Piloting A Theory-based Approach to Inferring Gender in Big Data

Jason Radford
Network Science Institute
Northeastern University
Boston, MA 02131
Email: j.radford@northeastern.edu

***Abstract*—Machine learning methods can be used to accurately predict core characteristics about people such as their gender, age, race, or political orientation. However, prediction models tend not to generalize, offer little explanation for particular corpora, produce weak theory, and suffer from latent biases. In this study, we present an alternative approach to demographic inference combining sociological theories of gender with machine learning to create high-dimensional measures of gender rather than predict sex. We create measurement models for gender across five corpora: blog posts, tweets, crowdfunding essays, movie scripts, and professional writing. We show these models validly measure gender in the corpora and then compare their ability to predict author gender to standard prediction models. We find that measurement models of gender are as accurate and sometimes more accurate than prediction models. Thus we show theory-based measurement models are not only interpretable but performant.**

1. Introduction

A wide variety of research demonstrates that big data can be used to make a variety of accurate inferences about the social phenomena producing it [1]. In the area of demographic inference, researchers use big data to infer basic characteristics about people such as their gender, age, and race [2]. For example, [3] uses Twitter data to infer an individual's political leanings.

However, these models have proven to be relatively brittle. They do not generalize across corpora [4], [5]. The models they produce are notoriously difficult to interpret. And, they often take advantage of biases in data rather than controlling for them [6]. The conceit of this study is that a measurement model approach based in social theory can lead to models that reliably and accurately capture the meaningful constructs that define our outcome of interest.

This study builds a theory-based model of gender in text which is used to engineer a set of theory-laden features. The model of gender is based on gender systems theory which posits that gender is constructed on three levels: individual, interactional, and institutional [7], [8]. These theory-based models are used to predict gender in five very different but commonly-used corpora; blog posts, tweets, crowdfunding essays, movie scripts, and professional writing; to test whether or not the model is valid. Descriptive results show that different levels of gender are more prominent in distinguishing men and women in different corpora, providing insight into when models do and do not generalize and why. For example, individual gender plays a large role in blogs but is non-existent in professional writing.

2. Standard 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 [9].

In the absence of a generic model for demographic inference, researchers build their own models based on corpus-specific feature sets. The most basic and frequently used approach involves using a bag of words feature set with a classification algorithm like a naive Bayes or Random Forest. Bag of words 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.

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 [10]. 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. Filter methods arguably have some moderate degree of theoretical interpretability. If filter methods work, they do so because members of one identity use words and phrases to distinguish themselves from other groups.

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

A third general approach which has shown promise are word embeddings [11]. The advantage to embeddings is, like topic models, they can condense a substantial amount

978-1-5386-2715-0/17/$31.00 ©2017 IEEE

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 such as average sentence length, punctuation usage, and complexity [12]. These non-linguistic features offer a completely different standard feature set for distinguishing members of different social groups. However, they are seldom used for demographic inference. Including them here not only provides another approach to inferring gender, but will provide a baseline for how much signal these features produce for gender. The model reported here uses 32 features from [12] which include words per paragraph and mean use of atypical punctuation like brackets or parentheses.

Model 4: Nonword. *32 Non-linguistic features of writing.*

3. Theory-laden Approach to Gender Inference

In sociology, gender systems theory posits that gender is constructed on three levels: individual, interactional and institutional [7], [8], [13]. 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. We capture this in two ways.

Model 1a: Raw Topic. *Topic models predicting gender segregation.*

Model 1b: Structure. *Chi-square filtering for Corpus Categories.*

First we use latent Dirichlet allocation to generate topic models within each corpus. Men and women should discuss these topics to different degrees [14]. Second, we use secondary metadata from each corpus to manually identify different roles and domains. For example, in the Movie Dialogue corpus men should use words and phrases typical of war while women be more likely to use words and phrases typical of love. We use filter methods on these secondary metadata to generate features for predicting the gender of authors.

The second form of gender is interactional. Often represented as "doing gender" by people like Judith Butler or West and Zimmerman [15], interactional gender represents the day-to-day actions we stereotype as masculine and feminine. Thus, while institutional gender sorts men and women into different domains, interactional gender determines how men and women act within those domains.

Model 2a: Subtopic. *Topic Models within Corpus Categories.*

Model 2b: Behavior. *Chi-square Filter for Gender within Categories.*

We measure this in two ways similar to Model 1. First, we run topic models for each of the texts in the corpus categories used for Model 1b. In the case of War movies in the IMDB corpus, we run a topic model on lines in war movies and use the topics to predict male and female characters in war movies. Second, we run filter methods on the lines spoken by male and female characters to predict the gender of characters within that category.

The third and final dimension is the individual dimension, what we call sex. Sex is identifying as or being identified as male or female. It is often surprisingly easy to infer the sex of authors because many people disclose it directly. People claim their sex, "As a man, I..." or "I am a mother who..." Using these self-declarative features can yield an accuracy of nearly 90 percent [2]. Model 3 is a dictionary classifier which looks for sentences which begin with variants of "I am a and "As a and contain a sex-definite noun like mother, father, King, and waitress in the following 15 words.

Model 3: Individual. *Dictionary methods detecting sex declarations.*

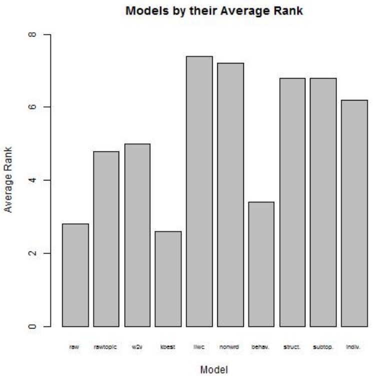
4. Data

*DonorsChoose Essays.*The DonorsChoose corpus contains 200,000 essays written by teachers across the United States to attract funding for classroom projects. I use school and teacher data like subject area, grade level and school location (e.g. urban or rural) to construct categories. Education is a female-dominated job in the United States. Thus the sample is highly imbalanced.

Blogger Corpus. The Blogger corpus was originally created by [5] for demographic inference. It is a gender-balanced sample of almost 30,000 blogs collected in August 2004 and the category data I use is the primary topic of their blog. Topics range from everyday subjects to specialty areas like the military and aeronautics.

Brown Corpus. I use the ubiquitous Brown corpus compiled in 1965 [16]. It contains 397 snippets of professional and hobbyist writing classified into fiction and non-fiction and then 20 subject areas including news, analysis, religion, romance, and comedy. Author gender is not given in the sample. Instead they are inferred from authors name, resulting in a sample that is 80 percent male authored.

Movie Dialogue Corpus. The Movie-Dialogue Corpus integrates dialogue between almost 10,000 characters from 600 movies with data for the genre of each film from IMBD [17]. 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) Average Rank

Figure 1: Model Results for All Corpora

a single document, meaning that demographic inference occurs at the character level rather than each utterance.

Congressional Twitter Corpus. The Twitter sample I use is all tweets from members of the 112th U.S. Congress. 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 is both to diversify the problem space and to increase the number of observations for female-authored tweets, given the low representation of women in congress.

5. Results

Figure 1a presents the average rank of the classifiers across all five corpora. There is substantial variation in the accuracy of classifiers both between *and* within corpora. The between corpora variance indicates that the general effectiveness of gender inference is very different in different contexts. Within corpora, the substantial variance reveals the degree to which different models capture different signals about gender. 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.

Looking at the average rankings in Figure 1, we see the pattern of performance we expected. 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 individual model had the highest average accuracy (89%); however it was very often missing in the DonorsChoose, Congressional Twitter, and Brown corpora. For example, only 77 of 218,000 DonorsChoose essays contained any key phrases.

6. Discussion

It appears that we can have models that are both substantively informative and algorithmically performant. However, the performance of individual models still varied substantially across corpora and, in some cases, even poor models with little theoretical motivation perform better than the best atheoretical or theory-driven models. This anomaly points to the need for better theory and demonstrates the potentially groundbreaking contributions machine learning and big data can make to social theory.

References

[1] D. Lazer and J. Radford, “Data ex machina: Introduction to big data,” vol. 43, no. 1, pp. 19–39. [Online]. Available: https://doi.org/10.1146/annurev-soc-060116-053457

[2] D. Ruths, “The promises and pitfalls of demographic inference on social media.”

[3] P. Barber, “Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data,” vol. 23, no. 1, pp. 76–91. [Online]. Available: http://pan.oxfordjournals.org/content/23/1/76.abstract

[4] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, “Gender, genre, and writing style in formal written texts,” vol. 23, no. 3, pp. 321–346. [Online]. Available: http://lingcog.iit.edu/doc/gendertext04.pdf

[5] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, “Mining the blogosphere: Age, gender and the varieties of self-expression,” vol. 12, no. 9.

[6] R. Tatman, “Gender and dialect bias in YouTubes automatic captions,” p. 53.

[7] C. L. Ridgeway and S. J. Correll, “Unpacking the gender system: A theoretical perspective on gender beliefs and social relations,” vol. 18, no. 4, pp. 510–531. [Online]. Available: http://gas.sagepub.com/cgi/doi/10.1177/0891243204265269

[8] B. J. Risman, “Gender as a social structure: Theory wrestling with activism,” vol. 18, no. 4, pp. 429–450. [Online]. Available: http://gas.sagepub.com/content/18/4/429.abstract

[9] M. A. Bashir, S. Arshad, W. Robertson, and C. Wilson, “Tracing information flows between ad exchanges using retargeted ads,” in *Proceedings of the 25th USENIX Security Symposium*.

[10] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” vol. 40, no. 1, pp. 16–28. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0045790613003066

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space.” [Online]. Available: http://arxiv.org/abs/1301.3781

[12] M. Corney, O. de Vel, A. Anderson, and G. Mohay, “Gender-preferential text mining of e-mail discourse,” in *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*. IEEE, pp. 282–289.

[13] B. F. Reskin, “Including mechanisms in our models of ascriptive inequality,” vol. 68, no. 1, pp. 1–21.

[14] D. Bamman, J. Eisenstein, and T. Schnoebelen, “Gender identity and lexical variation in social media,” vol. 18, no. 2, pp. 135–160. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/josl.12080/full

[15] C. West and D. H. Zimmerman, “Doing gender,” vol. 1, no. 2, pp. 125–151. [Online]. Available: http://gas.sagepub.com/content/1/2/125.abstract

[16] W. N. Francis, “A standard corpus of edited present-day american english,” vol. 26, no. 4, pp. 267–273. [Online]. Available: http://www.jstor.org/stable/373638

[17] C. Danescu-Niculescu-Mizil and L. Lee, “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*.