GenderQuant: Quantifying Mention-Level Genderedness

Ananya
University of California
Irvine, CA
aananya@uci.edu

Nitya Parthasarthi
Massachusetts Institute of Technology
Cambridge, MA

Sameer Singh
University of California
Irvine, CA
sameer@uci.edu

Abstract

Language is gendered if the context surrounding a mention is suggestive of a particular binary gender for that mention. Detecting the different ways in which language is gendered is an important task since gendered language can bias NLP models (such as for coreference resolution). This task is challenging since genderedness is often expressed in subtle ways. Existing approaches need considerable annotation efforts for each language, domain, and author, and often require handcrafted lexicons and features. Additionally, these approaches do not provide a quantifiable measure of how gendered the text is, nor are they applicable at the fine-grained mention level.

In this paper, we use existing NLP pipelines to automatically annotate gender of mentions in the text. On corpora labeled using this method, we train a supervised classifier to predict the gender of any mention from its context and evaluate it on unseen text. The model confidence for a mention's gender can be used as a proxy to indicate the level of genderedness of the context. We test this gendered language detector on movie summaries, movie reviews, news articles, and fiction novels, achieving an AUC-ROC of up to 0.71, and observe that the model predictions agree with human judgments collected for this task. We also provide examples of detected gendered sentences from aforementioned domains.

1 Introduction

Language can be extraordinarily gendered (Moulton et al., 1978). *Genderedness* in language is when we use words or phrases that are stereotypical or indicative of a particular gender (we only consider male vs female in this work) (Prior, 2017). It is important to detect this bias in language since not only is this bias propagated to the readers (Menegatti and Rubini, 2017), but also machine learning algorithms trained on gendered corpora tend to become biased (Zhao et al., 2018a; Rudinger et al., 2018), often aggravating the disparity (Zhao et al., 2017).

Bias in language and machine learning systems can lead to unfair treatment, e.g., early work by Moulton et al. (1978) shows that males have an advantage in contexts where they are referred to by a putative neutral term. Recent work on coreference resolution systems (Zhao et al., 2018a) shows that bias in machine learning systems originates from training on existing corpora, resulting in male-stereotyped professions like surgeon and president incorrectly resolved to males instead of females. Such biases in machine learning systems can lead to unintentional biases in downstream tasks producing effects like preferential treatment to male candidates over female candidates when selecting resumes (Dastin, 2018). Detecting these biases is the first step in finding a solution.

Most of the current works for related problems tend to be domain-specific (Fu et al., 2016), rely on techniques such as simple counting of gender occurrences (Ali et al., 2010), or use manually constructed lexicons and features for analysis (Trix and Psenka, 2003), and thus do not generalize well and require expensive manual supervision. Existing approaches also tend to either focus on the whole corpus/article being gendered (Schmader et al., 2007; Trix and Psenka, 2003) or a specific word being gendered (Caliskan et al., 2017; Bolukbasi et al., 2016; Zhao et al., 2018b), thus failing to capture the subtle occurrences of genderedness at mention-level or giving a quantifiable measure of how gendered the text is.

In this paper, we develop a method that eliminates the manual annotation requirement, and can generalize to words, phrases, sentences, articles, as well as whole corpora. We present a framework for automated data labeling by combining existing NLP pipelines to identify sentence boundaries and mentions (using NER tagger) and using a gen-

Proceedings of NAACL-HLT 2019, pages 2959–2969
Minneapolis, Minnesota, June 2 - June 7, 2019.

Female	Their client is [redacted] who has got gorgeous hair.
Male	[redacted] intends to marry lovely Gauri.
Female	Raja intends to marry lovely [redacted] .

Figure 1: **Examples of gendered sentences.** Our model predicts gender for mentions located at positions indicated by colored boxes.

der classifier for names to get the gender of the mentions. We build a classifier using this annotation to predict gender of a mention *only from its context* and quantitatively analyze the *genderedness* of various contexts using this model. Figure 1 shows example inputs and outputs for our model. Input to the model is the context sentence around a target mention (indicated by colored boxes), and the model prediction is the gender of this mention. For the first sentence, the model uses context information coming from 'gorgeous hair' to predict the gender of mention to be female, indicating a more gendered sentence. Similarly for the third sentence, model uses contextual information from the adjective 'lovely' and its proximity to the target mention to predict female. For the second sentence, the target mention is subject for the verb in phrase 'intends to marry' and the object to be married is 'lovely Gauri'. Our model uses this information to predict gender of target mention as male.

Since our data labeling pipeline is automated, we can easily annotate millions of documents and train complex classifiers that can accurately model the context. These classifiers can be used to predict the gender of a mention from given context and quantify genderedness. We present instantiations of this framework on four domains: news articles, novels, movie summaries, and movie reviews. Since we are the first to study the task of mention-level gender detection, we evaluate the difficulty of the task and introduce the first benchmark using a user study. We find that the task is challenging, and our model predictions corroborate with human predictions. We present qualitative results of our model showing genderedness at different granularities – word, phrase, sentence, and corpus.

2 Related Work

Gender Bias in Datasets A number of approaches have considered gendered language use. Blatt (2017) shows that *shivered, wept, screamed* are disproportionately used to describe women while *muttered, grinned* are used to describe men. Studies on gender bias in student evaluations for instructors (Eidinger, 2017; MacNell et al., 2015; Boring et al., 2016; Centra and Gaubatz, 2000), and recommendation letters (Trix and Psenka, 2003; Schmader et al., 2007) also show similar disparities in terms of harshness of evaluations, length of letters, descriptive words, and use of standout adjectives. Bias in language has also been studied for textbooks (Otlowski, 2003; Gharbavi and Mousavi, 2012; Macaulay and Brice, 1997), Wikipedia edits (Recasens et al., 2013), political text (Yano et al., 2010), media content (Ali et al., 2010; Len-Ríos et al., 2005; Smith, 1997), sports journalism (Eastman and Billings, 2000; Tyler Eastman, 2001; Kinnick, 1998; Fu et al., 2016) and in movie character portrayals (Ramakrishna et al., 2017; Sap et al., 2017). These approaches are domain-specific and rely on techniques like counting gender occurrences, manually annotating words or mentions, constructing list of keywords and lexicons, carrying out surveys, etc. Our approach works across domains, and does not require manual annotations.

There has been significant amount of work in detecting author's gender (Koppel et al., 2002; Herring and Paolillo, 2006; Sarawgi et al., 2011; Mukherjee and Liu, 2010; Burger et al., 2011) for text, speaker gender for dialogues (Schofield and Mehr, 2016) in films, and to detect and reduce biases in these (Tatman, 2017; Thelwall, 2018; Koolen and van Cranenburgh, 2017). While we do not focus on predicting the gender of the author, our framework can be used as a tool to compare the use of gendered language across various authors, or across various works by the same author.

Gender Bias in NLP Pipelines There has also been recent interest in examining the role of gender bias in existing NLP pipelines. Caliskan et al. (2017) and Bolukbasi et al. (2016) show that word embeddings exhibit gender stereotypes. Garg et al. (2018) build on this idea, using word embeddings to characterize the evolution of gender stereotypes during the 20^{th} and 21^{st} centuries. Subsequent works attempt to mitigate this bias in embeddings (Zhao et al., 2018b). Zhao et al. (2019) extend the idea to contextualized word embeddings (Peters et al., 2018), and quantify and propose ways to mitigate gender bias in them , while Gonen and Goldberg (2019) show that current approaches for debiasing embeddings are superficial.

Researchers have studied gender bias outside word embeddings as well. Zhao et al. (2017) show

that datasets for multi-label object classification and visual semantic role labeling are gender-biased and that models trained on these datasets amplify this bias, while Rudinger et al. (2017) find racial, religious and gender stereotypes in the SNLI corpus and Park et al. (2018) analyze gender bias in abusive language datasets. Zhao et al. (2018a) and Rudinger et al. (2018) detect bias in existing coreference resolution systems, and Webster et al. (2018) build a gender-balanced labeled corpus of ambiguous pronoun-name pairs to understand this bias. All of these either focus on whether the corpus as a whole is gendered or if a single word is gendered (in case of word embeddings). Instead, we train a classifier to detect and quantify gendered language at mention-level. Our framework can also be used to quantify genderedness at different levels – mention, sentence, document, or corpus.

3 Gendered Language Detection

In this section, we elaborate on gendered language, and describe our proposed architecture and training method to detect and quantify it.

Gendered Language Gendered language is the use of words and phrases that *discriminate*[1] the gender of a subject. In other words, the gender of mentioned person should be easy to predict from context if the text is gendered. Examples of gendered language can be found in the use of stereotypes like linking women to homemakers and men to programmers (Bolukbasi et al., 2016) or when pronouns, adverbs, adjectives, nouns are used carelessly, e.g., when a masculine pronoun mention 'he' is used to refer to both sexes (Cottier, 2018; Stout and Dasgupta, 2011) or when pronoun mentions are used exclusively to define professions by gender (using 'she' when talking about a nurse). Detecting gendered language is incredibly challenging since the ways in which gender is expressed can vary considerably across authors, domains, and time periods, making any approach that requires annotations to be corpus-specific.

Proposed Architecture We are interested in determining the extent to which language in context of the mention reveals the gender. Humans learn to detect gendered language based on a lifetime of reading and observing society, and learning language specific to each gender. We use this intuition to propose an automated framework for detecting mention-level genderedness for any corpus (an overview is shown in Figure 2).

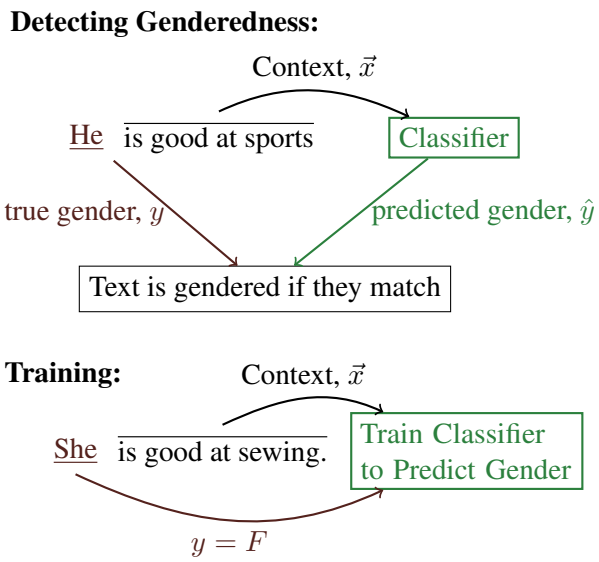

Figure 2: **Model overview.** Using an automatically labeled corpus, we train a classifier to predict the gender for each mention given its context (without the mention). For each target mention and its context sentence, we check whether the predicted gender matches the true prediction; the level of agreement indicates the *genderedness* of the text.

The input to the gender detector is the context (sentence) without the target mention and the output is the detector prediction for gender of that mention. For a mention i, let C_b^i be the context before the mention, C_a^i be the context after the mention and f_θ be the gender detector. Then,

$$p_i = f_\theta(C_a^i, C_b^i) \tag{1}$$

is the detector's probability (confidence) that the mention i is female, i.e. p_i close to 0 indicates high confidence for predicting male while close to 1 indicates high confidence for females.

We use the detector's probability of the *true* gender of the mention (g_i) as an estimate of how gendered the text is: a high probability indicates that gender is heavily reflected in the context. We define this as the gendered score, given by G_m^i here:

$$G_m^i = \begin{cases} p_i & \text{True gender } g_i \text{ is female} \\ 1 - p_i & \text{True gender } g_i \text{ is male} \end{cases} \tag{2}$$

We define gendered score for a document as the average of gender score for all of its mentions. i.e.

$$G_{\text{doc}} = \frac{1}{N} \sum_{i=1}^{N} G_m^i \tag{3}$$

where N is the number of person mentions.

[1] in the machine learning sense of the word

Dataset	Male Mentions	Female Mentions
Reviews	298,580	104,632
Summaries	405,368	186,626
News	19,012,473	3,902,510
Novels	18,433,400	6,982,348

Table 1: **Dataset details.** Number of male and female mentions in the different datasets are shown.

As detector, rather than relying on frequency-based linear models or Naive Bayes model, we can use simple as well as more complex classifiers that can accurately model semantics and syntax of the context. For example, we use bag-of-words models (with logistic regression classifier) and recurrent neural network models as described in Section 4.2.

4 Datasets and Classifiers

In this section, we give details about the datasets, our pipeline for automated data labeling, filtering and processing applied to contexts in order to remove obvious gender information, and the classifiers used to classify gender of a given mention.

4.1 Datasets and data processing

To illustrate the utility of our proposed approach, we analyze text from four different domains:

- New York Times articles from the Annotated Gigaword corpus (Napoles et al., 2012)
- Novels from Gutenberg corpus [2]
- IMDB Movie Reviews (Maas et al., 2011)
- Movie Summaries (Bamman et al., 2013)

These domains cover a variety of writing styles. While the novels represent fictional writing, news articles are non-fictional. Movie summaries dataset describes the plot of the movies, i.e., how gender is represented in the plots, and the movie reviews dataset provides the ways in which people express their views on the plots, i.e., how gender is represented in user perception of the movie.

We train classifiers for each domain to predict the gender of mentions from the context they appear in, and use the resulting classifiers to detect gendered language. Similar idea is explored in Choi et al. (2018) to detect the type of mention from the context. For news, data from the first 6 months for every year is used for training, next three for validation, and last three for testing. For novels and movie summaries, we divide the data randomly into 50 : 20 : 30 split for train, validation and test data. For movie reviews, we use the predefined split for train and test data, further dividing the training data into training and validation data.

[2]Project Gutenberg, from www.gutenberg.org

1	Miss Mary Briganza will go to Korea with her parents.
2	Miss <*female*> will go to Korea with her parents.
3	<*Title*> <*female*> will go to Korea with <*their*> parents.
4	<*Title*> _____ will go to Korea with <*Their*> parents.

Figure 3: **Annotation pipeline.** We show the annotation and data processing pipeline step-by-step. **1** is the original sentence. **2** is the sentence after detecting positions of all mentions in the sentence, and replacing them with a placeholder for gender of that mention. **3** is the sentence after removing obvious gender information such as replacing gendered-pronouns (*he*, *she*) with a gender-neutral pronoun (*them*), and titles *Mr., Mrs., Miss* with a placeholder title word. **4** is the final input to our classifier where _____ is the position of mention for which classifier needs to predict gender (male or female).

Labeling Gender for Mentions We illustrate our processing pipeline via the example sentence in Figure 3. For mention-level gender prediction, we need a dataset with identified person mentions and their genders. Since we do not have labeled data, we need to identify mentions in contexts, and assign gender labels to them. Along with pronouns 'he' and 'she', we use spacy[3] to tag all corpora with NER tags to identify the set of person mentions. We use the SSN baby names dataset[4] from 1880 to 2016 to assign gender to each name. If a name is associated with more than one sex, we exclude it if it is ambiguous (being less than 4 times more frequent for one sex), but otherwise assign it to the more frequent sex. If a name is absent from our list of names, we replace the mention with a placeholder <*Person*>. Table 1 shows the count of male and female mention-context pairs generated using this pipeline. Processed sentence after this step is sentence 2 in Figure 3.

Filtering and Input Context Processing To remove obvious, uninteresting gender information, we discard sentences that contain any word from a gender-specific lexicon as used by Bolukbasi et al. (2016) such as gender-specific occupation words and gender-specific familial relation words, e.g., 'man', 'woman', 'prince', and 'hostess'. Complete list is given in Appendix A. For contexts that contain gender-indicative pronouns ('him', 'her', 'his',

[3]https://spacy.io

[4]https://www.ssa.gov/oact/babynames/

'hers', 'himself', 'herself'), we replace them with a gender-neutral pronoun (*'them', 'their'*). All other mentions in the context (including 'he' and 'she') are replaced with a gender neutral word, and titles (*'Mr', 'Mrs', 'Miss'*) are replaced with a gender-neutral title word. Sentence 3 in Figure 3 is the result after this stage.

4.2 Classifier Details

The input to classifier is sentence 4 in Figure 3 and the target is 1 (for female). We extract such mention-context pairs from large text corpora to train classifiers that can predict the gender of individual mentions from their context using minimal manual supervision (as illustrated in Figure 2).

Bag-of-words and ngrams We construct bag-of-word classifiers by selecting the $50,000$ most frequent words from the training subset, and bag-of-ngrams models by selecting the $100,000$ most frequent n-grams (up to 3-grams), for each dataset. We explore a number of classifiers like logistic regression, support vector machines, random forest classifier, and choose logistic regression classifier since it consistently performs better than others.

LSTMs and CNNs We use both uni- and bidirectional LSTM recurrent neural networks for the context. In the 2-way LSTM model, we use two separate LSTMs: one for context before the mention, and the other for context after the mention. The direction of LSTM for latter part is reversed so that the model gives more importance to words closer to the target mention. This is followed by a sigmoid layer after the concatenation of the final hidden states. The input layers are initialized using the Glove vectors (Pennington et al., 2014), and are updated during training. We train the classifier with log-loss and Adam (Kingma and Ba, 2014) optimization algorithm, including dropout (Srivastava et al., 2014) and early stopping for regularization. Hyper-parameters are tuned for different domains separately. We experiment with ELMo embeddings (Peters et al., 2018), convolutional neural network (CNN)-based architectures, vanilla recurrent neural network (RNN), and gated recurrent unit (GRU) (Cho et al., 2014) models as well.

Performance Since our datasets are imbalanced, we use AUC-ROC as a performance metric. Table 2 shows AUC-ROCs for various models for all the datasets. Conventional bag-of-word/ngrams classifiers exhibit AUC-ROCs comparable to more complex LSTM and CNN classifiers. We use 2-way LSTM as the classifier for our final analysis.

	Reviews	Summaries	News	Novels
Bag-of-ngrams	0.64	0.62	**0.70**	**0.71**
Bag-of-word	0.63	0.62	**0.70**	**0.71**
Single LSTM	**0.67**	0.62	0.63	0.63
2-way LSTM	**0.67**	**0.66**	0.68	0.67
2-way LSTM + ELMo	**0.67**	0.65	**0.70**	0.69
2-way RNN	0.65	0.63	0.65	0.63
2-way GRU	**0.67**	**0.66**	0.69	0.66
CNN	0.66	0.64	0.68	0.64

Table 2: **AUC-ROCs for different models (evaluated on test data).**

5 Human Annotation Evaluation

To assess the difficulty of this task and to compare performance of our gendered language detector against a human baseline, we use Amazon Mechanical Turk to get human annotations for 500 random sentences from the test sets of each domain.

Task Description Turkers are shown sentences with missing mention, e.g., *'Sandwich maker ____ said mojo and fresh roasted pork are key to a great Cuban sandwich'*, and are asked to guess the gender of the missing mention. We sample the sentences such that the true labels (male/female) are balanced. For our study, we use two tasks that slightly differ from one another in the decisions turkers need to make. In one task, turkers are given only two options, *male* and *female*, forcing them to make a choice. In the second task, turkers are given five options on the Likert scale: *extremely likely male, likely male, neutral, likely female, extremely likely female* allowing for a finer scale of decision. We include examples in the instructions, and a few extremely easy examples as probes to verify quality (Munro et al., 2010). Each worker is shown 35 sentences from a single domain. On average, we collect 7 human annotations per sentence.

Do humans predict gender well? Sentences that do not have a clear majority are removed from our analysis. As a measure of inter-rater reliability, we compute pairwise and majority agreement, in Table 3. Percentage improvement over chance agreement is higher for 5-scale rating compared to 2-scale rating indicating that users tend to agree more when they are able to tag the borderline (possibly confusing) mentions as gender-neutral (chance agreement is 0.5 for 2-scale, and 0.2 for

Dataset	Pairwise		Majority	
	2-Scale	5-Scale	2-Scale	5-Scale
Reviews	0.62	0.32	0.74	0.52
News	0.65	0.38	0.77	0.55
Novels	0.60	0.33	0.73	0.52
Summaries	0.61	0.33	0.73	0.53
Combined	0.62	0.35	0.74	0.53

Table 3: **Pairwise and majority inter-annotator agreement for instances with clear majority.** 2-Scale indicates when users are asked to indicate male or female, while 5-Scale indicates gender on a scale of 5.

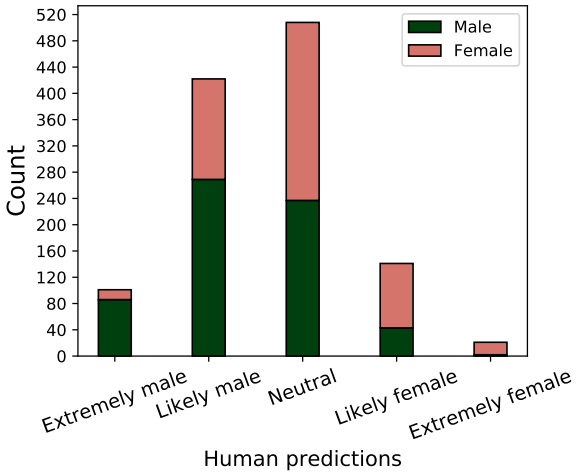

Figure 4: **Comparing Human Predictions to Truth.** x-axis represents the human prediction for mentions, while the green and pink bars represent counts of true male and female mentions respectively.

5-scale task). To analyze the kind of mistakes humans make, we show the true distribution of male and female mentions compared against human predictions in Figure 4. 42% of examples are predicted *'Neutral'* by humans showing that the task is pretty difficult for humans as they often find mention gender ambiguous. Further, *'Likely male'* and *'Likely female'* categories have around 30% wrong predictions as well. These high error rates explain low F1 of 0.52 for human annotations. *'Extremely male'* and *'Extremely female'* have the least error rates showing that humans are more precise when they are more confident about the predictions.

Does our model match humans? In order to compare human annotations against model predictions more concretely, we choose to use Kendall's τ_c statistic (Berry et al., 2009), because it allows us to compare two variables when their underlying scales have different numbers of values. Like correlation coeeficients, τ_c ranges from -1 (fully negative association) to +1 (fully positive association). τ_c between humans and our LSTM model predictions vary from 0.23 to 0.36 showing positive correlation. We also look at classifier probability distribution for human decisions shown in box and whisker plot in Figure 5, where x-axis is the classifier probability of the mention being female. The median value of classifier prediction for each category (shown in green line) shifts towards female prediction as we move from 'Extremely male' to 'Extremely female' category, corroborating the agreement between humans and model.

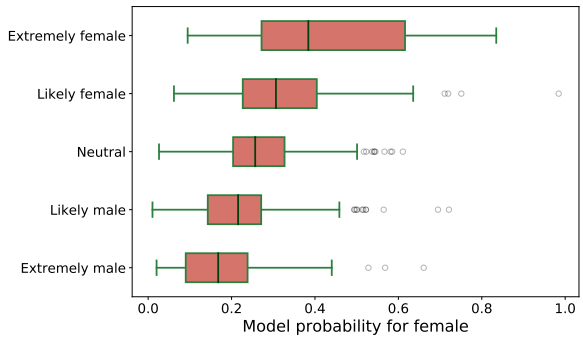

Figure 5: **Human and Model Predictions.** x-axis is the classifier probability where higher probability indicates female prediction. The points outside the range indicate outliers. Right shift of green line (representing median values) as we move from extremely male predictions to extremely female predictions corroborates the agreement between humans and classifiers.

6 Analysing Gendered Language

We first show aggregate word level and phrase level analysis, then show more complex and subtle sources of gendered language on sentence level.

Word-level Analysis Table 4 shows the top nouns and verbs extracted using bag-of-word classifier. We also train separate classifiers only on nouns, on adjectives, and on verbs in order to find out which are most informative for gender. Classifiers trained only on nouns performs best, indicating that nouns have most information. Top male-indicative nouns stem from typically male-dominated sports, while top female-indicative nouns are related to fashion and home industry.

Phrase-level Analysis Table 5 shows some of the top phrases for predicting males and females for different domains extracted using bag-of-ngrams models. We see phrases like *'clasped their hands'*, *'fashion show'*, *'jealous rage'*, *'was asked to'* for females, while *'clockwork orange'*, *'action hero'*,

Female-specific Nouns	Male-specific Nouns
Summaries: cherie, elisabet, crawlers, plastics, governess, cheerleader, prostitution, overdosing, bimbo, spinner	**Summaries:** quacker, platoon, tweety, shemp, cellmate, ham, nibbles, falstaff, pup, towel, mousehole, bullies
Novels: godmother, melvina, skirt, girlhood, lucile, womanly, eyebright, womanhood, shawl, dressmaker, demurely	**Novels:** disciples, yussuf, rifle, jr, pepe, cigar, colleague, followers, erasmus, judas, opponents
Reviews: comedienne, floriane, slut, adela, tch, topless, actress, tits, feminist, modeling, redhead, helen, vamp, bettie	**Reviews:** seagal, inventor, panther, opponent, sellers, ratso, comedian, lawman, yossi, creators, brutus, ted
News: gymnasts, dietitian, lpga, hingis, feminist, dowd, sorenstam, wie, receptionist, omnimedia, quilting, homemaker	**News:** spurs, astros, nicks, jets, sprewell, nets, vikings, clippers, lakers, holyfield, sonics, councilman, nba, bucs, pitches
Female-specific Verbs	**Male-specific Verbs**
Summaries: giggles, conceive, type, spurned, distorted, strokes, railing, rehearse, gag, disowned, plaguing, forgo	**Summaries:** commanding, barks, crack, credited, embezzled, executes, opposing, foils, relying, assassinate, engineered
Novels: sobbed, sew, blushed, wailed, pouted, scream, moaned, giggled, weeping, blushing, sob, shrieked, faltered	**Novels:** preached, elected, growled, states, yelled, roared, nominated, voted, grinned, slew, preach, fire, attack
Reviews: swims, bare, willed, raped, married, pouting, pleading, glows, kisses, liberated, seduces, fled, numbed	**Reviews:** direct, assuming, elected, defeat, casted, laid, mumbles, rule, directing, flicks, drinking, produce
News: fax, widowed, choreograph, raped, graduates, decorating, sobbing, majoring, giggling, married, cries, decorate	**News:** coach, pitching, batted, disarm, sacked, benched, fumbled, lightning, averaged, traded, sprained, vetoed

Table 4: Most important nouns and verbs for predicting male/female.

	News	**Reviews**	**Summaries**	**Novels**
Female gendered phrases	was pregnant	femme fetale	fashion show	suffrage association
	administrative assistant	nude scenes	jealous rage	clasped *their* hands
	their baby	strong willed	same way that	glass eye
	staff to vice	pop star	was asked to	corresponding secretary
	social worker	is hot	unborn child	dressing room
Male gendered phrases	defense secretary	action hero	construction worker	lieut col
	their locker	*their* screenplay	of the next	*their* pipe
	super tuesday	clockwork orange	school bully	*their* rifle
	offensive coordinator	court martial	Vietnam war	old fellow
	majority leader	nothing like the	make living	jimmy skunk

Table 5: **Most important phrases for predicting male/female.** *'their'* represents gender-neutral pronoun.

'construction worker' occur for males. Similarly, the term *'secretary'* occurs frequently with females, however phrases like *'defense secretary', 'treasury secretary'* are positive features for male mentions indicating male-domination of certain fields.

Sentence-level Analysis Our approach is the first to find gendered sentences from a huge corpora. We present examples of detected gendered language in Table 6; the first two columns show contexts for which a high level of gendering is estimated (most confident estimates), while the classifier has very low confidence for the examples in third column, indicating gender-neutral use of language. We see several interesting examples, e.g., male-gendered contexts from summaries show that society attributes roles like billionaire computer moguls and FBI-agents to males. The first female gendered example from novels depicts the way in which females are described and portrayed in fiction, which is in stark contrast to male descriptions.

Corpus-level Analysis Our approach allows us to automatically analyze large corpus of text, enabling high-level analysis of what documents are most, or least, gendered. Table 7 shows such an analysis, where the languages for movie reviews and summaries, genres for novels, and news desks for news are organized by their estimated genderedness. We see that children's history books are more gendered than their literature books. Books related to opera and one-act plays are among the least gendered ones, while those related to war, history, and philosophy use gendered language the most. Movie reviews for movies in Vietnamese, Turkish and Polish are among the most gendered, while Greek and Japanese are the least gendered. We see a similar pattern in movie summaries - summaries for movies filmed in Polish and Turkish are more gendered than for movies in Korean or Romanian. *Sports* is the most gendered category for news articles, while *Cultural*, *Leisure*, *Society*, and *Home* are among the least gendered ones. The table also contains some unexpected predictions, such as the low gendering of *Girls* and *Young women* novels.

	Female gendered	Male gendered	Gender neutral
News	– J.J.'s research brings *them* to find *Person*, a farmer who is slowly devouring a 747 to win the heart of lovely , editor of the local newspaper. – It's a film about *Person* directed by *Person*, and I play the soprano .	– USC is expected to announce that former New England patriots coach will be its next football coach. – It has fired up a torrid debate over President-elect 's $ 1.3 trillion tax cut proposal.	– *Person* had props to bolster *their* story, added. – In North Dakota, *Person* and - who are married to each other - were running against each other for foster county prosecutor.
Novels	– *Person* looked untidier than ever; wore a slatternly wrapper, and *their* hair was thrust unbrushed into its net. –"What is it?" asked *Person*, as folded and smoothed *their* best gown.	– If the collector will remember that, though is the present owner of *their* prints... – *Person* is not an orator; *person* is not a writer; is not a thinker.	– got up and went so – so unexpectedly. – The hand of the child admitted *them* to the chamber of death; the door closed, and stood motionless.
Reviews	– Lake's secretary, , is *Person*'s sweetheart . – Aeon played by the lovely in this adaptation, is dexterous as a line-dancer and deadly as a viper-snake.	– Herein, only old-time Broadway producer and *their* fey secretary *Person* maintain interest. – 's rolling in the sheets with *their* beautiful secretary Meredith.boris.	– 's hysterical and so are *their* backup singers. – is the only one worth seeing in this film , but *person* doesn't get to do much.
Summaries	– is raped by the estate owner, who then writes off *person*'s debt. – *Person*, an architect, is married to *their* sweetheart with two children.	– *Person* befriends a billionaire computer mogul and a cafe waitress.. – FBI-agent becomes an unwitting pawn of the white hand drug cartel.	– *Person* goes over to check on *them*, insisting *person* reopen the blinds, but denies doing it. – In order to stifle *their* theatrical aspirations, arranges a screen test.

Table 6: **Examples of gendered and non-gendered mention-context pairs from different domains.** *them/their* represents a gender-pronoun replaced with a gender neutral pronoun. *Person* represents mention. Note that in the first example for news (in female-gendered column), *J.J.* has not been identified as a mention. This is a weakness in the preprocessing.

Reviews Language	Summaries Language	Novels Genre	News News Desk
Vietnamese	Serbian	US Civil War	Sports
Turkish	Arabic	US Politics	Foreign
Polish	Thai	Mathematics	Financial
Cantonese	Czech	Military Science	Week in Review
Arabic	Polish	Evolution	National
Mandarin	Turkish	Dictionaries	Business
Korean	Khmer	Girls	Living
Latin	Korean	Young women	Travel
Portuguese	Sinhala	Sisters	Society
French	Hungarian	Family Life	Home
Greek	Punjabi	Marriage	Style
Japanese	Romanian	Italy	Performing Arts

Table 7: **High-level Analysis of Corpora.** Most and least gendered languages for movie reviews and summaries; genre for novels; news desk for news articles.

7 Discussion and Future Work

Sex vs Gender Since the current English language use is mostly limited to binary gender identities (both due to grammar and usage), we treat gender as a binary concept in this work. Inclusion of genderqueer and non-binary identities will require data annotated by humans with sufficient domain knowledge, which was out of scope for this work. We assume that mentions for which our labeling has associated the wrong 'gender' because of difference in sex/gender identities are sufficiently low in proportion that model is still able to learn relevant signals when trained on large corpora.

Facts vs Stereotypes In this work, we do not delineate between factual information (women get *pregnant*) and the intentional use of stereotypes (women are *sweethearts*). In some domains, such as news, ignoring this difference can be misleading, and exploring approaches that are able to better separate these different biases is important.

Extension to New Domains There remain a number of exciting avenues for future work. Although we analyze a variety of domains that differ from each other, our analysis focused on independently investigating each; it may be much more fruitful to compare and contrast the gendered language across multiple domains. When extending this work to other domains like Twitter, blogs, etc., the performance of the system can be affected by various factors like accuracy of NER system for the domain (e.g., it would be lower for tweets) and names to gender mapping (which can vary for different geographies and cultures).

8 Conclusions

We present a concrete implementation and evaluation of our gendered language detector. The main advantages of our pipeline and method are: (1) *Flexibility*, in application to different domains with minimal manual intervention, (2) *Mention-level* analysis, instead of article-level analysis in previous works, enabling more granular analysis, and (3) *Quantitative measure* of the extent of genderedness of context given a mention, allowing large-scale and detailed analyses and comparisons.

Our pipeline automatically extracts person mentions from a corpus, and by using an accurate gender predictor, trains a classifier to learn the ways in which language is gendered *for that corpus*. This automation provides multiple benefits; not only are there no humans in the loop to inject their biases about what is, and is not, gendered language, but further, collection of a large annotated corpus allows us to train sophisticated neural models that are able to capture semantic and syntactic constructions in the language. Evaluation suggests that our model is fairly accurate on this challenging task, and further, allows us to carry out analysis on multiple domains at varying levels of granularity, demonstrating potential applications of this work. The code to support such endeavours, and to reproduce the results, is available at https://ucinlp.github.io/GenderQuant.

Acknowledgements

We would like to thank Dheeru Dua, Matt Gardner, Robert L. Logan IV, and the reviewers for their feedback and suggestions. This work is supported in part by Allen Institute for Artificial Intelligence (AI2) and in part by NSF award #IIS-1756023.

References

Omar Ali, Ilias Flaounas, Tijl De Bie, Nick Mosdell, Justin Lewis, and Nello Cristianini. 2010. Automating news content analysis: An application to gender bias and readability. In *Workshop on Applications of Pattern Analysis*.

David Bamman, Brendan O'Connor, and Noah A Smith. 2013. Learning latent personas of film characters. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kenneth J Berry, Janis E Johnston, Sammy Zahran, and Paul W Mielke. 2009. Stuarts tau measure of effect size for ordinal variables: Some methodological considerations. *Behavior Research Methods*, 41(4).

Ben Blatt. 2017. *Nabokov's favorite word is mauve*. Simon & Schuster.

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 *Advances in Neural Information Processing Systems (NeurIPS)*.

Anne Boring, Kellie Ottoboni, and Philip B Stark. 2016. Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

John A Centra and Noreen B Gaubatz. 2000. Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, 71(1).

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*.

Eunsol Choi, Omer Levy, Yejin Choi, and Luke Zettlemoyer. 2018. Ultra-fine entity typing. In *Annual Meeting of the Association for Computational Linguistic (ACL)*.

Cody Cottier. 2018. From mouth to mind: How language governs our perceptions of gender.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women.

Susan Tyler Eastman and Andrew C Billings. 2000. Sportscasting and sports reporting: The power of gender bias. *Journal of Sport and Social Issues*, 24(2).

Andrea Eidinger. 2017. She's hot: Female sessional instructors, gender bias, and student evaluations.

Liye Fu, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Tie-breaker: Using language models to quantify gender bias in sports journalism. In *IJCAI Workshop on Natural Language Processing meets Journalism*.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16).

Abdullah Gharbavi and Seyyed Ahmad Mousavi. 2012. The application of functional linguistics in exposing gender bias in iranian high school english textbooks. *English Language and Literature Studies*, 2(1).

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4).

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Katherine N Kinnick. 1998. Gender bias in newspaper profiles of 1996 olympic athletes: A content analysis of five major dailies. *Women's Studies in Communication*, 21(2).

Corina Koolen and Andreas van Cranenburgh. 2017. These are not the stereotypes you are looking for: Bias and fairness in authorial gender attribution. In *ACL Workshop on Ethics in Natural Language Processing*.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4).

María E Len-Ríos, Shelly Rodgers, Esther Thorson, and Doyle Yoon. 2005. Representation of women in news and photos: Comparing content to perceptions. *Journal of Communication*, 55(1).

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT)*.

Monica Macaulay and Colleen Brice. 1997. Don't touch my projectile: Gender bias and stereotyping in syntactic examples. *Language*.

Lillian MacNell, Adam Driscoll, and Andrea N Hunt. 2015. What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4).

Michela Menegatti and Monica Rubini. 2017. Gender bias and sexism in language. *Oxford Research Encyclopedia of Communication*.

Janice Moulton, George M Robinson, and Cherin Elias. 1978. Sex bias in language use: Neutral pronouns that aren't. *American Psychologist*, 33(11).

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Robert Munro, Steven Bethard, Victor Kuperman, Vicky Tzuyin Lai, Robin Melnick, Christopher Potts, Tyler Schnoebelen, and Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, AKBC-WEKEX.

Marcus Otlowski. 2003. Ethnic diversity and gender bias in efl textbooks. *Asian EFL Journal*, 5(2).

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Jemma Prior. 2017. Teachers, what is gendered language?

Anil Ramakrishna, Victor R Martínez, Nikolaos Malandrakis, Karan Singla, and Shrikanth Narayanan. 2017. Linguistic analysis of differences in portrayal of movie characters. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *ACL Workshop on Ethics in Natural Language Processing*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Maarten Sap, Marcella Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. 2017. Connotation frames of power and agency in modern films. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ruchita Sarawgi, Kailash Gajulapalli, and Yejin Choi. 2011. Gender attribution: tracing stylometric evidence beyond topic and genre. In *Computational Natural Language Learning (CoNLL)*.

Toni Schmader, Jessica Whitehead, and Vicki H Wysocki. 2007. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, 57(7-8).

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Workshop on Computational Linguistics for Literature*.

Kevin B Smith. 1997. When all's fair: Signs of parity in media coverage of female candidates. *Political Communication*, 14(1).

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1).

Jane G. Stout and Nilanjana Dasgupta. 2011. When he doesn't mean you: Gender-exclusive language as ostracism.

Rachael Tatman. 2017. Gender and dialect bias in youtube's automatic captions. In *ACL Workshop on Ethics in Natural Language Processing*.

Mike Thelwall. 2018. Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42(3).

Frances Trix and Carolyn Psenka. 2003. Exploring the color of glass: Letters of recommendation for female and male medical faculty. *Discourse & Society*, 14(2).

Susan Tyler Eastman, Andrew C. Billings. 2001. Biased voices of sports: Racial and gender stereotyping in college basketball announcing. *Howard Journal of Communication*, 12(4).

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the gap: A balanced corpus of gendered ambiguous pronouns. In *Transactions of the Association for Computational Linguistics (TACL)*.

Tae Yano, Philip Resnik, and Noah A Smith. 2010. Shedding (a thousand points of) light on biased language. In *NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias smplification using corpus-level constraints. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

A Set of Gender-Specific Words

List from Bolukbasi et al. (2016) without pronouns:
man, women, men, woman, spokesman, wife, son, mother, father, chairman, daughter, husband, guy, girls, girl, boy, boys, brother, spokeswoman, female, sister, male, herself, brothers, dad, actress, mom, sons, girlfriend, daughters, lady, boyfriend, sisters, mothers, king, businessman, grandmother, grandfather, deer, ladies, uncle, males, congressman, grandson, bull, queen, businessmen, wives, widow, nephew, bride, females, aunt, prostate cancer, lesbian, chairwoman, fathers, moms, maiden, granddaughter, younger brother, lads, lion, gentleman, fraternity, bachelor, niece, bulls, husbands, prince, colt, salesman, dude, beard, filly, princess, lesbians, councilman, actresses, gentlemen, stepfather, monks, ex girlfriend, lad, sperm, testosterone, nephews, maid, daddy, mare, fiance, fiancee, kings, dads, waitress, maternal, heroine, nieces, girlfriends, sir, stud, mistress, lions, estranged wife, womb, grandma, maternity, estrogen, ex boyfriend, widows, gelding, diva, teenage girls, nuns, czar, ovarian cancer, countrymen, teenage girl, penis, bloke, nun, brides, housewife, spokesmen, suitors, menopause, monastery, motherhood, brethren, stepmother, prostate, hostess, twin brother, schoolboy, brotherhood, fillies, stepson, congresswoman, uncles, witch, monk, viagra, paternity, suitor, sorority, macho, businesswoman, eldest son, gal, statesman, schoolgirl, fathered, goddess, hubby, stepdaughter, blokes, dudes, strongman, uterus, grandsons, studs, mama, godfather, hens, hen, mommy, estranged husband, elder brother, boyhood, baritone, grandmothers, grandpa, boyfriends, feminism, countryman, stallion, heiress, queens, witches, aunts, semen, fella, granddaughters, chap, widower, salesmen, convent, vagina, beau, beards, handyman, twin sister, maids, gals, housewives, horsemen, obstetrics, fatherhood, councilwoman, princes, matriarch, colts, ma, fraternities, pa, fellas, councilmen, dowry, barbershop, fraternal, ballerina.