Statistical Analysis of Topic Bias in the New York Times Sunday Book Reviews

Just Review Team

.txtLab, Digital Humanities, McGill University, Montreal, QC

justreview.org

Abstract: The aim of this study was to confirm the existence of gendered topic bias in literary culture, and quantify the differences between the books written by men and women that are selected for review by the New York Times. As we expected, certain book themes were dominated by one gender or the other.

1.   Introduction

The Just Review Team performed a statistical analysis of the New York Times Book Reviews to confirm the existence of Gendered Topic Bias in literary reviewing culture. This dataset had author genders coded as a binary (m/f), leading to a lack of representation for non-binary identifying authors. As a result, this study can only analyze statistical differences in the genre of books reviewed as a comparison between men and women.

The aim of this study was to better understand and quantify the differences between the books written by men and women that are selected for review by the NYT.

2.   Overview of the Data

The dataset was collected by scraping the New York Times website and collecting the weekly Sunday reviews. We resolved each review to the book being discussed, and then used the GoodReads website to choose the top 5 themes of each book. We were able to fully resolve 5,693 books from their NYT review to their top themes. The themes assigned to books on **GoodReads.com** are crowdsourced, or voted on by the community. We made the assumption that these labelled themes accurately represented the books themselves.

In the final dataset, the books are not evenly distributed by year, so we avoided making any longitudinal claims based on this data (Fig. 1).
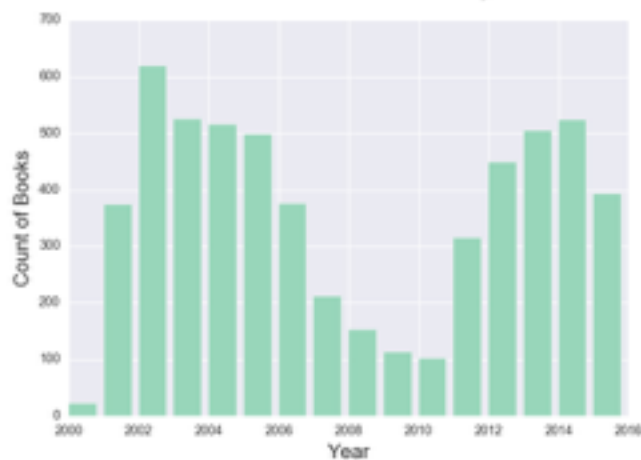

Fig. 1: Number of books in dataset by year.

In the final dataset, we observe that about two thirds of the books reviewed are written by men (Fig. 2). Before we even begin to investigate the themes of the books, we can see a serious discrepancy between the two genders in who is even afforded space in the New York Times reviews.

Because Goodreads themes are entered by users, there were a total of 212 unique themes over the 5,693 books. After taking a look at this
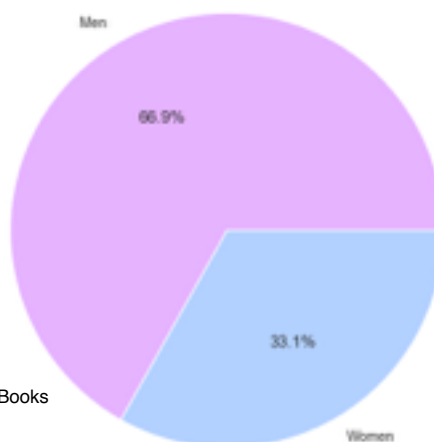

Fig. 2: Percentage of Books Reviewed by Gender

list, we noticed that many of them could be grouped into broader categories. For example, Social Movements, Social Issues, and Social Justice were grouped together into a broader 'Social Issues' Category. While this was a difficult process that involved much discussion, we were able to fit the themes into 33 distinct categories. We also flagged tags for removal as we did not feel they were descriptive enough of the content of the book. For example, we removed Contemporary, Novels, Classics, etc. You may find a detailed breakdown of each of our theme groups in the following public Google Sheet.

3.   Results: Fiction vs. Non-Fiction

A majority of the books were tagged as either Nonfiction or Fiction, as can be seen in Figure 3. While about 12 percent did not resolve to a F/NF label through Goodreads, we were able to perform analysis on the 88 percent that was labelled.

	Women	Men	All
Fiction	1111	1665	2776
Nonfiction	602	1615	2217
Total Identified	1713	3280	4993
Total Identified as % of all books	90.78%	86.2%	87.7%

Fig. 3: Overview of data regarding books labelled on Goodreads as fiction or non-fiction

Of all the books written by women, a majority are labelled as Fiction (59 percent), while only 32 percent are Non-Fiction (Fig. 4). Comparatively, the distribution across F/NF is far more even for books written by Men, with 42 percent Non-Fiction and 44 percent Fiction. This points to the first manifestation of bias in literary culture, the unequal representation of women's voices in reviewed Non-Fiction books.
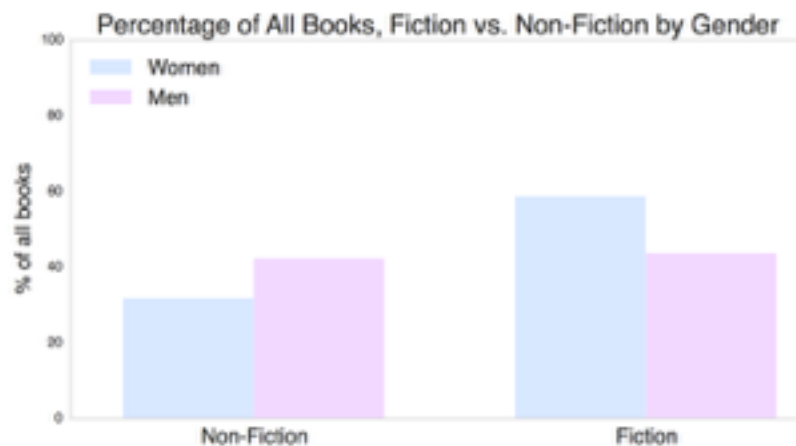

Fig. 4: Breakdown of books labelled as Fiction or Non-Fiction

4.   Results: Gender Across Genres

GoodReads.com is a community based website where people come together to review and discuss books. Each book has a unique webpage where users crowd-source information about the book. One of these data-points is the genre of the book, where people vote and the most popular choices float to the top of the page. We scraped the GoodReads website to compile the listed genres for all the books in our dataset. Of the 5,693 books in our dataset, we were able to resolve 5,461 of the books to its' corresponding top three themes.
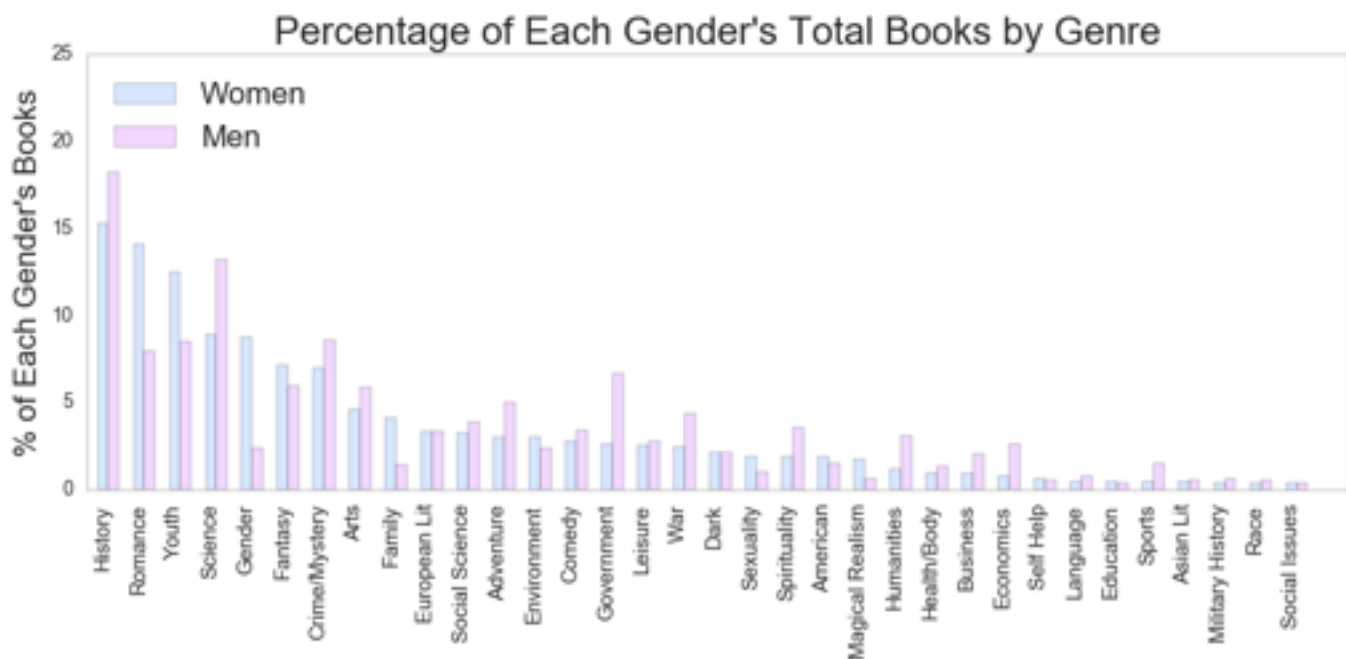

Fig. 5: Percent of books reviewed by each gender and its' corresponding genre.

The table on the next page (Fig. 6) presents, for each theme, the number of books by each gender and the percent by each within the category, the p-values and the direction of the statistical skew.

The p-values were found using a binomial test with a probability of 66.9% for male writers, as the dataset had more male than female writers. The purpose of the p-value is to measure how likely it is that the outcome would occur if the data was sampled completely randomly. That is, a perfectly random distribution for any given genre would 66.9% male and 33.1% female.

Theme	Women		Men		P-value	Direction of Skew
	#	%	#	%		
Gender	166	64	92	36	< 0.001	F
Romance	266	47	304	53	< 0.001	F
Government	50	16	254	84	< 0.001	M
Family	79	59	54	41	< 0.001	F
Economics	15	13	102	87	< 0.001	M
Science	168	25	504	75	< 0.001	M
Humanities	23	16	120	84	< 0.001	M
Youth	236	42	325	58	< 0.001	F
Magical Realism	33	57	25	43	< 0.001	F
Spirituality	36	21	138	79	< 0.001	M
Sports	9	13	58	87	< 0.001	M
War	47	22	166	78	< 0.001	M
Adventure	58	23	192	77	< 0.001	M
Business	18	18	81	82	0.001	M
Sexuality	37	47	41	53	0.011	F
History	289	29	693	71	0.013	M
Crime/Mystery	132	29	328	71	0.042	M
Arts	88	29	224	72	0.071	M
Fantasy	135	37	227	67	0.094	F
Comedy	53	29	132	71	0.212	M
Environment	57	38	93	62	0.225	F
American	36	39	57	61	0.271	F
Social Science	62	29	149	71	0.273	M
Military History	8	24	26	76	0.277	M
Health/Body	19	27	52	73	0.313	M
Education	10	38	16	62	0.540	F
Leisure	48	31	108	69	0.553	M
Self-Help	13	37	22	63	0.595	F
Race	8	28	21	72	0.694	M
Asian Literature	9	28	23	72	0.708	M
European Literature	63	33	129	67	1	M
Dark / Death	41	33	84	67	1	M
Social Issues	8	32	17	68	1	M

Statistically significant ↑ (marker between Crime/Mystery and Arts rows)

As can be seen in the table, there are many statistically significant themes that have a very strong gendered skew. Books about Gender, Romance and Family top the list for women-friendly themes, while Government, Economics and Science are male-dominated. Just Review is

seeking to change this gendered domination of certain subjects in order to allow all people to be able to write about all topics and be met with an appropriate response.