with fluctuations in IMDb ratings over time.

Female Dominant Movies: Female-dominant movies show a fluctuating pattern in average IMDb ratings over the years. Peaks and troughs in ratings indicate that the success of female-dominant movies may be influenced by various factors, such as evolving audience preferences or trends.

There is no clear pattern indicating that movies with a particular gender dominance consistently outperform others.

5 Methodology

5.1 Data Cleansing & Pre-processing

1. **Filtering Movie Data in the IMDB Dataset:** After loading the data into appropriate dataframes, we focused on the IMDB dataset. Given its diverse content types, the dataset needed to be filtered for movies. This step is crucial to prevent duplicate entries resulting from multiple title types with the same name. The primary dataset exclusively contains movies, ensuring a seamless join operation on the original dataset.

```
imdb_movie_title_df = imdb_movie_title_df[imdb_movie_title_df.titleType == 'movie']

imdb_df = imdb_movie_title_df.merge(imdb_movie_crew_df, on='tconst', how='left')
.merge(imdb_crew_name_df, left_on='directors', right_on='nconst', how='left')

titles_df['movie_name_upper'] = titles_df['movie_name'].str.upper()
imdb_df['primaryTitle_upper'] = imdb_df['primaryTitle'].str.upper()

movie_titles_directors_df = titles_df.merge(imdb_df,
    left_on=['movie_name_upper', 'movie_year'],
    right_on=['primaryTitle_upper', 'startYear'], how='left')
    [['movie_id', 'movie_name', 'movie_year', 'movie_rating', 'movie_genre', 'primaryName']]
```

2. **Identifying Genders of Movie Directors**

   - **Gender Identification Function :** To enhance our dataset, we implemented a gender identification mechanism for movie directors based on their Wikipedia biographies. The identify_gender(text) function was developed to assess a given text for the frequency of male and female pronouns, providing an inferred gender ('M', 'F', or 'Unknown').

```
def identify_gender(text):
    male_pronouns = ['he', 'him', 'his']
    female_pronouns = ['she', 'her', 'hers']

    words = text.lower().split()

    male_count = len([w for w in words if w in male_pronouns])
    female_count = len([w for w in words if w in female_pronouns])

    if male_count > female_count:
        return 'M'
    elif female_count > male_count:
        return 'F'
    else:
        return 'Unknown'

```

   - **Web Scraping Wikipedia Biographies :** The code iterates over entries in the `movie_titles_directors_df` dataframe, constructing Wikipedia URLs for each director, and retrieving the content of associated biographies. Special handling considers scenarios with only one paragraph on the Wikipedia page, fetching additional information by modifying the URL to include the director's profession.

```
for index, row in movie_titles_directors_df.iterrows():
  dg=[]
  paragraphs = []
  url = "https://en.wikipedia.org/wiki/" + row['director_name'].replace(" ", "_")
  response = requests.get(url)
  soup = BeautifulSoup(response.content, "html.parser")

  if len(soup.findAll("p")) == 1:
    paragraphs.clear()
    url = "https://en.wikipedia.org/wiki/" + row['director_name']
            .replace(" ", "_") + "_(director)"
    response = requests.get(url)
    soup = BeautifulSoup(response.content, "html.parser")

  paragraphs = (soup.findAll("p"))
  for (i, p) in enumerate(paragraphs):
    dg.append(identify_gender(p.text))

  if 'Unknown' in dg:
    dg = [i for i in dg if i != 'Unknown']

  if len(dg) == 0:
    dg.append('Unknown')

  movie_titles_directors_df.loc[index, 'director_gender'] = mode(dg)
  dg.clear()

```

- **Data Refinement:** Instances where the gender inference results in 'Unknown' are addressed by either removal or assignment of 'Unknown' if all results are inconclusive. Inferred genders are then incorporated into the `movie_titles_directors_df` dataframe under the `director_gender` column.
- **Considerations:**
  - Code accommodates varying Wikipedia page structures, adapting to different scenarios for reliable gender identification.
  - Special handling of 'Unknown' instances ensures completeness and accuracy in the final dataset.
  - Designed to scale for the entire dataset, providing a comprehensive gender analysis of movie directors.

3. **Cleaning the Movie Year Column:** The `movie_year` column, representing the year of movie release, must be numeric and in `YYYY` format. As part of pre-processing, the year part is extracted from rows where it is appended with a string.

```
column_name = 'movie_year'

# Filter non-numeric values in the specified column
non_numeric_values = final_df.loc[~pd.to_numeric(final_df[column_name], errors='coerce')
.notna(), column_name]

# Get unique non-numeric values
unique_non_numeric_values = non_numeric_values.unique()

# Display the unique non-numeric values
print(unique_non_numeric_values)

['1989/I' '1990/I' '1995/I' '1998/I' '2004/I' '2007/I' '1992/I' '2005/I'
 '2002/I' '1968/I' '1996/I' '2000/I' '2009/I' '2003/I']

# Extract numeric portion using regular expression
final_df[column_name] = final_df[column_name].str.extract('(\d+)', expand=False)
final_df[column_name] = final_df[column_name].astype(int)
```

```
array([1999, 1992, 2001, 1968, 1982, 1997, 1988, 1989, 1959, 1980, 1986,
       1984, 1981, 1932, 2000, 1998, 1991, 1975, 2003, 2006, 1996, 2004,
       1995, 1942, 1974, 1990, 1933, 1993, 2005, 1931, 2009, 2002, 1967,
       1971, 1979, 1987, 1940, 1961, 2007, 1953, 1934, 1983, 1994, 1985,
       1976, 1937, 1955, 1970, 1941, 1927, 1939, 1936, 1954, 1949, 1943,
       2010, 1977, 1972, 1963, 1945, 2008, 1960, 1964, 1950, 1973, 1966,
       1944, 1978, 1946, 1969, 1965, 1957, 1958])
```

4. **Removing Ambiguity From The Gender Column:** In this project, gender is categorized into two groups: "Male" (M) and "Female" (F). The initial data analysis revealed five different values in the `character_gender` column ('?', 'm', 'f', 'M', 'F'). Further investigation confirmed 'm' & 'M' as male (M) and 'f' & 'F' as female (F). Rows with '?' as gender are dropped since the `character_id` serves as the identifier column, and there is no `character_id` with multiple genders.

```
column_name = 'movie_year'
# Count of values of each gender type
final_df.character_gender.value_counts()
character_gender
m    145845
f     63284
?     59044
M     15399
F      4578

# Removing data ambiguity from gender
final_df = final_df[final_df.character_gender != '?']
final_df.character_gender = final_df.character_gender
.apply(lambda g: 'M' if g in ['m', 'M'] else 'F')
final_df.groupby('character_gender')['character_id'].nunique()

character_gender
F     948
M    2003
```

5. **Removing Non-numeric Values from the position_credits Column:** The `position_credits` column, representing the position where the character features in the credits, is numeric. Rows containing '?' are replaced with '-1' to indicate that the character did not appear in the credits. This ensures uniformity and supports downstream analyses.

```
non_numeric_position_credits = final_df.loc[~pd.to_numeric(final_df['position_credits'],
errors='coerce').notna(), 'position_credits'].unique()
#Examine the distribution of male & female with this junk data
filtered_df = final_df[final_df['position_credits'] == '?']
# Count the number of males and females
male_count = filtered_df[filtered_df['character_gender'] == 'M']
female_count = filtered_df[filtered_df['character_gender'] == 'F']
M    222
F    102

final_df['position_credits'] = final_df['position_credits'].replace('?', -1).astype(int)
```

5.2 Feature Engineering

1. **Pipeline for Numeric Features:** All numeric features, encompassing `dialog_length_median`, `dialog_word_count_median`, `character_id_count`, and `movie_year`, have been meticulously isolated from the ultimate dataframe for subsequent processing. This isolation is executed through a systematic methodology embedded within a dedicated data transformation pipeline. The pipeline incorporates the `MinMaxScaler`, strategically embedded within the settings of the Pipeline class. Serving as a facilitator, the `MinMaxScaler` ensures the conformity of all numeric values to a specified range, typically between 0 and 1. This procedural step assumes particular significance in the realm of machine learning, addressing challenges associated with variations in the scales of numeric details. In essence, the resultant numeric

transformer pipeline acts as a proficient aide, meticulously preparing these numeric features for seamless integration into broader modeling endeavors.

```
numeric_features = ['dialog_length_median', 'dialog_word_count_median'
,'character_id_count', 'movie_year']

# Created pipeline transformer for scaling numeric data
numeric_transformer = Pipeline(steps=[('scaler', MinMaxScaler())])
```

2. **Pipeline for Non-Numeric Features:** The features `director_gender` and `movie_genre` have been discerningly selected for ordinal encoding, signifying a recognition of their categorical nature. To adeptly handle these features, a bespoke `OrdinalEncoderTransformer` has been devised, leveraging the capabilities of the `OrdinalEncoder` with specific configurations. This transformer is engineered to effectively manage unknown values by assigning an encoded value and employing -1 as a placeholder for unknown values. During the fitting phase, the `OrdinalEncoderTransformer` undergoes training on the designated columns. In the subsequent transformation phase, it applies ordinal encoding, creating a duplicate of the data with numerical representations of categorical information. This tailored approach ensures the preservation of ordinal relationships between different categories. The resulting `label_transformer`, embodying the application of the `OrdinalEncoderTransformer` to the specified label features, emerges as a potent tool for seamlessly integrating categorical information into broader analytical or modeling processes.

   The dataframe column `position_credits` is identified as a nominal feature suitable for one-hot encoding. To address this nominal data effectively, a dedicated data transformation pipeline is established, housing a solitary transformer—the OneHotEncoder. This widely utilized technique converts categorical data into a binary matrix, where each category is represented by a binary column. Setting `ignore` in the `handle_unknown` parameter ensures that any unforeseen categories encountered during the transformation process are disregarded, preventing potential errors. Essentially, through the creation of this nominal transformer pipeline, the categorical nature of the `position_credits` feature is aptly accommodated, preparing it for seamless integration into our models.

```
label_features = ['director_gender', 'movie_genre']

class OrdinalEncoderTransformer(BaseEstimator, TransformerMixin):
    def __init__(self, columns):
        self.columns = columns
        self.ordinal_encoder = OrdinalEncoder(handle_unknown='use_encoded_value',
        unknown_value=-1)

    def fit(self, X, y=None):
        self.ordinal_encoder.fit(X[self.columns])
        return self

    def transform(self, X):
        X_copy = X.copy()
        X_copy[self.columns] = self.ordinal_encoder.transform(X[self.columns])
        return X_copy

# Create a pipeline transformer for ordinal data
label_transformer = OrdinalEncoderTransformer(columns=label_features)
```

3. **TF-IDF Vectorization :** The dataset, particularly the feature labeled `cleaned_dialogue_<lambda>`, is earmarked for further processing within the domain of natural language data. To facilitate this, a purpose-designed data transformation pipeline, `vectorizer_transformer`, is instituted. This pipeline orchestrates a two-step procedure. In the initial step, a custom transformer, `Converter`, flattens the input data frame into a one-dimensional array, ensuring compatibility with subsequent stages. The subsequent step incorporates a `TfidfVectorizer`, a specialized tool adept at transforming text data into a numerical format based on the Term Frequency-Inverse Document Frequency (TF-IDF) algorithm. TF-IDF captures the significance of each term within a document relative to a broader collection of documents, thereby facilitating the extraction of meaningful features for subsequent natural language processing tasks. Essentially, the `vectorizer_transformer` pipeline is crafted to

prepare the `cleaned_dialogue_<lambda>` feature for TF-IDF vectorization, marking a fundamental pre-processing step in the realm of text-based machine learning applications.

```
class Converter(BaseEstimator, TransformerMixin):
    def fit(self, x, y=None):
        return self

    def transform(self, data_frame):
        return data_frame.values.ravel()

vectorizer_features = ['cleaned_dialogue_<lambda>']

# Creating pipeline transformer to feed input to Vectorizer and then generating Tfid distribution
vectorizer_transformer = Pipeline(steps=[('con', Converter()),('tf', TfidfVectorizer())])
```

In the ultimate stage, a comprehensive pipeline named "Preprocessing" is created, orchestrating the application of all previously developed pipelines on the data. This cohesive and tailored approach handles diverse types of features, ensuring their appropriate transformation and integration into a unified and preprocessed dataset. This meticulously prepared dataset stands ready for deployment in subsequent machine learning tasks, embodying a harmonious amalgamation of numerical, ordinal, nominal, and textual features.

5.3 Model Building

In this project, we have built the following models to predict the gender of the speaker of the dialogue.

- MultinomialNB
- Logistic Regression
- Random Forest
- Support Vector Classification (SVC)

5.3.1 Model Training and Hyperparameter Tuning

Each model is meticulously fine-tuned for optimal performance using the RandomizedSearchCV technique. This approach involves a systematic exploration of the hyperparameter space through random sampling, complemented by cross-validation to ensure a robust evaluation. The primary objective is to automate the identification of the best hyperparameter configuration, thereby enhancing model efficiency and fine-tuning.

5.3.2 Code

```
model_name = 'Random Forest'

# Define the pipeline with the preprocessor and the Random Forest model
pipeline_rf = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier())
])

# Define the hyperparameter grid for RandomizedSearchCV
param_dist = {
    'classifier__n_estimators': [50, 100, 150, 200],
    'classifier__max_depth': [None, 10, 20, 30, 40, 50],
    'classifier__min_samples_split': [2, 5, 10],
    'classifier__min_samples_leaf': [1, 2, 4]
}

# Set up RandomizedSearchCV
model_rf = RandomizedSearchCV(
    pipeline_rf,
    param_distributions=param_dist,
    n_iter=10,
```

```
    cv=5,
    n_jobs = -1,
    random_state=42
)

model_rf.fit(X_train, y_train)

best_params = model_rf.best_params_

# Create a new pipeline with the best parameters
best_pipeline_rf = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(
        n_estimators=best_params['classifier__n_estimators'],
        max_depth=best_params['classifier__max_depth'],
        min_samples_split=best_params['classifier__min_samples_split'],
        min_samples_leaf=best_params['classifier__min_samples_leaf']
    ))
])

# Fit the model on the entire training dataset
best_pipeline_rf.fit(X_train, y_train)

# Predict on the test set
y_pred_rf = best_pipeline_rf.predict(X_test)

# Generate classification report
print("Classification Report for ", model_name)
print(classification_report(y_test, y_pred_rf))

# Map labels to binary format
label_mapping = {'F': 0, 'M': 1}
y_test_binary_rf = y_test.map(label_mapping)
y_pred_binary_rf = pd.Series(y_pred_rf).map(label_mapping)

# Print accuracy on the test set
accuracy_rf = accuracy_score(y_test_binary_rf, y_pred_binary_rf)
print(f"{model_name} Accuracy: {accuracy_rf:.3f}")

# Calculate AUC
y_pred_proba_rf = best_pipeline_rf.predict_proba(X_test)[:, 1]
auc_rf = roc_auc_score(y_test_binary_rf, y_pred_proba_rf)
print(f"{model_name} AUC: {auc_rf:.3f}")

# Plot ROC curve
fpr_rf, tpr_rf, _ = roc_curve(y_test_binary_rf, y_pred_proba_rf)
plt.figure(figsize=(8, 8))
plt.plot(fpr_rf, tpr_rf, color='darkorange', lw=2, label='ROC curve (area = {:.2f})'.format(auc_rf))
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title(f'ROC Curve for {model_name}')
plt.legend(loc="lower right")
plt.show()

# Calculate the confusion matrix
conf_matrix_rf = confusion_matrix(y_test_binary_rf, y_pred_binary_rf)

# Plot the confusion matrix
plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix_rf, annot=True, fmt='d', cmap='Blues', cbar=False,
            xticklabels=['F', 'M'], yticklabels=['F', 'M'])
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title(f'Confusion Matrix for {model_name}')
plt.show()
```

For brevity's sake, we have only included the code for our best model, the code structure remains the same

for all other models.

5.3.3 Performance Metrics

For each model, a comprehensive set of performance metrics is reported, providing a nuanced evaluation of their effectiveness. The metrics encompass Precision, Recall, F1-Score, Support, Accuracy, Area Under the ROC Curve (AUC), and the Confusion Matrix. This comprehensive reporting offers insights into the models' capacity to accurately classify gender labels and provides a holistic understanding of their strengths and limitations.

5.3.4 Addressing Class Imbalance

Initial model training is conducted on unbalanced data, characterized by a gender distribution of 2003 'M' (male) samples versus 948 'F' (female) samples. Recognizing the challenges posed by class imbalance, various re-sampling techniques, including RandomUnderSampler, RandomOverSampler, and SMOTE, are applied to generate balanced datasets. Subsequently, the models are retrained on each of these modified datasets to assess their performance under balanced conditions. This strategic approach aims to mitigate biases arising from imbalanced gender labels and improve the models' generalization capabilities.

Here is a comparative analysis of each technique employed:

1. **No Class Balancing:** In assessing the performance of four classification models — Naive Bayes, Logistic Regression, Random Forest, and Support Vector Machine (SVM) — for character gender identification using the given data, the key insights are as follows. The dataset's substantial class imbalance (2003 'M' vs. 948 'F' samples) posed challenges for all models, impacting their ability to accurately identify female characters. Naive Bayes demonstrated a reasonable balance in precision, recall, and F1-scores for both genders, achieving an accuracy of 0.689. Logistic Regression stood out as the top-performing model, achieving balanced metrics and the highest accuracy at 0.748. However, it faced a slight dip in precision for female identification. Random Forest excelled in identifying males but struggled significantly with female identification, resulting in imbalanced outcomes and an accuracy of 0.695. SVM exhibited a critical limitation in recognizing female characters despite perfect recall for males, yielding an accuracy of 0.679. Among the models evaluated, Logistic Regression emerged as the most reliable choice for character gender identification in the context of the data. It showcased balanced performance metrics and the highest overall accuracy, making it a promising candidate for this specific task. The model's ability to strike a balance between precision and recall, even in the face of class imbalance, positions Logistic Regression as the preferred choice for accurately identifying gender in movie dialogues. The central challenge encountered across all models pertained to the considerable class imbalance within the dataset, notably impacting precision and recall metrics for female characters. While the Random Forest excelled in pinpointing male characters, it struggled significantly with the identification of females. Conversely, the Support Vector Machine (SVM) exhibited a pronounced limitation in recognizing females, presenting a clear constraint. These limitations underscore the intricate nature of gender identification within datasets where one class markedly outweighs the other. In essence, the models' proficiency in characterizing the underrepresented class—females, in this scenario—was compromised due to the dominance of the opposite class. Mitigating the challenges posed by class imbalance is of utmost importance.

Metrics	Multinomial NB	Logistic Regression	Random Forest	SVC
Accuracy	0.689	0.748	0.692	1.000
AUC	0.709	0.799	0.789	0.568
Precision(F)	0.52	0.64	0.75	0.0
Precision(M)	0.74	0.78	0.69	0.68
Recall(F)	0.36	0.48	0.06	0.0
Recall(M)	0.84	0.87	0.99	1.0
F1-Score(F)	0.43	0.55	0.12	0.0
F1-Score(M)	0.79	0.82	0.81	0.81
Support(F)	190	190	190	190
Support(M)	401	401	401	401

Table 1: Performance results without class balancing

2. **Class Imbalance Resolution with RandomUnderSampler:** The dataset's substantial class imbalance (2003 'M' vs. 948 'F' samples) is now addressed using randomUnderSampler. After the rebalancing here are the latest statistics : (948 'M' vs. 948 'F' samples). Random under-sampling possesses its own set of challenges as it removes instances from the majority class randomly. This can lead to a significant loss of information, especially if the majority class has a limited number of instances to begin with. With a reduced dataset size, there is an increased risk of overfitting, especially if the model is complex. With this data, the best performance is seen with logistic regression with an accuracy of 70% Naive Bayes demonstrated a reasonable balance in precision, recall, and F1 scores for both genders, achieving an accuracy of 65%. Logistic Regression exhibits the highest accuracy 70.0% and AUC 77.9% among the models, indicating good overall performance and discrimination between classes. There is a marginal dip in precision in identifying the female dialogues. Random forest is also able to achieve reasonable metrics and the highest accuracy at 69.7%, however similar trend is seen with the precision values. Unlike with unbalanced data, SVM did reasonably well in achieving balanced metrics with an accuracy score of 57.1Unfortunately, the overall accuracy achieved in any of the models is not remarkable at any level. Hence we would like to see the performance of each model with other balancing techniques and examine if we can perform better predictions.

Metrics	Multinomial NB	Logistic Regression	Random Forest	SVC
Accuracy	0.650	0.700	0.697	0.571
AUC	0.688	0.779	0.794	0.676
Precision(F)	0.64	0.68	0.67	0.58
Precision(M)	0.66	0.72	0.74	0.57
Recall(F)	0.68	0.74	0.79	0.53
Recall(M)	0.62	0.66	0.61	0.61
F1-Score(F)	0.66	0.71	0.72	0.55
F1-Score(M)	0.64	0.69	0.67	0.59
Support(F)	190	190	190	190
Support(M)	190	190	190	190

Table 2: Performance results with Random UnderSampling

3. **Class Imbalance Resolution with RandomOverSampler:** To combat class imbalance, next we used the RandomOverSampler, a resampling technique tailored for imbalanced classification problems. In scenarios where one class is notably underrepresented, machine learning models may struggle to discern patterns related to the minority class. RandomOverSampler addresses this issue by randomly duplicating instances of the minority class until a more balanced distribution is achieved. By synthetically augmenting the representation of the minority class, this technique prevents model bias towards the majority class, thereby improving its ability to discern patterns in the minority class. RandomOverSampler,

a straightforward yet effective method, can be seamlessly integrated with various classification algorithms to enhance model performance, particularly when accurately predicting the minority class holds paramount significance.

Metrics	Multinomial NB	Logistic Regression	Random Forest	SVC
Accuracy	0.832	0.768	0.923	0.702
AUC	0.924	0.768	0.978	0.775
Precision(F)	0.80	0.78	0.91	0.68
Precision(M)	0.88	0.76	0.93	0.72
Recall(F)	0.89	0.75	0.93	0.75
Recall(M)	0.77	0.78	0.91	0.65
F1-Score(F)	0.84	0.76	0.92	0.72
F1-Score(M)	0.84	0.77	0.92	0.69
Support(F)	401	401	401	401
Support(M)	401	401	401	401

Table 3: Performance results with Random OverSampling

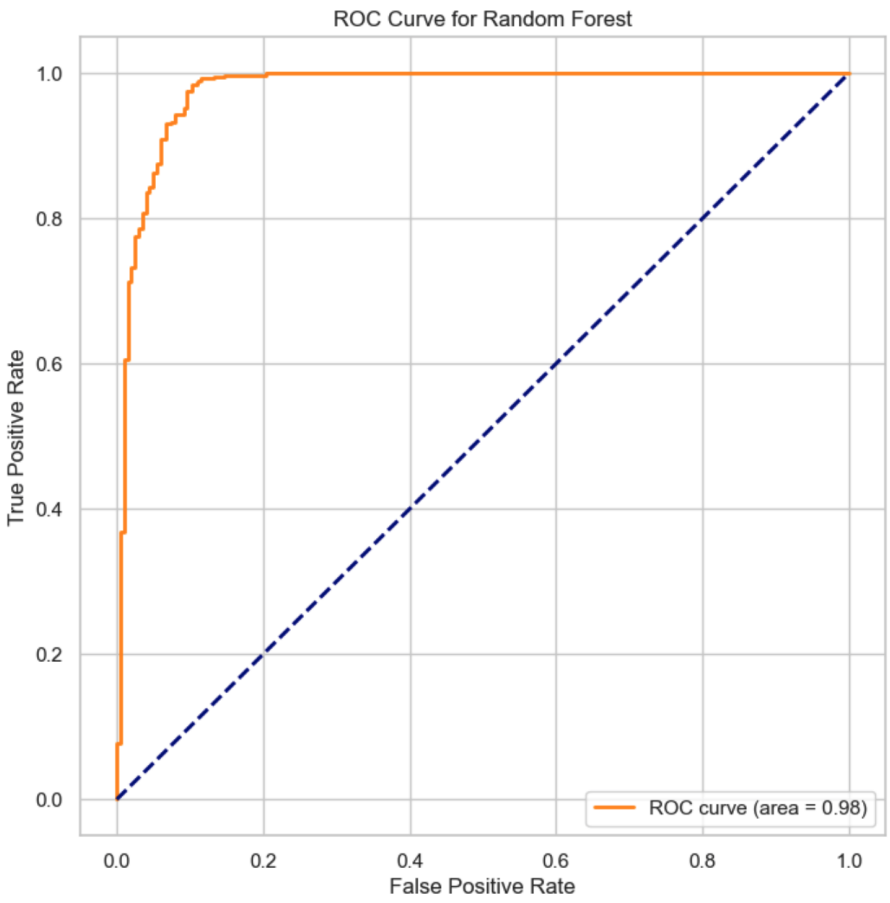
4. **Class Imbalance Resolution with SMOTE:** We have tested with the Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance, yielding a balanced dataset with 401 samples for both 'F' and 'M.' On this rebalanced data, Naive Bayes demonstrates an improved accuracy of 0.845, showcasing heightened precision, recall, and F1-scores for both genders. Logistic Regression maintains its reliability with an accuracy of 0.772, while Random Forest excels further with an impressive accuracy of 0.855 and a high AUC. The Support Vector Machine (SVM) exhibits improvement but continues to grapple with achieving balanced performance metrics, underscoring the persistent intricacies of gender identification tasks. The selection of models should carefully consider the specific requirements and constraints of the application, taking into account the trade-offs between precision, recall, and overall accuracy within the context of oversampled balanced data.

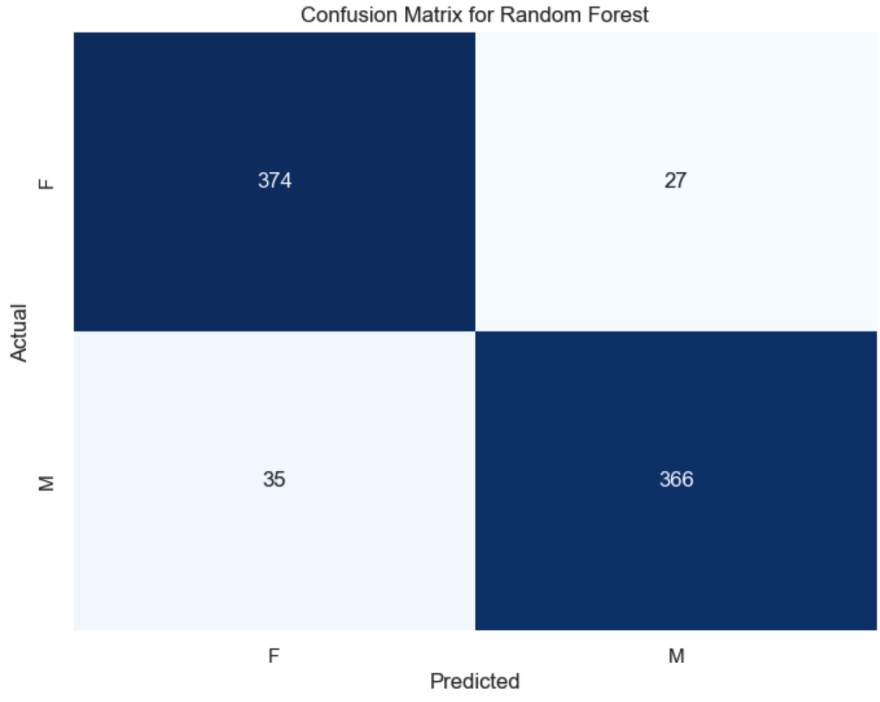
Metrics	Multinomial NB	Logistic Regression	Random Forest	SVC
Accuracy	0.845	0.772	0.859	0.732
AUC	0.928	0.854	0.938	0.808
Precision(F)	0.81	0.76	0.89	0.71
Precision(M)	0.89	0.78	0.83	0.76
Recall(F)	0.91	0.79	0.82	0.79
Recall(M)	0.78	0.76	0.90	0.68
F1-Score(F)	0.85	0.78	0.85	0.75
F1-Score(M)	0.84	0.77	0.86	0.72
Support(F)	401	401	401	401
Support(M)	401	401	401	401

Table 4: Performance results with SMOTE

5.3.5 Best Performing Model

After the complete analysis the best model is found to be random forest classifier with RandomUnderSampler of data. The model has a high accuracy of 92.3%. The model has high precision (F : 0.91, M : 0.93) and recall (F : 0.93, M : 0.91) , indicating that it is performing well on both positive and negative classes. This means it can accurately identify the gender and also has a low rate of false negatives and false positives. The high AUC score of 0.978 indicates that the model has a high measure of separability. It is capable of distinguishing between dialogues from male and female characters.


ROC Curve for Random Forest
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.98)

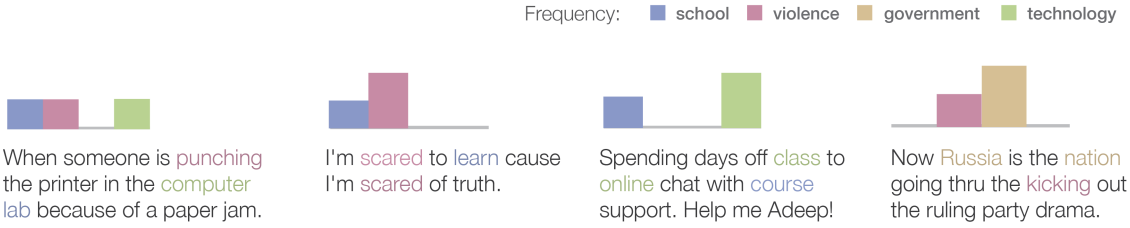

Confusion Matrix for Random Forest
Actual
Predicted
F
M
374
27
35
366

6 Decoding Gender Depiction: Analyzing Dialogue Dynamics in Film

In our project, our primary aim was to thoroughly analyze the intricacies of dialogue content across both male and female characters. We sought to investigate the evolution of conversational topics, emotional expressions, and overall sentiment within the dialogues of each gender. A significant aspect of our exploration involved understanding how character gender influences dialogue dynamics, with a specific focus on whether female characters tend to employ more tentative language compared to their male counterparts, who may project a more assertive tone.

Our overarching objective was not only to uncover contemporary dialogue patterns but also to gain insights into the variations in the depiction of female and male characters over the years, as reflected in their dialogues. To achieve this, we utilized Empath, a tool for text analysis across lexical categories, similar to LIWC, to generate new lexical categories for our analysis.

Empath is a linguistic analysis tool that derives its insights from a substantial dataset of fiction, totaling 1.8 billion words. It begins by constructing word embeddings to capture the subtle usage patterns of words within this dataset. These embeddings are then organized into a vector space, which facilitates the measurement of word similarity. To categorize words, Empath uses predefined seed terms and expands these categories by identifying new, related words. The final step involves crowdsourcing to filter and refine these categories, ensuring that they accurately reflect collective human understanding and consensus on word meanings and associations. This process allows Empath to systematically organize words into meaningful categories based on their contextual relationships in literature.



6.1 Topic of Conversation

Our primary project objective entailed conducting a thorough analysis of dialogue topics. To achieve this goal, we carefully curated a set of categories that not only aligned with the specific topics we intended to analyze but also complemented the capabilities of the Empath tool. Following is the list of categories we examined for when analysing the topic of conversation in the dialogue

```
categories=["help", "office", "money", "domestic_work", "occupation", "government",
"blue_collar_job",
"real_estate", "business", "office", "cooking", "leader", "politics", "economics",
"technology", "white_collar_job"]
```

Subsequently, we implemented a function:

```
def analyze_dialogue(dialogue):
    return lexicon.analyze(dialogue, categories)

character_conversations_df['dialogue_topics'] =
character_conversations_df['speaker_dialogue'].apply(analyze_dialogue)
```

This function was applied to the 'speaker_dialogue' column of our dataset, resulting in the addition of a new column, 'dialogue_topics', in the dataframe. The function analyzed each dialogue, predicted the classes to which they belonged, and assigned a score between 0 and 1 for each category. In instances where a dialogue did not fit into any predefined category, the function assigned an empty dictionary.

Subsequently, we performed data cleaning by eliminating rows in the dataframe where the dialogue topic resulted in an empty dictionary. This refinement process yielded **24,572** dialogues with discernible topics of conversation out of the initial **229,106** dialogues in our dataset. We further segregated the filtered dataframe based on character gender for subsequent analysis.

Upon this division, we obtained **6,435** dialogues featuring female characters engaging in some form of conversation topics and **18,137** dialogues involving male characters. To delve deeper into the analysis, we employed the following function:

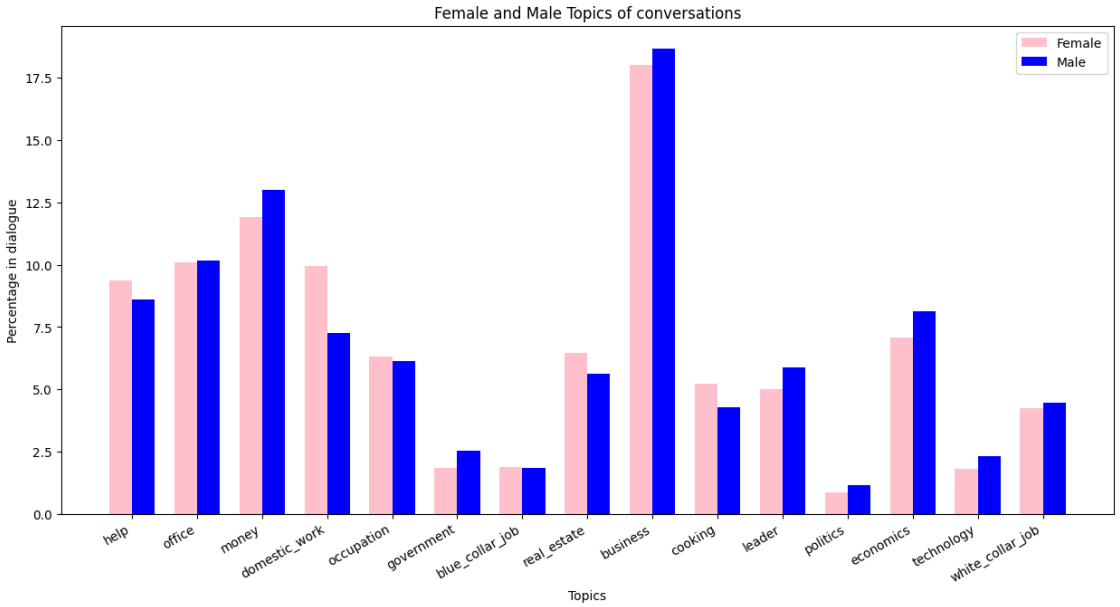
```
def topic_count_analyser(topics_in_dialogue):
    topics_in_dialogue = Counter(topics_in_dialogue)
    female_topic_counter.update(topics_in_dialogue)
    return None
```

This function utilized a Counter to tally the occurrences of topics in all dialogues, updating the female_topic_counter accordingly.

Next, we converted these raw counts into percentages using the following function:

```
total_dialogue_count = sum(female_topic_counter.values())
percentage_occurrences =
{topic: (occurrences / total_dialogue_count) * 100 for topic,
occurrences in female_topic_counter.items()}
female_conversation_topic_percentage =
dict(zip(percentage_occurrences.keys(), percentage_occurrences.values()))
print(female_conversation_topic_percentage)
```

This process generated a dictionary, female_conversation_topic_percentage, containing the percentage occurrences of each category across all the dialogues involving female characters.



Thus upon plotting we were able to reveal clear distinctions in the dialogue topics between female and male characters. It becomes evident that female characters are more frequently engaged in conversations revolving around **love, family, and relationships**, while their male counterparts are prominently associated with dialogues about **politics, business, and technology**. This pattern suggests a portrayal of gender-specific interests within film narratives, reflecting a tendency to assign traditional societal roles to females and males.

To visually convey these trends, we employed a color-coded approach — using pink bars for female dialogues and blue for male dialogues.

6.2 Valence of emotion

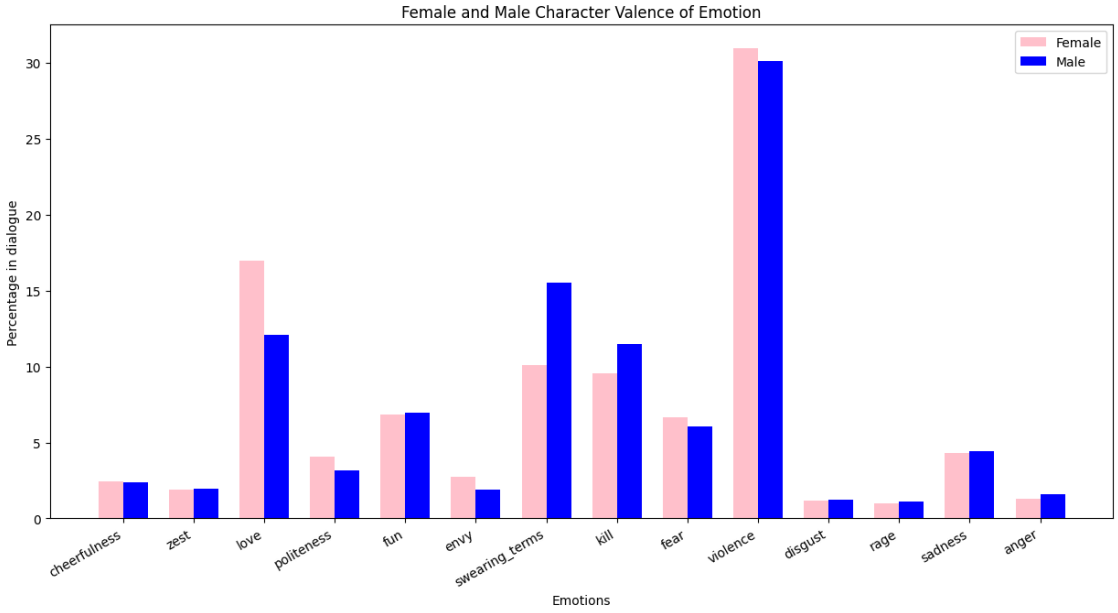
Having concluded the analysis of conversation topics within dialogues, our subsequent step involved examining the valence of emotions expressed in the dialogues. To achieve this, we employed the Empath tool once again, utilizing the previously established process. This time, the focus was on exploring the valence of emotions, and the list of categories we sought to investigate included:

```
valence_of_emotion_categories = ["cheerfulness", "zest", "love", "politeness", "fun",
"aggression", "envy", "swearing_terms", "kill", "fear", "violence", "disgust", "rage",
"sadness", "anger"]
```

Subsequently, we changed the function to:

```
def analyze_dialogue(dialogue):
    return lexicon.analyze(dialogue, valence_of_emotion_categories)

character_conversations_df['valence_of_emotion_dialogue'] =
character_conversations_df['speaker_dialogue'].apply(analyze_dialogue)
```

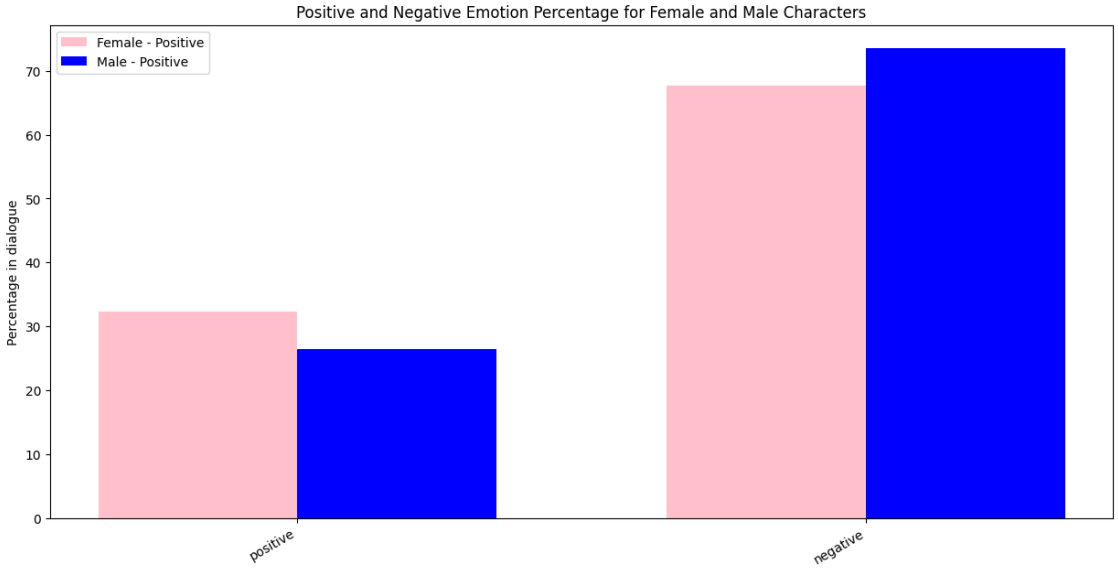
We replicated the methodology employed in the topic of conversation analysis for examining the valence of emotions within dialogues. In this iteration, the filtered dataframe comprised **24,979** dialogues that exhibited some form of valence of emotion from our finalized categories. Following the same approach, we conducted a gender-based split, distinguishing between male and female characters, resulting in **7,227** dialogues for female characters and **17,752** for male characters. The subsequent step involved tallying the occurrences of each emotion category across all male and female dialogues and subsequently converting these counts into percentages.



Upon visualizing the data through graphs, it really brought out some interesting insights into the emotions expressed by female and male characters in the movies. For female characters, emotions like **"love"** and **"violence"** were quite prevalent, hinting at the challenges, including instances of violence, that these characters face in the movie narratives.

On the flip side, in dialogues featuring male characters, we observed a higher frequency of emotions like **"kill," "sadness," and "anger."** This points to a different emotional landscape, suggesting that the storylines for male characters often involve themes related to conflict, sorrow, and intense emotional experiences.

These distinct emotional portrayals provide a deeper understanding of how gender-specific experiences are depicted in the movies we analyzed.



Moving forward, we extended our analysis to encompass the overall sentiment conveyed in the dialogues of both female and male characters. Through visualization, we uncovered a noteworthy distinction: the sentiment of dialogues involving female characters tended to be positive, while dialogues featuring male characters exhibited a more negative sentiment than their female counterparts. This observation provides valuable insights into the emotional tone and characterization of gender-specific dialogues in the movies we analyzed.

6.3 Assessing Assertiveness and Tentativeness in Female and Male Character Conversations

After completing our in-depth analyses of dialogues concerning the topic of conversation, valence of emotion, and sentiment, our investigation extended to assess the levels of assertiveness and tentativeness. We maintained continuity by using the Empath tool and focusing on specific assertiveness categories:

```
assertive_categories = ["dominant_heirarchical", "pride", "dominant_personality",
"nervousness", "confusion"]
```

To implement this analysis, we introduced a dedicated function designed to add a new column to the dataframe. This new column, named 'assertive,' contains dictionaries indicating the assertive categories attributed to each dialogue. The function is defined as follows:

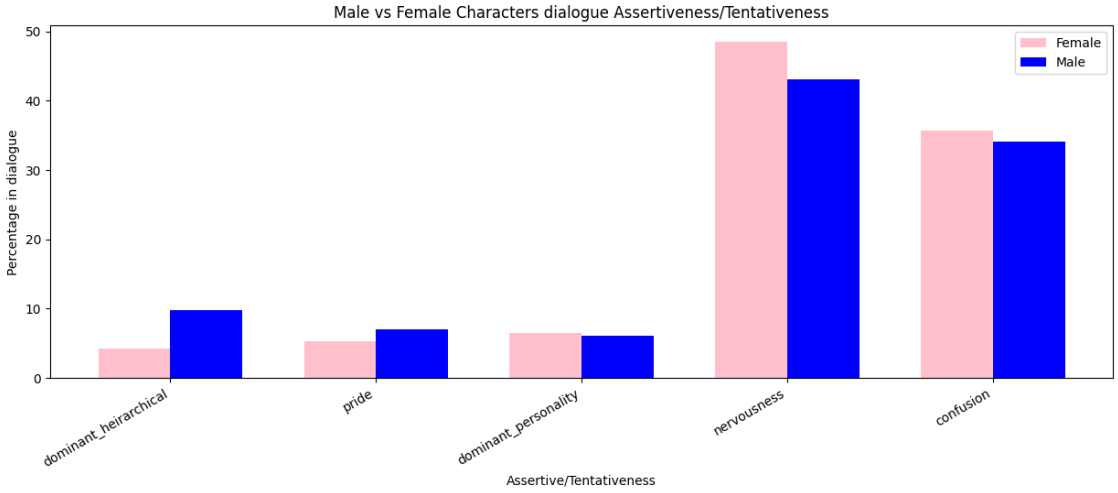
```
def analyze_assertive_category(assertive_categories):
    selected_classes = {}
    if assertive_categories:
        selected_classes =
        {key: value for key, value in assertive_categories.items() if value > 0.0}

    return selected_classes
```

This function was applied to the 'assertive_category' column in our dataframe, effectively capturing and categorizing assertive characteristics within each dialogue.

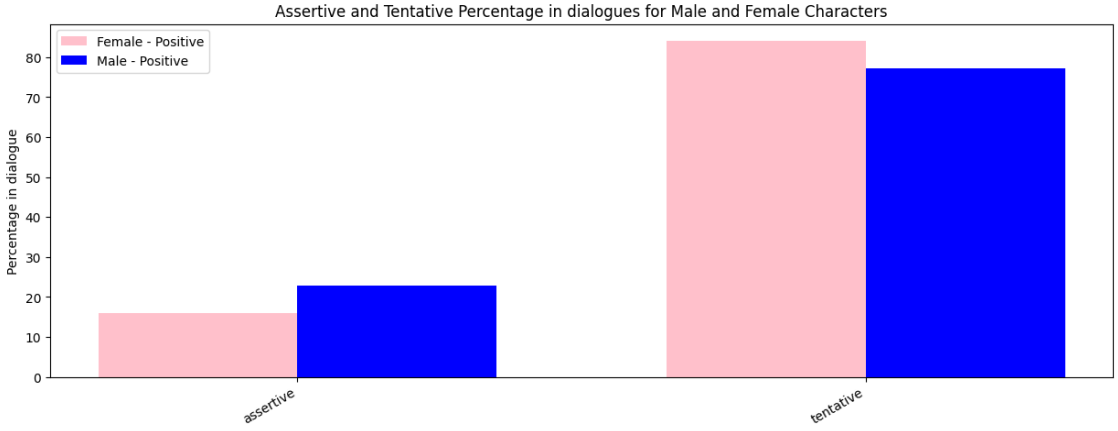
```
character_conversations_df['assertive'] =
character_conversations_df['assertive_category'].apply(analyze_assertive_category)
print(character_conversations_df['assertive'])
```

Upon filtering for dialogues exhibiting either tentativeness or assertiveness, a total of **8,108** dialogues were identified. Following a gender-based split, this subset comprised **2,385** dialogues involving female characters and **5,723** dialogues featuring male characters. This refined dataset allows for a focused examination of the nuanced variations in tentativeness and assertiveness across different gender portrayals within our analysis. We proceeded to compute the percentage occurrence of each category across all dialogues involving female and male characters. The results were then visually depicted using a bar plot.

As we delved into the analysis of assertiveness and tentativeness within dialogues, fascinating gender-specific patterns emerged. In dialogues featuring female characters, there was a notable prevalence of emotions associated with nervousness and confusion, indicating a nuanced portrayal of vulnerability and uncertainty. Conversely, dialogues involving male characters exhibited a tendency toward assertive qualities, with dominant_hierarchical and pride being more pronounced. This distinction provides valuable insights into the diverse emotional dynamics attributed to different genders within the dialogues, enriching our understanding of gender portrayals in the dataset.



Subsequently, we conducted a broader analysis of dialogues based on character gender to explore whether there were discernible patterns such as female characters having more tentative dialogues and male characters exhibiting greater assertiveness. The findings aligned with our expectations, revealing a clear trend: female characters tended to express more tentativeness in their dialogues, while male characters displayed a higher degree of assertiveness. This outcome supports our initial hypothesis and offers valuable insights into the consistent portrayal of gender-specific communication styles within the dataset. The nuanced exploration of dialogue dynamics enhances our understanding of the intricacies involved in character interactions and contributes to a more comprehensive interpretation of gender roles in film narratives.



7 Conclusion & Future Work

In the United States, less than a third of all speaking roles in top-grossing films are assigned to female characters, exacerbating disparities for women of color, older women, and those from the LGBTQ+ community. This under-representation in films carries far-reaching consequences:

- Fewer roles for female actors, perpetuation of damaging stereotypes, and a scarcity of positive role models for children.
- Frequent portrayal of female characters as one-dimensional stereotypes, hindering the nuanced representation and development of women in films.
- Depicting women in leadership roles can contribute to dismantling real-world gender barriers, inspiring young girls to seek their own opportunities.

With our work we hope to make an impact on the following:

- Addressing gender biases in films can lead to a more inclusive and equitable industry, allowing a diverse range of stories to be told.
- Increased awareness of gender biases can spark cultural change, influencing attitudes and behaviors towards a more accepting and egalitarian society.
- Research and analysis on gender biases pave the way for targeted interventions, not only in films but also across academia, workplaces, healthcare, finance, and HR.
- Collective efforts against gender bias foster progress, empowerment, and a more inclusive world, empowering individuals regardless of gender.

The utilization of Empath allowed us to uncover subtle gender dynamics, affirming our initially predicted analysis regarding male and female characters within the context of conversation. We examined the valence of emotions and assessed the assertiveness and tentativeness of characters in movies based on their dialogues. These discoveries provide valuable insights into the representation of diverse perspectives in cinematic storytelling, highlighting the influence of societal expectations on character interactions.

In extending our NLP analysis beyond the speaker, pinpointing the subject of conversations, and redefining female-centric with metrics like the Bechdel Test, we deepen our understanding. This approach broadens the scope, ensuring a comprehensive exploration of gender biases in movie dialogs.

The ongoing work in combating gender bias, coupled with a broadened NLP analysis scope and exploration across domains, promises a transformative future. As we strive for inclusivity and diversity, the impact extends beyond films, shaping a world where everyone can thrive, unfettered by gender-based constraints.

References

[1] https://www.cs.cornell.edu/∼cristian/Cornell_Movie-Dialogs_Corpus.html

[2] Haris, M.J., Upreti, A., Kurtaran, M. et al. Identifying gender bias in blockbuster movies through the lens of machine learning. Humanit Soc Sci Commun 10, 94 (2023). https://doi.org/10.1057/s41599-023-01576-3

[3] K. Khadilkar, A. R. KhudaBukhsh, and T. M. Mitchell, "Gender bias, social bias, and representation in Bollywood and Hollywood," Patterns, vol. 3, no. 2, 2022, doi: 10.1016/j.patter.2021.100409.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4171–4186, doi: 10.18653/v1/N19-1423.

[5] Yu, Yulin & Hao, Yucong & Dhillon, Paramveer. (2022). Unpacking Gender Stereotypes in Film Dialogue. 10.1007/978-3-031-19097-1_26.

[6] Data courtesy of IMDb. https://developer.imdb.com/non-commercial-datasets/?ref_=pe_2610490_199225680

[7] Fast, E., Chen, B., & Bernstein, M. S. (n.d.). Empath: Understanding Topic Signals in Large-Scale Text. Stanford University.