Detecting Gender Stereotypes: Lexicon vs. Supervised Learning Methods

Jenna Cryan*
University of Chicago
Chicago, IL
jennacryan@cs.uchicago.edu

Shiliang Tang*
University of California, Santa Barbara
Santa Barbara, CA
shiliang_tang@cs.ucsb.edu

Xinyi Zhang
University of California, Santa Barbara
Santa Barbara, CA
xyzhang@cs.ucsb.edu

Miriam Metzger
University of California, Santa Barbara
Santa Barbara, CA
metzger@comm.ucsb.edu

Haitao Zheng
University of Chicago
Chicago, IL
htzheng@cs.uchicago.edu

Ben Y. Zhao
University of Chicago
Chicago, IL
ravenben@cs.uchicago.edu

ABSTRACT

Biases in language influence how we interact with each other and society at large. Language affirming gender stereotypes is often observed in various contexts today, from recommendation letters and Wikipedia entries to fiction novels and movie dialogue. Yet to date, there is little agreement on the methodology to quantify gender stereotypes in natural language (specifically the English language). Common methodology (including those adopted by companies tasked with detecting gender bias) rely on a lexicon approach largely based on the original BSRI study from 1974.

In this paper, we reexamine the role of gender stereotype detection in the context of modern tools, by comparatively analyzing efficacy of lexicon-based approaches and end-to-end, ML-based approaches prevalent in state-of-the-art natural language processing systems. Our efforts using a large dataset show that even compared to an updated lexicon-based approach, end-to-end classification approaches are significantly more robust and accurate, even when trained by moderately sized corpora.

Author Keywords

Gender Bias; Gender Stereotypes; Machine Learning; Natural Language Processing; Lexicon

CCS Concepts

•**Social and professional topics** → **Gender;** •**Computing methodologies** → *Machine learning; Lexical semantics;*

*Both authors contributed equally to this research.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6708-0/20/04 ...$15.00.
http://dx.doi.org/10.1145/3313831.3376488

INTRODUCTION

The values of our society are both reflected in and reinforced by our use of language. In that context, sexism and gender discrimination is often perpetrated and reproduced through lexical choices in everyday communication. Recent studies have identified descriptions that reflect gender stereotypes in different types of articles, such as biographical pages of notable people [35], recommendation letters [19], fictional stories [14] and movie dialogue [26].

These issues are further exacerbated today by the ubiquitous usage of machine learning tools in language processing. We know that machine learning algorithms often translate and incorporate gender biases from training data [31], and such biases have been proven in popular techniques including word embeddings [5, 6], coreference resolution [37] and sentence encoders [20].

While the case is obvious for accurately identifying gender stereotypes in language, today's tools for this process are woefully lacking. Gender stereotypes are traditionally captured by a gender word inventory: a pre-compiled word lexicon which contains items describing social traits and behaviors that supposedly differentiate male or female genders. The original lexicon was a bag of words hand-picked from a survey of psychology students in 1974 [3]. These word banks have been further extended in later studies to detect gender bias in job postings [15, 33]. Though widely utilized, items in traditional gender word inventories are less endorsed by women in recent years [34, 12], and their efficacy in detecting gender stereotypes in language remains unclear.

Meanwhile, today's natural language processing tasks are dominated by an end-to-end approach using deep neural network models. Instead of using a pre-compiled word lexicon, the end-to-end approach trains a neural network model that produces the desired output using labeled raw text as input. This approach has shown great success in most NLP tasks such as sentiment analysis [17] and hate speech detection [1], which

were traditionally solved by a lexicon approach [32, 9]. Unfortunately two risks remain with this approach: a) these models often require tens of thousands of samples to train an accurate model, and b) bias can seep into training data and affect detection results.

The goal of our work is to empirically analyze different approaches to the problem of detecting gender stereotypes in natural language, in order to understand the best methodology for ongoing and future studies. More specifically, we are interested in three general questions. *First*, can we update the traditional lexicon-based approach to reflect gender stereotypes in modern society, with modern machine learning tools at our disposal? *Second*, can we build a gender stereotype detection model using the end-to-end approach? Given the dependence of deep learning models on large training sets, how accurate can this approach be given moderately sized datasets? *Third*, how do these two approaches (lexicon-based and end-to-end deep learning) compare in practice? What accounts for the differences in their accuracy results? Our ultimate goal is to develop methodology guidelines for identifying gender stereotypes moving forward.

Our study consists of two key components: building a gender stereotype lexicon (an update to the lexicon approach), and a careful empirical comparison between the lexicon approach with the end-to-end deep learning model. First, to build the gender stereotype lexicon, we begin by extracting verbs and adjectives that are used to describe humans from English Wikipedia. We select a set of frequently used words, and ask users to evaluate the masculinity and femininity of each word. We then apply a supervised learning approach, using user-labeled words to generate scores for all remaining words, resulting in a gender stereotype lexicon that contains stereotype scores of over 10,000 words. Second, we build an end-to-end deep learning model by training an NLP BERT model with a dataset of online articles marked by crowdworkers as consistent with or contradictory to common gender stereotypes. We compare the end-to-end and lexicon approaches, and find the end-to-end models significantly outperform. Finally, we manually analyze results to understand the underlying causes of misclassifications for the lexicon-based approach.

Our work makes three key contributions:

- We develop a robust gender stereotype lexicon reflecting modern language and interpretations, using a combination of data-mining, crowd-sourcing, and supervised learning using linear classifiers. We also empirically show that existing unsupervised methods fall short in comparison on accuracy measures.
- We collect the first human labeled text corpus (4,333 articles) for gender stereotypes, and use it to train an end-to-end, deep learning classification model based on the BERT language representation tool.
- We evaluate both approaches using a secondary user study, and find that our end-to-end approach significantly outperforms our lexicon approach in its ability to recognize gender stereotypes. We manually study errors made by the lexicon classifier, and identify key underlying reasons for those errors.

We hope that our results will inform best practices moving forward for detecting gender stereotypes in text. We will share our datasets and models with the research community, and our lexicon dataset should help improve the accuracy and robustness of currently deployed lexicon-based gender detection systems.

RELATED WORK

Gender Stereotypes in Language

Gender stereotypes are common beliefs about what men and women's physical and personality traits are and should be like. According to traditional gender stereotypes, women should display *communal* traits (*e.g.,* nice, caring, warm) and men should display *agentic* traits (*e.g.,* assertive, competent, effective) [13, 8].

Gender stereotypes emerge in language choices used in written and verbal communication [21]. It has been found that a category label used to refer to a group automatically activates the characteristics stereotypically associated with the group, even in supposedly unprejudiced people who do not explicitly endorse the stereotype [18, 21]. This also applies when the category label is one's gender. For example, after primed by words consistent with gender stereotypes (*e.g.,* "nurse"), people are faster to associate gender pronouns (*e.g.,* "she") with the corresponding gender (*e.g.,* "female") [2].

As a result, gender stereotypes are common in contemporary languages, both in written and spoken communication. For example, in fiction writing, traditional gender stereotypes such as dominant men and submissive women are common throughout nearly every genre, regardless of the gender of the author [14]. On Wikipedia, articles about notable women emphasize more on romantic relationships or family-related issues compared to articles about notable men [35]. When writing recommendation letters for faculty positions, women are often described as more communal and less agentic than men [19]. Additionally, in movie dialogue, male characters use more words related to achievement than female characters [26].

Gender Word Inventory

Stereotypes can be captured by *gender word inventories* – precompiled lists of items describing social traits and behaviors that differentiate males and females [3, 27]. Gender word inventories are historically extracted from self-reported characteristics through questionnaires given to college students to measure their self-concept and valuation of feminine and masculine characteristics. The Personal Attributes Questionnaire (PAQ [29]) and Bem Sex Role Inventory (BSRI [3]) are two of the most representative questionnaires in early studies. The items extracted for the BSRI and PAQ typically associate females with more communal attributes (*i.e.,* gentle, warm) and men with more agentic attributes (*i.e.,* aggressive, competitive), which are highly consistent with traditional perceptions regarding gender stereotypes. Other studies generalized these words into *expressive* and *instrumental* traits [28]. Tying these together, aggregated lists of masculine and feminine characteristics have been compiled from previous studies, particularly through gendered wording in job advertisements [15].

These gendered word inventories are traditionally used as a way to measure gender role self-concepts, *i.e.,* whether people see themselves as masculine or feminine. Among them, BSRI is considered as a golden standard in gender role evaluation, and has been used in thousands of studies in the more than 40 years since it was developed [10]. However, perceptions captured by BSRI are less endorsed by women in recent years [34, 12]. These works reviewed a large collection of studies that apply BSRI, and tracked how user responses change over a long period of time. Women's femininity scores have decreased significantly over the years, indicating that societal gender norms may require an update of masculine and feminine stereotyped characteristics.

Given previous results showing that existing gender word inventories may not properly reflect these concepts in the modern world, we seek to develop a lexicon that captures people's perceptions of gender stereotypes in contemporary society.

Gendered Stereotype Studies in NLP

There are few tools or algorithms to determine if any piece of text perpetuates modern gender stereotypes. One class of tools shares some similarity to ours are tools used to detect gender biased language in job advertisements [33]. These tools leverage precompiled lists of gender biased words aggregated from previous psychology studies [15] that may affect decisions of job applicants, and calculate gender bias of a job advertisement based on the number of occurrence of these words.

Although detecting gender stereotypes in natural language is still an under-explored area, the NLP community has increasingly focused on issues of fairness and bias in NLP models. Many projects focus on identifying and removing biases in algorithms, e.g. in word embeddings [5]. Prior studies also observed performance discrepancies across genders in systems including coreference resolution [37], image caption generation [16], and sentiment analysis [24]. Such biases can be mitigated by creating an augmented dataset that counters gender bias in the original training dataset [24], or adding constraints during model training to enforce gender neutral prediction [38].

METHODOLOGY

To detect gender stereotypes in articles, we implement and compare two approaches: a traditional lexicon-based method that operates on individual words, and an end-to-end method that operates directly on text paragraphs. Both approaches are data-driven and apply machine learning models to scale up our evaluation to arbitrary words/articles.

An overview of the approaches can be seen in Figure 1. Specifically, our lexicon-based method starts from breaking down the article to word level tokens, then uses crowdsourced workers to score the perceived masculinity and femininity of a set of most frequently used words. We train supervised models using this data, and apply the result to build the full modern gender stereotype lexicon. The lexicon scores individual words, which when combined, given the overall gender score of an input article. Our end-to-end approach takes crowdsourced text samples illustrating gender stereotypes, and uses them to fine-tune the BERT deep learning language model. The result is a deep learning model capable of gender scoring arbitrary paragraphs and articles.

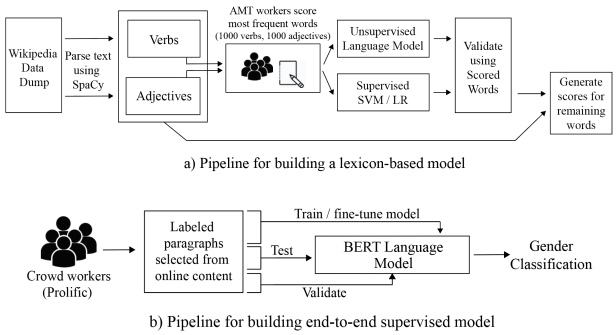


Figure 1: Building gender-stereotype detection models.

Crowdsourced Gender Scores

Both the lexicon and end-to-end approaches require datasets that exemplify gender stereotypes at the word or paragraph levels. Currently, no such databases exist that reflect modern perspectives on gender stereotypes in language. Thus, we build our own datasets using crowdsourcing.

Our goal is to create datasets that represent current language use, with minimal bias, and can easily scale up when additional resources become available. We approach this goal in three steps. First, we leverage a large corpora of existing text samples to reflect typical use of language. Second, we use human crowdsourcing to label the data. Finally, these datasets can be iteratively updated and expanded through the methods described, thereby providing practical scalability.

Mitigating Bias

To avoid potential bias, rather than ask respondents to brainstorm original content, we instead focus on gathering data that reflects perceptions of existing written content. To reduce potential biases due to variations across cultures, we limit our respondent pool to US residents. By gathering data points from assorted modern perspectives, we believe this provides representative perspectives of the current US population. For our end-to-end approach, we chose a model not previously trained for any particular language task, which allows us to examine and search for common patterns of language use that may be associated with gender stereotypes.

Next, we describe each approach in detail.

DETECTING GENDER STEREOTYPES: A LEXICON-BASED APPROACH

In this section, we introduce our lexicon approach for detecting gender stereotypes in written articles. A lexicon-based approach first analyzes how people associate particular words with common gender stereotypes, and then aggregates these scores to derive a gender score for the entire article.

While gender lexicons have been the preferred approach for detecting bias or stereotypes [3, 27, 29], they have a number of limitations. First, because they are manually constructed and

labeled, they are limited in size and coverage. Second, they cannot produce gender scores of arbitrary words, and can lose relevance over time as language describing gender stereotypes evolve. One potential solution is to apply (unsupervised) language models to automatically estimate gender score of words without human labels. We explore the empirical efficacy of this approach on a dataset of ground-truth labeled data (see below), and report results later in this section.

We also propose a supervised learning solution for computing gender scores of arbitrary words. This solution includes four steps. First, we use existing text corpora (*i.e.* Wikipedia) to identify frequently used, descriptive words as our gender lexicon dataset. Second, we generate our ground truth data by apply human crowdsourcing to label a subset of this dataset. Third, we use our ground truth data to train a supervised learning model that derives *gender scores* (a score reflecting a word's perceived masculinity or femininity) for arbitrary words. We apply this model to label our larger gender lexicon dataset. Finally, we use this gender lexicon dataset to compute the gender score of an article and evaluate how consistent or contradictory the article is with gender stereotypes.

Our Gender Lexicon Dataset

We begin by identifying a large set of words that are are potentially related to gender stereotypes. Here, we restrict our selection to *verbs* and *adjectives*, as stereotypes often manifest in people's behavior and how they are described [14].

We extract candidate words from Wikipedia Datadump (https://dumps.wikimedia.org/). We choose Wikipedia because it is large and diverse, and thus likely to include most of the commonly used English words. We downloaded a snapshot of Wikipedia text on March 4th, 2019, removing all images and links. In total, our dataset includes 5,817,125 documents, 42,653,358 paragraphs and 2,076,621,930 words.

To extract verbs and adjectives that characterize humans, we analyze the word dependencies in each sentence using a parse tree implemented using SpaCy (https://spacy.io/). For verbs, we extract "subject-verb" relationships in each sentence, where the subject is a human-related word like "he," "she," "man," "woman" *etc*. For example, in the sentence "He ran away from her," the subject is "he" and the verb is "ran." So the word "ran" is extracted. We lemmatize all words: we merge variants of a single noun or verb down to its stem, *e.g.* past tense "ran" becomes the lemma "run." For adjectives, we consider two different types: *predicate* and *attribute*. Adjectives are predicates when they are connected to their subject words by a verb, usually "be," *e.g.,* "he is *handsome*." Attribute adjectives are used as modifier before the subject, as in "a *handsome* man." In both cases, we keep the adjective if it is used on a human-related word.

We then filter out words that cannot be found using the Oxford Dictionary API (https://developer.oxforddictionaries.com/documentation). The removed words are mostly non-English words, non-existent words, or those with the wrong part-of-speech (*e.g.,* a word extracted as an adjective but only used as a noun).

After this process completes, we obtain our final lexicon dataset of 6,178 verbs and 4,424 adjectives (10602 total).

Ground Truth Gender Lexicon via Crowdsourcing

From our lexicon dataset, we choose the most frequently used 1,000 verbs and 1,000 adjectives, and use a user survey to label them. The resulting labeled words will serve as our ground truth dataset. We manually checked all words to ensure that they are suitable candidates, removing words that depend strongly on context (*e.g.,* "next", "final"). Also, we remove references to race, country, or religion because those biases remain outside the scope of this work.

Survey Design

Like previous studies [3, 29], our survey asks the participants to rate the extent to which they associate each word with a typical man or woman. Specifically, the participants are shown a list of words, and asked to evaluate the statement "I feel that _____ is commonly associated with the characterization of a typical man in US society" or "of a typical woman in US society." The evaluation uses a 7-point Likert Scale, from "strongly disagree (1)" to "strongly agree (7)." The participant can also select "I don't understand the word." Each participant rates 50 adjectives and 50 verbs "of a typical man" (*male rating*) and another 50 adjectives and 50 verbs "of a typical woman" (*female rating*). We also collect their demographic information at the end. The survey takes on average 15 minutes to complete, and each participant received $3 as compensation.

To ensure that the participants pay attention during the survey, we randomly insert in each survey 4 words that do not exist in English. The participants are expected to select "I don't understand the word" for these words. We include another quality control question when collecting demographics information, which is a multiple choice question asking them to choose both A and D. We removed all the responses that failed these quality check questions.

We recruited our participants on Amazon Mechanical Turk. To reduce potential differences across cultures, we limit our participant pool to US residents over the age of 18. Each participant can answer our survey up to 10 times, but will rate different words each time. We collected a total of 1097 qualified response sessions (HITs), among which 619 are from male participants, 476 are from female participants, and 2 chose not to disclose their gender. Over 99% of the words have more than 50 male ratings and 50 female ratings.

Gender Score Calculation

The ground truth score of a word is measured by the difference between the ratings associating the word with men and the ratings associating the word with women. For example, if a word is perceived as strongly associated with typical men but not associated with typical women, then we evaluate the word as carrying a strong masculine stereotype.

Specifically, we use the T-statistic in a two sampled T-test to measure the difference between masculine ratings and the feminine ratings of a word. The T-statistic reflects the extent to which the average value differs across samples. Like other statistical tests, the T-statistic also maps to a p-value which

indicates how likely the average value of the two samples are identical. A small p-value indicates a statistically significant difference between samples [30].

We plot the distribution of T-statistic of each word in Figure 2. Most words score around 0, indicating gender neutrality. Figure 3 shows examples of words with different T-statistics. Except a few words related to appearance (*e.g.,* hairy, beautiful), our highly stereotypical words are consistent with recent work demonstrating that stereotypically men are perceived as strong, active and violent, and women are perceived as weak, emotional and kind [14].

Reliability of User Responses

We perform two tests to examine the reliability of survey responses. The first is *split-half reliability* [23] that measures how likely the data collection is *reproducible*. To do so, we randomly split all our participants into two equal-sized halves, and calculate two sets of T-statistics for each word independently using the two halves. We calculate the Pearson Correlation between the T-statistics of the two sets, resulting in 0.85 for adjectives and 0.82 for verbs. This result means that repeating the data collection process is unlikely to significantly change results.

In our second test, we determine to what extent responses from male and female participants agree with each other. We split our responses by the gender of the participant, calculate two sets of T-statistics for each word independently using the two splits. From the two sets, we observe a correlation of 0.82 for adjectives and 0.80 for verbs, similar to the correlations in the split-half reliability. This means no significant difference exists between responses of male and female participants. Thus, in the following analysis, we aggregate all responses (both genders) together for our calculations.

Labeling Gender Lexicon via Word Embedding

Using our ground truth dataset, we first examine whether existing (unsupervised) language models (*i.e.* word embedding) can be used to automatically label gender lexicon without human input. For this, we consider four metrics used by prior work to calculate gender information of words.

- *Odds ratio.* It calculates how likely a verb or an adjective is used to characterize a man rather than a woman. If a word is more likely to be used on a man, it may indicate masculinity of the word. Specifically, given a word, odds ratio is calculated as:

$$\frac{\#\,this\ word\ on\ man\ /\ \#\,this\ word\ on\ woman}{\#\,other\ words\ on\ man\ /\ \#\,other\ words\ on\ woman} \quad (1)$$

  Here, "#" denotes "number of times." Odds ratio reflects gender stereotypes in large language corpus [14]. We calculate the odds ratio using the Wikipedia data.
- *Distance to gender specific words.* It has been shown that word embeddings contain gender biases due to stereotypes in the language [7]. Such biases can be captured by calculating word distance to gender specific words. Specifically, given a word, we calculate its average distance to a set of male specific words (*e.g.,* "he", "man") and its average distance to a set of female specific words (*e.g.,* "she", "woman"), then calculate the difference between the two distances. We test 3 commonly used word embeddings: word2vec [22], GloVe [25], and FastText [4]. Here we do not train our own word embeddings, but apply widely used pre-trained models for each of the three embedding methods: *word2vec* from Google News [1], *GloVe* from the 6 billion token Wikipedia dataset [2], and *FastText* from English Wikipedia [3].
- *Projection on gender direction.* One way to reduce gender bias in word embeddings is to extract a gender direction and remove the vector projection on the direction [5]. Here, the gender direction is the direction parallel to $\overrightarrow{she} - \overrightarrow{he}$ or $\overrightarrow{woman} - \overrightarrow{man}$. For each word, we take its projection on the gender direction as its gender stereotype score.
- *Values on gender dimensions.* Another way to reduce gender bias in embeddings is to encode gender information in a reserved dimension during training [38]. Here, we use the magnitude of the gender dimension as a way to quantify the gender stereotype associated with a word. We use the pre-trained word embeddings provided by [38].

Word Embedding Method	Adjectives	Verbs
Odds ratio	0.09	0.29
Distance + word2vec	0.44	0.37
Distance + GloVe	0.47	0.41
Distance + FastText	0.47	0.41
Gender direction	0.40	0.33
Gender dimension	0.20	0.08

Table 1: Pearson Correlation of gender scores between predictions from word embedding methods and ground truth.

To evaluate these methods, we use them to calculate the gender score for each word in our ground truth data, and compute the Pearson Correlation between the calculated gender score and the ground truth. As shown in Table 1, while the scores derived by all these methods are positively correlated with human defined gender stereotypes, the magnitude of the correlation is no larger than 0.47.

Labeling Gender Lexicon via Supervised Learning

Our results show that automated lexicon labeling via word embedding produces gender scores with mediocre results. Next, we propose to apply supervised learning to train gender score prediction models using our ground-truth dataset.

Since our labeled training dataset only contains around 2000 words, we cannot use deep neural network models that require large training datasets. Instead we use two classical machine learning models: Support Vector Machine (SVM) and Linear Regression (LR). Our models use word embeddings of each word as features, and the pre-trained word2vec, GloVe and FastText as model inputs.

We train our model using a random subset of 80% of the words from our ground truth dataset, then use the model to predict the score for the remaining 20%. We calculate the Pearson Correlation between the model predicted score and the ground

[1] https://code.google.com/archive/p/word2vec/
[2] https://nlp.stanford.edu/projects/glove/
[3] https://fasttext.cc/docs/en/pretrained-vectors.html

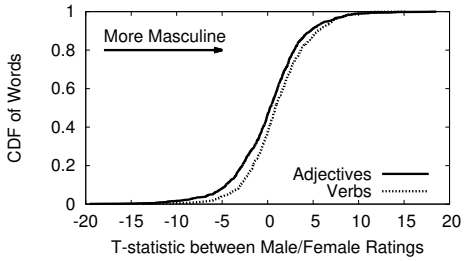


Figure 2: CDF of T-statistic between male ratings and female ratings. A higher positive number indicates people more easily associate the word with a typical man than a typical woman.

(a) Adjectives (b) Verbs

Figure 3: Wordle of adjectives and verbs. Red denotes feminine words and green denotes masculine words. Larger font size indicates stronger gender associations (larger T-statistic magnitude).

Supervised Learning Method	Adjectives	Verbs
LR + word2vec	0.63	0.57
SVM + word2vec	0.62	0.57
LR + GloVe	0.53	0.52
SVM + GloVe	0.58	0.55
LR + FastText	0.52	0.45
SVM + FastText	0.58	0.53

Table 2: Pearson Correlation of scores calculated by supervised learning methods and ground truth.

truth. We repeat the experiment 100 times, and compute the average correlation value. Results in Table 2 are higher than those produced by word embedding (Table 1). Since LR with word2vec produces the highest correlation value, we will use this configuration as our strongest supervised learning approach.

Computing Gender Score of Articles

Finally, we leverage our gender lexicon to detect gender stereotypical language in articles. Similar to previous work [33], we assign a gender score to an article based on word usage, first extracting all verbs and adjectives in the article, then adding the scores of these words together to get an article's gender score. If the total score is positive, the paragraph is labeled as consistent with masculine stereotypes or contradictory to feminine stereotypes, otherwise it is labeled as consistent with feminine stereotypes or contradictory to masculine stereotypes.

We evaluate the performance of this approach against an end-to-end deep learning model in Section 6.

DETECTING GENDER STEREOTYPES: AN END-TO-END APPROACH

We now introduce our end-to-end approach. Different from the lexicon approach, the end-to-end approach operates directly on paragraphs without breaking them down to individual words. Here we take a *supervised* learning approach: first gathers human perceptions of gender stereotypes at the paragraph level to build a moderately sized training dataset, then uses it to train a deep learning classification model based on the BERT language representation tool [11]. The resulting classification model can detect gender stereotypes on arbitrary paragraphs and articles.

We take two important steps to avoid potential bias in our trained model. *First*, we collect our training (and testing) data by crowdsourcing articles with different representation of gender stereotypes. These are existing articles aiming at describing a man (or woman) while the actual human perception can be either consistent with or contradictory to the original intent. Using this ground truth dataset, we formulate our gender stereotype detection problem as *two binary classification problems*: determining whether the description of a man is consistent with masculine stereotypes, and whether the description of a woman is consistent with feminine stereotypes.

Second, our model is built on BERT [11], a unsupervised language representation tool that converts text articles into vectors. Since BERT does not target any particular language task, we can use it to examine and search for common patterns of language use that may be associated with gender stereotypes. Specifically, we use our training dataset to fine-tune BERT by adding one additional output layer to implement the above mentioned binary classification tasks.

Collecting Labeled Articles via Crowdsourcing

To collect the articles for our tasks, we perform a survey study. Our survey ask users to search the Internet, and copy & paste articles (or a few paragraphs of an article) that meet the following requirements: it describes a man (or woman), and the description is consistent with (or contradictory to) common gender stereotypes. We ask participants to briefly state the reason for choosing each article. Each participant is asked to provide 4 articles, with 4 different combinations of requirements (man or woman, consistent or contradict). The entire survey takes about 25 minutes.

Domain	Number	Domain	Number	Domain	Number
wikipedia.org	385	npr.org	58	huffpost.com	46
nytimes.com	178	forbes.com	57	washingtonpost.com	45
theguardian.com	78	dailymail.co.uk	54	biography.com	39
cnn.com	78	foxnews.com	48	cnbc.com	37
people.com	63	time.com	47	vogue.com	37

Table 3: Top domains and number of articles from each domain.

	Consistent	Contradict
Masculine	championship, ceo, gun, league, player, businessman, top, service, mountain, fight, basketball, win, drive	gay, makeup, gender, singer, fashion, comfortable, mom, youtube, cosmetic, dress, feel, wear, caregiver, beauty, sexuality
Feminine	cook, child, home, beautiful, beauty, care, clean, fighter, daughter, makeup, family, mother, dress, kid, mom	field, champion, history, sport, athlete, fight, martial, force, training, team, technology, institute, lesbian, rank, tech

Table 4: Top keywords that distinguish consistent and contradicting stereotypes.

We recruit survey participants from two different sources, Prolific[4], and undergraduate students from our university. Prolific is a crowdsourcing service aiming at providing high quality data that empowers research. The Prolific participants are compensated $3, and the students are compensated with 0.5 research course credits.

In total, we received results from 980 distinct Prolific workers and 110 students. Again, we limit participation to US residents to reduce potential bias due to cultural differences. Among the 980 Prolific workers, 508 (51.8%) are male participants, 457 (46.6%) are female participants, and 15 (1.5%) chose not to disclose their gender. Among the 110 college students, 52 (47.3%) indicated male and 58 (52.7%) female.

We received 4360 articles (4 per participant), and filtered out 27 articles that do not contain any pronouns, named entities or gender specific words, indicating that these articles are not likely to be descriptions of people. When looking at the sources of the articles, most of the articles are from biography pages (*e.g.,* Wikipedia), or news sites (*e.g.,* New York Times). Table 3 lists the most frequently used domains.

To understand the content of these articles, we extract top keywords in each category using Chi-square statistics [36], which measures how strongly a word can be used to distinguish articles in different categories, *i.e.,* consistent or contradictory. We calculate Chi-square statistics for masculine stereotypes and feminine stereotypes separately, and list the top keywords in Table 4. We see that our survey participants commonly choose sports and business related terms for men and domestic related terms for women as exemplifying gender stereotypes. Further, some similarities appear between men who contradict stereotypes and women who are consistent with stereotypes (and vice versa).

Building the Classification Model

Our classification model will run two tasks: determining whether the description of a man is consistent with masculine stereotypes, and whether the description of a woman is consistent with feminine stereotypes. Thus we use the articles describing men for the first task and those describing women for the second task. We randomly split up the data into chunks of 8:1:1 for training:validation:testing. We use our training data to fine-tune the BERT model, and use the validation set to identify the optimal hyper-parameters, which is 2e-5 learning rate for 3 epochs. In the following section we use the test data to examine the model performance and compare it to the lexicon approach.

[4]https://prolific.ac/

EMPIRICAL EVALUATION

We now evaluate and compare the lexicon approach and the end-to-end approach using the above mentioned test data. We apply each approach on the test data to predict whether each test article is consistent with or contradictory to its intended gender stereotypes. We then compare these results to the ground truth provided by humans.

Overall, our study shows that the end-to-end approach largely outperforms the lexicon approach, in terms of detection accuracy and robustness. A closer look at these results also offers insights into some fundamental problems facing the lexicon approach. We further confirm that the end-to-end approach does not require a large training dataset to perform well. Finally, we test the end-to-end approach on a practical task of detecting gender bias in job advertisements, which outperforms the industry state-of-the-art.

Validating Testing Dataset via User Survey

To ensure that our evaluation (using the testing dataset) is sound, we performed another user study to understand whether the per-user contributed labels in the test dataset can accurately capture public perception of gender stereotypes. Specifically, each survey participant is given 10 articles randomly selected from our testing dataset, and is asked to score on an 7-point Likert Scale, where 1 indicates "strongly contradictory" and 7 indicates "strongly consistent".

We ran the study on Prolific, and each user was compensated $1.10. In total, we received results from 203 distinct Prolific workers, of which 108 (53.2%) are male participants, 92 (45.3%) are female participants, and 3 (1.5%) participants chose not to disclose their gender.

Each article in the testing dataset received at least 4 ratings, from which we computed the average rating and compared it to the actual label of the article. Overall, the new multi-user rating is reasonably consistent with the original rating, indicating that our testing dataset offers a consistent, public view of gender stereotypes.

	Accuracy (M)	AUC (M)	Accuracy (F)	AUC (F)
Lexicon (Full Set)	0.67	0.70	0.68	0.71
Lexicon (Ground truth Set)	0.58	0.61	0.62	0.64
End-to-End	0.77	0.85	0.80	0.87

Table 5: Accuracy / AUC of lexicon and end-to-end approaches among articles describing male (M) and female (F).

	PAQ	BSRI	Gaucher et al.
% words overlap	0.22	0.38	0.48
% overlap and matching labels	N/A	0.83	0.85

Table 6: Comparison of lexicon coverage against prior work. PAQ does not provide gender labels, thus no direct comparison.

Comparing Lexicon and End-to-End Approaches

We evaluate how accurately the lexicon and end-to-end approaches can predict gender stereotypes in written articles, by computing prediction accuracy and Area Under the Curve (AUC). The results in Table 5 show that the end-to-end approach is much more accurate.

In this table we also show the results of the lexicon approach when using the ground-truth lexicon (labeled by our user survey) and the full set lexicon (expanded via supervised learning). We see that the use of full set lexicon effectively improves the detection accuracy, but still cannot match that of the end-to-end approach. Although the two approaches are trained on different data, both datasets are curated from commonly used language in current bodies of text, then evaluated by multiple crowdworkers to generate ground truth labels. As such, we believe these comparisons between the two approaches are fair.

Understanding the Lexicon Approach

To understand why the lexicon approach generates less satisfactory prediction results, we manually examine *all* the incorrect predictions the lexicon approach makes in the test set. We summarize the possible reasons behind the misclassifications along with examples in Table 7. For each reason, we also calculate how many times the lexicon approach makes incorrect predictions ("Lexicon Wrong" column) and how many times the end-to-end approach makes incorrect predictions among these cases ("Also E-to-E Wrong" column). The detailed explanations are as follows:

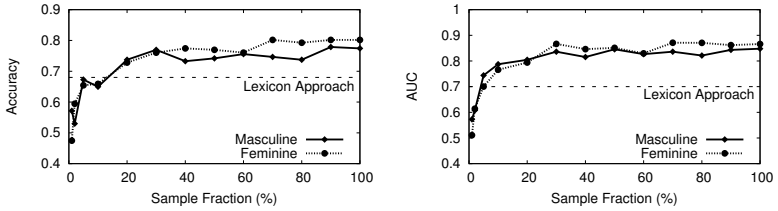
- *Lexicon Coverage:* Our lexicon only covers adjectives and verbs, and gender stereotypes can be expressed by words outside of our lexicon. For example, "PhD" could be a word associated with masculine stereotypes.
- *Phrase:* The stereotype is expressed by a multiple-word description, which can not be captured by single words in the lexicon.
- *Non-human:* The word that indicates strong gender stereotypes is used on a non-human object. For example, in "tiny shadows", "tiny" is labeled as a feminine word but "tiny shadows" does not indicate femininity.
- *Consistent and contradictory:* The article contains a description of a person who has some characteristics that are consistent with stereotypes and some other characteristics that are contradictory. Although the article may focus on one more than other, the lexicon approach can not identify the general focus by word count.
- *Multiple people* The article describes more than one person, usually one person as the main character while the others are supporting characters. The lexicon approach can not isolate the descriptions of the correct person.
- *Subtle stereotyping, insufficient information:* Some users may have different understandings of stereotypes, or insinuate gender stereotypes not explicitly written in the text. For example, an article about American actor Peter Dinklage is labeled as contradicting masculine stereotypes because he is a dwarf, but the fact is not found in the article.
- *Data noise:* These are the low quality response including cases when the users provide responses that do not fit our task requirement (*e.g.,* the paragraph is not a description of a person, article is consistent with stereotypes when we ask for contradiction).

We also compare our lexicon to those from previous works (PAQ, BSRI, Gaucher), and the overlap with previous lexicons is less than half (see Table 6). We found that many of the terms are not often found in current language, and the sparsity of their occurrence in our data makes any comparison of results marginally meaningful. For instance, we found that words such as ?aggressive,? ?assertive,? ?dominant? or ?forceful? are labeled as masculine items in the BSRI, but are not commonly used enough to be included in our lexicon. While PAQ [29] and BSRI [3] include a mix of words and short phrases, our lexicon only considers single words. Phrases such as ?acts as a leader,? ?defends own beliefs,? makes decisions easily? or ?willing to take risks? do not directly translate to single words so a direct comparison to such items was not possible. The lexicon generated by Gaucher et al. [15] includes several truncated words, which we found to be an oversimplification, and resulted in some conflicting labels (*e.g.,* for the base "respon*": "response" scored as feminine, but "respond" scored as masculine). Also, since we evaluate adjectives and verbs separately, some words score as feminine in one tense but masculine in the other (*e.g.,* "yield" scores masculine as a verb, but feminine as an adjective). Of those words that overlap, most of our labels are consistent with BSRI and Gaucher et al., with several exceptions (*e.g.,* previous works labeled "loyal" and "communal" as feminine, and "confident" and "individualistic" as masculine, but our scores are opposite).

Being data-driven, our work is able to evaluate most commonly used verbs and adjectives in *current* bodies of text. The differences in word coverage between our lexicon and previous works may indicate that many of the words from previous lexicons are not commonly used to describe people in modern society. Although some of the terms appear to exemplify strong gender connotations (*e.g.,* "feminine," "masculine"), such words do not often appear in descriptive language and therefore are not necessary to include in the lexicon. Moreover, our method demonstrates how contextual information,

Reason	Lexicon Wrong	Also E-to-E Wrong	Example
Lexicon Coverage	8	0	The first woman I invited to co-author a **publication** was in 2015, four years after completing my **PhD**.
Phrase	10	0	... who **paints his fingernails**, **braids his hair** and **poses for gay magazines**...
Non-human	6	0	Katie Bouman has already worked on looking around corners by analyzing **tiny** shadows ...
Consistent and contradictory	27	4	Even as I regularly work out and **lift weights**, I am a rather **fragile** excuse for a woman, constantly getting sick...
Multiple people	10	3	**My wife** had more earning potential and so I volunteered to concentrate on family and home.
Subtle stereotype, insufficient information	50	123	*American actor Peter Dinklage is labeled as contradicting masculine stereotypes because he is a dwarf, which is not discussed.*
Data noise	30	18	*Random response or failure to meet task requirement.*

Table 7: Reasons for lexicon approach making wrong classification. The "Lexicon Wrong" column is the number of cases when the lexicon approach makes a wrong prediction, and the "and E-to-E Wrong" column is the number of cases the end-to-end approach is also wrong among these cases. Bold words are words that are closely related to the reasons provided by the survey participants. Italic words are not exact content from our data, but summarize participant explanations.



(a) Accuracy (b) AUC

Figure 4: End-to-end approach performance with different training data size, compared to lexicon approach (similar number for masculine task and feminine task).

	Textio	Unitive	BERT fine-tune
% of females	0.59	0.54	0.77
Attractiveness to female applicants	0.64	0.54	0.80

Table 8: Pearson correlation between user responses and gender bias scores.

including part-of-speech, plays a significant role in people's perceptions of language, an aspect unaccounted for in previous works.

Impact of Training Data on End-to-End Approach

End-to-end learning approaches often require a large amount of training data. Here, we seek to quantify how much training data is needed to outperform the lexicon approach. We vary the size of the training data, train the BERT fine-tuning model, and evaluate the performance on the same test set. The results are shown in Figure 4. We have two observations. First, the performance plateaus when the training data size reaches 40% of our full training data, beyond which further increase in training data size yields no performance gain. Second, the end-to-end approach can outperform the lexicon approach even when the training data is only 10% of current size, which is about 150 articles. This indicates that the end-to-end approach does not need a large corpus to learn typical gender stereotypes.

Application: Gender Bias in Job Postings

Finally, we apply our end-to-end approach to detect gender biased language in job postings, and compare the performance to state-of-the-art tools from industry. The data is publicly shared data from a 2017 survey study [33]. In the survey, participants were asked to read 30 job postings and answer questions about the extent to which the job posting is biased by gender. Here, we use their answers on 2 questions to quantify the gender bias in the job postings: 1) the percentage of women they presume are currently in the position (0%-100%) and 2) how likely the job posting attracts more female applications or male applicants (7-point Likert Scale, from 7 indicating the post attracts mostly male applicants to 1 indicating the post attracts mostly female applicants).

To convert our male and female stereotype detection models into a single gender bias indicator, we take the job posting text and use it as input in both the masculine stereotype classifier and feminine stereotype classifier. We take the probability output from both classifiers, and calculate the difference in the two probabilities. For example, if a job posting is predicted as 90% likely to be consistent with masculine stereotypes and 60% likely to be consistent with feminine stereotype, the score of the job posting is 0.3, which indicates a masculine bias.

We calculate the score for all the job advertisements, along with two state-of-the-art services cited by prior work: *Textio* and *Unitive* (now renamed Talent Sonar), both of which are specifically designed to detect gender bias in job posts using a lexicon-based approach [33]. Table 8 shows the correlation between the scores and user responses. This shows that although

the models in our end-to-end approach are not specifically trained for job advertisements, they still outperform the best lexicon approaches designed for this task.

CONCLUSION

Our work seeks to reconcile the traditional lexicon-based approaches for detecting gender stereotypes in language, with modern natural language processing tools almost entirely based on end-to-end deep learning models. The high level question is: what approach should researchers and practitioners take moving forward, an updated version of lexicon-based models (which we developed in this work), or an end-to-end deep learning model built on existing language models (BERT) and further trained with paragraph-length text samples? Our work finds that despite our best efforts to update and strengthen the lexicon-based models, end-to-end models based on BERT provide substantially stronger results, even when trained on our moderately-sized, crowdsourced dataset. In fact, when applied to the context of gender bias in job listings, our end-to-end model significantly outperforms models used by industry services.

Our work has several limitations. First, we simplified the task of gender stereotype detection as binary classifications for masculinity or femininity. A correct and more inclusive model would include non-binary and trans labels. Also, we only collect a moderate dataset with a few thousand training samples, which is small relative to some popular datasets in the NLP community that contain hundreds of thousands of samples. It is unclear whether the performance can be significantly improved if the training data was orders of magnitude larger. Further, we attempt to mitigate effects from cultural biases by limiting our participant pool to US workers, but acknowledge that some biases may remain. Lastly, our end-to-end approach comes from fine-tuning the current state-of-the-art BERT model. It is possible that a more task-specific model can generate a better performance. We hope to address these limitations in our ongoing work.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their feedback. This work is supported in part by the National Science Foundation grants CNS-1923778 and CNS-1705042. Any opinions, findings, and conclusions or recommendations expressed in this material do not necessarily reflect the views of any funding agencies.

REFERENCES

[1] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 759–760.

[2] Mahzarin R Banaji and Curtis D Hardin. 1996. Automatic stereotyping. *Psychological science* 7, 3 (1996), 136–141.

[3] Sandra L Bem. 1974. The Measurement of Psychological Androgyny. *Journal of Consulting and Clinical Psychology* 42, 2 (1974), 155–162.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. of NIPS*. 4349–4357.

[6] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 14 (April 2017), 183–186.

[7] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017b. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.

[8] Amy JC Cuddy, Susan T Fiske, and Peter Glick. 2004. When professionals become mothers, warmth doesn't cut the ice. *Journal of Social issues* 60, 4 (2004), 701–718.

[9] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Eleventh International AAAI Conference on Web and Social Media*.

[10] M Lee Dean and Charlotte Chucky Tate. 2017. Extending the legacy of Sandra Bem: Psychological androgyny as a touchstone conceptual advance for the study of gender in psychological science. *Sex Roles* 76, 11-12 (2017), 643–654.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[12] Kristin Donnelly and Jean M Twenge. 2016. Masculine and feminine traits on the Bem Sex-Role inventory, 1993–2012: a cross-temporal meta-analysis. *Sex Roles* (2016), 1–10.

[13] Alice H Eagly, Wendy Wood, and Amanda B Diekman. 2000. Social role theory of sex differences and similarities: A current appraisal. *The developmental social psychology of gender* 12 (2000), 174.

[14] Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *Proc. of ICWSM*.

[15] Danielle Gaucher, Justin Friesen, and Aaron C Kay. 2011. Evidence that gendered wording in job advertisements exists and sustains gender inequality. *Journal of Personality and Social Psychology* 101, 1 (2011), 109.

[16] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proc. of ECCV*. 793–811.

[17] Rie Johnson and Tong Zhang. 2016. Supervised and Semi-Supervised Text Categorization using LSTM for Region Embeddings. In *International Conference on Machine Learning*. 526–534.

[18] Anne Maass and Luciano Arcuri. 1996. Language and stereotyping. *Stereotypes and stereotyping* (1996), 193–226.

[19] Juan M Madera, Michelle R Hebl, and Randi C Martin. 2009. Gender and letters of recommendation for academia: agentic and communal differences. *Journal of Applied Psychology* 94, 6 (2009), 1591.

[20] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On Measuring Social Biases in Sentence Encoders. In *Proc. of NAACL*.

[21] Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. (2017).

[22] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[23] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 174–184.

[24] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proc. of EMNLP*. 2799–2804.

[25] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[26] Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Proc. of ACL*, Vol. 1. 1669–1678.

[27] Paul Rosenkrantz, Susan Vogel, Helen Bee, Inge Broverman, and Donald M Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of Consulting and Clinical Psychology* 32, 3 (1968), 287.

[28] Stephen A Schullo and Burton L Alperson. 1984. Interpersonal phenomenology as a function of sexual orientation, sex, sentiment, and trait categories in long-term dyadic relationships. *Journal of Personality and Social Psychology* 47, 5 (1984), 983.

[29] Janet T. Spence, Robert L. Helmreich, and Joy Stapp. 1974. The Personal Attributes Questionnaire: A measure of sex role stereotypes and masculinity-femininity. *JSAS Catalog of selected documents in psychology* 4, 43 (1974).

[30] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

[31] Latany Sweeney. 2013. Discrimination in online ad delivery. In *arXiv:1301.6822*.

[32] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37, 2 (2011), 267–307.

[33] Shiliang Tang, Xinyi Zhang, Jenna Cryan, Miriam J. Metzger, Haitao Zheng, and Ben Y. Zhao. 2017. Gender bias in the job market: A longitudinal analysis. In *Proc. of CSCW*.

[34] Jean M Twenge. 1997. Changes in masculine and feminine traits over time: A meta-analysis. *Sex roles* 36, 5-6 (1997), 305–325.

[35] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proc. of ICWSM*.

[36] Yiming Yang and Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. In *ICML*, Vol. 97. 35.

[37] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proc. of NAACL*, Vol. 2.

[38] Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning Gender-Neutral Word Embeddings. In *Proc. of EMNLP*. 4847–4853.