Gender Stereotypes in News Media: Empirical Evidence

Maria Grebenshchikova

Abstract

Over the past decades, women have achieved progress in labor force participation, breaking the glass ceiling, and narrowing the gender wage gap. And even though women constitute approximately 50 percent of the world population, they are still underrepresented in many mathematically-intensive occupations and decision-making positions. Since media played an important role in shaping people's beliefs and creating unconscious biases, I examine whether women are portrayed in news media in accordance with traditional stereotypes and gender roles or not. The methodological idea is to test the association between textual and visual data of the major US newspaper with the help of machine learning tools, advanced text analysis, and econometric models. This is the first attempt to combine both high-dimensional text analysis and large-sample visual analysis. I document a negative association between the tendency of an article being professional and the share of women depicted on article images. In particular, women are less likely to appear on images of science, politics, and economics articles and more likely in appearance, fashion, and family articles. This result sheds more light on factors influencing educational and occupational choices of women.

Keywords: Media Economics; Gender Stereotypes; Gender Data Gap.

Contents

1 Introduction **3**

2 Literature review **4**

2.1 Literature on the role of media 4

2.2 Literature on representation of women in the media 5

2.3 Literature on gender stereotypes and occupational segregation 6

2.4 Literature on methods 6

3 Data **7**

3.1 Textual Data 7

3.2 Visual Data 7

4 Empirical strategy **8**

4.1 Text analysis 8

4.2 Topic modelling 9

4.3 Text Regression 12

4.4 The effect of the empowerment movement 15

5 Conclusion **16**

6 Appendix **20**

1 Introduction

Over the past decades, women have achieved progress in labor force participation, breaking the glass ceiling, and narrowing the gender wage gap. And even though women constitute approximately 50 percent of the world population, they still represent only 21.3% of computer programmers, 26.9% of chief executives, 39.9% of financial analysts, 40.2% of physicians and surgeons. On the contrary, there are more than 80% of female physician assistants, nurses, elementary school teachers, and teacher assistants (BLS, 2019). In 2018, women CEOs of the Fortune 500 companies were outnumbered by men named James running such top companies (Miller et al., 2018).

There is an ongoing debate regarding the reasons behind the underrepresentation of women in different occupations, especially mathematically-intensive. Kahn and Ginther (2017) highlights that cultural and family's gender stereotypes influence educational and occupational choices. Such gender stereotypes can be broken by exposure to experts of the same gender. Relevant role models can also impact the aspirations of women and their sense of competitiveness in the labor market of male-dominated occupations. And, therefore, narrow the gender gap of aspirations (Beaman et al., 2012). Although research papers in economics more and more focus on gender stereotypes, there is still a plethora of opportunities to study.

This paper aims to address the topic of gender stereotypes from the perspective of media economics. News media provide role models and shape attitudes of people towards political and social issues, economic and everyday decisions. Anecdotal evidence suggests that despite the evolution of gender norms, women are still portrayed in news media in accordance with traditional stereotypes as home-makers and caregivers. Conversely, images of men appear more frequently in articles about success and business, authority and power. All of these factors may facilitate discrimination on the labor market and influence the occupational choices of women. However, such observations have not been tested statistically yet.

In this paper, I use sophisticated and modern statistical tools to document that women, contrary to men, are less likely to be portrayed as workers and professionals in articles published by the New York Times – the large news media in the US. The main findings indicate that women are underrepresented in articles related to business, politics, science, and technology. On the contrary, women are overrepresented in stories related to fashion, appearance, home, and well-being. Moreover, the text analysis revealed common patterns of gendered language. Article texts with pictures of women mostly describe body, family, and activism, while articles with men images relate to success, power, and career. Besides, I check whether this pattern changed after the explosion of the empowerment movement #MeToo in October 2017. To this date, results suggest minor, but statistically significant changes.

For economists, the data encoded in the text can be complementary to the administrative and structured data conventionally used in research. Because of that, there is an increasing number of empirical economics papers focusing on texts. Existing studies in media economics have primarily focused on textual analysis, while the most popular media consumed today is nonverbal such as videos and images. Besides, nonverbal information is more persuasive than verbal information is.

Up to this time, academic papers have commonly relied on hand collection and coding of nonverbal media content. Therefore, studies have been restricted to small samples, number of sources, types of content. In this research, I expand data limitations and use the state-of-the-art Machine Learning algorithm to automatically process a large amount of visual data, identify the gender of people on images, and proceed with quantitative analysis. This allows me to construct a 100-times larger dataset than in previous studies.

The methodological idea of this paper is to look at the relationship between textual and visual data. In addition to the modern facial recognition techniques, I use a lasso logistic regression model to deal with high-dimensional data – digital texts with over 1,000 features. Methods used in this study advance the mainstream methodological approach employed by researchers in media economics.

This paper is organized as follows. Section 2 presents the existing literature on media, gender stereotypes, and methods of dealing with high-dimensional data. Section 3 describes the data collection process, dataset, and features of variables. Section 4 presents the empirical methods and discusses the findings. Section 5 briefly concludes.

2 Literature review

The research contributes to several strands in the literature on media economics and gender stereotypes. The literature review presents the main theoretical and empirical studies addressing the issue of attitude formation through media channels as well as the consequences of media bias in terms of stereotypes and occupational segregation. Besides, the literature on methods and analytical techniques used in this paper is discussed.

2.1 Literature on the role of media

Literature suggests the conventional way of defining media bias. Gentzkow and Shapiro (2006) define media bias as a deliberate choice of newsmakers to slant information "by selective omission, choice of words, and varying credibility ascribed to the primary source", which leads to a completely reverse impression of what happened. More technically, media bias occurs when a media report "deviates from its best guess of the true state".

One of the most relevant theoretical studies aimed at building a theoretical model in which media bias occurs when newsmakers slant their content in accordance with the beliefs of the audience (Gentzkow and Shapiro, 2006). Such decisions are based on the desire of media to create a reputation of providers of high-quality information. The model was further tested empirically on the US daily newspapers by Gentzkow and Shapiro (2010).

According to the studies on media economics, content from the news can have a significant impact on beliefs and outcomes. Gentzkow and Shapiro (2004) study how beliefs depend on news media exposure and education. Most empirical studies are focused on the role of media bias in the formation of specific beliefs. In particular, the effect of media on political attitudes has been studied extensively. Gentzkow (2006) and Gerber et al. (2009) document that news media influence preferences and participation of voters. Few recent studies looked

at the wider picture. DellaVigna and Gentzkow (2010) investigate persuasion aimed at voters, donors, investors, and consumers and found the strongest effect on the first two targets.

However, existing empirical studies mainly focus on the preferences, attitudes, and behavior of consumers and voters. Therefore, there is a lack of empirical evidence of media bias effects on stereotypes. Gilens (1996) discuss the connection between racial bias in the coverage of poverty by media and public misperceptions of the poor. However, Gilens does not study gender stereotypes – the main focus of this paper. To my knowledge, economic studies have not documented any relationship between media bias and gender stereotypes leading to various labor market outcomes.

2.2 Literature on representation of women in the media

Concerns about the misrepresentation of women in the media have played a prominent role in academic debates. The first studies related to the topic were published 40 years ago. Luebke (1989) highlighted that images of men and women in newspapers correspond to gender role stereotypes. For instance, men are portrayed as professionals and sportsmen, while women as spouses and caregivers. Further, Tuchman (2000) extended the search and analized not only newspapers, but also television and magazines. The author argue that women are mostly represented by mass media as victims, consumers, homemakers, and mothers. As for the paid workplace, women are mostly depicted as clerical, care-oriented ("pink-collar") jobs. Therefore, media, according to Tuchman, do not encourage women to get education, pursue training or make decisions that lead to more independent and powerful positions.

Despite the evolution in gender norms, roles, and attitudes in society, women are still stereotypically depicted in media. A few recent papers not only document a persistence of women misrepresentation in the media but also explore channels through which the media effect works (Simon and Hoyt, 2012) and propose policy recommendations (Sharma, 2013). Simon and Hoyt conducted two experiments to understand whether media images portraying non-stereotypical roles for women, in comparison to those that show stereotypical roles for women, influence women's attitudes toward gender roles and their responses to leadership tasks. Scholars report that exposure to non-stereotypical images reduces negative self-perception and increases the leadership aspiration of female participants comparing to those who received only stereotypical images. Sharma point to the importance of changes in the media images approach to make a difference, rather than reinforce the representation of traditional roles of men and women.

The concentration of male and female workers differs between occupations. There are many reasons for such occupational segregation. Firstly, an employer discrimination problem in one occupation may be harsher than in another. Secondly, a self-selection problem may occur due to social norms or institutional constraints regarding the allocation between occupations and gender groups. Finally, occupational choices can be affected by differences in comparative advantages which are results of pre-labor market human capital investments and non-labor activities. All these factors may partially result from media bias and gender stereotypes.

2.3 Literature on gender stereotypes and occupational segregation

One of the first frameworks aimed to address occupational segregation was introduced by Bergmann (1974). The model accounts for "occupational exclusion" – the situation in which one group (e.g., female workers) is clustered in a subset of the occupations in the labor market. From the framework's analysis follows that as occupational exclusion reduces via institutional constraints or social norms, a shift of the female workers' labor supply from the"women-concentrated occupation" to the male ones occurs. Moreover, as women get better at men's jobs, women's wages rise and average men go down further. Besides, as women get better at women's jobs, both women and men gain.

According to Goldin (2014), as human capital investments of female and male workers have been converging over time, women still underrepresented in decision-making positions and occupations. Many factors were introduced so far. Among them, social norms, gender stereotypes and attitudes, employers discrimination, competitiveness in the occupation, and so on. However, these arguments take into account differences between occupations. Goldin introduces the empirical evidence that the larger part of the current earnings inequality comes from within occupational differences, rather than between them.

2.4 Literature on methods

Existing academic papers relied a lot on manual collection of data. However, there is a growing commitment to use machine learning algorithms and other programming tools to extend samples and increase the number of resources. In the recent working paper, Boxell (2018) employs the Internet Archive's Wayback Machine to collect historical data on the Internet Usage and Microsoft's Emotion API to estimate emotional content of images. These tools allow for an automatic collection of nonverbal data and creation of the dataset with 1 million observations, which is 100 times more in comparison to conventional methods of data collection.

There is also extensive literature on text analysis. The most applicable paper (Gentzkow and Kelly, 2017) discusses how to employ the text as a data for economic research. In the paper, authors offer an overview of existing statistical methods, practical techniques, and applications.

To summarize, this paper contributes to the emerging literature on media economics by addressing the following arguments. First, there is a strong relationship between news media coverage and gender stereotypes. Second, gender stereotypes may be partially created by the misrepresentation of women in the media. Third, gender stereotypes may influence educational and occupational choices of women. As literature review shows, all three arguments have never been covered in economic papers. By employing text mining techniques as well as visual data analysis, this paper may facilitate a further usage of machine learning algorithms in Applied Economics and increase the overall quality of empirical research.

3 Data

The data on gender stereotypes and images in the media comes from the New York Times Application Programming Interface (API). As a single repository for media data, NY Times API provides researchers with articles on a month-by-month basis starting from 1851. The New York Times is a leading American newspaper that covers a variety of topics and has many subscribers. Therefore, the New York Times has a significant impact on public opinion which makes it a suitable media platform for the purposes of my research.

3.1 Textual Data

The process of building the dataset was performed in two steps: parsing and web scrapping. The raw data has been parsed and cleaned using Python[1]. In the first step, I collect the raw data from the New York Times Developer Network, which provides a convenient NYT API for free non-commercial uses. In particular, I access Archive API returning all NYT article metadata for a given month. Since #MeToo movement became extremely popular in news media in early October 2017, I extract all articles from 2015, two years before the event, to April 2020. The raw dataset consists of 434,164 observations, where each observation represents an article. Articles are attributed to different sections (blogs, news, multimedia) and may contain only textual, only visual, or both types of data. I restrict the sample to 274,710 observations that include text and images simultaneously. In the second step, I build the automated web scraper that requests a full text for each article. The scraper is a perfect solution to speed up the data collection process for text analysis. The final dataset includes the following variables: URLs of articles and article images, full article texts, headline, snippet, section, news desk, date of publication, author, article type. The description and summary statistics of all variables used in estimations are presented in Appendix (Table 6).

3.2 Visual Data

Article metadata collected through NYT API do not contain information about images except for URLs. I use Microsoft Azure which provides the artificial intelligence service that identifies faces in images and, for each face, detects attributes such as gender. However, this method has several limitations: image resolution, size and angle of faces. Consequently, there are some faces that cannot be detected by the facial recognition algorithm resulting in 89,818 processed images. Figure 1 presents the example of the algorithm output.

I build up the final dataset by quantifying image attributes and calculating a number of women, men, and the share of women portrayed in each image. On each article i, I define $Female_i = 1$ if the image depicts only women and $Female_i = 0$ if it depicts only men. Under this type of classification, 21,788 images labelled as $Female_i = 1$, 45,309 as $Female_i = 0$ and no label is attributed to 22,721 images portraying multiple faces of women and men. To classify the latter category, I use the lasso penalazed regression method suitable for high dimension data

[1]To access the code, feel free to contact me via email: *gmaria1904@gmail.com*

and regress labelled images on text from a corresponding article. The model is presented in the section 4.3. As a result, 18,692 images are classified as $Female_i = 0$ and 4,029 as $Female_i = 1$.

Figure 1: The result of applying the face recognition algorithm on the NYT article image depicting Simone de Beauvoir.

4 Empirical strategy

Analysis of Visual Data

The descriptive analysis suggests that women, on average, appear less on images of articles related to Business, Science, Politics, and Sport. On the contrary, they are depicted more in news desks related to Fashion and Style, Parenting and Home. It is noticeable that women slightly more appear in the category "Neediest" which "supports a global community of people less fortunate". This category also needs a detailed examination of racial stereotypes in further research. Good news: women, on average, appear less in Obituaries. Table 7 presents a full summary statistics for the share of women images grouped by all news desks.

It is clear that women are underrepresented in categories mostly related to professional articles and overrepresented in desks about appearance, entertainment, and home. However not only a number of women or even their presence is important. The context is also of paramount importance. In the next sections I analyze text of articles to address the question of *how* women are discussed in news media. Are they less likely to be portrayed as professionals?

4.1 Text analysis

Text from news and social media provides rich data about economic, social, and political activities. With the development of new statistic methods and an overall enhancement of its capacities, high-dimensional statistic methods are more and more applied to economics. Researchers in media economics employ modern statistical tools to collect and preprocess high-dimensional digital texts for causal inference and descriptive analysis.

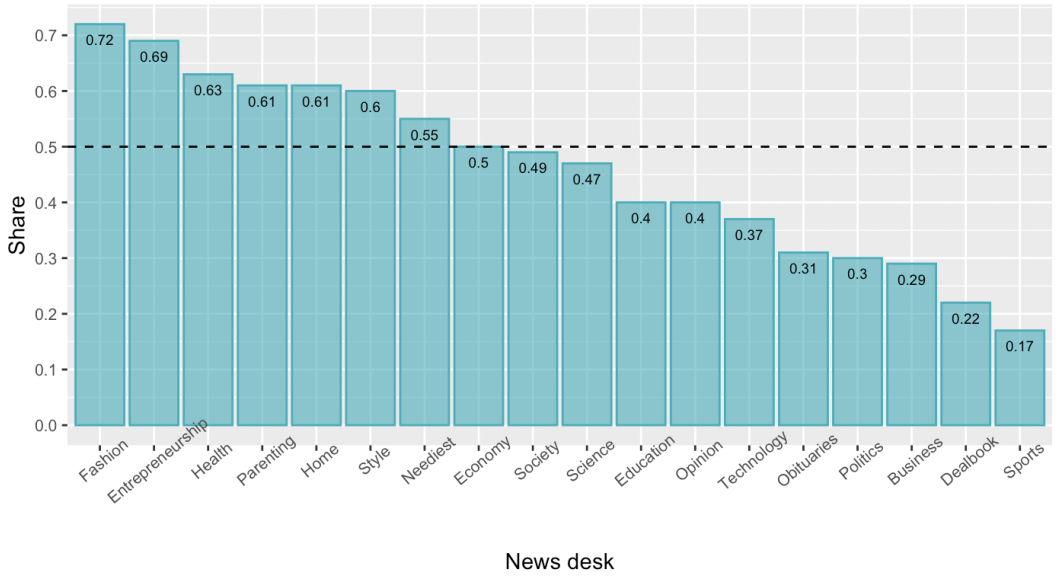

Figure 2: Mean shares of women depicted in article images grouped by news desks.

In this paper, the text analysis follows a three-steps procedure. First, the raw text is represented as a matrix W, where each row w_i is a vector of word counts. Second, W is mapped to predicted outcomes of interest Y. To perform a mapping, I use two methods suggested by Gentzkow and Kelly (2017): the dictionary-based method and supervised text regression. In the former, W represents a bag-of-words of a topic so that $Y = f(W)$ is specified. In the latter, W is divided into submatrices W^{train} and W^{test} in such a way that elements in W^{train} correspond to labeled observations Y^{train} of Y, W^{test} – to unobserved Y^{test}. Then a fitted model is used to predict Y. Finally, models are used to reveal gender stereotypes expressed via gendered language in news media.

4.2 Topic modelling

To address the question of whether women are less portrayed as workers and experts than men do, I begin with the topic modeling. Conventionally, the topic is defined as "a distribution over a fixed vocabulary"(Blei et al., 2003). That is, words associated within a given article 'i' constitute a latent topic.

Since this study focuses on professional and non-professional (personal) topics, I manually created two bag-of-words corresponding to each topic from the most frequent words in the dataset. The first "bag" includes 244 "professional" words describing business, finance, policy, education, and work. The second "bag" has 212 words mainly representing appearance, attractiveness, home, family, and relationships. Some words can not be unambiguously classified (e.g., "partner"), and some other words are irrelevant for analysis (e.g., "December"). Table 1 shows example of words related to each topic.

Table 1: Topic Modelling: Example of Words

Professional	Personal
Analyst, approach, attorney, business, career, conference, chief, economy, employee, executive, financial, government, industry, job, leadership, manager, minister, owner, politician, professor, research, statement, science, success, worker	Appearance, beautiful, child, couple, dream, emotion, face, family, fashion, feeling, friend, happy, home, husband, love, marriage, private, race, religion, sex, sexual, touch, trust, victim, violence, young

Words from both sets can be present within an article. To address that, I construct a proxy variable measuring how professional text of an article 'i' is (in relative terms):

$$Topic\_slant_i = \frac{\sum_{j=0}^{244} Prof\_words_{ji} - \sum_{k=0}^{212} Pers\_words_{ki}}{\sum_{m=1}^{N} Total\ words_{mi}}$$

Table 7 presents a summary statistics of the $Topic\_slant_i$ variable grouped by news desks. Positive (negative) slant indicates that an article is relatively professional (personal). For instance, news desks "Business", "Politics", and "Science" have negative mean values, while "Fashion", "Parenting", and "Neediest" have positive mean values. This result is in line with the rationale behind a variable construction.

Are women less frequently depicted in images of articles with a positive topic slant than men? A natural way to start answering the question is to look at $Topic\_slant_i$ distribution of "female" and "male" articles. Figure 2 shows that the distribution of articles with women on images is shifted to the left and has a mean value less than 0. The opposite situation with articles which images portray men. This evidence provides a positive answer to the question.

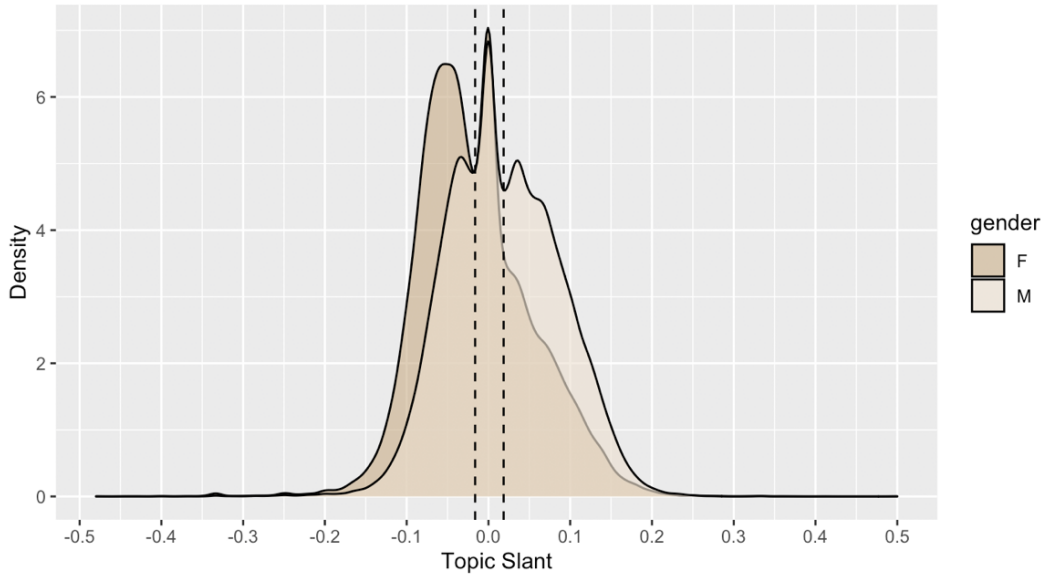

Figure 3: The distribution of the Topic Slant by gender

A more statistically convincing way to approach the question is to estimate a regression of the topic slant on the share of women:

$$Fem\_share_i = \alpha + \beta\ Topic\_slant_i + X_i\boldsymbol{\gamma} + u_i,$$

where Fem_share_i denotes the share of women depicted on image of article 'i', $Topic\_slant_i$ is the main variable of interest, X_i is a set of controls that includes news desks and article types, u_i is an error term.

Table 2 reports the main findings. Columns (1)-(3) shows results for specifications with the main explanatory variable $Topic\_slant_i$, but with different combinations of control variables. For instance, news desk controls account for unobserved sources of heterogeneity such as preferences of editors. Controlling for article types allows to address factors particular to articles, blogs, and multimedia content. The results reveal a negative relationship between the tendency of an article being professional and the share of women depicted in an article image. The more professional article is, less women appear on article images. This finding is persistent in all estimated specifications. Including a stronger set of controls slightly attenuates the main coefficient of interest, while keeping it statistically significant at 1% level.

Columns (4) and (5) reports estimates from alternative specifications with specific words counts for each article. On average, "professional" and "occupational" words decreases the share of women, while "personal" words increases. Table 6 shows that results are also robust to alternative specifications, where $Female_i$ is the dependent variable that takes a value of 1 if the image of the article 'i' depicts women, and 0 otherwise.

To summarize, regression estimates indicate a negative and statistically significant relationship between the share of women and the topic slant. The analysis provides evidence that women appear less on articles related to professional topics such as finance, politics, science, and technology. On contrary, an extremely large shares can be found in news desks such as fashion and style, home, and neediest.

Table 2: Main Results

	Dependent Variable: Share of Women				
	(1)	(2)	(3)	(4)	(5)
Topic Slant	-1.037***	-0.838***	-0.836***		
	(0.020)	(0.026)	(0.026)		
Professional				-0.001***	
				(0.000)	
Personal				0.001***	0.001***
				(0.000)	(0.000)
Occupations					-0.003***
					(0.000)
Constant	0.368***	0.408***	0.405***	0.403***	0.407***
	(0.001)	(0.009)	(0.009)	(0.010)	(0.010)
News Desk F.E.		Y	Y	Y	Y
Article Type F.E.			Y	Y	Y
R-squared adj.	0.031	0.093	0.093	0.091	0.092
Observations	89647	87730	87730	87900	87900

Notes: OLS regressions where the dependent variable is the share of women depicted on article image. Columns (1)-(3) report results for the specification with Topic Slant as the main explanatory variable. Topic Slant measures how professional article is. Negative values indicate the tendency of an article being personal. Positive values indicate professional article. Columns (4)-(5) report results for the alternative specification with a number of "professional", "personal" words and words related to occupations in a given article. Robust standard errors are reported in parentheses. Significance levels: $* \ p < 0.1, ** \ p < 0.05$, and $*** \ p < 0.01$

4.3 Text Regression

A natural way to examine which words are associated with images portraying women and men in the New York Times articles is to estimate text regression. Since text is high dimensional, I estimate a linear model with L1 penalization, as proposed by Gentzkow and Kelly (2017). In particular, I use the lasso logistic regression for propensity score estimation. The model identifies words with the most predictive power of nonverbal "gender" of article and sheds more light on gender stereotypes in news media. Since the dictionary of articles is rich and includes more than 50,000 unique words, I created a bag-of-words[2] W representing raw text as the unordered set of its the most common 10,000 words.

In the text regression model, Y is the outcome of interest and equals to 1 if article image depicts only women, 0 otherwise. The model also resolves the issue with unambiguously labeled images of both men and women by using train and test sets of original data. Let w_{ij} denotes the element of a matrix W representing a number of times word j occurs in article i. Since the dataset has 89,818 observations, the dimension of matrix W is 89,818 $\times$ 10,000 and the dimension of outcome Y is 89,818 $\times$ 1. Next, assume the posterior probability that image in

[2]The model is commonly used in natural language processing

article i depicts women given set of words used in article is:

$$Pr(\mathbf{y}_i|\mathbf{w}_i) = \frac{1}{1 + e^{-(\alpha + \mathbf{w}_i'\boldsymbol{\beta})}}$$

Assuming the independence of N observations, the unregularized objective function can be represented as the negative log likelihood in logistic regression:

$$l(\alpha,\boldsymbol{\beta}) = -log(\prod_{i=1}^{N} Pr(\mathbf{y}_i|\mathbf{w}_i)) = -\sum_{i=1}^{N}[(\alpha + \mathbf{w}_i'\boldsymbol{\beta})\boldsymbol{y}_i - log(1 + e^{\alpha + \mathbf{w}_i'\boldsymbol{\beta}})]$$

Then, the estimator of penalized lasso regression of the Y on the bag-of-words can be obtained from:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left[l(\alpha,\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{10{,}000} |\beta_j| \right]$$

where $\lambda_1 > 0$ is a lasso L1 penalty parameter. This model shrinks the coefficients of all insignificant words in terms of predictive power to zero, while keeping all significant words in the set for further analysis. There are 1,114 that are left with non-zero coefficients.

Next, I follow Wu (2018) and calculate a marginal effect of the word j from the post-estimation set:

$$ME_j = Pr(Y_i = 1|\mathbf{w}_{i,-j}, (\mathbf{w}_{i,j} + 1)) - Pr(Y_i = 1|\mathbf{w}_{i,-j}, (\mathbf{w}_{i,j}))$$

Table 3 presents the main findings and lists words with the largest marginal effects. "Female" words are mostly related to family, emotions, and appearance. Among them are "married", "felt", "wanted", "body", "wearing". However, a small fraction of words can be identified as a professional. For instance, "profession", "organization", "complex". "Male" words essentially relate to expertise, competence, and qualification. For example, "analysis", "ability", "leader", "politician", "success". The "male" set also includes more words of actions, decisions, and initiations: "signed", "committed", "introduced", "provide", "denied". Eventually, comparing two sets gives the following patterns: women receive, men earn and spend. Women are perfect and wanted, men are famous and successful. At least, women might be independent and inspired activists, and men might be wrong. However, to construct proper associations, in further research the word embedding method can be employed. This method accounts not only for the statistical significance of words but also for the context within the text. For instance, a male nurse could appear more frequently than a female nurse, as implicit bias creates an image of a nurse to be female by default.

Table 3: Words with Largest Marginal Effects on Gender of Articles

"Female"		"Male"	
Word	ME	Word	ME
mother	0.041	chairman	0.047
color	0.035	suggested	0.026
spring	0.034	signed	0.026
married	0.028	father	0.023
hair	0.024	pushed	0.023
fashion	0.022	bought	0.023
profession	0.020	spending	0.023
surprise	0.020	anysis	0.022
voice	0.020	beat	0.021
collection	0.019	committed	0.021
perfect	0.019	introduced	0.020
representative	0.017	fellow	0.020
complex	0.017	greater	0.020
emotion	0.016	approach	0.020
colleague	0.016	famous	0.020
felt	0.016	direction	0.019
violence	0.016	wrong	0.018
wearing	0.016	spokesman	0.018
learned	0.015	ability	0.017
senator	0.014	interested	0.017
pop	0.014	leader	0.017
role	0.013	politician	0.017
affair	0.013	minister	0.017
star	0.013	responsibility	0.016
wanted	0.013	produced	0.016
body	0.013	respect	0.016
receive	0.013	founder	0.015
fun	0.013	provide	0.014
experience	0.012	denied	0.013
hope	0.012	defense	0.013
organization	0.012	club	0.012
drama	0.012	league	0.012
focused	0.012	management	0.012
inspired	0.012	position	0.012
secretary	0.012	earned	0.012
independent	0.012	governor	0.012
argued	0.011	war	0.012
movement	0.011	great	0.012
activist	0.011	successful	0.011
touch	0.011	success	0.011

4.4 The effect of the empowerment movement

The previous analysis focused on the overall relationship between the topic slant and the nonverbal slant of articles averaged over the entire period of 5 years. Does the pattern of representation of women in news media change over time? Did the empowerment movement #MeToo change the way women are portrayed in media? In the analysis, I focused on the #MeToo movement as it became the loudest case in the professional field and was covered by all news media. Figure 4 plots the shares of women on article images averaged by months and grouped by news desks.

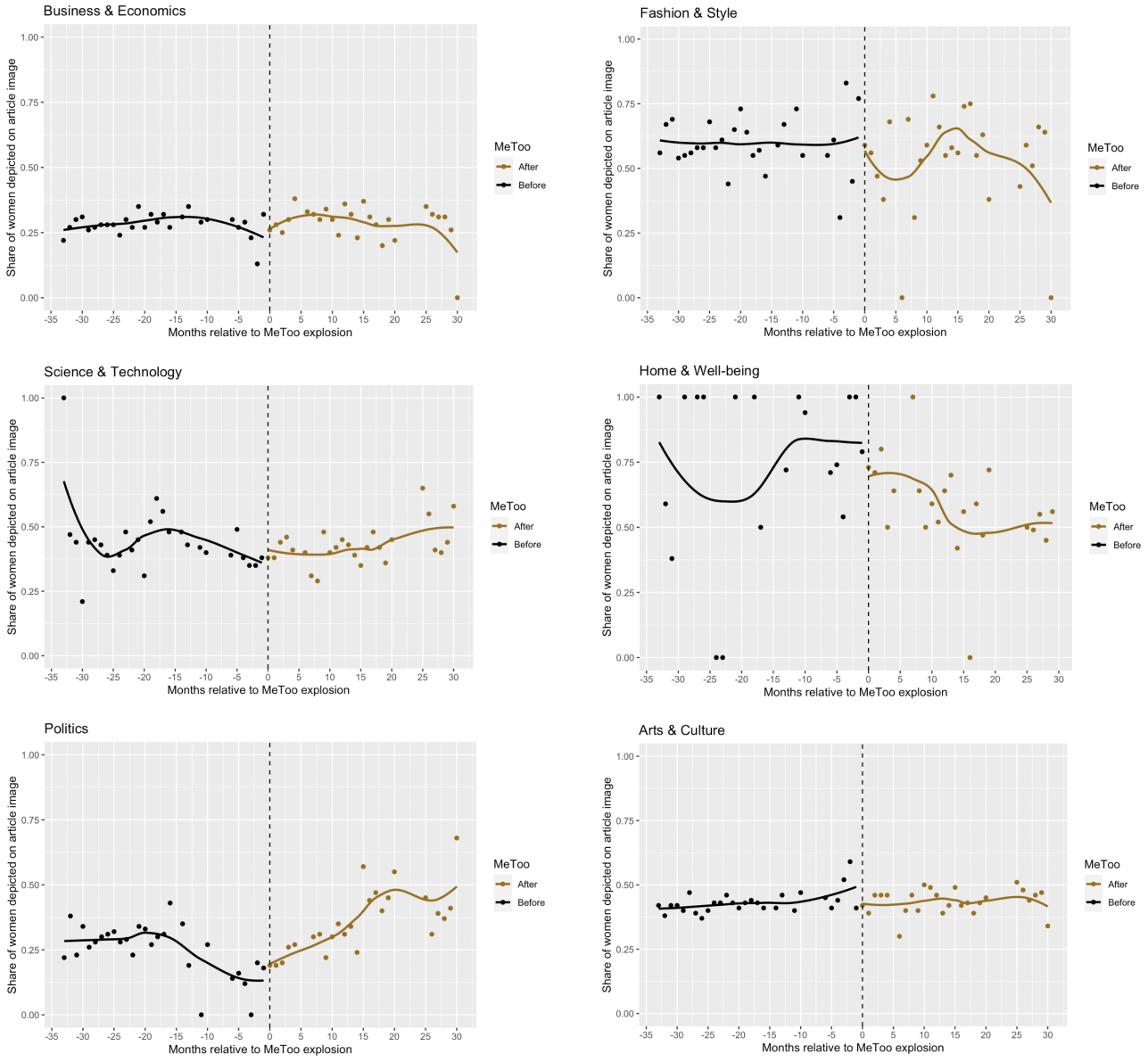

Figure 4: Shares of women depicted in article images before and after the MeToo explosion on October 2017

Although it is hard to conclude that it is a causal impact of the movement, results suggest that after the #MeToo explosion all presented news desks except for economics and business tend to converge to the equal share of both men and women representations.

Another way of addressing the question is to estimate the same specifications as in the Topic modelling section, but on subsamples before and after the movement. I also report the results for the specification with the dummy variable indicating whether the #MeToo movement

happened (=1) or not (=0). The cross product of the dummy and news desks controls for the effect heterogeneity: some categories, such as "Arts" and "Theater", might have a stronger effect because the movement originated within creative industries.

Columns (1)-(4) of table 4 indicates that the movement #MeToo deteriorated the improvements of women representation in media. Conversely, columns (5)-(6) shows a slight improvement. Regression results, together with 4, yield ambiguous results. Thus, this topic may be addressed in further research by including other events and empowerment movements in the analysis.

Table 4: The Effect of the Empowerment Movement #MeToo

	Before MeToo		After MeToo		Full sample	
	(1)	(2)	(3)	(4)	(5)	(6)
Topic Slant	-0.998***	-0.770***	-1.119***	-0.964***	-0.836***	-0.832***
	(0.024)	(0.031)	(0.035)	(0.047)	(0.026)	(0.026)
MeToo dummy					0.029***	0.053**
					(0.003)	(0.021)
Constant	0.355***	0.392***	0.393***	0.433***	0.396***	0.389***
	(0.002)	(0.011)	(0.002)	(0.018)	(0.009)	(0.011)
News Desk F.E.		Y		Y	Y	Y
Article Type F.E.		Y		Y	Y	Y
MeToo×News Desk F.E.						Y
R-squared adj.	0.029	0.103	0.034	0.075	0.094	0.095
Observations	59729	58413	29918	29317	87730	87730

Notes: OLS regressions where the dependent variable is the share of women depicted on article image. All specifications include Topic Slant as the main explanatory variable. Topic Slant > 0 indicates the tendency of an article being professional. Columns (1)-(2) report results on a subsample of articles published before the explosion of #MeToo on October 2017, columns (3)-(4) – after. Columns (5)-(6) report results of estimation on the full sample. Specification (5) includes a dummy on whether the explosion of #MeToo happened. Specification (5) controls for specific treatments of #MeToo explosion on different news desks. Robust standard errors are reported in parentheses. Significance levels: $* \ p < 0.1, ** \ p < 0.05$, and $*** \ p < 0.01$

5 Conclusion

This paper addresses the issue of women's misrepresentation in the news media and discusses potential consequences for the labor market outcomes. First of all, I document a negative and statistically significant relationship between coverage of professional topics in the New York Times articles and the share of women in the article. This result indicates that women are underrepresented in news desks related to science and technology, economics and politics, business, and finance. If we are to accept the idea of news media influencing educational and occupational choices of women through role models, then role models based on gender stereotypes may discourage women from starting a career of a financial analytic, computer scientist, and so on.

The analysis of high-dimensional digital text revealed common patterns of gendered language for both men and women. Articles with women images are focusing more on appearance, family, and violence issues, while with men – on career and professionalism. This creates a representation of women as caregivers, homemakers, and activists. The lack of strong verbs from the "male" set such as "signed", "suggested", "denied" may facilitate a further underrepresentation of women in decision-making occupations including politics.

Moreover, the study addresses the question of how gender stereotypes evolved over time by showing the dynamics of the share of female images in different news desks. Except for business and economics, all news desks converge to an approximately equal share of both men and women while choosing visual content for articles. A more statistically convincing approach reveals a slight deterioration in the progress to the more equal media world. Therefore, the impact of the empowerment movement #MeToo on women representation in the major American news media is ambiguous.

The methodological idea of the paper is to analyze a large dataset combining textual and visual data with the help of modern machine learning tools, advanced text analysis techniques, and econometric models. In this dimension, this paper is expected to contribute to the literature on media economics and facilitate economists to employ more advanced methods in their empirical studies. Although gender stereotypes are the main focus of this paper, the empirical model can be applied in further research to address other types of discrimination including racial. This topic has not been covered by economic scholars so far.

As the main results shed more light on the gender stereotypes in news media, I sincerely hope to advance the discussion of gender equality in media. If we are to design a world that is meant to work for everyone, we cannot afford to underrepresent women because at best, we will be left with only half of the truth. Failing to include women in professional articles strengthens the unconscious bias which, in turn, is responsible for much of discrimination (Bohnet, 2018). Accordingly, there are to be many policy implications. By introducing a more equal representation of women, the government and firms may expect an improvement in the aspirations of women. Seeing is believing, and for younger generations, such changes may influence both occupational and educational choices towards more mathematically-intensive and professional.

References

Beaman, L., Duflo, E., Pande, R., and Topalova, P. (2012). Female leadership raises aspirations and educational attainment for girls: A policy experiment in india. *science*, 335(6068):582–586.

Bergmann, B. (1974). Occupational segregation, wages and profits when employers discriminate by race or sex. *Eastern Economic Journal*, 1(2):103–110.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

BLS (2019). Women in the labor force: a databook : Bls reports.

Bohnet, I. (2018). *What works: gender equality by design*. The Belknap Press of Harvard University Press.

Boxell, L. (2018). Slanted images: Measuring nonverbal media bias. MPRA Paper 89047, University Library of Munich, Germany.

DellaVigna, S. and Gentzkow, M. (2010). Persuasion: Empirical evidence. *Annual Review of Economics*, 2(1):643–669.

Gentzkow, M. (2006). Television and Voter Turnout. *The Quarterly Journal of Economics*, 121(3):931–972.

Gentzkow, M. and Kelly, B. T. (2017). Text as data. *SSRN Electronic Journal*.

Gentzkow, M. and Shapiro, J. (2006). Media bias and reputation. *Journal of Political Economy*, 114:280–316.

Gentzkow, M. and Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, 78(1):35–71.

Gentzkow, M. A. and Shapiro, J. M. (2004). Media, education and anti-americanism in the muslim world. *Journal of Economic Perspectives*, 18(3):117–133.

Gerber, A. S., Karlan, D., and Bergan, D. (2009). Does the media matter? a field experiment measuring the effect of newspapers on voting behavior and political opinions. *American Economic Journal: Applied Economics*, 1(2):35–52.

Gilens, M. (1996). Race and poverty in america: Public misperceptions and the american news media. *The Public Opinion Quarterly*, 60(4):515–541.

Goldin, C. (2014). A grand gender convergence: Its last chapter. *American Economic Review*, 104(4):1091–1119.

Kahn, S. and Ginther, D. (2017). Women and stem. Working Paper 23525, National Bureau of Economic Research.

Luebke, B. F. (1989). Out of focus: Images of women and men in newspaper photographs. *Sex Roles*, 20(3-4):121–133.

Miller, C. C., Quealy, K., and Sanger-katz, M. (2018). The top jobs where women are outnumbered by men named john.

Sharma, B. (2013). Image of women in media. *SSRN Electronic Journal*.

Simon, S. and Hoyt, C. L. (2012). Exploring the effect of media images on women's leadership self-perceptions and aspirations. *Group Processes Intergroup Relations*, 16(2):232–245.

Tuchman, G. (2000). The symbolic annihilation of women by the mass media. *Culture and Politics*, page 150–174.

Wu, A. H. (2018). Gendered language on the economics job market rumors forum. *AEA Papers and Proceedings*, 108:175–79.

6 Appendix

Table 5: Summary statistics

Variable	Description	Mean	Std. Dev.	Min.	Max.	N
female	No. of women detected on image	0.92	1.87	0	53	89817
male	No. of men detected on image	1.58	2.73	0	87	89817
fem_share	Share of women on image	0.36	0.42	0	1	89817
professional	No. of "professional" words	47.74	51.52	0	1963	89817
personal	No. of "personal" words	43.09	44.73	0	1443	89817
occupations	No. of words related to occupations	10.62	14.29	0	523	89817
token_count	No. of words in a text	502.53	452.02	0	15871	89817
topic_slant	Professional slant of the text	0.01	0.07	-1	0.5	89647
metoo	Dummy variable indicating whether #metoo happened (October 2017)	0.33	0.47	0	1	89817
gen_label	Dummy variable indicating whether the propensity score model predicted article to be "female" = 1 or "male"= 0	0.29	0.45	0	1	89817
news_desk	Categorical variable indicating the news desk: Business, Fashion, Sport, etc.					
doc_type	Categorical variable indicating the type of article: article, blogpost, etc.					
date	Date of publication					

Notes: Descriptive statistics for main variables used in estimations. Unit of observation is an article.

Table 6: Main Results: Sensitivity Analysis

	Dependent Variable: Indicator for Female Image			
	(1)	(2)	(3)	(4)
	OLS	Logit	OLS	Logit
Topic Slant	-1.429***	-7.542***	-1.072***	-5.997***
	(0.022)	(0.117)	(0.028)	(0.149)
Constant	0.299***	-0.903***	0.362***	-0.659***
	(0.002)	(0.008)	(0.011)	(0.044)
News Desk F.E.			Y	Y
Article Type F.E.			Y	Y
Observations	89647	89647	87730	87717

Notes: OLS and logit regressions where the dependent variable is the indicator variable of whether women is present on article image and predicted by the text regression model to be female. Columns (1)-(2) report results for the benchmark specification Columns (3)-(4) report results for the specification with a set of fixed effects for news desks and type of an article. Robust standard errors are reported in parentheses. Significance levels: $* \ p < 0.1, ** \ p < 0.05$, and $*** \ p < 0.01$

Table 7: Summary statistics for *Fem_share* and *Topic_slant* by *news_desk*

	Fem_share		*Topic_slant*		
News desk	mean	sd	mean	sd	N
Arts	.458	.443	-.060	.040	2250
Blogs	.434	.437	-.028	.099	359
Book	.459	.473	- .033	.046	1893
Business	.287	.409	.063	.060	7856
Climate	.227	.367	.081	.047	104
Culture	.420	.439	- .043	.043	10199
Dealbook	.218	.336	.106	.043	26
Economy	.500	.577	.063	.064	4
Education	.404	.416	.031	.077	1479
Entertainment	.434	.436	- .033	.045	930
Entrepreneurship	.693	.413	- .043	.037	5
Fashion	.721	.426	- .013	.053	137
Fashion men	.048	.182	- .057	.107	110
Fashion women	.920	.244	-.036	.103	391
Food	.393	.446	- .028	.037	536
Foreign	.259	.371	.038	.050	8264
Gender	.946	.145	- .013	.040	13
Health	.626	.435	- .057	.054	244
Home	.607	.487	- .035	.047	14
Investigative	.363	.414	.074	.061	72
Media	.267	.435	.037	.076	5
Multimedia	.527	.448	- .029	.062	106
National	.322	.395	.054	.058	6055
Neediest	.547	.460	- .005	.050	27
Obituaries	.310	.427	- .002	.054	1053
Opinion	.404	.438	- .015	.085	193
Other	.322	.408	.017	.063	10425
Parenting	.611	.419	- .066	.030	21
Politics	.300	.369	.061	.050	3818
Real Estate	.520	.431	- .043	.063	507
Science	.470	.457	.010	.060	650
Society	.490	.208	.065	.059	3332
Sports	.169	.323	- .014	.044	7745
Style	.600	.439	- .046	.058	5471
T Magazine	.504	.464	- .039	.069	1595
Technology	.371	.413	.030	.086	25
Theater	.612	.440	- .040	.067	64
Transport	.337	.406	.029	.067	4426
Travel	.513	.451	- .046	.042	475
U.S.	.263	.371	.070	.066	1932
Universal	.367	.397	.001	.034	126
Upshot	.343	.389	.052	.060	531
Weekend	.465	.439	- .059	.036	3138
World	.240	.359	.023	.054	1101
Your Money	.625	.413	- .022	.066	23
Total	.358	.417	.008	.070	87900