Are Online Reviews of Physicians Biased Against Female Providers?

Avijit Thawani THAWANI@USC.EDU *University of Southern California*
Los Angeles, CA, USA

Michael J. Paul MICHAEL.J.PAUL@COLORADO.EDU *University of Colorado, Boulder*
Boulder, CO, USA

Urmimala Sarkar URMIMALA.SARKAR@UCSF.EDU *University of California, San Fransisco*
San Francisco, CA, USA

Byron C. Wallace BYRON@CCS.NEU.EDU *Northeastern University*
Boston, MA, USA

Abstract

Patients increasingly seek out information regarding their healthcare online. Online reviews of caregivers in particular may influence from whom patients seek treatment. Are these sources biased against female providers? To address this question we analyze a new dataset of online patient reviews of male and female healthcare providers with respect to numerical ratings and language use. We perform both regression and (data-driven) qualitative analyses of language via neural embedding models induced over review texts. In both cases we account for provider specialty. To do so while learning embeddings, we explicitly induce specialty, sex, and rating embeddings from review meta-data via a 'matched-sampling' training regime.

We find that females consistently receive less favorable numerical ratings overall, even after adjusting for specialty. To analyze language use in reviews of male versus female providers, we induce neural embeddings (distributed representations) of gender and qualitatively characterize the 'distributional semantics' that this induces. We observe differences in language use, e.g., analysis of average vector similarities over repeated runs reveal that many of the words closest to the coordinates in embedding space associated with positive sentiment and female providers describe interpersonal characteristics (*sweet*, *considerate*, *caring*, *personable*, *compassionate*): such descriptors do not seem as similar to the point corresponding to positive sentiment regarding male providers. To facilitate research in this direction we publicly release data, embeddings, and all code (including Jupyter notebooks) to reproduce our analyses and further explore the data: https://github.com/avi-jit/RateMDs.

1. Introduction

Individuals are increasingly turning to the web to gather information relevant to their healthcare. Online reviews of physicians are important examples of this: a relatively recent survey (Fox and Duggan, 2013) found that 72% of internet users have looked online for health information in the past year. One in five of these users have looked for reviews of either particular treatments or doctors.

Patient-generated reviews are also important as a data source because they provide a direct, unmediated window into the patient's experience. Further, these reviews may influence other individuals' opinions of (potential) physicians, in turn affecting patient care. Indeed, Li *et al.*(Li et al., 2015) concluded via a randomized trial that exposure to negative reviews "led to a reduced willingness to use the physician's services." Much of the previous work on examining online reviews of physicians has been qualitative in nature (López et al., 2012; Gao et al., 2012; Kilaru et al., 2016). Extending upon some quantitative prior work (Wallace et al., 2014; Paul et al., 2013), we adopt a large-scale, data-driven approach in this study.

Our focus in this paper concerns investigating differences in online reviews of male versus female physicians across several specialties, with respect to both patient satisfaction (ratings) and language use. We aim to complement the relatively robust body of evidence that strongly suggests female physicians "don't get the credit they deserve"(Roter and Hall, 2015). For instance, Hall et al. (2011) performed a meta-analysis of studies looking at patient satisfaction and concluded that female physicians "are not evaluated as highly by their patients, relative to male physicians, as one would expect based on their practice style and patients' values." We assess whether the same holds in online reviews.

In initial inspection of reviews data we observed that specialties serve as potential confounders: neither genders nor ratings are equally represented across the specialities. We thus control for this confounding variable by adjusting both our regression analyses and our models for learning representations (embeddings) of physician gender, based on the text in reviews.

Our contributions are as follows:

1. We create and share a new, large dataset comprising reviews scraped from RateMDs.com of medical practitioners, along with their meta-data such as provider gender, specialty, and ratings, but not the physician names[1]. This may facilitate further analyses.
2. We quantify the differences in ratings assigned to doctors due to gender and specialty within these ratings, finding that female providers consistently receive less favorable reviews than their male counterparts, even accounting for specialty.
3. We use NLP models to characterize language use in online reviews of female versus male doctors. For this we use both a log-linear model relating words, gender, and sentiment (Paul et al., 2016), and – more qualitatively – the 'distributional semantics' induced by neural embedding representations of the same.

Ideally we would conduct a non-binary, more gender-inclusive study, but here we use the simple dichotomized gender retrieved from the corresponding RateMDs.com review.

1.1. Technical Significance

We adopt modern natural language processing (NLP) technologies for our qualitative lexical analysis of reviews, using both count-based (log-linear) models of words, and via 'distributional semantics'. We extend *doc2vec* (Mikolov et al., 2013) to learn embeddings for specialties, genders, and ratings from review texts (in the same space as word embeddings).

1. Despite our intention to anonymize the dataset, some mentions of physician names may have crept into the review texts.

To ensure that 'male' and 'female' embeddings reflect gender but not specialty (a potential confounder), we propose a simple *matched sampling* training regime to learn these representations.

1.2. Clinical Significance

Patients are increasingly seeking health information online (Fox and Duggan, 2013). Online reviews of physicians constitute a source of potential transparency (Lee, 2017), and may afford insight into aspects of patient care (Agarwal et al., 2018). It has been shown that exposure to reviews of caregivers may directly impact patient decision-making with respect to source of care (Li et al., 2015). It follows that systematic biases in popular online sources of physician reviews may have large aggregate effects on population health. Understanding and characterizing such biases is therefore important.

2. Materials and Methods

2.1. Dataset

We use a subset of a corpus of reviews downloaded from RateMDs.com, a website of healthcare provider reviews written by patients. We crawled the website for reviews of doctors across all specialties and for both 'Male' and 'Female' physicians (this information is explicitly stored by RateMDs.com).

We sampled reviews for our dataset as follows. For each specialty (57) and each gender (2), we selected up to 100 doctors from the first 10 search result pages retrieved on RateMDs.com. Subsequently we collected up to 10 reviews for each such doctor identified. At most this would have yielded a list of 11400 doctors and 114000 reviews. However, owing to an uneven distribution of reviews and doctors found per specialty (some are sparse), this process yielded a dataset of 6495 unique doctors (3713 male, 2782 female) and their corresponding 48567 reviews (29873 for male physicians and 18694 for female). Table 1 reports statistics and examples of reviews.

A subset of the data just described , after cleaning, lower-casing, and correcting for misspellings, is used in the experiments that follow. We create a balanced corpus of reviews spanning the 17 specialties grouped according to Table 3. This grouping was performed by one of the authors (an MD). One thousand reviews from each of these 17 specialties were selected and subsequently analyzed. Furthermore, for our ratings and lexical analyses we used $9,000$ additional reviews in an 'other' category which served as a baseline.

2.2. Rating Analysis

Each review includes numerical scores assigned by the author with respect to four aspects of care: *knowledgeability*, *helpfulness*, *punctuality* and *staff*. These are provided on a five-point Likert scale, where higher implies greater satisfaction. As a simple first analysis, we quantify the correlation between the gender and specialty of physicians and these ratings.

Indexing reviews by i and indexing a specific target (e.g., *knowledgeability*) by t with mean rating y_i^t, we perform regressions of the following form (one per target):

$$y_i^t = \boldsymbol{\beta}_0^t + \boldsymbol{\beta}_{female}^t \cdot g_i + \boldsymbol{\beta}_{specialty}^t \cdot \mathbf{S}_i \tag{1}$$

Table 1: Description of a single entry (a single review) in our dataset along with relevant statistics of the attributes

	Description	Example	Statistics
Doctor Name	Name of physician	'FIRSTNAME-LASTNAME.html'	6945 unique docs
Specialty	Specialty of physician	'Gynecologist'	57 specialties
Gender	Gender of physician	'f'	2 genders
Text	Review text	"She is a life-saver! From my... "	48567 reviews
Scores	[S,P,H,K][1]	[4,5,3,5]	-
Upvotes	# times upvoted	0	-
Date	Review publish date	December 6, 2017	-

[1] Staff, Punctuality, Helpfulness, Knowledgeability

g_i is an indicator function set to 1 if physician i is female (as per RateMDs.com categorization) and 0 otherwise. Likewise, $\mathbf{S}_i$ encodes the doctor specialty via a one-hot encoding over the specialties (thus at most one column per doctor will be 1). For instance, $\mathbf{S}_i$ may be $< 1, 0, 0, 0... >$ for the specialty *surgeon* and $< 0, 1, 0, 0... >$ for *gynecologist*. If the categorical variable specialty can take k possible values, then $\mathbf{S}_i$ must be a $k-1$ sized encoding, wherein the last value of 'specialty' is represented by all zeros $< 0, 0, 0, 0... >$. We code this such that '*others*' is the baseline category (i.e., specialties which belong to none of the Table 3 groups); correlations with ratings for reviews of physicians in this catch-all category are thus captured by the model intercept.

Here $\boldsymbol{\beta}^t_{female}$ is a coefficient capturing the correlation between being female and the review scores one receives for target t, and $\boldsymbol{\beta}^t_{specialty_j}$ captures the correlation between the j^{th} specialty and these scores. $\boldsymbol{\beta}^t_0$ is the background (intercept) capturing overall average rating estimates.

2.3. Lexical Analysis

We analyze variations in review texts as a function of the gender of the physician being reviewed. Similar to the open vocabulary approach of Schwartz et al. (2013) for examining demographic differences of social media users, we seek to identify words in the corpus that are most strongly associated with reviews of male and female physicians. For this we consider a few modeling approaches, discussed below.

2.3.1. Lexical Regression

We use a log-linear regression model, following the approach used in an earlier work (Paul et al., 2016). We model the (log) frequency of a word being used in a review of a physician of a given gender as a function of: (i) Background word frequency; (ii) A general gender intercept (adjusting for differences in the overall volume of text for doctors of that gender); (iii) A general specialty intercept (adjusting for differences in overall volume of text for doctors of different specialties); (iv) Gender-specific word coefficients; (v) Specialty-specific

word coefficients, and, finally; (vi) Specialty-gender interactions. With y_{gsw} denoting the number of doctors of gender g in specialty s for which word w was used in a review, our model is defined as:

$$\log y_{gsw} = \beta_0 + \beta_w + \beta_g + \beta_s + \beta_{gw} + \beta_{sw} + \beta_{gs} \tag{2}$$

The word counts y_{gsw} from reviews are counted as indicator values by doctor, such that the word count for a particular doctor is at most 1. This is done so that each doctor's reviews contribute roughly evenly to the word counts for their gender. Otherwise, doctors with many reviews might skew the results. We take the log of y_{gsw} so that the linear coefficients of the model represent relative, multiplicative differences rather than absolute differences in frequency. We fix the gender-independent word intercepts β_w to the log-frequency of word w over the entire corpus. All other coefficients are learned by fitting a standard least squares model to the data.

To provide insight into what this lexical regression model learns, consider each individual variable, its significance and the highest values from the actual training. For instance, $y_{f,gyn,fertility}$ is the log of the number of female gynecologists (or reproductive endocrinologists) in whose reviews the word *fertility* occurs. During training the model tries to learn to predict a word's frequency in this setting, i.e., its frequency in terms of number of doctors in whose reviews it appears (of the given specialty and given gender). β_0 represents a base estimate for any word's doctor-wise frequency for a given specialty and gender. This variable has a low value of -3.27 as expected owing to the sparse distribution of words across specialties (the word *bone* rarely occurs with *gynecologists* or *psychologists*). β_g is found to be -0.492, possibly accommodating for the fewer reviews associated with female doctors. β_w is trained to find a base estimate for that word alone.

The β_{gw} term accounts for word frequencies specifically for the female gender. The words with highest value of β_{gw} are *podiatrist*, *herself*, *her*, and *she* and those with the lowest are *he's*, *him*, *his*, and *man*. Likewise, the specialties most associated with females (in terms of number of doctors) is learned in the parameter β_{gs}. β_{sw} is similar to β_{gw} except it is for specific specialties, not genders. Elaborate results can be seen in Table 5, and for an even larger list, please refer to Appendix C.

2.3.2. Embeddings

As a complementary analysis, we consider the distributional semantics Firth (1961) over words, genders, and specialties induced by neural models. For this we extend doc2vec, a standard method for learning document embeddings that itself is an extension of the original word2vec family of models (Mikolov et al., 2013). Before presenting our extension, we first review two word2vec and doc2vec variants.

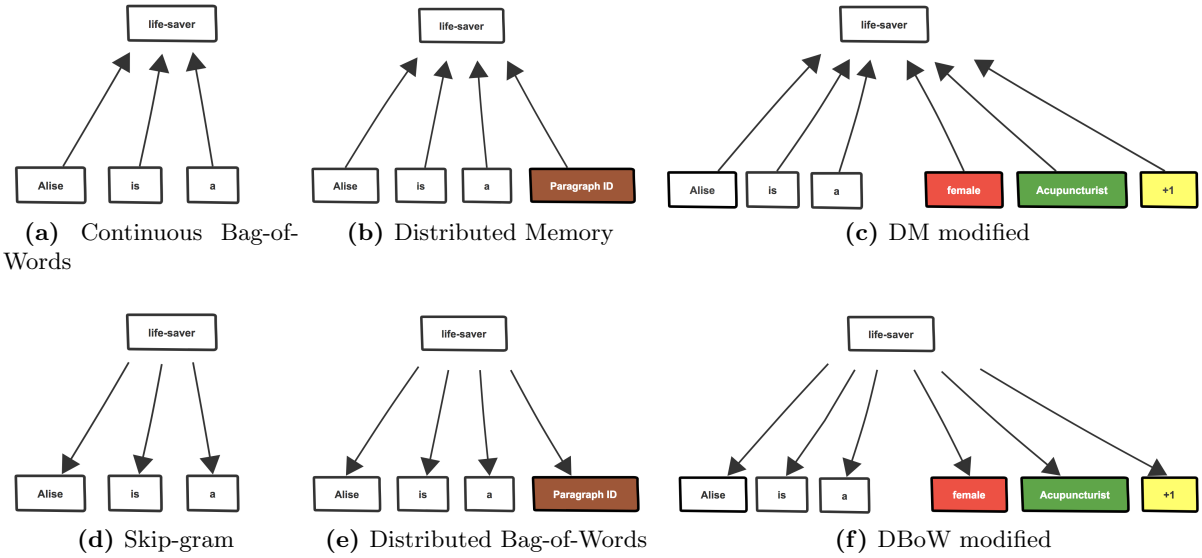
- **Word2vec** refers to a few related models for inducing (comparatively) low-dimensional representations of words using a shallow neural network. Specifically, embeddings of words are learned in such a way that they capture the context in which they appear. This is an instantiation of *distributional semantics* (Firth, 1961), which is the notion that meaning of a word is defined by the company it keeps (i.e., collocations). There are two popular variants of word2vec, which differ in how they model the relationship between a target word and its neighbors.

The **Skip-gram** variant is based on the task of trying to predict the context (the words within a small fixed window size) of a particular word (Figure 1(*d*)). Words with similar themes (at least in terms of contexts) should realize nearby representations in a skip-gram model. For instance, words like *tooth*, *crown* and *canal* will tend to be associated with vectors that are near each other in the embedding space because they will all tend to appear in the context of *dentistry*. By contrast, the **Continuous Bag-of-Words** (CBoW) version defines the objective of predicting the center word given nearby words/context (Figure 1(*a*)). Thus, words that can be replaced with one another will tend to be nearby each other in the embedding space induced by a CBoW model. For example, words like *psychiatrist*, *orthodontist* and *chiropractor* will be close to one another because any of these could fill the following blank: "He/She is the best _____ in this city."

- **Doc2vec** associates paragraphs or documents with their own vectors (effectively these are treated as special tokens or words), with the context being the entire text, i.e., paragraph or document. The learned embedding is thus unique to a given document. There are two common variants for doc2vec as well, corresponding to the word2vec models just described. **Distributed Bag of Words (DBoW)** is analogous to the skip-gram word2vec variant. Document vectors are obtained by training a neural network on the task of predicting a probability distribution of words in a document given a randomly-sampled word from the document (Figure 1(*e*)). And **Distributed Memory (DM)** is the doc2vec model analog to Continuous-Bag-of-Words word2vec. Document vectors are obtained by training a neural network on the task of inferring a center word from context words and a context document (Figure 1(*b*)). A paragraph (document) serves as a context for all words that it contains, and a word in a document can have that document as a context.

Our model is a simple extension of the doc2vec model that aims to learn embeddings not for documents or reviews, but for specialties, genders, and ratings. Because these embeddings will occupy the same space as word vectors, this affords inspection of words nearby the learned gender embeddings, providing insight into language use via distributional semantics. Operationally, the context that was provided by the document embedding in doc2vec is now provided by a specialty vector, a gender vector, and a rating vector, for each review. To distinguish these from word vectors, we refer to them as 'class embeddings'. More specifically, we introduce: two gender embeddings; three rating embeddings, for positive (*1*), negative (*-1*) and neutral (*0*) reviews; and one embedding for each specialty. Given four dimensions of rating (*Staff*, *Punctuality* and so on) we average out the rating given by a reviewer in a single review and create three bins: [1,3.5) is *-1*, [3.5,4) is *0*, and [4,5] is *1*. These bins were so decided to ensure a somewhat uniform distribution of reviews in each category.

Next, out of the two broad embedding approaches: skipgram/DBoW/modified-DBoW and CBoW/DM/modified-DM, we must settle on one. Before final analysis, we learn embeddings for all 57 specialties. As a comparison, the top 10 words most similar to the specialty embedding *dentist* (along with their similarity values), for both *modified-DBoW* model (Figure 1(*f*)) and the *modified-DM* model (Figure 1(*c*)) are given in Table 2. Note that the specialty *dentist* is ultimately excluded from our final subset of data to be analyzed



(a) Continuous Bag-of-Words **(b)** Distributed Memory **(c)** DM modified

(d) Skip-gram **(e)** Distributed Bag-of-Words **(f)** DBoW modified

Figure 1: Depictions of what is being learned in different embeddings. Arrows point towards the variable being predicted and away from the input variables in each learning step.

(as suggested by its absence from Table 3). Both models capture aspects of meaning that are dissimilar and each would undoubtedly produce some interesting results. That being said, in our experiments we found that the distributed bag of words model is more suited to our needs for us than the distributed memory model. This is because the DBoW model groups co-occurring words together while the DM model groups together words that can replace each other given the context.

We use a matched-sampling approach during training to estimate embeddings. For each training instance, i.e. review, we first draw a specialty at uniform random. Suppose *surgeon* is selected, as an example. We then draw two reviews: one for a male surgeon and another for a female surgeon. Both of these reviews are selected at IID random, with replacement (the same review may be matched multiple times) and added to the training batch. This matched sampling procedure ensures an equal number of reviews for males and females, as well as a uniform distribution across specialties. The aim is to uncover embeddings that correlate to males and females, accounting for their disparate prevalence across specialties.

We used the Gensim version 0.12.4 implementation of doc2vec for our experiments (Řehůřek and Sojka, 2010). As an additional detail, we pretrained word embeddings for 100 epochs first, before introducing/learning class embeddings. Hyperparameter tuning led us to use dm-concat = True, negative- sampling = 7, sample = 0 (no subsampling of highly frequent words), window size = 20, no hierarchical sampling, and 10000 draws (described in the previous paragraph) to train on-line over 20000 reviews. In order to accommodate for variance, we repeat the entire learning process 20 times from scratch and report average values, except in Figure 2 which is simply a t-SNE visualization based on one of these 20 learned embeddings.

Table 2: Words closest to the embedding of the specialty 'dentist' for dbow and dm models, along with their cosine similarity scores

DBoW		**DM**	
word	*score*	*word*	*score*
dental	0.62	optometrist	0.59
periodontist	0.53	neurologist	0.59
filling	0.50	chiropractor	0.58
endodontist	0.49	orthodontist	0.58
hygienist	0.49	periodontist	0.57
canal	0.49	allergist	0.56
tooth	0.49	psychologist	0.55
dentistry	0.47	podiatrist	0.54
crown	0.46	gastroenterologist	0.53
orthodontist	0.45	doctor	0.53

Table 3: Grouped specialties (for 17 most frequent specialties)

Group	**Constituent Specialties**
int	Internist-Geriatrician
gastro	Gastroenterologist
endo	Endocrinologist
surg	Surgeon-General
bone	Pain-Management, Orthopedics
skin	Dermatologist, Cosmetic Plastic-Surgeon
gyn	Gynecologist, Reproductive Endocrinologist
brain	Neurologist, Neurosurgeon
eye	Ophthalmologist
family	Family-general physician
ent	Ear-Nose-Throat-ENT
psych	Psychiatrist
pedia	Pediatrician

3. Results

3.1. Rating Results

Parameter estimates from rating regressions (we trained four regression models independently, one for each rating scale: *Staff*, *Helpfulness*, *Punctuality*, and *Knowledgeability*) are presented in Table 4. The female coefficient is significantly negative for all aspects. That is, reviews of male physicians are more reliably favorable than those of female doctors, even after accounting for specialty.

3.2. Lexical Results

In Table 5 we present a qualitative analysis of language used. Broadly, there are three questions we aim to answer:

1. **Which words are most associated with male and female doctors, respectively?**
   The *Embeddings* column lists words with learned embeddings closest to the gender class embeddings. The *Lexical Regression* column lists the words with highest (for female) and lowest (for male) values of β_{gw}.

2. **Which words are most associated with doctors of specific specialties?**
   The first column lists words whose embeddings are closest to the class embeddings for corresponding specialties. The second column lists the words with highest values of β_{sw} where s is the required specialty.

3. **Which specialties are most associated with male and female doctors, respectively?**
   The first column lists specialties whose class embeddings are closest to the female (and male) embeddings. The second column lists words with highest (for female) and lowest (for male) values of β_{gs}.

One useful property of word embeddings is that they afford algebraic manipulation of words via their vector representations, which can provide insights into relationships as captured in the distributional semantics. Here we exploit this property to infer adjectives that are frequently used for males but not as often for females (and vice versa), ignoring stopwords. We summarize the results of this exercise in Table 6, which reports the nearest words to the vector corresponding to positive reviews (**1**), closest to the vector induced by adding the female embedding to this and subtracting the male vector. While the female vector seems more associated with words like *sweet*, *considerate*, and *personable*, the male vector counts among its neighbours words such as *professional*, *knowledgeable*, and *helpful*.

Lastly, Figure 2 is a t-SNE plot as a summary of the embeddings learned. t-SNE (Maaten and Hinton, 2008) is a popular tool to visualize vectors of multiple dimensions by mapping them into 2D or 3D point coordinates. We see how the *-1* embedding is closest to negative words like *rude* and *worst*, whereas *1* embedding is close to *professional* and *caring*. *However* is the closest word to the *0* embedding, an in-between rating wherein the target ratings on average fell within this range. One can imagine the associated review texts to look like "Staff not so great, *however* the doctor is very good."

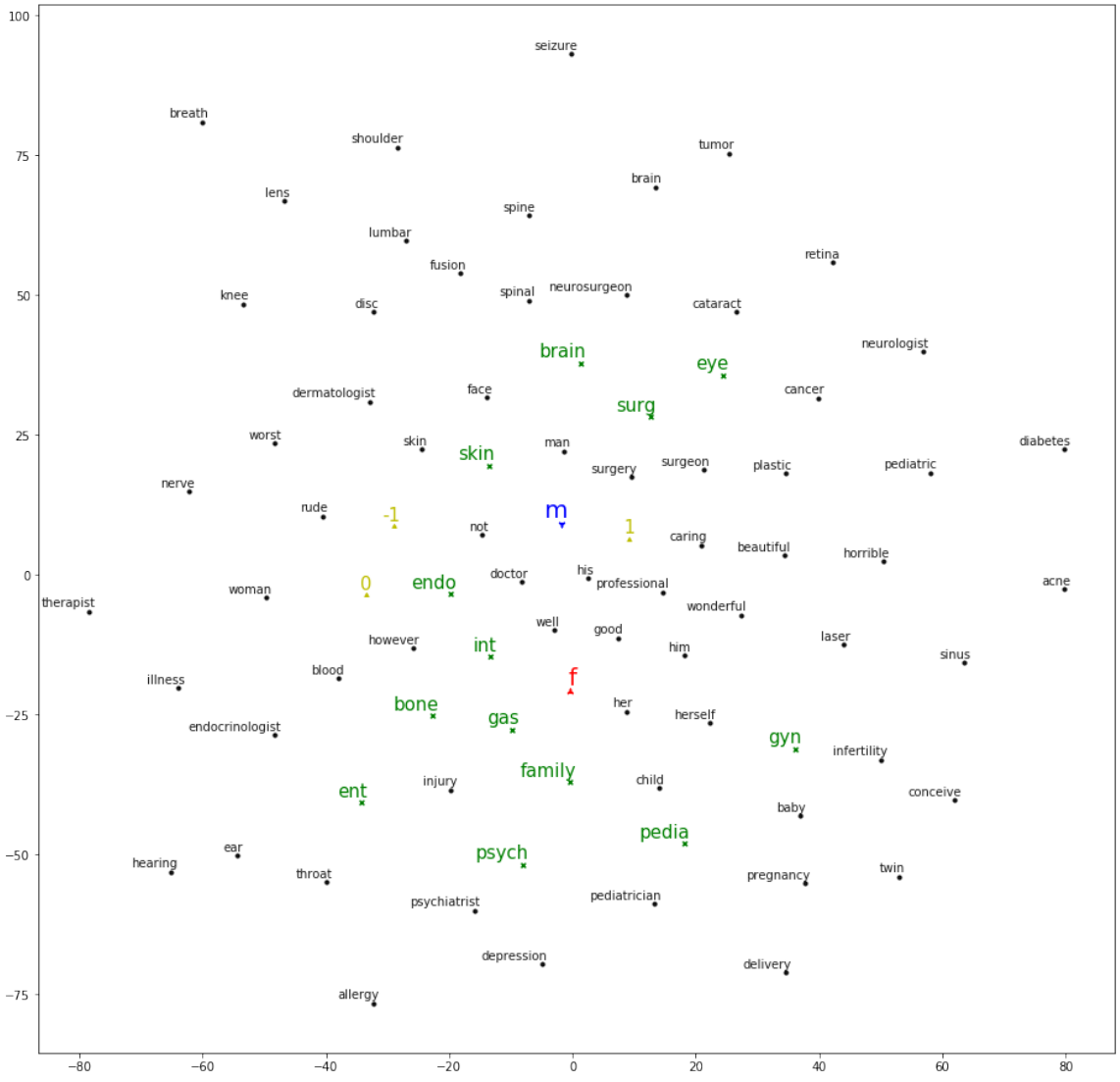


Figure 2: t-SNE plot for embeddings, adjusting for specialties. Points in green are class embeddings for specialties. For clarity, we show only top 10 words closest to the class embeddings (in black).

Table 4: Regression analysis of ratings. The female coefficients are bolded. SE denotes Standard Error. The extremely low p-values for β_{female} shows that the parameter is a useful predictor of ratings.

	Coeff	**SE**	**P>|t|**	**95.0% CI**		**Coeff**	**SE**	**P>|t|**	**95.0% CI**
		Staff					*Helpfulness*		
β_{skin}	0.2535	0.027	0.000	0.201 , 0.306	β_{skin}	0.2268	0.030	0.000	0.168 , 0.286
β_{gyn}	0.1368	0.027	0.000	0.084 , 0.190	β_{gyn}	0.1758	0.030	0.000	0.117 , 0.235
β_{bone}	0.1426	0.028	0.000	0.088 , 0.198	β_{bone}	0.1123	0.031	0.000	0.051 , 0.174
β_{brain}	-0.2257	0.031	0.000	-0.287 , -0.164	β_{brain}	-0.2676	0.035	0.000	-0.336 , -0.199
β_{family}	0.1191	0.036	0.001	0.048 , 0.190	β_{family}	0.2337	0.041	0.000	0.154 , 0.314
β_{eye}	0.2314	0.037	0.000	0.159 , 0.304	β_{eye}	0.2071	0.041	0.000	0.126 , 0.288
β_{psych}	-0.0196	0.038	0.609	-0.095 , 0.056	β_{psych}	-0.0667	0.043	0.120	-0.151 , 0.017
β_{pedia}	-0.0010	0.038	0.979	-0.076 , 0.074	β_{pedia}	0.1254	0.043	0.004	0.041 , 0.210
β_{ent}	0.0842	0.039	0.030	0.008 , 0.160	β_{ent}	0.0813	0.043	0.061	-0.004 , 0.166
β_{int}	-0.0370	0.039	0.345	-0.114 , 0.040	β_{int}	0.0926	0.044	0.035	0.007 , 0.178
β_{surg}	0.1177	0.040	0.004	0.039 , 0.197	β_{surg}	0.0414	0.045	0.359	-0.047 , 0.130
β_{endo}	-0.6135	0.041	0.000	-0.694 , -0.533	β_{endo}	-0.6607	0.046	0.000	-0.750 , -0.571
β_{gas}	-0.0763	0.041	0.064	-0.157 , 0.005	β_{gas}	-0.1345	0.046	0.004	-0.225 , -0.044
$\boldsymbol{\beta_{\text{female}}}$	-0.2953	0.020	0.000	-0.334 , -0.256	$\boldsymbol{\beta_{\text{female}}}$	-0.3436	0.022	0.000	-0.387 , -0.300
β_0	4.6389	0.012	0.000	4.616 , 4.662	β_0	4.6512	0.013	0.000	4.625 , 4.677
		Punctuality					*Knowledgeability*		
β_{skin}	0.2593	0.028	0.000	0.205 , 0.314	β_{skin}	0.2301	0.022	0.000	0.186 , 0.274
β_{gyn}	0.1634	0.028	0.000	0.109 , 0.218	β_{gyn}	0.1767	0.022	0.000	0.134 , 0.220
β_{bone}	0.1221	0.029	0.000	0.065 , 0.179	β_{bone}	-0.2026	0.022	0.000	-0.246 , -0.160
β_{brain}	-0.2336	0.032	0.000	-0.297 , -0.171	β_{brain}	0.1177	0.022	0.000	0.075 , 0.161
β_{family}	0.1868	0.037	0.000	0.113 , 0.260	β_{family}	0.0769	0.030	0.010	0.018 , 0.136
β_{eye}	0.1001	0.038	0.009	0.025 , 0.175	β_{eye}	-0.5507	0.030	0.000	-0.610 , -0.492
β_{psych}	-0.0143	0.039	0.717	-0.092 , 0.063	β_{psych}	0.0064	0.030	0.832	-0.052 , 0.065
β_{pedia}	-0.0155	0.040	0.696	-0.093 , 0.062	β_{pedia}	0.2635	0.030	0.000	0.204 , 0.323
β_{ent}	0.0785	0.040	0.049	0.000 , 0.157	β_{ent}	0.0373	0.030	0.220	-0.022 , 0.097
β_{int}	0.0793	0.040	0.049	0.000 , 0.158	β_{int}	0.2085	0.030	0.000	0.150 , 0.267
β_{surg}	0.0530	0.042	0.202	-0.028 , 0.134	β_{surg}	0.0656	0.030	0.029	0.007 , 0.124
β_{endo}	-0.5842	0.042	0.000	-0.667 , -0.502	β_{endo}	-0.2317	0.030	0.000	-0.291 , -0.173
β_{gas}	-0.0946	0.042	0.026	-0.178 , -0.011	β_{gas}	-0.0052	0.030	0.863	-0.064 , 0.054
$\boldsymbol{\beta_{\text{female}}}$	-0.3134	0.020	0.000	-0.353 , -0.273	$\boldsymbol{\beta_{\text{female}}}$	-0.3309	0.016	0.000	-0.362 , -0.300
β_0	4.5997	0.012	0.000	4.576 , 4.624	β_0	4.6895	0.009	0.000	4.671 , 4.708

In Appendix B, we show another t-SNE plot visualizing only the adjectives in our learned vocabulary, filtering out those with fewer than 1000 occurrences.

4. Discussion and Limitations

We have presented an exploratory analysis of online reviews written about male versus female caretakers. For this we have created a new corpus, which we make publicly available to facilitate further research. We performed regressions over ratings and lexical analyses using both regression and neural embedding methods. For the latter we used a variant of the doc2vec objective in the spirit of related recent efforts for inducing embeddings for

Table 5: Top associations from two methods of lexical analysis. Embeddings results are aggregated over 20 independent runs to reduce variance.

	Embeddings	**Lexical Regression**
Most associated words (with respective genders) - Stopwords removed		
f	doctor, time, office, patient, care, well, caring, see, staff	podiatrist, adjustment, unprofessional, supplement, cold
m	doctor, patient, staff, time, great, office, year, care, best	man, guy, 2017, healing, recovery, successful, reassuring
Most associated words (with respective specialties)		
int	listens, primary, doctor, physician, practice, test, holistic	primary, blood, lifestyle, lab, diet, sick
brain	neurologist, m, migraine, spinal, seizure, lumbar	neurologist, brain, migraine, spinal, chiropatic
bone	pain, shoulder, knee, fusion, spine, hip, joint, injury	knee, hip, ankle, injury, shoulder, replacement
gastro	efficient, knowledgeable, thorough, friendly, bedside	scan, disease, facility, procedure, suggestion
pedia	child, pediatrician, kid, parent, son, baby, sick	pediatrician, kid, parent, sick, child, birth
surg	surgeon, surgery, hernia, neurosurgeon, skilled	orthodontist, breast, cancer, lee, surgeon
ent	sinus, nasal, deviated, septum, correct, hearing, polyp	sinus, ear, allergy, nose, dental, infection
gyn	pregnant, fertility, u, infertility, conceive, pregnancy	pregnant, fertility, pregnancy, ob, cycle
family	family, truly, listens, care, intelligent, compassionate	pour, nous, tr, en, temp, et, listener, est
psych	psychiatrist, depression, across, feedback, therapist	psychiatrist, depression, mental, psychologist
skin	procedure, hair, skin, result, transplant, dermatologist	dermatologist, acne, skin, plastic, lift
eye	eye, cataract, vision, glass, smooth	vision, cataract, glass, lens, optometrist, eye
endo	endocrinologist, diabetes, thyroid, diabetic, sugar	diabetes, thyroid, hormone, blood, control
Most associated specialties (with respective genders)		
f	brain, int, bone	family, skin, pedia
m	psych, bone, ent	brain, surg, bone

heterogeneous types (e.g., users, meta-data) jointly with words (Xing and Paul, 2017; Amir et al., 2017, 2016; Benton et al., 2016).

Our regression analysis shows that female caregivers are consistently given lower ratings than their male counterparts across all aspects of care, even after accounting for specialty. The lexical analyses are more exploratory and, while interesting, do not lend themselves to straightforward interpretation. We make these exploratory scripts available in our repository for an interested reader to further tinker.[2] Further, in Appendix A, we report results from some explorations to show some examples of apparent biases that are difficult to quantify with the relatively simple language technologies we have used here, yet are qualitatively apparent in this dataset.

A potential factor that we were unable to accommodate for was the gender of the review author. It may be that male patients author online reviews more than females, and bias in ratings may reflect this. As RateMDs.com does not collect or store reviewer information such as name, gender, or location we were unable to explore whether this relationship holds. We may only extrapolate that such a gender disparity is somewhat less likely, given that the United States has more females visiting healthcare providers than males (*Source*: Summary Health Statistics: National Health Interview Survey, 2017 [3]).

In any case, regardless of whether the gender bias can or cannot be explained away by certain factors including specialty of the physician or the gender of the review author, the bias may nonetheless have consequences on consumers of reviews. Whatever the reasons may be, a person reading online reviews is simply more likely to see poor ratings for female

2. See our github page: https://github.com/avi-jit/RateMDs
3. https://www.cdc.gov/nchs/fastats/physician-visits.htm

Table 6: Top results for (i) positive words in general, i.e. words closest to the positive rating vector, (ii) the vector obtained by adding it to the female vector and subtracting the male vector, (iii) the vector obtained by adding it to the male vector and subtracting the female vector, and (iv) words associated with fewer ratings in general, (v) the vector obtained by adding it to the female vector and subtracting the male vector, and finally (vi) the vector obtained by adding it to the male vector and subtracting the female vector.

1	**1+f-m**	**1+m-f**	**-1**	**-1+f-m**	**-1+m-f**
great	her	his	not	she	he
knowledgeable	she	he	doctor	her	his
recommend	earth	him	patient	herself	him
caring	greatest	great	never	never	he's
doctor	herself	he's	he	not	staff
professional	awesome	professional	his	doctor	team
highly	sweet	doctor	office	told	doctor
care	breast	best	one	office	grateful
kind	recommend	care	him	go	time
thorough	skilled	staff	staff	breast	thank
wonderful	tremendously	knowledgeable	time	patient	patient
best	considerate	man	after	rude	feel
excellent	caring	kind	care	care	pain
personable	exceptionally	team	year	another	man
always	personable	helpful	like	annual	great
compassionate	compassionate	amazing	go	take	best
pleased	fabulous	thorough	well	medical	one
awesome	hesitation	highly	medical	please	kind
helpful	green	excellent	see	woman	year
friendly	knowledgeable	grateful	first	health	after

physicians than male, and with the absence of indication about who wrote the review (male or female), the impression the reader will likely form is liable to be biased against female physicians. In a world where online reviews carry such impact, such a bias can in theory have real repercussions for healthcare.

There remains ample scope for future work on quantifying language differences in terms of sentiment displayed (e.g., by using sentiment lexicons) or aspects discussed. Our hope is that this initial effort and the new corpus spurs further research into quantifying and characterizing the perception of male versus female healthcare providers. We also hope that recognizing that the ratings (and perhaps the language within them) in online reviews of physicians is biased against female providers will spur efforts to mitigate the consequences of this, or at least increase awareness of the issue.

References

Anish K Agarwal, Kevin Mahoney, Amy L Lanza, Elissa V Klinger, David A Asch, Nick Fausti, Christopher Tufts, Lyle Ungar, and Raina M Merchant. Online ratings of the patient experience: Emergency departments versus urgent care centers. *Annals of emergency medicine*, 2018.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*, 2016.

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mário J Silva, and Byron C Wallace. Quantifying mental health from social media with neural user embeddings. *arXiv preprint arXiv:1705.00335*, 2017.

Adrian Benton, Raman Arora, and Mark Dredze. Learning multiview embeddings of twitter users. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 14–19, 2016.

John Rupert Firth. *Papers in Linguistics 1934-1951*. Oxford University Press, 1961.

Susannah Fox and Maeve Duggan. Health online 2013. *Health*, pages 1–55, 2013.

Guodong Gordon Gao, Jeffrey S McCullough, Ritu Agarwal, and Ashish K Jha. A changing landscape of physician quality reporting: analysis of patients online ratings of their physicians over a 5-year period. *Journal of medical Internet research*, 14(1):e38, 2012.

Judith A Hall, Danielle Blanch-Hartigan, and Debra L Roter. Patients' satisfaction with male versus female physicians: a meta-analysis. *Medical care*, 49(7):611–617, 2011.

Austin S Kilaru, Zachary F Meisel, Breah Paciotti, Yoonhee P Ha, Robert J Smith, Benjamin L Ranard, and Raina M Merchant. What do patients say about emergency departments in online reviews? a qualitative study. *BMJ quality & safety*, 25(1):14–24, 2016.

Vivian Lee. Transparency and trustonline patient reviews of physicians. *New England Journal of Medicine*, 376(3):197–199, 2017.

Siyue Li, Bo Feng, Meng Chen, and Robert A Bell. Physician review websites: effects of the proportion and position of negative reviews on readers willingness to choose the doctor. *Journal of health communication*, 20(4):453–461, 2015.

Andrea López, Alissa Detz, Neda Ratanawongsa, and Urmimala Sarkar. What patients say about their doctors online: a qualitative content analysis. *Journal of general internal medicine*, 27(6):685–692, 2012.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

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

Michael J Paul, Byron C Wallace, and Mark Dredze. What affects patient (dis) satisfaction? analyzing online doctor ratings with a joint topic-sentiment model. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*, 2013.

Michael J. Paul, Margaret S. Chisolm, Matthew W. Johnson, Ryan G. Vandrey, and Mark Dredze. Assessing the validity of online drug forums as a source for estimating demographic and temporal trends in drug use. *Journal of Addiction Medicine*, 10(5):324–330, 2016.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

Debra L Roter and Judith A Hall. Women doctors dont get the credit they deserve. *Journal of general internal medicine*, 30(3):273, 2015.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, and Lyle H Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.

Byron C Wallace, Michael J Paul, Urmimala Sarkar, Thomas A Trikalinos, and Mark Dredze. A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, 21(6):1098–1103, 2014.

Linzi Xing and Michael J Paul. Incorporating metadata into content-based user embeddings. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 45–49, 2017.

Table 7: Reviews for male physicians with the word 'handsome' and for female physicians with the word 'beautiful'

'handsome'	**'beautiful'**
family physician 10 year he not ***HANDSOME*** caring physician he date thorough he great demeanor always top his game love doctor health care thanks taking great care health care need	exquisite cataract surgeon warm caring inspiring human always ***BEAUTIFUL*** smile her face she truly miracle worker she performed miracle eye never thought possible see well after cataract surgery she ophthalmologist five year whenever mention her name health professional tell lucky gotten her counting lucky star she not control hospital staffing scheduling complain cut her slack not her fault hospital
tear thinking needed brain surgery first word need surgery he calm explained situation detail comforting thankful referred him see he's knowledgeable caring young ***HANDSOME*** his assistant nice based experience highly recommend	recently went new aegis clinic blown away staff excellent made feel relaxed doctor she absolutely amazing never doctor caring compassionate not mention fact she ***BEAUTIFUL*** natural not something see everyday her field highly recommend her anyone wish she still practiced family medicine switch her right away
highly recommend friendly ***HANDSOME*** doctor	lovely doctor ***BEAUTIFUL*** kind know listen her patient trust her procedure she always provided amazing outcome looking recommend friend her always thanked top notch clinic best doctor town
	absolutely fabulous doctor extremely ***BEAUTIFUL*** woman she top notch professional kind informative forever grateful
	endocrinologist model her patient she smart nonsense open minded friendly ***BEAUTIFUL*** obvious she practice she preaches ...
	one nicest ***BEAUTIFUL*** doctor ever seen
	excellent job brain she ***BEAUTIFUL*** person care her patient she angel sent heaven
	... 9 more such reviews (not shown for brevity) ...

Appendix A.

Using exploratory scripts, we filtered reviews (i) for male physicians with the word 'handsome' and (ii) for female physicians with the word 'beautiful'. We obtained just 15 occurrences of 'handsome' as opposed to the 266 occurrences of 'beautiful'. For brevity, we show here the reviews with these physical descriptors highlighted, after discounting those with the word 'office' (so we can be sure that the review is describing the physician and not the hospital/office), and after excluding gynecologists, dermatologists, and surgeons (because physical descriptors were often used for babies and plastic surgeries as well) in Table 7.

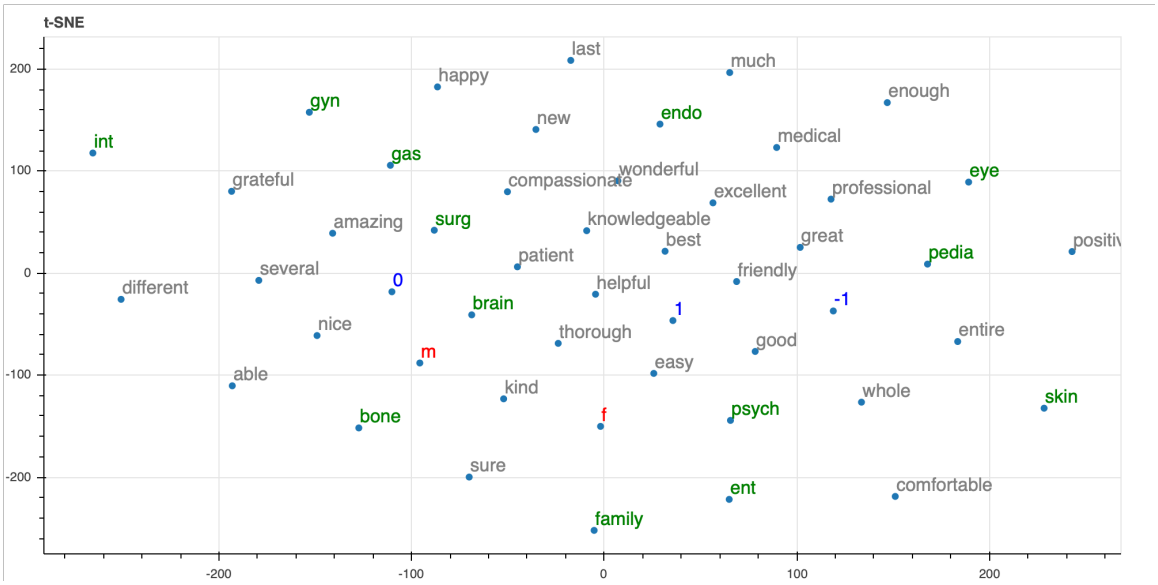


Figure 3: t-SNE plot for class embeddings, along with 31 most frequent adjectives

Table 8: The full list of most associated words with respective genders using two methods of lexical analysis (including stopwords). Embeddings results are aggregated over 20 independent runs to reduce variance.

	Embeddings	**Lexical Regression**
f	her, she, doctor, not, time, office, patient, care, take, go, never, well, caring, one, always, see, staff, herself, recommend.	podiatrist, herself, her, she, adjustment, unprofessional, supplement, cold, optometrist, brace, rude, psychologist, woman, ankle, toe, refused, tooth, chiropractor, worst, intelligent
m	he, his, him, not, doctor, patient, staff, time, great, office, year, care, first, after, like, he's, best, know, always.	he's, him, his, man, guy, himself, 2017, healing, recovery, successful, reassuring, honest, discomfort, god, expect, team, living, free, post, complication.

Appendix B.

Figure 3 visualizes the most frequent occurring adjectives in our dataset, as mined using the SpaCy library.

Appendix C.

In Table 8, we present a larger list than that in Table 5, of words associated with male and female genders, as learnt by our two methods.