The Ghost in the Machine: A Theory-laden Approach to Demographic Inference in Machine Learning

Blinded for Review

Abstract

A variety of studies have shown that machine learning methods like convolutional neural nets and random forests can be used to accurately infer characteristics of people online such as their gender, age, race, or political orientation. However, these methods are atheoretical: producing models which fail to generalize across context and offer little insight why some models perform better in some contexts and not others. This study compares the performance of state-of-the-art, atheoretical models to models informed by social theory on the task of inferring gender in text. Theory-laden models are developed using gender systems theory in sociology. Texts come from five corpora: blog posts, tweets, crowdfunding essays, movie scripts, and professional writing. The results show that models of gender built from theory are as accurate or more accurate than state of the art models. However, performance still varied substantially across corpora and, in some cases, even poor models with little theoretical motivation perform better than the best models. The success of this model suggests the presence of anomalous gender differences with little theoretical explanation.

A wide variety of research demonstrates that large, unstructured data from social media and the web can be used to make a variety of accurate inferences about social phenomena (Lazer and Radford 2017). One particular subfield within this computational social science is demographic inference wherein big data is used to infer demographic characteristics about the person or people generating the data (Ruths 2014). For example, (Barber 2015) uses follower networks on Twitter to infer an individual's political leanings. (Schwartz et al. 2013) use Facebook posts to infer individuals' personality traits, age, and gender. In general, the robustness of these studies suggests that it may be difficult to find demographic characteristics that cannot be inferred through some form of everyday big data.

Some advocates of machine learning and data science more broadly argue that our ability to translate large amounts of data into accurate predictions about human characteristics and behaviors eclipses the need to theorize how that data is produced and why our models work for particular problems(Anderson 2008). For example, we do not need to know what about blog posts exemplifies different personality, political, or racial differences because we'll always predict them accurately whether we understand why or not.

This "end of theory" argument was first posed in light of the fact that non-subject matter experts routinely outperformed subject matter experts on the data science competition website *Kaggle*. Such competitions routinely show that state of the art machine learning algorithms like XGBoosting or Convolutional Neural Nets can make more accurate predictions about most human characteristics than models built by subject matter experts.

Theory is not going away for two reasons. First, there is a growing use of machine learning by subject matter experts who are proving adept at model building and are using algorithms in unconventional ways. Second, because these models are invariant to the subject matter at hand, it is difficult to interpret their predictions and explain and correct their errors. For example, a statistician can say that the variance is high, but only the survey methodologist can say that it was because the question was ambiguous.

If we are to productively put these models to the purpose of explaining human behavior, then the gold-standard solution is to build models which meet or exceed benchmarks set by state of the art algorithms *and* produce theoretically interpretable model behavior. In this study, theories of gender are used to construct theoretically-informed models for how gender is expressed in text. These models are compared to theoretically-naive, state of the art models on both their interpretability and performance on a diverse set of corpora. The results show that, in most cases, theoretically-informed models are as performant as and often more accurate than state of the art models.

Interestingly, I find that performance for both naive and theory-laden models varies substantially across corpora and, in some cases, even poor models with little theoretical support perform better than state of the art models. This inter-corpus variation suggests that our theories and models of gender miss critical ways in which gender structures text. I conclude by discussing the role of theory in model building and suggest possible explanations for the unexpected findings.

State of the Art Approaches to Demographic Inference

Demographic inference is a foundational task in machine learning which involves using a set of features to predict the demographic characteristics of subjects. Demographic inference is at the heart of many contemporary big data systems including content recommendation, ad targeted, price personalization, and content filtering (Bashir et al. 2016; Bashir, Arshad, and Wilson 2016; Bakshy, Messing, and Adamic 2015). However, extant research on demographic inference has yet to develop a systematic approach to identifying features that predict demographic characteristics across domains (Argamon et al. 2003; 2007; Cohen and Ruths 2013). Features predicting gender or age on one website or web service do not predict them on others.

In the absence of a generic model for demographic inference, researchers build their own models based on context-specific feature sets and training data. The most basic and frequently used approach involves using a bag of words feature set with a classification algorithm like a naive Bayes, Random Forest, or the aforementioned XGBoost. In the bag of words approach, researchers take every possible feature in the data set or the most frequent 10,000. They then train large numbers of decision trees or, in the case of naive Bayes, feature weights on these words. This represents a maximally naive approach. No theory is used to select features and the models are aggregations of many weak learners. The theoretical interpretation is typically whether or not there is any signal at all.

Importantly, this naive approach has proven robust and few advances in feature engineering or estimation have proven to outperform this model across problems (Chandrashekar and Sahin 2014). In fact, with large amounts of text data, even naive Bayes estimators perform nearly equivalent to the most advanced algorithms (Ruths 2014).

Model 1: Raw. *Bag of words using the top 10,000 unigrams and bigrams.*

A category of approaches similar to the bag of words model are filter methods (see (Kohavi and John 1997), reviewed by (Chandrashekar and Sahin 2014). With filter methods, texts are preprocessed where words or phrases, called ngrams, go through a preliminary selection process wherein the ngrams with the highest preliminary correlation with the predicted outcome are included in the overall model. The advantage of these models is that they typically reduce the number of features needed, decreasing computational needs as well as noise. While there is no prevailing interpretation of filter methods, one argument is that they represent keyword differences. For example, in an ongoing and unpublished study, filter methods successfully separate politicians into liberal and conservative much better than any other method, largely because words and phrases like "Obamacare," "killery" and "tcot" (which stands for "top conservative on Twitter") are signals of partisan membership. In this way, this filter method arguably has some moderate degree of theoretical interpretability. If filter methods work, they do so because words and phrases are used to distinguish members of one identity from those of another.

Model 2: KBest. *Chi-square filtering using the top 5,000 unigrams and bigrams.*

A third general approach which has shown promise are word embeddings (Mikolov et al. 2013). Embeddings are a category of neural networks designed for deep learning. They use word sequence prediction to generate vectors of words. These vectors tend to act like synonyms, such that the vector for "King" subtracted by the vector "male" gives the vector for "Queen" (Mikolov, Yih, and Zweig 2013). This approach is new and generic models being developed will provide better vectors to use for text analysis (Jozefowicz et al. 2016). However, they can still be used for individual corpora. The advantage to embeddings is, like topic models, they can condense a substantial amount of variance into a small, meaningful subset. However, unlike topic models, text-level representations of word embeddings have no clear linguistic meaning. Although individual words can be translated into vectors, texts are aggregates of word-level vectors, typically 300-500 vectors, which have no clear linguistic meaning.

Model 3: Word2Vec. *Word Embeddings using 300 vectors.*

Finally, while nearly all forms of text analysis focus on words and phrases, a separate set of stylistic features is available. These non-linguistic features were originally developed in author identification tasks where researchers used stylistic markers like the frequency of commas or average size of paragraphs to determine whether some specific person was the author of a given text. Yet, they offer a range of non-linguistic features which most any text contains but are seldom used for demographic inference (see (Corney et al. 2002) for an exception in gender inference). There is little potential theoretical insight about gender to be gleaned from using this model as we have no apriori theory for why men or women would write with different styles. The point of including this model is twofold: 1) to provide another commonly used set of features wholly different from the other three and 2) to provide more evidence for whether or not these features produce any useful signal for our models of gender. The model reported here uses 32 features from (Corney et al. 2002) which include words per paragraph and mean use of atypical punctuation like brackets or parentheses.

Model 4: Nonword. *NNN Non-linguistic features of writing.*

This study treats the bag of words model as the gold standard for naive models of demographic inference. Filter methods and word embeddings represent more recently developed, atheoretical approaches which should outperform the bag of word model. Finally, non-linguistic features are added to test the inferences from structure rather than content. This is one of the few studies to test all four methods together across a shared set of corpora. In the next section, I discuss models based in sociological theory and one theoretical model derived elsewhere but often applied to gender inference.

Substantive Approaches to Inferring Gender

One assumption of this study is that, with a proper theory of gender, we can create models that reliably and accurately infer gender across contexts. In sociology, gender systems theory prevails (Ridgeway and Correll 2004; Risman 2004; Reskin 2003). In this view, gender is constructed on three levels: individual, interactional and institutional. Generally, the institutional level corresponds to large-scale gender segregation wherein males and females do different kinds of work and are responsible for different social domains (i.e. public and private). In text, this means that male authors will talk about male-typical roles and responsibilities: finance, politics, guns, sports, etc.; while women will talk about female-typical roles and responsibilities often referred to as the five F's: food, fashion, family, feelings, and feminism.

As a form of segregation, institutional gender should correspond to the topic structure of most corpora. Females should write articles in different sections of the newspaper while males should be more likely to tweet about different sports or video games they play. Prior research has shown that topic modeling is effective at capturing gender differences (Bamman, Eisenstein, and Schnoebelen 2014; Bamman, O'Connor, and Smith 2014). This study compares topic modeling to other state of the art measures as well as other theoretically-motivated measures.

Model 5a: Raw Topic. *Topic models predicting gender segregation.*

Topics are not the only way to identify structural segregation within a corpus. For example, a corpus of news articles already contains subjects and keywords which often do not intersect with the topics in a topic model. For example, women make up 23 percent of characters in war movies and 37 percent of characters in romance. Thus, gender segregation may occur along dimensions that are captured by context-specific social structures not revealed by topic models.

A second approach to institutional gender then is to use filtering methods on words and phrases that best predict these secondary categories. For example, in the Movie Dialogue men should use words and phrases typical of war while women be more likely to use words and phrases typical of love. This second approach to using institutional gender for demographic inferences is difficult to generalize because it requires finding a secondary set of categories for whatever corpus is being analyzed. However, this additional work may be worth it if it helps accurately classify author gender.

Model 5b: Structure. *Chi-square filtering for Corpus Categories.*

The second form of gender is interactional. Often represented as "doing gender" by people like Judith Butler or West and Zimmerman (West and Zimmerman 1987; 2009; Butler 1999), interactional gender represents the day-to-day actions that males and females tend to perform which we associate with masculinity and femininity. Thus, while institutional gender sorts men and women into different domains, interactional gender determines how men and women act within those domains. For example, cooking may be considered a feminine domain; however certain forms of cooking are considered masculine, like grilling, or feminine, like baking.

Gender systems theory thus posits a hierarchical structure to gender where men and women are segregated institutionally and, within these institutions, are expected to behave differently. Two different approaches are used to estimate this sub-level process. First, topic models are estimated for texts within the categories used in Model 5b. This acts as straightforward, subtopic model.

Model 6a: Subtopic. *Topic Models within Categories.*

The second approach to an interactional model involves pulling out the words and phrases which most distinguish males and females within institutional categories. In the example of cooking above, if we can accurately classify texts as about cooking, we should estimate a separate model for gender within cooking that captures differences between "grilling" and "baking." Because this represents a filter strategy applied within a topic or category, it theoretically suggest that what differentiates male and female authors within a given institution are keywords

Model 6b: Behavior. *Chi-square Filter for Gender within Categories.*

The third and final level of gender systems theory is the individual level. This level represents self-identification with being male or female. Specifically, do authors make statements about their sex? It is often surprisingly easy to infer the gender of authors because many people disclose it directly. Computationally speaking then, we should build classifiers which can detect when people make such declarations. Beyond looking for phrases like "As a man, I..." or "I am a mother who..." many languages force individuals to pick a gender for themselves. For example, to describe oneself in Spanish or French, you must gender the adjective: "yo soy bonito/a" "je suis beau/belle." Ruths 2014 has shown that building a classifier using these features in languages like French can yield an accuracy of nearly 90 percent. This study only involves English language texts. Thus, individual gender is only measured through direct disclosure rather than conjugations.

Model 7: Individual. *Dictionary methods detecting gender declarations.*

Gender systems theory offers a three-level model of gender which can be measured in a number of different ways. Here, five are put forward for examination. Theoretically, models for each level should be considered independent measures. Just because males talk about male-typed roles, does not mean that males within those roles use more masculine interaction-level words or phrases. For example, some argue that men in female-typical roles are more likely to display masculinity in that role in order to protect their status (Williams 1995).

Finally, one other theory-laden approach has been used to study gender extensively: the Linguistic Inquiry and Word Count (LIWC) corpus (Pennebaker and King 1999;

	Low Theory	High Theory
High Performance	Models 1-3	Models 5-7
Low Performance	Model 4	Model 8

Table 1: Outline of Model Performance

Tausczik and Pennebaker 2010). LIWC is a dictionary mapping words to features that correspond to psychometric characteristics. The features include parts of speech, "social processes" like family and friends, and "personal concerns" like money and religion. The theory behind LIWC is that the dictionary of words accurately measures these psychometric characteristics. While some may argue that men and women have different psychometric characteristics, LIWC was not created to detect gender in text. However, LIWC is still frequently used to generate feature sets for gender inference.

Newman et al. 2008 apply LIWC to 14,000 text files across 70 studies to identify whether or not men and women used LIWC-features differently across texts. Their results show many gender differences. However, most signals were weak with an average correlation (d) of roughly .10. Furthermore Newman et al. do not investigate whether LIWC-based feature sets improve gender classification in texts across corpora. One contribution of this paper is to systematically evaluate the performance of LIWC-based classifiers across standard online corpora and compare them to other approaches.

Model 8*: LIWC-based features for gender inference.*

The goal of this study is to compare the performance of these eight models on a shared set of corpora to determine whether or not theory-laden models can predict gender more accurately and reliably than state of the art, atheoretical models. The ideal models for demographic inference should provide both accurate inference *and* theoretical insight. In table 1, I arrange the eight models on these two dimensions.

Methods

To compare theory-laden and atheoretical models, I evaluate their ability to accurately predict the gender of the author or speaker of texts across five commonly-used corpora. I chose these corpora for their diversity and because they each contained category data which could be used for Models 5 and 6. features The corpora themselves are not directly comparable. Some like the DonorsChoose corpus contain vastly more texts than others while other like the Twitter corpus have very few words per text. The purpose of selecting such a diverse set of corpora from a diverse range of communication is to test the models on a diverse range of information spaces. The bag of words model does as well as cutting edge models in large corpora, but is this true for large but highly sparse corpora?

The corpora I chose typically included ground truth information about gender. However, some did not and others did not fully utilize the information available. In the case of the

Corpus	Num Texts	Balanced	Categories	Infer Gender
Donors Choose Essays	218,763	Yes	School and Teacher Variables	No
Blogger	19,090	No	Blog Topic	No
Brown	397	Yes	Non/Fiction and Genre	Yes
Movie Dialogue	6,915	No	Movie Genre	Some
Congress Twitter	709,302	Yes	Political Party	No

Table 2: Summary Statistics for Corpora

Brown corpus and Movie-Dialogue corpus I used the U.S. name database developed using Social Security data from the OpenGenderTracker project. Gender inference based on names is highly accurate given the extreme bimodal use of names (Liu and Ruths 2013).

Additionally, given the extent of gender segregation across social domains, many corpora are substantially imbalanced. For example, the DonorsChoose Corpus is roughly 87 percent female and the Congressional Twitter corpus is 81 percent male. These large imbalances cause prediction algorithms to converge to the dominant class. The result is that no matter which features one uses, the estimate is always the dominant class. To address this, I balance the sample by taking a random subset of equal numbers of males and females for the DonorsChoose, Twitter, and Brown corpora. Blogger data is already balanced and the Movie-dialogue data was only moderately imbalanced.

Corpora

*DonorsChoose Essays.*The DonorsChoose corpus contains over 800,000 essays written by teachers across the United States to attract funding for classroom projects. The essays are part of a larger data set with metadata on the teachers, students, schools, projects, and individual donations made by donors over the decade long history of the website.

For this study, I use 218,763 essays from the original data dump in 2011 which contained all of the data for the website through April 2011. I use school and teacher level metadata to construct categories including the subject area teachers taught in, their grade level, whether the school was a traditional public school or a nontraditional school such as a charter or magnet school and whether the school was in an urban, suburban, or rural location. Education is a female-dominated job in the United States and this is reflected on DonorsChoose as roughly 86 percent of the essays are written by females. Thus the corpus was genderbalanced.

Blogger Corpus. The Blogger corpus was originally created by (Argamon et al. 2007) for demographic inference. It is a gender-balanced sample of almost 30,000 blogs collected in August 2004 and includes data on the authors age, the primary topic of their blog, and their astrological sign. Topics range from everyday subjects to specialty areas like the military and aeronautics. The blogger corpus represents relatively long-form, expository writing about a variety of subjects.

Brown Corpus. I use the ubiquitous Brown corpus which was compiled in in 1965 by W. Nelson Francis as one of the first corpora for the computational study of texts (Francis 1965). It contains 397 snippets of professional and hobbyist writing from newspapers, books, pamphlets, and magazines. The sample is broken down into fiction and non-fiction writing and, within these, there are just over 20 subject areas including news, analysis, religion, romance, and comedy. Author gender is not given in the sample, but the names of document writers are provided for most of the texts. I thus infer gender using author name resulting in a sample that is 80 percent male authored. Finally, the sample is balanced on gender.

Movie Dialogue Corpus. The Movie-Dialogue Corpus integrates dialogue between almost 10,000 characters from 600 movies with movie metadata from IMBD (Danescu-Niculescu-Mizil and Lee 2011). The original corpus includes gender for 3,000 characters and I expand this using gender classification on character first names. I aggregate individual lines into a single document, meaning that demographic inference occurs at the character level rather than each utterance.

This is the only corpus in which a fictional character rather than the author is being inferred. Theoretically, relies on the assumption that the groups of writers involved in creating script and actors translating it to screen can realistically affect character gender in movies. This is not a stretch given that most movies are considered tenable representations of human action. However, there may still be substantial, unconscious divergences.

Congressional Twitter Corpus. The Twitter sample I use is a curated list of all tweets from members of the 112th congress which was seated from January 3, 2011 to January 2, 2013. The category used in this sample is the party of the member of congress. In this sample, I only attempt to predict the gender of tweets, rather than twitter accounts. The reason both to diversify the problem space I apply my models to and because there are so few women in congress overall.

Data Analysis

This study is not a test of the contribution estimation methods make to demographic inference. To control for estimation methods, I use the same two, random forests and naive Bayes, to calculate the estimated gender for individual texts across all corpora. The results indicate that random forests generally produce more accurate predictions than naive Bayes, but the models that perform best do not change with the estimation method. Other estimation models are possible, but no study suggests one type of estimation method works better with certain kinds of feature engineering models than others.

For each corpus, I used hold-p-out cross-validation to estimate the gender of each text. In the case of large samples like DonorsChoose essays, I used 10 percent of the sample to train the classifier and held out the remaining 90 percent as my test sample. For small corpora such as Brown, I used 90 percent to train the model and tested on the final 10 percent.

For topic modeling, I generally estimated 1 topic for every ten texts in a sample with a maximum of 100 topics. Thus, for example, I estimated 100 topics for DonorsChoose, but 30 topics for Brown. I visually inspected the topics to ensure they substantively captured topics consistent with the corpora and tweaked parameters when they did not.

Results

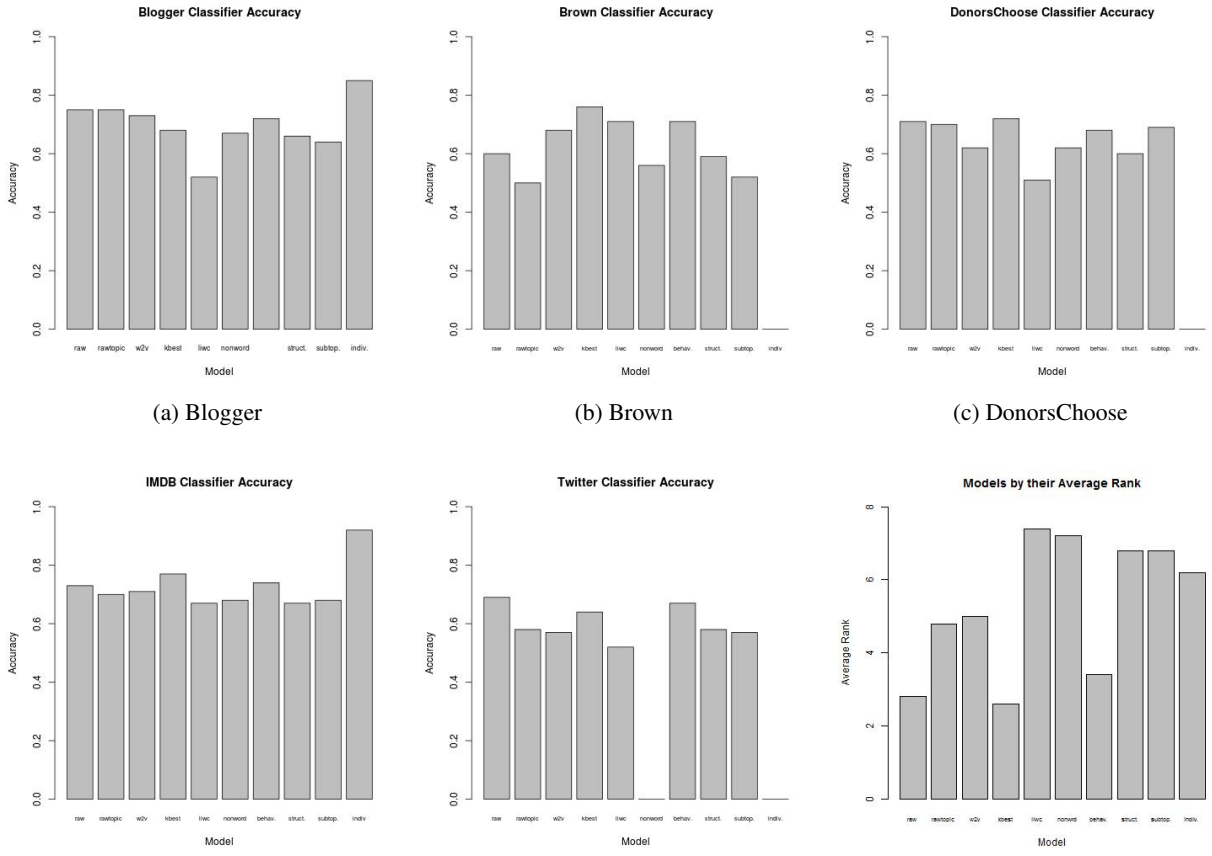
The results for each model in each corpus are given in Figure 1 as well as their average rank across all five corpora (1f). There is substantial variation both between *and* within corpora. The between corpora variance indicates that the general effectiveness of gender inference is very different in different contexts. The average accuracy for the individual tweets is only 60 percent, while in movie dialogue corpus the average accuracy is 71 percent.

Within corpora, the substantial variance reveals the degree to which different models capture different signals about gender. Most models perform equally well in the movie dialogue corpus, except the individual model (Model 7) which performs substantively better. On the other hand, the Brown, DonorsChoose, and to a lesser extent the Blogger corpora exhibit substantial internal variation in model performance. In these corpora different features give very different results. In short, model choice matters in some cases and, in others, it matters much less.

Such substantial variation makes it difficult to make clear statements about which models are better across all corpora. In fact, every model finishes fifth or worse in at least one corpus. And even a model that does well across most corpora like topic modeling, performs abysmally on the Brown corpus with an accuracy of 50 percent, equivalent to noise. The first result of this study then is that, even including theory-laden models, we have a long way to go to accurately infer gender across corpora. Furthermore, none of our models perform well enough to be a general model for all cases.

Looking at the average rankings in Figure 1, we see the pattern of performance we expected in Table 1. The bag of words model and chi-square filtering produced consistently highly ranked estimates while LIWC and nonword models produced the worst ranked estimates on average. Notably however, the behavior model (Model 6b) performed almost as well as the atheoretical bag of words and chi-square filter and better than word embeddings.

The behavior model specifically performed better than bag of words in Brown and Movie-quotes corpora while outperforming raw chi-square filtering in Twitter and Blogger. In addition, topic models performed as well as models based on word embeddings on average. Though they never per-



(a) Blogger (b) Brown (c) DonorsChoose

(d) Movie Dialogue (e) Congress Twitter (f) Average Rank

Figure 1: Model Results for All Corpora

formed better than bag of words, they did outperform raw filtering on the Blogger corpus.

These results suggest that topic models and the behavior model (using chi-square filtering for gender within each secondary category) capture valid gender differences across corpora and can act as models for measuring institutional and interaction-level gender. The other two theoretical models for institutional and interactional gender perform much worse. Both the structure model and subtopic model have an average rank of 6.8 and never finish higher than the fourth most accurate models.

Individual Model. The model with the highest average accuracy was the individual model (Model 7). However, the model was almost always inestimable. Texts which did not have one of the key phrases like "I am a mother" are excluded from the analysis. This missingness is so pervasive in the Congressional Twitter, DonorsChoose, and Brown corpora that they are not even estimated. In fact, only 77 of 218,000 DonorsChoose essays and 24 of the 700,000 tweets actually contained any key phrases.

Furthermore, even in corpora where such phrases can be found, few texts contain them. In the Movie Dialogue corpus, 208 characters (3 percent) make a direct claim about their gender. In Blogger, 2167 authors (11 percent) make such claims. In corpora where the model is estimable, the average accuracy is 89 percent well above the best model on any corpus (77 percent).

Poor Models. Models based on LIWC and non-linguistic features perform consistently worse than most all other models. Their average rank is 7.4 and 7.2 respectively. In general, the four worst models, LIWC, non-word, structure, and subtopics, have a median rank of seventh or worse and finish no higher than fourth. Astonishingly however, LIWC performs second-best on the Brown corpus with an accuracy of 71 percent. It also does well on movie dialogues with 67 percent accuracy. This indicates that even the weakest models may produce substantial information in some contexts.

Discussion

There are several important findings in this study. First, no model is king. Models created for performance do better on average than theoretical models. State-of-art models based in deep learning and filter methods produce middling results in several corpora. And theory-laden models with low performance expectations prove to have strong signals in other corpora. Thus, not only do we not have a generic model of

gender, but we also do not have a universal approach to modeling gender across different kinds of corpora.

On the positive side, the results show that theory-laden models can perform just as well and sometimes better than atheoretical, state-of-the-art models. This finding is a critical proof of concept that substantive theory can actually improve on the traditional big data approaches to inference which simply seek to exploit any signal in as much data as possible. The theory-laden models are able to capture these diverse signals in a theoretically principled way, make them performant while simultaneously retaining theoretical interpretability of model performance and the corpus itself. Thus, there does not seem to necessarily be a trade-off between model performance and theoretical contribution.

It is also important to note that this study does not set theory-laden and atheoretical methods as orthogonal approaches. The theory-laden and atheoretical models are very similar in construction and principle. Theory helps guide and customize the use of typically atheoretical models. Thus, for example, I try to use filtering and topic modeling methods to generate theory-laden, high performance models. In some cases, a feature set of one hundred topics performs as well or better than the 10,000 most frequent ngrams or the 5000 most correlated ngrams. As this study shows, using theory to customize state-of-the-art models appears to be able to capture the best of both worlds.

Beyond the methodological contribution, this study does provide a range of theoretical insights into gendered language. First, the performance of topic models and the behavioral model suggest that these two best capture gender differences in language use corresponding to institutional and interactional gender. Specifically, topic models are better at recovering the society-wide effects of gender segregation than filtering on words and phrases about social structure. That is, reference to the topic of cooking is typically more indicative of gender than just using words and phrases typical of cooking.

For measuring interactional gender, the best approach was using chi-square filtering within secondary categories like teachers' subject area or the genre of professional writing. We may be able to guess that men are more likely to write professional essays about religion in the Brown corpus, but there is substantially more signal to be found within those writing about religion. Surprisingly, this did not hold true for subtopic models. Men and women did talk about different topics within already defined categories, but these topical differences were much weaker than ngram differences. Perhaps other topic models could do better. For example, hierarchical topic models may produce better estimates because they arguably estimate institutional and interactional gender simultaneously (Bamman, Eisenstein, and Schnoebelen 2014). More research is needed.

Finally, individual gender was the most powerful, but least consistent measure of gender. Consistent with Ruths 2014, I find that individual declarations of gender are the most accurate sources of information for demographic inference. Yet few people made these declarations and the rate of declaration varied substantially among corpora.

It seems that people do disclose their gender in general, informal venues like blogs and movie dialogue. However, formal and official written communication such as congressional tweets, professional writings, and teachers proposal essays almost never mentioned their gender. In future research, we can expect individual classifiers to work in such informal settings but we must rely on other methods to infer gender indirectly.

Practically speaking, while English writers cannot undo the biographical processes which lead them to work in gender segregated fields and favor certain gender-stereotyped behaviors; they can choose when to disclose the gender we identify with in online discourse. This is not true for those using languages with many gender conjugations. Laboratory studies show that individuals typically do not apply gender stereotypes to another until they know the other's gender (Blair and Banaji 1996; Blair 2002) Thus, we should expect models using individual disclosure of gender to be more performant in these languages. And online interactions occurring in those languages should be substantially different than those in more gender neutral languages.

Finally, there is one anomaly to discuss: the high performance of LIWC on the Brown corpus. Like the nonlinguistic model, expectations for LIWC were low. Its theoretical motivation was only tangentially related to gender and it was only included because of how often it had been used to identify what psychometric features may differentiate men and women in text. Furthermore, it is composed of only a few features which are relatively idiosyncratic to the general process of writing. However, it performs second best in distinguishing men and women in the Brown corpus, equal to the behavioral model, better than word embeddings, bag of words, and topic models.

One possible explanation is that the corpus contains writing from an era when gender was constructed differently. Texts in 1961 are the product of different segregating processes and more stringent behavioral stereotypes. This theory could be tested by looking at other writing from that era and earlier eras to investigate whether or not LIWC features reliably distinguish gender better or worse during different historical eras. Another possible explanation is that gender is less visible in long-form professional writing. Men and women can write a romance novel and the differences between a male-authored and female-authored romance novel may be more subtle than social structures which typically construct gender for a society. Either way, this is a finding that lacks a clear expectation and deserves further study.

The general takeaway should be that even poorly motivated and often ineffective models can identify meaningful and often powerful signals. In this case, nearly every model performed better than random on every corpus. When studying gender computationally, it behooves us to use many different feature engineering strategies to not only understand the many ways gender comes through in text, but also to help us develop better theories of gender. Each strategy represents a lens through which we can see different dimensions of gender.

References

Anderson, C. 2008. The end of theory. *Wired magazine* 16(7):16–07.

Argamon, S.; Koppel, M.; Fine, J.; and Shimoni, A. R. 2003. Gender, Genre, and Writing Style in Formal Written Texts. *Text* 23(3):321–346.

Argamon, S.; Koppel, M.; Pennebaker, J. W.; and Schler, J. 2007. Mining the Blogosphere: Age, Gender and the Varieties of Self-Expression. *First Monday* 12(9).

Bakshy, E.; Messing, S.; and Adamic, L. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*.

Bamman, D.; Eisenstein, J.; and Schnoebelen, T. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18(2):135–160.

Bamman, D.; O'Connor, B.; and Smith, N. A. 2014. Learning latent personas of film characters. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 352.

Barber, P. 2015. Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data. *Political Analysis* 23(1):76–91.

Bashir, M. A.; Arshad, S.; and Wilson, C. 2016. "Recommended For You": A First Look at Content Recommendation Networks. 17–24. ACM Press.

Bashir, M. A.; Arshad, S.; Robertson, W.; and Wilson, C. 2016. Tracing information flows between ad exchanges using retargeted ads. In *Proceedings of the 25th USENIX Security Symposium*.

Blair, I. V., and Banaji, M. R. 1996. Automatic and controlled processes in stereotype priming. *Journal of Personality and Social Psychology* 70(6):1142–1163.

Blair, I. V. 2002. The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review* 6(3):242–261.

Butler, J. 1999. *Gender Trouble: Feminism and the Subversion of Identity*. New York: Routledge.

Chandrashekar, G., and Sahin, F. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40(1):16–28.

Cohen, R., and Ruths, D. 2013. Classifying Political Orientation on Twitter: It's Not Easy! In *ICWSM*.

Corney, M.; de Vel, O.; Anderson, A.; and Mohay, G. 2002. Gender-preferential text mining of e-mail discourse. In *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*, 282–289. IEEE.

Danescu-Niculescu-Mizil, C., and Lee, L. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Francis, W. N. 1965. A Standard Corpus of Edited Present-Day American English. *College English* 26(4):267–273.

Jozefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; and Wu, Y. 2016. Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Kohavi, R., and John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence* 97(12):273 – 324. Relevance.

Lazer, D., and Radford, J. 2017. Introduction to Big Data. *Annual Review of Sociology* 43.

Liu, W., and Ruths, D. 2013. What's in a Name? Using First Names as Features for Gender Inference in Twitter. In *AAAI spring symposium: Analyzing microtext*, volume 13, 01.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T.; Yih, W.-t.; and Zweig, G. 2013. Linguistic Regularities in Continuous Space Word Representations. In *HLT-NAACL*, volume 13, 746–751.

Newman, M. L.; Groom, C. J.; Handelman, L. D.; and Pennebaker, J. W. 2008. Gender Differences in Language Use: An Analysis of 14,000 Text Samples. *Discourse Processes* 45(3):211–236.

Pennebaker, J. W., and King, L. A. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology* 77(6):1296.

Reskin, B. F. 2003. Including Mechanisms in Our Models of Ascriptive Inequality. *American Sociological Review* 68(1):1–21.

Ridgeway, C. L., and Correll, S. J. 2004. Unpacking the Gender System: A Theoretical Perspective on Gender Beliefs and Social Relations. *Gender & Society* 18(4):510–531.

Risman, B. J. 2004. Gender as a Social Structure: Theory Wrestling with Activism. *Gender & Society* 18(4):429–450.

Ruths, D. 2014. The Promises and Pitfalls of Demographic Inference on Social Media.

Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE* 8(9):e73791.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.

West, C., and Zimmerman, D. H. 1987. Doing Gender. *Gender & Society* 1(2):125–151.

West, C., and Zimmerman, D. H. 2009. Accounting for Doing Gender. *Gender & Society* 23(1):112–122.

Williams, C. L. 1995. *Still a Man's World: Men who do "Women's Work"*. Berkeley: University of California Press.