Differences in portrayal of movie characters

Introduction

We tried to observe the difference in the portrayal of movie characters due to the significant influence that movie has on the society. Some claim that movies reflect the cultural values of the society whereas the others claim that movies update the existing social boundaries. In either case it's quite important to study the portrayal of movie characters in a movie. Here we have considered the language usage of the characters along with the attributes of the characters that include gender, age and race.

The reason for which we have considered language usage of the characters as a measure is that language is used to identify the speaker's psychological and emotional state. We have tried to find out the differences in portrayal of movie characters using linguistic and graph measures. The linguistic metrics include psycholinguistic normatives and LIWC tool. The graph metrics include betweenness centrality and degree centrality.

Data

Primarily we have collected the scripts of the movies from IMSDB. We've the distinguished the scripts that are unusable for our work. After finalizing the scripts of the movies, we developed a parser that takes the script as an input file and returns an output file that has dialogues with the respect to the utterance from the character of a particular scene. So now the data set is of form where we have the dialogue as a key and the attributes that it has include the name of the character and the scene indicator (this is used to distinguish each scene from the other which would help us in centrality measures).

After the development of parser and usage of the parser on the movie scripts that we've collected we have collected few other movie details for each and every movie from IMDB. Various details that have been collected include year of release, director name, writer name and producer name. We have then mapped the character names to the original name of the actors. For each actor, we have collected the age, gender and race. The methods used to determine the actor attributes are:

- Gender: We have designed a gender classifier
- Age: We have collected the date of birth of the actor and the year of release to determine the age of the actor. However, this may be different for the age of the character. So we had to classify into fifteen-year age groups before our analysis.
- Race: We have collected ethnicity information crawling the web from the website ethnicelebs.com.

Analysis

Our analysis primarily comprised of observing the differences based on linguistic and graph measures. Linguistic measures include psycholinguistic normatives and LIWC. The graph measures include degree centrality and betweenness centrality.

Linguistic Measures

Psycholinguistic normatives

These provide the emotional constructs of the speaker such as arousal, valence, concreteness, intelligence levels, etc. and are computed using the dialogues that have been retrieved using the parser that had been developed. The manual annotations of the word ratings are limited to only very small extent due to the time taking procedure. Here we use a linear regression model to compute psycholinguistic normative ratings for a word based on its similarity to a set of concept words. The computed normative score for the word is equal to the sum of a regression coefficient and the sum of product of regression coefficient and the similarity between the given word a concept word.

K=Sum of (R1 * Similarity (word, concept word)) for each concept word

Normative score (word)= R2+ K

Here R1, R2 are regression coefficients

Similarity(word) is the cosine of the binary context vectors with window size 1

Psycholinguistic normatives used are valence, arousal, age of acquisition and gender ladenness. Valence is the degree of positive or negative emotion evoked by the word. Arousal is a measure of excitement in the speaker. Age of acquisition is the average age at which the word is learned and it denotes sophistication of language use. Gender ladenness is a measure of masculine or feminine association of a word. We compute the normative scores only on the content words from each dialogue. So, we remove all the words other than nouns, verbs, adjectives and adverbs.

Linguistic inquiry and word counts

LIWC is a text processing application that processes text and returns the percentage of words that belong to linguistic, affective, perpetual and other dimensions. The data is processed word by word and the corresponding counter is incremented. Finally, the percentage of words belonging to different dimensions are returned. LIWC metrics used are Achievement, Religion, Death, Sexual, Swear.

Graph Measures

We have developed a character network graph. Each node represents a character. We place an edge between two nodes if there is at least one scene in which one character speaks to the other character. Character importance had been measured using degree centrality and betweenness centrality.

Degree centrality: The number of edges incident on a node. Degree centrality measures the number of unique characters that interact with a given character

Betweenness centrality: The number of shorted paths that go through the node. Betweenness centrality measures how much would the plot be disrupted if said character was to disappear completely, i.e., how important is a character to the overall plot.

Results

Character Statistics: The ratio is considerably skewed with male actors having more than twice as many roles and dialogues as compared to female actors.

	MALE	FEMALE
#CHARACTERS	4899	2008
#DIOLOGUES	375711	154897

Production team statistics: Same as above

	MALE	FEMALE
Writers	1326	169
Producers	2866	870
Casting Directors	135	135
Directors	544	46

Character – Writer Ratio: Female writers and directors appeared to produce movies with relatively balanced gender proportions (still slightly skewed towards the male side) as compared to male writers and directors.

Characters Writers gender	FEMALE	MALE
FEMALE	249 (41.2%)	356 (58.8%)
MALE	1541 (27.6%)	4040 (72.4%)

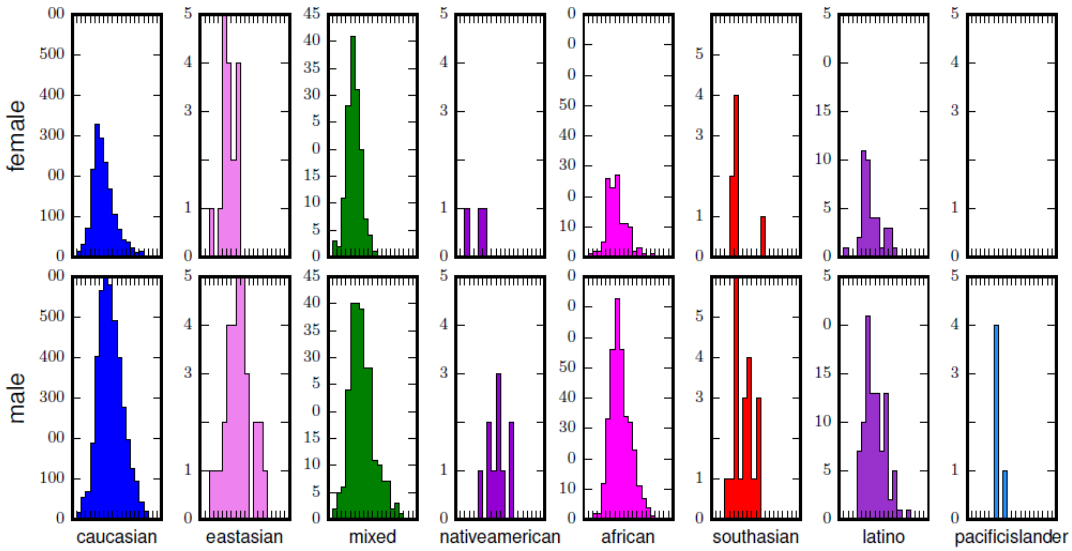
Character – Director Ratio: Same as above

Characters Directors	FEMALE	MALE
FEMALE	114 (39.3%)	176 (60.7%)
MALE	1676 (28.4%)	4220 (71.6%)

Character-Casting Director Ratio: Casting directors appear to have no influence on gender of the characters

Characters Casting Directors	FEMALE	MALE
FEMALE	1374 (29.1%)	3350 (70.9%)
MALE	416 (28.5%)	1046.5%)

Histogram of age for actors belonging to different gender and racial categories



The distribution of age for each category appeared to be approximately normal, except for the native american and pacific islander character groups which were skewed due to a small sample size. The mode of the distribution for female actors appeared to be at least five years less than the mode for male actors.

Results based on Linguistic measures

Gender

Mann–Whitney U test is a non-parametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample. This can be used to determine whether two independent samples were selected from populations having the same distribution. Mann-Whitney U tests between male and female characters along **nine dimensions** have been considered (shown in the table in the next slide) out of which **six** have been proven to be **significant**. Our results also indicate that female character utterances tend to be more positive in valence compared to male characters. Male characters seem to have higher percentage of words related to achievement. Male characters appear to be more frequent in using words related to death as well as swear words compared to female characters.

Mann-Whitney U tests taking gender into consideration

Dimension	MALE	FEMALE	P value
Age of acquisition	**−0.1590**	**−0.1715**	< 10^−5
Arousal	0.0253	0.0246	0.41
Gender	**−0.0312**	**−0.0055**	< 10−5
Valence	**0.2284**	**0.2421**	< 10−5
Sex	0.00015	0.0000	0.08
Achieve	**0.0087**	**0.0080**	< 10−5
Religion	0.0025	0.0022	0.10
Death	**0.0025**	**0.0016**	< 10−5

Swear	**0.0037**	**0.0015**	< 10−5

Race

Kruskal-Wallis test (a generalization of Mann-Whitney U test for more than two groups) on each of the nine metrics with **race** as the independent variable is considered to study the differences in portrayal of the racial categories. For **gender ladenness**, caucasian and mixed race characters have higher medians than african and native american characters. In **sexuality**, latino and mixed race characters were found to have higher median than at least one other racial group with significance indicating a higher degree of sexualization. In **swear word usage**, african characters use higher percentage of swear words.

Age

Linear regression models with each dimension as the dependent variable and age as the independent variable is considered to examine the relationship between **age** and different metrics. According to the results, an increase in sophistication of word usage with age had been observed. **Arousal,** on the other hand, has a significant negative coefficient indicating a decrease in activation as age increases. Usage of words related to **achievement** and **religion** increase with age.

Coefficients of age for linear regression models along each dimension along with p-values

Dimension	Beta1 (X 10^{-3})	p-value
Age of acquisition	3.9	$<10^{-10}$
Arousal	-1.1	$<10^{-10}$
Gender	−2.5	$<10^{-10}$
Valence	0.078	0.7
Sex	-0.25	$<10^{-5}$

Achieve	0.26	$<10^{-10}$
Religion	0.12	0.001
Death	-0.039	0.2
Swear	-0.34	$<10^{-5}$

Results based on Graph measures

Gender

Mann Whitney U tests between male and female characters for different genres have been taken into consideration. Male characters were found to have higher values in both metrics Betweenness and Degree centrality compared to female characters. Significant differences were observed in movies of the **horror genre** in which the median degree centrality or females (0.221) was much higher than median degree centrality of males (0.166). This is in accordance with the prior studies which outline **women** to have prominent roles in horror movies, mainly as victims of violent scenes.

Race

To examine differences in major roles across the racial categories, we perform **Kruskal-Wallis tests.** Latino characters were observed to hold more **non central roles** based on the fact that they possessed significantly lower degree centralities compared to caucasian and south asian races. Caucasian characters were found to have median **betweenness centralities** significantly **higher** than at least one other race. We noticed that characters of the native american race exhibited considerably lower medians in both degree and betweenness centralities than caucasian, african and mixed characters.

Age

We built a **linear regression** model on the two centralities with age as the primary independent variable. In the case of **degree centrality**, the regression coefficient β was found to be equal to 0.003, indicating a positive dependence. In **betweenness centrality**, the regression coefficient was also positive, given by $\beta = 8.41 \times 10{-}4$. As both of these metrics signified a **positive correlation** for character importance with age, we could conclude that as the age of characters' progresses, there was an increase of interaction with other characters in the movie as well as higher importance in the movie plot.