An Unfair Affinity Toward Fairness: Characterizing 70 Years of Social Biases in Bollywood

Kunal Khadilkar¹ and Ashiqur R. KhudaBukhsh¹
¹Carnegie Mellon University
{kkhadilk, akhudabu}@andrew.cmu.edu

Abstract

Bollywood, aka the Mumbai film industry, is one of the biggest movie industries in the world. With a current movie market share of worth 2.1 billion dollars and a target audience base of 1.2 billion people, Bollywood is a formidable entertainment force. While the entertainment impact in terms of lives that Bollywood can potentially touch is mammoth, no NLP study on social biases in Bollywood content exists. We thus seek to understand social biases in a developing country through the lens of popular movies. Our argument is simple – popular movie content reflects social norms and beliefs in some form or shape. We present our preliminary findings on a longitudinal corpus of English subtitles of popular Bollywood movies focusing on (1) social bias toward a fair skin color (2) gender biases, and (3) gender representation. We contrast our findings with a similar corpus of Hollywood movies.

1 Introduction

On a diachronic corpus of popular entertainment, what types of social biases can we analyze and detect? In this paper, we analyze a curated corpus of film subtitles from the Bollywood and Hollywood film industry, spread over 70 years (1950 to 2020) and present our preliminary findings on three aspects (1) social bias toward a fair skin color (2) gender biases, and (3) gender representation. While studies analyzing gender stereotypes across different languages [Lewis and Lupyan, 2020] and detecting bias in word embeddings [Garg et al., 2018] exist, barring few lines of work focusing on a single Bollywood movie [Chatterjee, 2016] or a small subset of movies [Khan and Taylor, 2018], no existing work has focused on analyzing social biases and under-representation in Bollywood films. We perform a first ever large-scale study comprising 1400 films, to uncover implicit social and racial biases in the entertainment industry, using computational science and natural language processing techniques.

In Figure 1, what do the prominent presence of negative adjectives such as wanton in old Bollywood movies (see, Figure 4(a)) juxtaposed with the presence of positive verbs such as respect (see, Figure 4(b)) tell us? In this paper, we explore a wide array of NLP techniques to analyze our research questions through the lens of popular movies. We use

1. Simple count-based statistics relying on highly popular lexicons [Warriner et al., 2013] and gender representation studies [Twenge et al., 2012; Sendén et al., 2015];
2. Cloze test, an analysis technique that have a solid grounding in psycholinguistics literature [Taylor, 1953; Smith and Levy, 2011]. To the best of our knowledge, for the first time, we explore a recent technique [Petroni et al., 2019] previously used to mine political insights [Palakodety et al., 2020] in the context of uncovering social biases. Through a series of cloze tests on a language model [Devlin et al., 2019] fine-tuned on our data sets, we present our findings.
3. Analysis of aligned diachronic word embedding spaces using recently proposed techniques [Hamilton et al., 2016].

2 Data Set

We construct the following two data sets.

1. Bollywood movies, $D_{bolly}$: We consider movies spanning seven decades (1950–2020). For each decade, we retrieved subtitles [Lison and Tiedemann, 2016] of 100 movies (700 total films).
2. Hollywood movies, $D_{holly}$: Similar to Bollywood movies, we considered 100 top-grossing movies from each of the seven decades (700 total films). Overall, $D_{bolly}$ and $D_{holly}$ consist of 1.1M dialogs (6.2M Tokens) and 1M dialogs (5.4M Tokens), respectively.
For a subset of our analyses, we divide our corpus into three buckets: (1) Films from 1950 to 1969 ($D_{\text{old}}^{\text{bolly}}$ and $D_{\text{old}}^{\text{holly}}$); (2) Films from 1970 to 1999 ($D_{\text{mid}}^{\text{bolly}}$ and $D_{\text{mid}}^{\text{holly}}$); and (3) Films from 2000 to 2019 ($D_{\text{new}}^{\text{bolly}}$ and $D_{\text{new}}^{\text{holly}}$).

3 Our Analyses

For our first two analyses of gendered pronouns and cloze tests, we restrict ourselves to the temporally extreme subsets of old and new movie sub-corpora.

3.1 Pronouns as a Proxy for Representation:

Following extensive literature on gendered pronouns’ relative distributions and their implications [Twenge et al., 2012; Senden et al., 2015], we first consider a simple measure of gender representation: relative occurrence of pronouns of each gender (Men: he, him. Women: she, her). Let $N_w$ denote the number of times a token $w$ appears in a corpus. We define Male Pronoun Ratio (MPR) as follows:

$$MPR = \frac{N_{\text{he}} + N_{\text{him}}}{N_{\text{he}} + N_{\text{him}} + N_{\text{her}}} \times 100.$$  

Figure 2 plots MPR of our decade-wise movie data sets and contrasts with MPR computed using google n-grams. While compared to 1950s both Bollywood and Hollywood demonstrate considerable decrease in MPR, our results indicate that even now, both Bollywood and Hollywood exhibit comparable skew in gendered pronoun usage.

![Figure 2: MPR in $D_{\text{bolly}}$ and $D_{\text{holly}}$](image)

3.2 Cloze Test Using Language Models:

When presented with a sentence (or a sentence stem) with a missing word, a cloze task [Taylor, 1953] is essentially a fill-in-the-blank task. For instance, in the following cloze task: In the [MASK], it is very sunny, summer is a likely completion for the missing word. Given a cloze test, $\text{BERT}$, a well-known language model [Devlin et al., 2019], outputs a series of token ranked by probability. In fact, in the above cloze test, the top three tokens (ranked by probability) predicted by $\text{BERT}_{\text{base}}$ are: summer, winter and spring. Recent lines of research has explored $\text{BERT}$’s masked query prediction for (1) knowledge base extraction [Petroni et al., 2019] and (2) mining political insights [Palakodety et al., 2020].

Following [Palakodety et al., 2020], we fine-tune $\text{BERT}$ on four sub-corpora: $D_{\text{bolly}}^{\text{old}}, D_{\text{holly}}^{\text{old}}, D_{\text{bolly}}^{\text{new}}$ and $D_{\text{holly}}^{\text{new}}$, with 100 movies in each corpora. We denote the pretrained $\text{BERT}$ model as $\text{BERT}_{\text{base}}$ and a fine-tuned $\text{BERT}$ on corpus $D$ as $\text{BERT}_D$. Out of a thorough analysis with several cloze tests with phrase variations, in Table 1 we report three such results for the following cloze tests: (1) A beautiful woman should have [MASK] skin. (cloze1); (2) A woman should be a [MASK] by occupation. (cloze2); and (3) A man should be a [MASK] by occupation. (cloze3).

For all cloze tests, we evaluate robustness to phrase variations to ensure that our results of the masked words are qualitatively similar. For instance, one of the phrase variations of cloze2 we consider is A woman should be a [MASK] by profession. Similarly, one of the phrase variations for cloze3 we consider is For a woman to be beautiful, she should have [MASK] skin. Overall, our findings remain qualitatively similar across the phrase variations.

### Summary of findings

Given cultural and social bias toward fair skin prevalent in India for years [Karan, 2008], it is not surprising that Bollywood exhibits an affinity toward fairness. However, our analysis indicates Hollywood also shows bias toward lighter skin color. Further, we observe cloze completions for both genders across both movie industries improve over time. As compared to men, the completions for women in older Bollywood movies exhibit stronger negative terms such as slave and prostitute.

<table>
<thead>
<tr>
<th>Probe</th>
<th>$\text{BERT}_{\text{base}}$</th>
<th>$\text{BERT}_{\text{old}}^{\text{bolly}}$</th>
<th>$\text{BERT}_{\text{old}}^{\text{holly}}$</th>
<th>$\text{BERT}_{\text{mid}}^{\text{bolly}}$</th>
<th>$\text{BERT}_{\text{mid}}^{\text{holly}}$</th>
<th>$\text{BERT}_{\text{new}}^{\text{bolly}}$</th>
<th>$\text{BERT}_{\text{new}}^{\text{holly}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>cloze1</td>
<td>soft, beautiful, pale, tanned, smooth</td>
<td>fair, no, pale, tanned, tan</td>
<td>fair, tanned, golden, smooth, pale</td>
<td>fair, tanned, golden, smooth, pale</td>
<td>fair, tanned, golden, smooth, pale</td>
<td>fair, pale, blue, golden, gold</td>
<td>fair, pale, tanned, golden, dark</td>
</tr>
<tr>
<td>cloze2</td>
<td>woman, man, widow, doctor, slave, soldier, bachelor, merchant, farmer, lawyer, servant [4.8]</td>
<td>prostitute, servant, woman, slave, bachelor, doctor, lawyer, man, widow, maid, worker [4.64]</td>
<td>doctor, woman, servant, lawyer, maid, man, nurse, teacher, gardener, lady, hindu [5.7]</td>
<td>woman, slave, servant, lawyer, maid, man, doctor, lawyer, peasant, maid, wife [5.3]</td>
<td>woman, lawyer, doctor, nurse, teacher, man, writer, secretary, prostitute, professional [5.7]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Cloze test results. Predicted tokens are ranked by decreasing probability. Positive and negative words are color coded with blue and red, respectively. The number in the bracket represents the average valence score (obtained from [Warriner et al., 2013]) calculated for the answers to the cloze test.
Table 2: Percent increase in average valence score for cloze test completions between old movies and new movies.

<table>
<thead>
<tr>
<th></th>
<th>Bollywood</th>
<th>Hollywood</th>
</tr>
</thead>
<tbody>
<tr>
<td>Women</td>
<td>22%</td>
<td>1%</td>
</tr>
<tr>
<td>Men</td>
<td>5%</td>
<td>15%</td>
</tr>
</tbody>
</table>

We now present a quantitative evaluation of our completion results. We consider a well-known lexicon of emotional valence ratings [Warriner et al., 2013] of nearly 14,000 English words to quantify the change of cloze test completions over time. The valence score of these words are presented in a scale of 1 to 10 with 10 indicating highly positive and 1 indicating highly negative. For example, the emotional valence score of happy and sad are 8.47 and 2.10, respectively. For a given data set and a cloze test pair, we compute the average valence score of the completions (listed in square brackets in Table 1). We further note that comparing completion words across the two genders can be technically difficult simply because we observe that there can be potential bias in the scores. For instance, the valence scores for man and woman are 5.42 and 7.09, respectively. Hence we restrict ourselves to comparing within a specific gender for a given movie industry.

Table 2 lists the percentage of increase in the valence score of the completion for a particular gender across different movie industries. We note that for both Bollywood and Hollywood, the valence scores for both genders improved over time. However, for Bollywood, we notice that rate of increase for women is substantially more pronounced than that for men. This observation aligns with the continual fight for gender equality in India [Nayak and Mahanta, 2012] and major movements that have mobilised voices for women’s right to work [Chen, 1995], financial independence [Goyal and Parkash, 2011], and marital laws [Nigam, 2005].

### 3.3 Aligning Diachronic Word Embeddings

The meaning of words and the context in which they are used change over time [Xie et al., 2019]. The language spoken in a community is representative of the cultural norms and customs followed in that region. Existing hypotheses [Bybee, 2006] indicate that word frequency may play a role in changing the meaning of the words over time. The meanings of less-frequent words are more susceptible to drifts than that of highly frequent words. [Hamilton et al., 2016] provide a robust multilingual approach to align diachronic word embeddings using orthogonal procrustes. We follow the same method to align different sub-corpora for Bollywood and Hollywood.

We train word2vec [Mikolov et al., 2013] with SGNS (Skip-gram with Negative Sampling), to create embeddings for each of the buckets mentioned above. Let \( \mathcal{V} \) be the matrix of word embeddings learnt for period \( t \) for vocabulary \( \mathcal{V} \). Following [Hamilton et al., 2016], we align the word embeddings using the top 10,000 common tokens present across time periods \( t \) and \( t + 1 \) by optimizing:

\[
R^t = \arg \min_{Q^{t+1}} \| Q \mathcal{W}^t - \mathcal{W}^{t+1} \|_F, \tag{1}
\]

where \( R^t \in \mathbb{R}^{d \times d} \).

### Affinity toward Fairness

Figure 3 visualizes the nearest neighbors of beautiful in our aligned embedding spaces of Hollywood and Bollywood sub-corpora. As shown in Figure 3, the age-old affinity toward lighter skin in Indian culture [Dlova et al., 2015; Chat-topadhyay, 2019; Madhukalya, 2020] is reflected through the consistent presence of fair among the nearest neighbors of all three Bollywood sub-corpora. Although our cloze tests indicate Hollywood also exhibits bias towards lighter skin color, our diachronic word embedding analysis reveals that possibly the bias is less pronounced than in Bollywood.

### Portrayal of Women and Men

We next focus on the portrayal of women and men. We observe that the valence scores for both genders across both movie industries show a similar pattern. The scores are the lowest during the 1970-1999 time period. The valence scores for the newer movies are better than the scores for the older movies. The dip in the valence scores during the time-period of 1970-1999 in India can be ascribed to a social and cultural crisis influenced by an unstable political climate (assassinations of two prime ministers [Hardgrave Jr, 1985; Kaarthikkenyan and Raju, 2008], two major wars between India and Pakistan [Schofield, 2010; Bose, 2009], and a large overlap with a pre-economic liberalization period [Pedersen, 2000].

### Social Evil

Our next focus is on a social evil discriminatory toward women. We consider the ‘dowry’ system that has plagued the Indian society for a long time [Dalma and Lawrence, 2005]. Dowry refers to a transaction of tangible financial objects in the form of durable goods, cash, and real or movable property between the bride’s family gives and the bridgroom, his parents and his relatives as a condition of the marriage. Although, legally, dowry has been prohibited in India since 1961 [Rao, 1973], this practice has continued well after its legal prohibition and has a strong link to social crises such as female foeticide [Ghansham, 2002], domestic abuse and violence [Banerjee, 2014; Rastogi and Thery, 2006], and dowry deaths [Ahmad, 2008]. However, while the practice continued abated, recent studies have reported positive changes in the society where the general attitude towards the system has become negative [Srinivasan and Lee, 2004].
We were curious to analyze how word embedding spaces reflect the attitude toward dowry. As shown in Figure 5, we observe that while nouns such as money, debt, jewellery, and loan are the nearest neighbors in older films indicating compliance to this practice, modern films exhibit non-compliance (e.g., guts and refused) and indicate some of the consequences of such non-compliance (e.g., divorce and trouble) in the form of nearest neighbors. We find that indeed, films provide a snapshot of cultural values of a particular country thus allowing us to gauge the progress of a nation over time.

4 Future Directions

Our future lines of work include (1) tracking moral sentiment [Xie et al., 2019]; (2) exploring debiasing techniques [Manzini et al., 2019]; (3) analyzing representation of other minorities (e.g., Muslims, dalits and the LGBTQ community); and (4) exploring a newly-proposed machine-translation data set distance measure [KhudaBukhsh et al., 2020] to uncover contrasting social aspects.

While our current focus is on characterizing gender bias and representation, our rich data set has untapped potential in analyzing a wide range of research questions. For instance, in line with the cloze tests conducted in [Palakodety et al., 2020], Table 3 and 4 list the results for two additional cloze tests: (1) The biggest problem of India is [MASK]. (referred as cloze4) - For Bollywood and (2) The biggest problem of America is [MASK]. (referred as cloze5) - For Hollywood.

We observe that the dynamic political conditions are reflected in the completion results (e.g., Kashmir, Pakistan and Russia) [Schofield, 2010; Bose, 2009; LaFeber and Abbott, 1972]. We also note that the list of ongoing problems in the U.S. contains the major issue on which the 2020 election is being fought over: racism.
References


[Madhukalya, 2020] Anwesa Madhukalya. These dialogues from Bollywood blockbusters are so sexist that you’ll want to pull your hair out, 2020.


