They Said What About Whom?
Automatic Classification of Subject Gender from News Text

Hope B. J. Schroeder
hopes@stanford.edu

Louis B. Lafair
lafair@stanford.edu

Pratyusha Javangula
pjavan@stanford.edu

Abstract

Countless classifiers exist to predict the author gender of a text. We explored another related task. We built a linear classifier to predict the subject gender of a text. We used a data set of approximately 143,000 news articles from 15 major publications. We split the articles into sentences, and filtered to those containing exclusively male or female pronouns within one sentence. With this method, our training data set consisted of around 870,000 examples. Our zero rule baseline yielded a macro-averaged F1 score of 0.425 on the development set. A unigram baseline yielded an F1 score of 0.625. Our oracle (human guessing) yielded an F1 score of 0.743. Our best model was a stochastic gradient descent classifier trained on part of speech unigrams that achieved a macro-average F1 score of 0.495, a modest increase of 0.07 over our zero rule baseline. By examining the most predictive features for this and other tested classifiers, we achieved our goal of uncovering stylistic differences between sentences about males versus females.

1 Introduction

Many existing systems to classify the subject of text do not seek to classify the gender of the subject, despite mounting evidence from prior research that authors of various texts use different language to describe men and women (Watson, 1987; Deaux, 1973; Trix & Psenka, 2003). Furthermore, the existing systems that do seek to capture these differences typically do so with data sets of very long texts, like movie scripts and plays (Agarwal et al. 2015;Sobhan et al. 2016). In this project, we classified the gender of the subject of a text by drawing on insights from previous sociolinguistic research showing the ways language written about men and women systematically differs, computational techniques to classify text with respect to non-gendered topics, and computational techniques to classify text with respect to the gender of the author. Furthermore, we sought to classify the gender of the subject in much shorter texts than have been investigated in the existing literature. Specifically, we designed and implemented a system for subject gender classification of individual sentences taken from 143,000 news articles.

To accomplish this goal, we investigated using lexical, syntactic, and lemma feature sets together with a linear classifier (logistic regression and stochastic gradient descent with a hinge loss function). We conducted experiments on the very large, wide-scope data set All the News from Kaggle (Thompson, 2017).

2 Related Work

We built on the findings from recent research on the ways in which the language used to describe men and women is strikingly and predictably different. We draw inspiration from two areas of research as it relates to the present work: sociolinguistics and natural language processing (author gender prediction, character gender prediction, and topic classification).

2.1 Findings from Sociolinguistic Research

A number of sociolinguistic studies have focused on the text of professional evaluations, teaching evaluations, and other written work in which a person is evaluated or characterized using written words. These studies concluded that there are specific, systematically pervasive patterns in certain features of texts depending on whether the subject being written about was a man or a woman. In the context of the project, we shifted focus from analysis of a set of performance evaluations to using this news corpus when it became unfeasible to procure the former data set in a timely manner. This shift led us to consider prior research on evaluative text to be somewhat less relevant, but still potentially relevant due to recent interest in the idea that writers may describe men and women differently, even in text such as news writing that is, to some degree, intended to be objective. We used findings from sociolinguistics as a starting point to designing some feature sets.

Schmader et al. 2007 found in an analysis of letters of recommendation in the fields of chemistry and biochemistry that men are more likely to be described with standout adjectives in academic letters. The researchers used Linguistic Inquiry and Word Count (LIWC) software to analyze data for mentions of grindstone traits (words like "conscientious" and "hardworking," which they found to be more common in reviews of women), ability traits (words like "talented" and "gifted," which they found to be more common in reviews of men), and standout adjectives (such as "extraordinary," which they found to be more common in reviews of men), as well as the standard LIWC lexicons.

Smith et al. observed similar findings in a very recent investigation into the evaluations of female and male military leadership. The research team found that managers used more positive words to describe the performance of men and more negative words to describe the performance of women. Furthermore, positive adjectives most likely to describe men included "analytical" and "competent" whereas positive adjectives for women included "compassionate" and "enthusiastic." These findings align with previous work in the field which found that language used to describe female leaders is usually more "communal" and language used to describe male leaders is more likely to be "agentic." Because women in the news are frequently politicians or celebrities, we speculated that at least for the politician contingent, these different descriptive adjectives of female leaders may be informative features.

2.2 Predicting Author Gender

Author gender prediction is a well-established task in natural language processing (Peersman et al. 2011; Argamon et al. 2007). Insights from these experiments show that the systematic differences in the ways that men and women differ as the writers can be operationalized in terms of feature sets for machine learning classifiers. We applied the feature sets and classifier types used in this work to the current work and data set as a means of testing the hypothesis that the ways that men and women differ as the subject of text can be operationalized in the same manner.

Koppel et al. 2002 were able to achieve a maximum classification accuracy of 77.3% with 1,081 stylometric features on a data set of novel texts taken from the British National Corpus. The features they used included 405 function words, the 100 most frequent bigrams, and various part-of-speech (POS) n-grams. Based on these findings, it is clear that there are purely stylometric features that yield a high predictive accuracy when applied to a gender-based classification task. However, the data set used by Koppel et al. contained only 566 samples, each of which was very long (mean length of 34,420 tokens) and comprised of more formal prose-style text. We aimed to build a classifier that detects stylometric differences in text used to describe different genders that uses shorter and more current texts from news articles.

Cheng et al. 2011 predicted author gender for texts with fewer tokens taken from the Enron email corpus. In addition, they were able to train a classifier on more samples than Koppel et al. (2002) with more current techniques (Bayesian logistic regression, AdaBoost decision tree, and SVM). In contrast to Koppel et al. (2002), Cheng et al. incorporated both stylistic and semantically oriented features. Critically, they extracted LIWC features, which we also sought to do. Cheng et al. 2011 achieved a maximum classification accuracy of 85.1% and determined that function words, word-based features, and structural features were most predictive. The present work leveraged these conclusions by including LIWC-based and related features.

2.3 Predicting Character Gender

Because our task was to detect the gender of a text with respect to the subject of text rather than the author, we combined approaches from author gender prediction and character gender detection. While the former task yielded feature sets well-suited to detecting the differences between genders in current texts, the latter task provided useful nuances as it relates to detecting which gender a text is about rather than which gender wrote the text.

Agarwal et al. 2015 combined a sensitivity for gender-oriented language with the classic task of topic detection by building a classifier that predicts, given a dialogue between two female characters in a film, whether the topic of that dialogue was a man or not. Their data set consisted of 457 movie screenplays from the Internet Movie Script Database (IMSDB). The researchers used four broad classes of features (bag of words with 18,889 features, 4 linguistic features, distribution of conversations over topics resulting in a 72-dimensional vector, and 43 social network analysis features) to train a logistic regression classifier, as well as SVM classifiers with both linear and RBF kernels. They achieved a final macro-averaged F1 score of 0.80 on the unseen test set using a linear SVM classifier. We implemented a bag of words model in the current work, and derived topic-related features by designing hand-crafted category lexicons.

Sobhan et al. 2016 predicted the gender of various characters in Shakespeares text using a variety of lexical, syntactic, and lemma features. Using Sequential Minimal Optimization (SMO), they were able to achieve a maximum classification accuracy of 82% for character gender. Potentially more interesting is that particular combinations of features predictive of character gender (part-of-speech and lemma n-grams) have also been shown in previous texts to be predictive of author gender-establishing a crucial link between who is writing and who is being written about. As a result we extracted part of speech and lemma unigrams as feature sets.

3 Data

3.1 Obtaining Article Data set

We obtained our data set of 142,570 articles via Andrew Thompson on Kaggle (Thompson, 2017). The articles are from 15 different American publications across the political spectrum, published both in print and online. Most articles were written between January 2016 and July 2017. Thompson scraped the text with BeautifulSoup, by accessing a year and a half of content on the Internet Archive. Note that the articles were all linked on the given publications' home pages or in their RSS feeds. There are different quantities of articles from different publications, as seen in Figure 1. Ultimately, the wide range of publications helped to ensure that a particular political ideology was not overly represented in the data set.

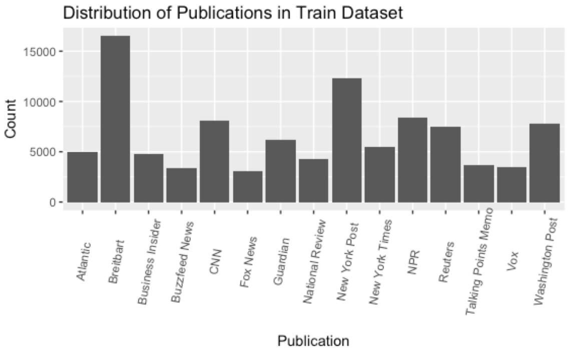

Figure 1: Distribution of AllTheNews data set over publications.

We split the data set into train (70%), development (dev) (10%), and test (20%) sets. There were 99,799 train articles, 14,257 dev articles, and 28,514 test articles. We split at the article level because we did not want to both train and test on sentences from the same article. After splitting, we tokenized the articles into sentences. Each sentence served as an individual example. We removed especially short (under 20 characters) and long (over 1,000 characters) sentences.

3.2 Labeling Sentence Gender

We counted the number of male (he, he*, him, his, himself) and female (she, she*, her, hers, herself) pronouns, including lowercase and uppercase, in each sentence. We labelled sentences male or female if they had exclusively male or female pronouns. Then we filtered to these sentences and replaced gendered pronouns with the gender-neutral "it". Because "her" corresponds with both "his" and "him", we did not convert possessive pronouns to "its," and instead kept any possessives as "it" as well. We note that as a result, we were not just predicting the gendered subject of the sentence, but also the gendered objects and possessors in the sentence. In other words, we were building a classifier for sentences that contained exclusively male or female pronouns.

Advantages of this approach included that by labeling according to pronouns, we diminished the probability of misgendering a sentence. However, an important disadvantage was that pronouns alone do not necessary mean that a sentence is *about* a man or woman. Sentences with exclusively female pronouns may contain a male name (e.g. "She talked to Chris.") As such, sentences labelled male or female may not actually be exclusively about males or females. However, male or female pronouns served as the cleanest line to draw for labeling the data set in an efficient manner.

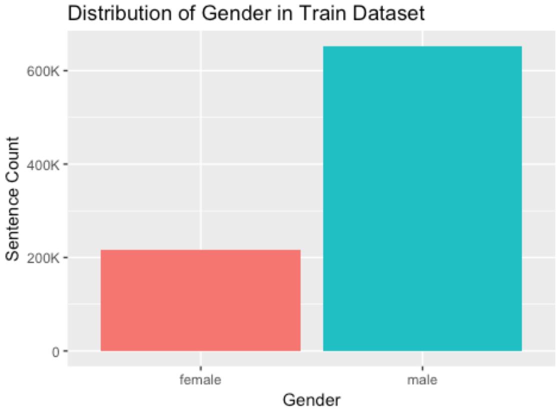

Figure 2: Distribution of gender over samples in training data set.

Female sentences (sentences with exclusively female pronouns) made up under one quarter (24.9%) of our training examples. Specifically, there were 653,071 male sentences in the train data set and 216,463 female sentences. This fact is, in itself, an interesting (if not unexpected) finding. Articles from 15 major publications were predominantly about men. Many of these articles, due to Thompson's scraping techniques, are featured, front-page articles. Basic explanations for uneven gender distribution in news coverage include the fact more world leaders are men and men's sports are more publicized than women's.

4 Methods

4.1 Evaluation Metrics

Since we have an uneven number of male and female examples, classification accuracy alone is under-informative. Instead, we used the macro-averaged F1 score. In particular, a macro-average (as opposed to micro-averaged) F1 score will penalize classifiers that only do well on male examples. We wanted our classifier to be good at predicting both male and female examples, despite the much smaller number of female examples. We had to hold it accountable accordingly.

4.2 Oracle

In order to set an estimated upper bound for our classifier, we used our own human predictions as an oracle. We randomly selected 300 samples from the development set, then split them evenly among our team. The three of us each predicted pronoun gender for 100 one-sentence samples. Only the text of the samples was shown, without any other information (e.g. publication or author name). As in the train data set, gendered pronouns had been replaced with the gender neutral "it."

	Pred. Male	Pred. Female	Total
Male	205	37	242
Female	17	41	58
	222	78	

Table 1: Confusion matrix illustrating precision and recall for human-performance oracle. Pred. stands for Predicted.

Table 1 shows our cumulative performance on the oracle task in terms of precision and recall. Cumulatively, we performed with 92% precision and 85% recall for male samples and 53% precision and 71% recall for female samples. Our male F1 score was 0.88, and our female F1 score was 0.60. Our macro-averaged F1 score was 0.743. If our classifier comes close to this range, it will be performing at a human level. Note that we did not anonymize named entities from the samples. In many cases, the researchers were able to derive gender from names. For our oracle, representing the best possible expected performance, having names as a crutch was reasonable. For a true gender prediction task, however, including named entities felt like cheating. We will discuss decisions related to named entities further in the unigram baseline section.

4.3 Zero Rule Baseline

The oracle served as our upper performance bound. We established a lower performance bound via a simple, "always choose majority class" baseline. As already mentioned, male samples comprised approximately 75% of the data set. Therefore, we imagined a hypothetical classifier that predicted a class label of male for every sample in the development data set. Once again, here is the corresponding confusion matrix.

We computed per-class F1 scores as well as a

	Pred. Male	Pred. Female	Total
Male	94,472	0	94,472
Female	31,503	0	31,503
	125,975	0	

Table 2: Confusion matrix illustrating precision and recall for zero rule baseline. Pred. stands for Predicted.

macro-averaged F1 score. The hypothetical classifier that predicted male for every sample yielded an undefined F1 score for samples about females, which we will consider 0.0, and an F1 score of 0.85 for samples about males. The macro-averaged F1 score for this hypothetical classifier was 0.425.

4.4 Unigram Baseline

For a stronger baseline, we ran a logistic regression classifier that used the 5,000 most frequent unigrams as features, after cleaning punctuation. This logistic regression classifier, trained with the top 5,000 unigrams on the training data set, and evaluated on the development data set, yielded 69% precision and 27% recall for female pronouns, and 80% precision and 96% recall for male pronouns. There was an F1 score of 0.38 for samples about females and 0.87 for samples about males. The macro-averaged F1 score was 0.625.

	Pred. Male	Pred. Female	Total
Male	90,693	3,779	94,472
Female	22,997	8,506	31,503
	113,690	12,285	

Table 3: Confusion matrix illustrating precision and recall on more sophisticated baseline with logistic regression classifier and 5,000 most frequent unigrams. Pred. stands for Predicted.

Table 3 presents a confusion matrix of the true and false positives of a logistic regression classifier trained on the 5,000 most frequent unigrams and tested on the development set. In summary, a macro-averaged F1 score of 0.743, from our oracle, is an estimated upper bound. An F1 score of 0.425, from our zero rule baseline, is our estimated lower bound. Finally, an F1 score of 0.625, from our 5,000 unigram baseline, is a more realistic lower bound.

4.5 Named Entities

An early iteration of development of our logistic regression classifier with every unigram in the data set revealed that the most predictive unigram features were named entities, which do not yield any interesting insights about how men and women are written about differently in news. 35 of the top 50 most negative weighted features (i.e. highly predictive of female pronouns) were names like Sharapova and Ginsburg, for example. Likewise 28 of the top 50 most positive weighted features (i.e. highly predictive of male pronouns) were also names.

When the majority of informative features were names, we missed out on other, more interesting features. Ultimately, the goal of the present research was not only to build a successful classifier, but also to discover informative features that reveal gender bias patterns (or lack thereof) in news articles. Therefore, we removed named entities from our data set to the best of our capabilities and focused the remainder of our experiments on other stylistic features, beyond names.

4.6 Feature Extraction

4.6.1 Text Preprocessing

All feature sets were collected after the text was stripped of all punctuation and lowercased. Text was tokenized using `nltk`'s built-in `word_tokenize` method.

4.6.2 Feature Sets

We used syntactic (part of speech), lexical, and lemma features with unigram and bigram combinations in an attempt to capture the differences in stylistic choices that journalists make when writing about men versus women.

We collected all 115,933 unigrams and all 3,146,374 bigrams in the training set and used them independently and together as separate feature sets. In addition, we extracted the 5,000 most frequent unigrams and 10,000 most frequent bigrams, and also used those independently and together as feature sets. We also collected all part of speech (PoS) unigrams and used those as feature sets alone and with our previously collected set of all unigrams. In order to collect PoS unigrams, we used the `pos_tag` method from `nltk` in order to tag each token in a sample with respect to its part of speech. Then, we collected unigrams from the tagged samples' parts of speech rather than the original tokens themselves. Lastly, we collected all lemma unigrams. In order to collect lemma unigrams, we used the `WordNetLemmatizer()` method from `nltk` to convert each token in a sample to its lemma prior to collecting unigrams and bigrams from the sample.

We created a feature set based on word lists adapted from LIWC (Pennebaker et al., 2001) and sociolinguistic findings about language differences in describing men and women. The feature set was adapted from the lexicons that accompany the LIWC software. We made features for word list categories like positive emotion, money, and family. For example, one LIWC word list is called "posemo", standing for positive emotion, and it contains word stems like "happy" and "enjoy." Another word list is "money", which contains word stems like "bank" and "dollar." Another is a list of all prepositions. We included all 64 standard LIWC word lists as features.

We also collected words from word categories identified in sociolinguistic research as informative about language used to describe men and women. We curated lists of words for communal and agentic language based on Madera et al. 2007, grindstone language and standout adjectives based on Schmader et al. 2007, and the most informative words discussed in Smith et al. for describing male and female leaders. This provided an additional nine features.

We turned lists of these word stems into regular expressions, and checked if any word in the sample sentence matched a word in a word list. If it was found in the sample, we gave it a binary feature named after the word list.

4.7 Matrix Re-weighting: Term Frequency Inverse Document Frequency

After collecting these feature sets, we reweighted our original term-document matrices by converting them to term frequency-inverse document frequency (tf-idf) matrices. In this case, "term" is a general purpose term that could apply to unigrams, bigrams, lemma unigrams and bigrams, or PoS unigrams, and "document" refers to a single sentence of news text. The purpose of applying this reweighting is to better capture the informativeness of a single term within a single sample as well as across the whole data set. We applied tf-idf reweighting to all feature sets.

4.8 Text Classification

We implemented a binary logistic regression model to classify samples as being about a man or a woman. We also utilized a stochastic gradient descent classifier with a hinge loss function that served as a scalable substitute for a linear SVM in light of the large size of our data set. Both of these models were instantiated using the Python `sklearn` package. All hyperparameters for all models were kept standard to the defaults given by the package except for class_weight, to which we passed a dictionary {`male:0.25, female: 0.75`} as a means of compensating for the vast imbalance between samples about males and samples about females.

5 Results

Table 4 summarizes the performance of our different feature sets and classifiers, given by the per-class and macro-average F1 score for the logistic regression and stochastic gradient descent classifiers. One of the most interesting observations about the F1 scores yielded by these different models is that they are extremely similar to each other. With a minimum F1 score of 0.46 and a maximum F1 score of 0.495, it is clear that our choice of feature sets did not differ very much in how informative they were to accurate classification of samples as being about men or women.

The highest average F1 score came from a stochastic gradient model with a hinge loss function trained on part of speech unigrams. This was an unexpected result, but may indicate that our other feature sets caused our models to overfit to the idiosyncrasies of the training set. Furthermore, these results indicate, in line with the previous literature as well as our primary hypothesis that there are syntactic differences in the way men and women are written about in the media, without accounting for the semantics of the words chosen. Furthermore, the fact that the SGD model performed the best replicates the findings of previous work that had similar success with a linear SVM, for which our SGD model wad a scalable analog.

By contrast, the lowest average F1 score was derived from an SGD model trained on every unigram and bigram in the training set. We suspect that this result occurred due to massive overfitting to the training set that yielded poor generalizability. The fact that each sample was composed

Feature Set	Logistic Regression			SGD		
	Female F1	Male F1	Avg F1	Female F1	Male F1	Avg F1
5,000 most frequent unigrams	**0.31**	0.66	0.485	0.32	0.64	0.48
5,000 most frequent unigrams and 10,000 most frequent bigrams	**0.31**	0.67	0.49	0.32	0.63	0.475
Training set unigrams	0.30	0.68	0.49	0.32	0.63	0.475
Training set unigrams and bigrams	0.29	**0.69**	0.49	**0.33**	0.59	0.46
PoS unigrams	**0.31**	0.66	0.485	0.27	**0.72**	**0.495**
PoS unigrams and word unigrams	0.30	0.68	0.49	**0.33**	0.61	0.47
Lemma unigrams	0.30	0.68	0.49	0.32	0.63	0.475
Category word lists	0.29	**0.69**	0.49	**0.33**	0.62	0.475
Category word lists and word unigrams	0.30	0.68	0.49	0.32	0.62	0.47

Table 4: Per-class and macro-averaged final F1 scores from an unseen test set.

of so few tokens also means that including bigrams would increase the risk of overfitting without providing any meaningful new information to the model.

Table 4 also includes the per-class F1 scores for each model and feature set. In general, while we were able to achieve significant gains over the zero rule baseline F1 score for female samples, we were unable to achieve any gains at all over the unigram baseline for either male or female samples. The model that yielded the highest male F1 score was the SGD model trained on PoS unigrams, which we did not anticipate. By contrast, the models yielding the highest female F1 scores were SGD models trained on unigrams and bigrams, PoS unigrams and word unigrams, and category word lists. These results suggest that female F1 scores benefited most from an augmentation of the information provided to the model considering how many fewer female samples were in the data set compared to male samples. In addition, the success of the category word lists lends computational credence to the findings from sociolinguistic research that qualitatively classified the types of words authors chose when writing about women.

6 Discussion

6.1 Feature Analysis

As discussed previously, by removing named entities ourselves, we outlined a fairly difficult classification task. The purpose of the current research was not only to build a classifier that can classify the pronoun gender of sentence from the news, but also to investigate whether there are stylistic differences in text written about men and women that a classifier can detect. Therefore, in addition to considering our classifier accuracy, it is imperative to examine the most predictive features to a particular classification for a given model. We examine three different models that were trained here.

For the stochastic gradient descent model trained on the 5,000 most frequent unigrams in the training data set, the most predictive unigrams for sentences about either gender included gendered terms like “woman” and “guy.” However, the second most predictive unigram for female sentences was “husband”, and other predictive unigrams included “server,” “secretary,” and “children”. In addition, the logistic regression model trained on category word lists weighted the list about family as the most predictive category for female sentences, while the list about work was the most predictive category for male sentences.

These features suggest that women are written about in the news with reference to low-status careers, as well as having families. For men, predictive unigrams included terms from athletics like “Yankees” and high-status careers like “quarterback” and “businessman.” It is difficult to conclude to what degree these findings simply reflect an imbalanced world or if they actually serve to reinforce imbalance in the world.

The model that yielded the highest average F1 score was the SGD model trained on PoS unigrams. For women, the most predictive part of speech was “UH,” or an interjection. By contrast, for men, the most predictive part of speech was a noun phrase. It is difficult to know exactly why, but our hypothesis is that this suggests that sentences about women may more filler words, whereas sentences about men seek to characterize them in terms of the people, places, and things they interact with or are similar to.

Ultimately, an investigation into the most highly weighted features to each category provided specific information about how men and women are portrayed in media. Thus, we see that while our classifiers did not achieve high F1 scores, they did provide salient insights into the question about how journalistic bias falls along lines of gender. This is promising for a planned application of these findings onto a data set of performance evaluations.

6.2 Limitations of the Current Research

There are a number of limitations in our current work. One of these issues arose from the tool we to remove named entities from the data set. We utilized a publicly available tool called Deidentify, which performed poorly on a large data set. It also had poor recall despite using Stanford's NER system, which we speculate was due to its use of an outdated version of the package. These were issues we noticed too late in the timeline to devise our own solution. In the future, we hope to make use of the most updated version of Stanfords NER classifiers that are integrated with `nltk` in order to build a de-identifying tool that is more robust to large data sets.

Additionally, a consideration for the present work is that a single sample constituted only one sentence from a news article. Although we made an effort to eliminate very short and very long sentences, the fact remains that a single sentence does not capture the gender-based differences that arise across paragraphs and across whole articles. Furthermore, the small size of each sample predisposed a model to overfitting if n-grams of more than size 1 were included as feature sets, which limits the extent to which this work can comment on how language use differs when talking about men versus women. Intuitively, it is rather difficult to infer a conclusion about gender based on a single de-identified sentence, and our results indicated this despite our best efforts. Future research in this area should likely consider the use of paragraphs as samples in order to capture more of the long-range stylistic differences between text about men and text about women.

News is also a different domain than the previous work in performance evaluations we were referencing. We might have been less prepared to create features for a news corpus than a performance evaluation corpus as we had originally planned.

6.3 Future Directions

As previously discussed, we hope to apply these findings to a new data set of performance evaluations from a major technology company. Currently, no classifier exists for predicting the gender of the subject of a workplace evaluation.

Studying bias in performance evaluations in the workplace (versus previous research in academia) is critical because performance reviews inform management decisions. At each stage of a career-recruitment, retention, and promotionwomen in STEM fields are disadvantaged compared to men. By studying how women are discussed in their reviews, we can gain insights into why women are less likely to be retained or promoted. Computational language analysis of these reviews is a critical next step to a robust understanding of unconscious bias in technology companies.

7 Conclusion

Computational investigations into the relationship between language use and gender have largely been limited to research into the stylistic differences made by authors of different genders. The current research sought to examine a different aspect of this relationship by introducing the task of gender classification of the subject of text as a means of probing the stylistic differences made by journalists when writing about different genders. We created a new data set of news text that eliminated unfairly predictive named entities and gendered pronouns. Unfortunately, we were only able to achieve modest gains over our zero rule baseline. However, in examining the most predictive features to our linear classifiers, we discovered that there are lexical and stylistic differences in the ways men and women are written about in media.

8 Notes

We will submit a version of the All the News Kaggle data set that is stripped of gendered pronouns and named entities to Kaggle. Due to issues with named entity extraction, we will look into more accurate methods of NER before making it public.

Additionally, our code for this project is public, though have left out LIWC word lists due to the protection of those lists under license.

Acknowledgments

We want to thank Professor Chris Potts for his clear, consistent feedback and support of this project.

References

Apoorv Agarwal, Jiehan Zheng, Shruti Kamath, Sriramkumar Balasubramanian, and Shirin Ann Dey. 2015. Key female characters in film have more to talk about besides men: Automating the bechdel test. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

Na Cheng, R. Chandramouli, and K.p. Subbalakshmi. 2011. Author gender identification from text. *Digital Investigation*, 8(1):7888.

Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and Linguistic Computing*, 17(4):401–412.

Juan Madera, Michelle (Mikki) Hebl, and Randi Martin. 2007. Gender and letters of recommendation: Agentic and communal differences. *PsycEXTRA Dataset*.

Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents*, SMUC '11, pages 37–44, New York, NY, USA. ACM.

James Pennebaker, M E. Francis, and R J. Booth. 2001. Linguistic inquiry and word count (liwc): Liwc2001. 71.

T Schmader, J Whitehead, and V H Wysocki. 2007. A linguistic comparison of letters of recommendation for male and female chemistry and biochemistry job applicants. *Sex Roles*, 57(7-8):509514.

David G Smith, Judith E Rosenstein, Margaret C Nikolov, and Darby A. Chaney. The power of language: Gender, status, and agency in performance evaluations. *Sex Roles*, page 113.

Raj Sobhan, Shlomo Hota, Rebecca Argamon, and Chung . 2016. Gender in shakespeare: Automatic stylistics gender character classification using syntactic, lexical and lemma features.

Andrew Thompson. 2017. All the News. `https://www.kaggle.com/snapcrack/all-the-news`.