# Analyzing Gender Perceptions in Dating App Data and Movie Preferences

Abhishek Sharma
*Master in Data Science*
*ashar58@ur.rochester.edu*
*University of Rochester*

Bhargav Sai Bhuvanagiri
*Masters in Data Science*
*bbhuvana@ur.rochester.edu*
*University of Rochester*

***Abstract*—This research investigates the intricate dynamics of gender-specific perceptions and preferences within dating apps, focusing on the qualities each gender finds appealing in the other. It aims to understand how individuals perceive and assess attributes of the opposite gender in online dating while examining how media shapes personal preferences. Using the fpgrowth algorithm, frequent itemsets were obtained, and simple lemmatization methods were applied to extract significant words for each category.**

**To enhance this analysis, the study integrates Myers-Briggs Type Indicator (MBTI) personality prediction models by utilizing user-generated movie reviews. This involves identifying key genres for each personality type, extracting characters from top films in those genres, and deriving the personalities of these characters. The research aims to correlate these personality traits with the earlier extracted words, measuring their similarity. A metric will be employed to gauge the closeness of these two personality indicators and assess their correlation, offering insights into the alignment between these distinct methodologies in understanding how media strategically targets audiences by featuring the most favored men or women.**

## 1. Introduction

In today's digital era, the realms of online dating and media consumption intersect significantly influencing how individuals perceive and connect with potential partners. The study of gender-specific perceptions and preferences within dating apps has garnered immense interest, particularly in unraveling the intriguing qualities each gender finds appealing in the other.
As individuals navigate the virtual landscape of dating apps, they are presented with a myriad of profiles and potential matches, each reflecting a blend of personal traits and characteristics. Understanding how these traits are perceived and evaluated within the context of gender-specific preferences stands as a captivating puzzle waiting to be deciphered. This project mainly answers the following question.

Does the media exploit its target audience by showcasing characters that align with their preferred personality types?

This project embarks on a fascinating journey to unravel this query, seeking to establish whether the extracted personality traits from dating app data and movie preferences align. The analysis not only delves into understanding gender-specific perceptions but also ventures into exploring the correlation between these divergent data sources. Ultimately, this exploration offers invaluable insights into how media strategically targets audiences by portraying the most favored male or female personalities, bridging the gap between dating preferences and cinematic portrayals.

## 2. Related Work:

Most of the studies that are related to this study are about just personality prediction based in just one direction [1]. There have been existing studies on charactercentric narrative understanding. While many of them (Massey et al., 2015; Srivastava et al., 2016; Brahman et al., 2021) work on summaries of stories or summaries of characters. Their scopes thus have a different assessment purpose from ours and have the challenge of understanding long narrative inputs greatly reduced.
A recent study in a related field has focused on analyzing social media data from Twitter and Facebook to comprehend users' personalities using the Big Five personality traits model [4]. Their methodology included gathering and annotating Twitter data with the aid of psychological experts to characterize user personalities.

## 3. Data

This project utilized two primary datasets: the OkCupid dataset and the MovieLens dataset. The OkCupid dataset comprises around 60,000 entries featuring diverse columns related to individuals' health, preferred body type, relationship status, occupation, and more. Particularly, each individual describes themselves across eight essays, including the ninth essay detailing their desired partner's personality, which is the focus of this study. On the other hand, the

MovieLens dataset consists of 1 million data points encompassing limited columns such as occupation, gender, and age. The movie dataset includes genre information, movies within those genres, and their respective ratings. The subsequent sections will elaborate on how these datasets were processed to achieve the desired outcomes.

## 3.1. Data Acquisition

To acquire the OkCupid dataset [6], an Application Programming Interface (API) provided by the platform was utilized. This API granted access to a comprehensive dataset comprising nearly 60,000 data entries, each containing a multitude of columns featuring diverse information.
On the other hand, the MovieLens dataset was acquired directly from the official MovieLens website [3]. This dataset boasted a vast collection of approximately 1 million data points. It primarily included key demographic information such as occupation, gender, and age, which played a crucial role in the context of this research. Additionally, the MovieLens dataset offered an extensive array of genre-related data, encompassing various genres, the movies contained within those genres, and their corresponding ratings. Acquiring these datasets through distinct methods ensured a varied and comprehensive dataset, which was essential for meeting the objectives of this study.

## 3.2. Data Preprocessing

During the data preprocessing phase, we conducted an alignment process between the OkCupid and MovieLens datasets by retaining only the shared variables. Subsequently, we eliminated surplus columns in the OkCupid dataset to ensure uniformity across both datasets. Additionally, we restricted our analysis within the OkCupid dataset to individuals identifying as heterosexual since other sexual orientations weren't represented in the MovieLens dataset. To facilitate uniformity in age representation, we bucketed the continuous age variable in the OkCupid dataset, aligning it with the pre-existing bucketed age ranges within the MovieLens dataset.

Moreover, we standardized the approach for age representation by bucketizing age ranges in both datasets to ensure consistency. Further refinement involved extracting the most prevalent occupation for specific bucketed age and gender categories, creating a modified OkCupid dataset that aligned with the movie dataset's structure. This modified dataset incorporated essential attributes like age, gender, and occupation. Additionally, we focused on consolidating relevant information from the OkCupid dataset, particularly the ninth essay section, which pertained to individuals' descriptions of their preferred personality types in potential partners. Similarly, we tailored the MovieLens dataset to mirror this refined structure, ensuring uniformity across age, gender, and occupation categories for subsequent analyses. This meticulous data preprocessing strategy aimed to harmonize both datasets, enabling seamless integration for subsequent phases of the project.

## 3.3. Data Visualization

### OKCupid:

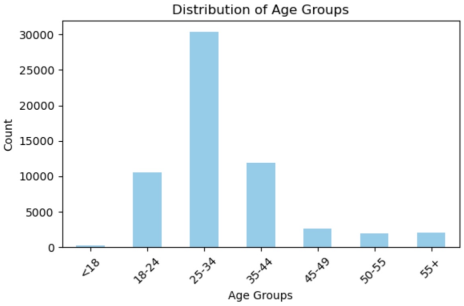


Figure 1. Age distribution of the OkCupid dataset

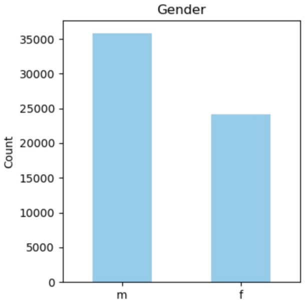
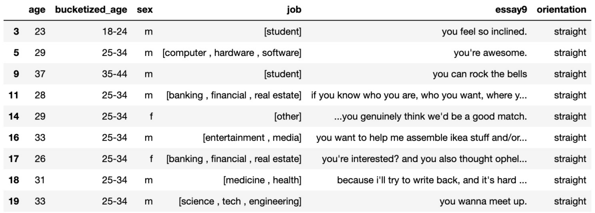


Figure 2. Gender Distribution in the dataset



| | age | bucketized_age | sex | job | essay9 | orientation |
|---|---|---|---|---|---|---|
| 3 | 23 | 18-24 | m | [student] | you feel so inclined. | straight |
| 5 | 29 | 25-34 | m | [computer , hardware , software] | you're awesome. | straight |
| 9 | 37 | 35-44 | m | [student] | you can rock the bells | straight |
| 11 | 28 | 25-34 | m | [banking , financial , real estate] | if you know who you are, who you want, where y... | straight |
| 14 | 29 | 25-34 | f | [other] | ...you genuinely think we'd be a good match. | straight |
| 16 | 33 | 25-34 | m | [entertainment , media] | you want to help me assemble ikea stuff and/or... | straight |
| 17 | 26 | 25-34 | f | [banking , financial , real estate] | you're interested? and you also thought ophel... | straight |
| 18 | 31 | 25-34 | m | [medicine , health] | because i'll try to write back, and it's hard ... | straight |
| 19 | 33 | 25-34 | m | [science , tech , engineering] | you wanna meet up. | straight |

Figure 3. Post-Processed dataset

### Movielens

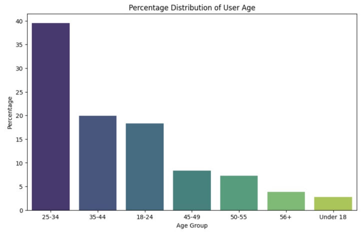


Figure 4. Age distribution of the data

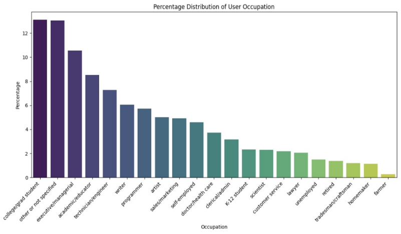


Figure 5. Distribution of user occupation

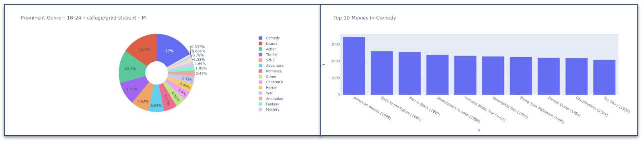
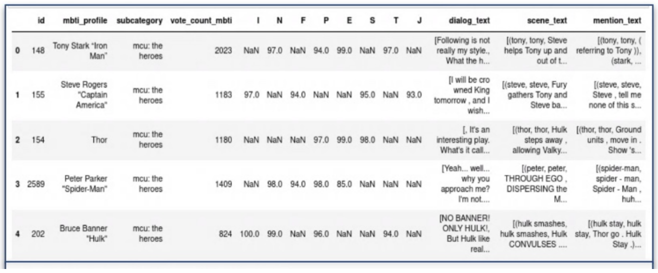


Figure 6. Example visualization of popular genre and movies by classes - (18-24, Male, College student)

**MTBI Personality Prediction:** Will be discussed in further sections.



| | id | mbti_profile | subcategory | vote_count_mbti | I | N | F | P | E | S | T | J | dialog_text | scene_text | mention_text |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 148 | Tony Stark "Iron Man" | mcu: the heroes | 2023 | NaN | 97.0 | NaN | 94.0 | 99.0 | NaN | 97.0 | NaN | [Following is not really my style., What the h... | [(tony, tony, Steve helps Tony up and out of t... | [(tony, tony, ( referring to Tony )), (stark, ... |
| 1 | 155 | Steve Rogers "Captain America" | mcu: the heroes | 1183 | 97.0 | NaN | 94.0 | NaN | NaN | 95.0 | NaN | 93.0 | [I will be cro wned King tomorrow , and I wish... | [(steve, steve, Fury gathers Tony and Steve ba... | [(steve, steve, Steve , tell me none of this s... |
| 2 | 154 | Thor | mcu: the heroes | 1180 | NaN | NaN | NaN | 97.0 | 99.0 | 98.0 | NaN | NaN | [, It's an interesting play. What's it call... | [(thor, thor, Hulk steps away , allowing Valky... | [(thor, thor, Ground units , move in . Show 's... |
| 3 | 2589 | Peter Parker "Spider-Man" | mcu: the heroes | 1409 | NaN | 98.0 | 94.0 | 98.0 | 85.0 | NaN | NaN | NaN | [Yeah... well... why you approach me? I'm not.... | [(peter, peter, THROUGH EGO , DISPERSING the M... | [(spider-man, spider - man, Spider - Man , huh... |
| 4 | 202 | Bruce Banner "Hulk" | mcu: the heroes | 824 | 100.0 | 99.0 | NaN | 96.0 | NaN | NaN | 94.0 | NaN | [NO BANNER! ONLY HULK!, But Hulk like real... | [(hulk smashes, hulk smashes, Hulk CONVULSES .... | [(hulk stay, hulk stay, Thor go . Hulk Stay .)... |

Figure 7. The post-processed data set ready to get personality prediction

## 4. Methodology

In this study, we investigate the correlation between individuals' dating preferences and movie choices using three datasets: OkCupid profiles, the Movielens 1M dataset, and the Myers-Briggs Type Indicator (MBTI) dataset. Data preprocessing for the OkCupid dataset involved stratification into classes based on user age groups and gender, following the taxonomy established by the Movielens dataset. The methodology employs a two-part parallel approach, with the first part focusing on OkCupid dataset analysis, and the second part on Movielens and MBTI dataset analysis. To validate our findings, tests were conducted to assess the consistency of observations between the OkCupid and Movielens/MBTI segments.

Understanding demographic differences and linguistic nuances is necessary to provide insight into personal preferences and behaviors. We investigated the use of two different methods, the Fp-growth algorithm and Natural Language Processing (NLP), to disclose important details regarding the most common types of jobs for specific age and gender groups as well as important words and language patterns in these groups. By combining these methods, we hope to promote a deeper understanding of the relationship between language preferences and demography and provide significant new information in this area.

### 4.1. FP-Growth Algorithm:

The FP-Growth algorithm was employed to discern the frequent occurrences of occupations associated with distinct age and gender categories. This approach facilitated the identification of substantial trends and prevalent occupations within specific demographic groups.

The FP-Growth algorithm, known for its efficiency in mining frequent itemsets, proved instrumental in analyzing the dataset. By generating frequent patterns directly from the transaction data, FP-Growth excelled in identifying patterns without the need for candidate generation and maintaining a compact data structure (the FP-Tree) for expedited processing.

Utilizing this technique revealed compelling insights into the predominant occupations prevalent among various age and gender segments, providing a comprehensive understanding of the occupational landscape within distinct demographic cohorts.

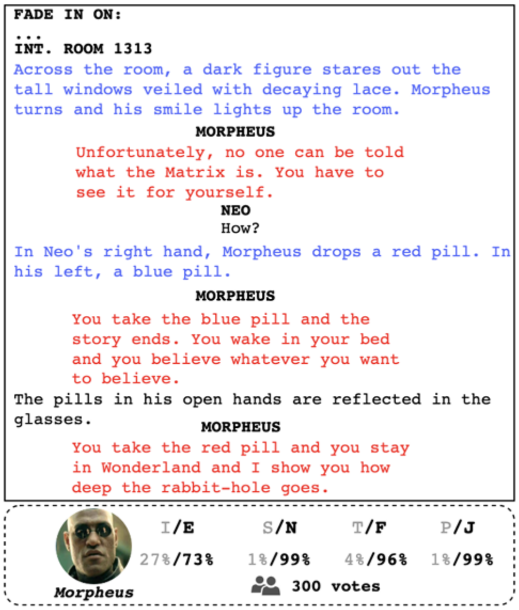
### 4.2. Lemmetization

The methodology involved in utilizing lemmatization to extract important words comprises several key steps [5]. Initially, the textual data is prepared, undergoing preprocessing steps like cleaning and organizing. Subsequently, tokenization breaks down the text into individual words or tokens, followed by applying lemmatization to reduce words to their base forms. Further analysis includes optional removal of stopwords and conducting frequency analysis to identify the most frequent lemmas. These lemmas are ranked based on significance and relevance within the dataset to extract crucial terms. Finally, interpreting these important words provides insights and patterns, aiding in drawing meaningful conclusions from the dataset. This iterative process ensures the refinement and validation of the extracted words for a comprehensive understanding of the data.

### 4.3. MTBI Personality prediction

The Myers–Briggs Type Indicator (MBTI) is a well-known tool for evaluating personality traits, focusing on four primary dimensions: Extraversion/Introversion (E/I), Sensing/Intuition (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). Extraversion signifies being active and objective, whereas Introversion implies a more passive and subjective nature. Sensing involves attention to sensory experiences, while Intuition relates to perceptive insights. Thinking pertains to logical reasoning, contrasting with Feeling, which represents subjective and interpersonal tendencies. Judging refers to prompt decision-making, while Perception involves a more patient approach, awaiting additional information before making decisions. An individual's

MBTI type is determined based on their predominant inclination within each of these dimensions. For example, if a person is identified as an ENFJ type, signifying extraversion, intuition, feeling, and judging as his dominant preferences.



```
FADE IN ON:
...
INT. ROOM 1313
Across the room, a dark figure stares out the
tall windows veiled with decaying lace. Morpheus
turns and his smile lights up the room.
                    MORPHEUS
        Unfortunately, no one can be told
        what the Matrix is. You have to
        see it for yourself.
                       NEO
                       How?
In Neo's right hand, Morpheus drops a red pill. In
his left, a blue pill.
                    MORPHEUS
        You take the blue pill and the
        story ends. You wake in your bed
        and you believe whatever you want
        to believe.
The pills in his open hands are reflected in the
glasses.             MORPHEUS
        You take the red pill and you stay
        in Wonderland and I show you how
        deep the rabbit-hole goes.
```

```
          I/E        S/N       T/F       P/J
          27%/73%    1%/99%    4%/96%    1%/99%
Morpheus              300 votes
```

Figure 8. An example excerpt from the movie script of “The Matrix.” Blue utterances are Morpheus’s character’s scene descriptions; Red are his dialogues.Morpheus’s personality was rated as ENFJ by 300 user votes. (Credit to Yisi Sang)

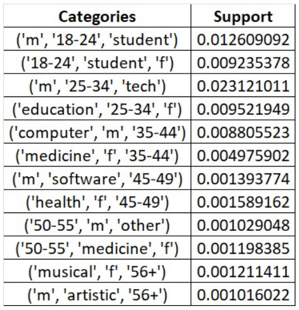
Building upon the groundwork laid by Yisi Sang [2], we adopted their resulting dataset for character extraction from movie scripts. We applied an MBTI personality prediction dataset to extract personality traits from the movie scripts obtained from our chosen films. This dataset utilized the storyline to forecast the MBTI personality types of the fictional characters portrayed in the movies. We aimed to decipher and comprehend the characters’ personalities by analyzing their behaviors, dialogues, and interactions within the movie plots. Leveraging the Movielens dataset, we curated a list of the most popular movies within each predefined class. Employing the pre-trained model, we processed the movie scripts for each selected movie, identifying characters and their associated MBTI personality types. Simultaneously, we manually recorded the gender of each character. The resultant dataset included prominent movies, the characters within, their associated personality types, and gender information for all classes. For each class, we meticulously tallied the personality types for relevant genders (e.g., for a male class, the relevant gender was female, and vice versa). Subsequently, we identified the most prevalent personality types within each class. Finally, we compared these identified personality types with our observations derived from the OkCupid dataset.

## 5. Project SetUp

Initially, we standardized the attributes of the post-processed data to ensure consistency between the two datasets. This involved extensive removal of attributes from the OkCupid dataset to match it with the MovieLens dataset. However, this alignment procedure introduced a limitation in our project. To organize the data effectively, we categorized it into distinct groups based on age, gender, and occupation. To reconcile the disparity in how age was represented—continuous in OkCupid and binned in MovieLens—we normalized the continuous age data by binning it into comparable ranges in both datasets. Leveraging the Fp growth algorithm, we identified significant occupations within each age and gender subgroup, extracting the most influential class that encapsulates associations between age, gender, and occupation. Consequently, we generated multiple classes for both the MovieLens and OkCupid datasets using a similar methodology. We now divided this initial setup into three parts and we work through each part.

### 5.1. Part1

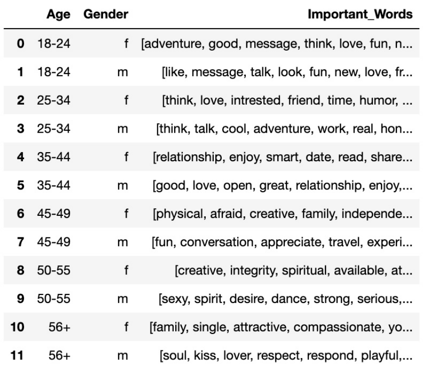
This section’s objective is to extract significant words from the essays where users describe their ideal partners. By segmenting the entire dataset into distinct classes, we filtered the data for each class, focusing on these specific essays.



| Categories | Support |
|---|---|
| ('m', '18-24', 'student') | 0.012609092 |
| ('18-24', 'student', 'f') | 0.009235378 |
| ('m', '25-34', 'tech') | 0.023121011 |
| ('education', '25-34', 'f') | 0.009521949 |
| ('computer', 'm', '35-44') | 0.008805523 |
| ('medicine', 'f', '35-44') | 0.004975902 |
| ('m', 'software', '45-49') | 0.001393774 |
| ('health', 'f', '45-49') | 0.001589162 |
| ('50-55', 'm', 'other') | 0.001029048 |
| ('50-55', 'medicine', 'f') | 0.001198385 |
| ('musical', 'f', '56+') | 0.001211411 |
| ('m', 'artistic', '56+') | 0.001016022 |

Figure 9. Frequent patterns of Age,gender,Occupation using fpgrowth

Using lemmatization, we simplified words to their base forms, aiming to unify variations. After this text transformation, we analyzed to pinpoint essential terms within the datasets. Our focus was on extracting significant words specific to each category, allowing us to identify and isolate the distinctive vocabulary associated with individual classes. This process enabled the extraction of crucial, class-specific terms, providing valuable insights into the unique language and expressions used within each dataset. After postprocessing the output data frame that we got from the OkCupid dataset was.

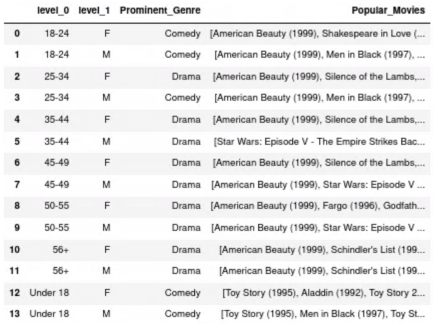


| | Age | Gender | Important_Words |
|---|---|---|---|
| 0 | 18-24 | f | [adventure, good, message, think, love, fun, n... |
| 1 | 18-24 | m | [like, message, talk, look, fun, new, love, fr... |
| 2 | 25-34 | f | [think, love, intrested, friend, time, humor, ... |
| 3 | 25-34 | m | [think, talk, cool, adventure, work, real, hon... |
| 4 | 35-44 | f | [relationship, enjoy, smart, date, read, share... |
| 5 | 35-44 | m | [good, love, open, great, relationship, enjoy,... |
| 6 | 45-49 | f | [physical, afraid, creative, family, independe... |
| 7 | 45-49 | m | [fun, conversation, appreciate, travel, experi... |
| 8 | 50-55 | f | [creative, integrity, spiritual, available, at... |
| 9 | 50-55 | m | [sexy, spirit, desire, dance, strong, serious,... |
| 10 | 56+ | f | [family, single, attractive, compassionate, yo... |
| 11 | 56+ | m | [soul, kiss, lover, respect, respond, playful,... |

Figure 10. The resulting okcupid dataset

### 5.2. Part2

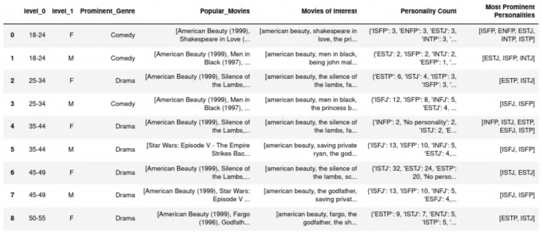
In this phase, we utilized the MovieLens dataset, comprising 1,000,209 anonymous ratings of around 3,900 movies by 6,040 MovieLens users who joined in 2000. The initial step involved integrating three datasets within MovieLens. The user dataset ”user.dat” contained userID, age group code, gender, occupation code, and zipcode. It is important to note that users voluntarily provide all demographic information, which has not been verified for accuracy; only those who provided demographic information were included. The second dataset, ”movies.dat,” contained MovieID, Title, and Genres of the movies. The third dataset, ”ratings.dat,” included UserID, MovieID, Rating, and Timestamp, with ratings observed on a 5-star scale, ensuring each user had at least 20 ratings. Merging these datasets created the MovieLens dataframe, incorporating userID, user age bucket, occupation, movie title, movie genre, and the corresponding rating. The dataset was then stratified into 14 classes based on user age bucket and gender. We identified the prominent genre and the top 10 popular movies for each class based on average user ratings. The resulting dataframe captured the defined classes and a list of the most influential movies for each class.



| | level_0 | level_1 | Prominent_Genre | Popular_Movies |
|---|---|---|---|---|
| 0 | 18-24 | F | Comedy | [American Beauty (1999), Shakespeare in Love (... |
| 1 | 18-24 | M | Comedy | [American Beauty (1999), Men in Black (1997), ... |
| 2 | 25-34 | F | Drama | [American Beauty (1999), Silence of the Lambs,... |
| 3 | 25-34 | M | Comedy | [American Beauty (1999), Men in Black (1997), ... |
| 4 | 35-44 | F | Drama | [American Beauty (1999), Silence of the Lambs,... |
| 5 | 35-44 | M | Drama | [Star Wars: Episode V - The Empire Strikes Bac... |
| 6 | 45-49 | F | Drama | [American Beauty (1999), Silence of the Lambs,... |
| 7 | 45-49 | M | Drama | [American Beauty (1999), Star Wars: Episode V ... |
| 8 | 50-55 | F | Drama | [American Beauty (1999), Fargo (1996), Godfath... |
| 9 | 50-55 | M | Drama | [American Beauty (1999), Star Wars: Episode V ... |
| 10 | 56+ | F | Drama | [American Beauty (1999), Schindler's List (199... |
| 11 | 56+ | M | Drama | [American Beauty (1999), Schindler's List (199... |
| 12 | Under 18 | F | Comedy | [Toy Story (1995), Aladdin (1992), Toy Story 2... |
| 13 | Under 18 | M | Comedy | [Toy Story (1995), Men in Black (1997), Toy St... |

Figure 11. The resulting Movielens data frame

Upon discovering a substantial overlap exceeding 90% between the curated list of prominent movies and those included in the MTBI personality dataset, a thorough examination ensued. Detailed records were maintained for shared movie titles, encompassing the characters featured within each film and their corresponding personalities as outlined in the MTBI dataset. Notably, the manual effort invested in cataloging the gender of each character added a layer of precision to the process. Employing this comprehensive dataset, a systematic exploration unfolded. For each class, an exhaustive iteration through overlapping movies transpired, facilitating a count of encountered personality types. Crucially, the count was specific to relevant personalities; for males, it involved female characters' personalities in the movies, and vice versa for females. The subsequent step involved identifying and selecting the most prevalent personality types specific to each class. These noteworthy personality types were seamlessly appended to our preexisting data frame, enhancing the richness of our dataset.



| | level_0 | level_1 | Prominent_Genre | Popular_Movies | Movies of Interest | Personality Count | Most Prominent Personalities |
|---|---|---|---|---|---|---|---|
| 0 | 18-24 | F | Comedy | [American Beauty (1999), Shakespeare in Love (... | [american beauty, shakespeare in love, the pri... | {'ISFP': 3, 'ENFP': 3, 'ESTJ': 3, 'INTP': 3, '... | [ISFP, ENFP, ESTJ, INTP, ISTP] |
| 1 | 18-24 | M | Comedy | [American Beauty (1999), Men in Black (1997), ... | [american beauty, men in black, being john mal... | {'ESTJ': 2, 'ISFP': 2, 'INTJ': 2, 'ESFP': 1, '... | [ESTJ, ISFP, INTJ] |
| 2 | 25-34 | F | Drama | [American Beauty (1999), Silence of the Lambs,... | [american beauty, the silence of the lambs, fa... | {'ESTP': 6, 'ISTJ': 4, 'ISTP': 3, 'ISFP': 3, '... | [ESTP, ISTJ] |
| 3 | 25-34 | M | Comedy | [American Beauty (1999), Men in Black (1997), ... | [american beauty, men in black, the princess b... | {'ISFJ': 12, 'ISFP': 8, 'INFJ': 5, 'ESTJ': 4, ... | [ISFJ, ISFP] |
| 4 | 35-44 | F | Drama | [American Beauty (1999), Silence of the Lambs,... | [american beauty, the silence of the lambs, fa... | {'INFP': 2, 'No personality': 2, 'ISTJ': 2, 'E... | [INFP, ISTJ, ESTP, ESFJ, ISTP] |
| 5 | 35-44 | M | Drama | [Star Wars: Episode V - The Empire Strikes Bac... | [american beauty, saving private ryan, the god... | {'ISFJ': 13, 'ISFP': 10, 'INFJ': 5, 'ESTJ': 4,... | [ISFJ, ISFP] |
| 6 | 45-49 | F | Drama | [American Beauty (1999), Silence of the Lambs,... | [american beauty, the silence of the lambs, sc... | {'ISTJ': 32, 'ESTJ': 24, 'ESTP': 20, 'No perso... | [ISTJ, ESTJ] |
| 7 | 45-49 | M | Drama | [American Beauty (1999), Star Wars: Episode V ... | [american beauty, the godfather, saving privat... | {'ISFJ': 13, 'ISFP': 10, 'INFJ': 5, 'ESFJ': 4,... | [ISFJ, ISFP] |
| 8 | 50-55 | F | Drama | [American Beauty (1999), Fargo (1996), Godfath... | [american beauty, fargo, the godfather, the sh... | {'ESTP': 9, 'ISTJ': 7, 'ENTJ': 5, 'ISTP': 5, '... | [ESTP, ISTJ] |

Figure 12. The resulting final dataframe incorporating our insights from MovieLens dataset and MTBI Personality dataset

### 5.3. Part3

In evaluating the similarity between provided sentences and a designated set of words, we explored various metrics to measure this likeness. The choice between count, average similarity, and weighted average similarity metrics offers distinct perspectives. The count metric stands out by focusing on the presence of relevant words within a sentence, offering a straightforward indication of the number of words that surpass a specified similarity threshold. Conversely, the average similarity metric presents an overview of the overall similarity between the sentence and the word list, providing an average measure of their likeness. Finally, the weighted average similarity metric enhances this assessment by assigning different weights to words based on their similarity scores, offering a more nuanced evaluation that considers both presence and intensity.

We compared a bunch of sentences that describe different personalities with a specific list of codes. The goal was to see how well these words matched or related to each personality description. For every sentence, we counted how many words from the provided list seemed to be closely connected or similar in meaning to the words used in that sentence. This count helped us understand which sentences had more words that closely matched the given list, giving us an idea of how well the words related to each personality description. It was a way to measure how much these words reflected or related to the traits described in each personality.

# 6. Results

We present visualizations generated from the data processed using the word lists obtained from Cupid Dates and personality types derived from the Movie Lens dataset, categorized based on age and gender. The analysis included several plots, a selection of which is showcased below.
We used a special tool to understand the text better. It did this by finding the main form of words, like changing ”running” to ”run” so the data was easier to compare. Then, it looked at how different personality types were described in a dataset. It checked if the words used to describe personalities were similar to the words people used in their dating profiles. This comparison helped to see if there were any connections between personality descriptions and the words people use when looking for a date.
Using a threshold value(0.5) for similarity, the code counted words that crossed the specified threshold within each personality description. This approach provided an insightful metric for identifying commonalities between the Cupid Dates’ word lists and the Movie Lens personality types. The resulting counts of similar words served as an indicator of alignment or shared vocabulary across different personality categories.

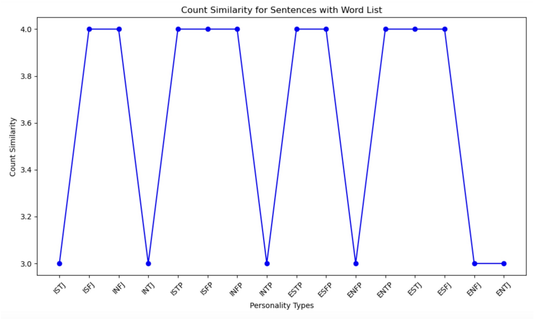


Figure 13. Class ’18-24’,female

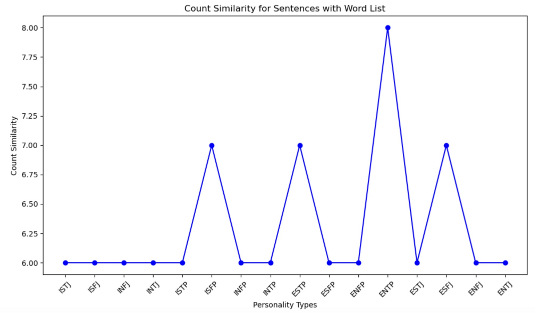


Figure 14. Class ’25-34’,female

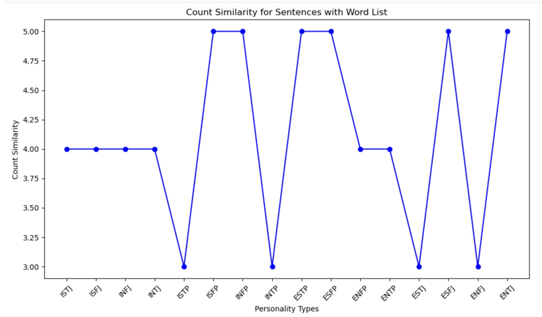


Figure 15. Class ’35-44’,female

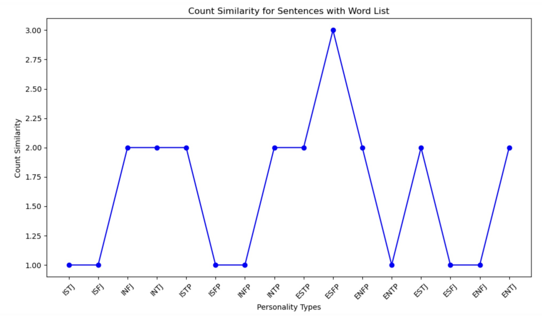


Figure 16. Class ’25-34’,male

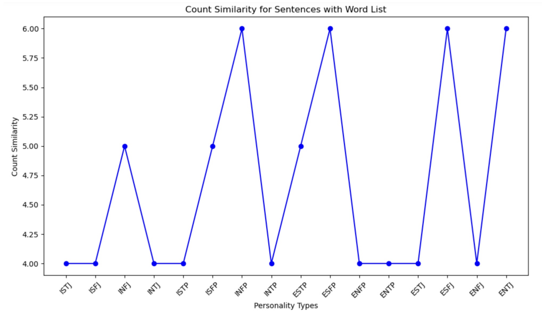


Figure 17. Class ’50-55’,male

Essentially, the plots depict the correlation between personality types and the count of words they share with the important terms extracted for each class. By examining the similarity of words, we aimed to identify which personality types exhibit greater commonality with the extracted important words. This comparison is drawn in relation to the results obtained from all the code words for each class through the MBTI personality prediction.

# 7. Conclusion

The line charts generated from the analysis yielded results in line with our expectations. This substantiates the strategic targeting of audiences by the media, showcasing the most favored personalities. Additionally, the comparison between words extracted from Cupid Dates and the Movie

Lens dataset corroborated anticipated similarities, signifying a meaningful association between personality descriptions and media preferences.

## 8. Limitations

While our analysis couldn't definitively validate the hypothesis, utilizing higher-quality datasets could substantially enhance the accuracy of our conclusions. Further improvements in our metrics could also refine the analysis, enabling a more precise assessment of the relationships between personality descriptions and media preferences.If the dataset contained additional attributes, our outcomes and findings could have potentially improved.

## References

[1] Luoying Yang, Zhou Xu, Jiebo Luo, "Measuring Female Representation and Impact in Films over Time," ACM Transactions on Data Science, in press.

[2] Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. 2022. MBTI Personality Prediction for Fictional Characters Using Movie Scripts. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6715–6724, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

[3] MovieLens dataset - F. Maxwell Harper and Joseph A. Konstan. 2015. The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems (TiiS) 5, 4, Article 19 (December 2015), 19 pages.

[4] Hans Christian, Derwin Suhartono, Andry Chowanda, Kamal Z. Zamli ,"Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging".

[5] Emerging Trends:Word2Vec,Church, Kenneth Ward,Natural language engineering, 2017, Vol.23 (1), p.155-162

[6] OkCupid Profiles : https://www.kaggle.com/datasets/andrewmvd/okcupid-profiles