He Said, She Said: Comparing Gender Classification in Various Contexts

Kaylee Bement
Computer Science
Stanford University
kbement@stanford.edu

Henry Lin
Symbolic Systems
Stanford University
henryln1@stanford.edu

Abstract

In recent years, there has been a growing interest in understanding what nuances, if any, distinguish male and female speakers in a variety of settings. The majority of previous works, however, have been concentrated within a single dataset or genre, and there has been limited work in the area of a general classifier that can be applied to all forms of natural language, ranging from blog posts to movie lines to everyday conversations. For our work, we explore the use of several different common classifiers, including Naive Bayes, Logistic Regression, and Stochastic Gradient Descent, and their performance across the Cornell Movies dataset (Danescu-Niculescu-Mizil and Lee, 2011), Switchboard dataset (Boulis and Ostendorf, 2005), a Twitter dataset (Crowdflower, 2016), and a blog post dataset (Schler et al., 2006). We find that there is no single model that performs better across all datasets, and there is a significant decrease in performance when using multiple datasets in training and testing. Further, we find interesting sociolinguistic implications from our highest feature weights for each gender, including findings that contradict those of previous works.

1 Introduction

One of the most persistent problems plaguing society today is the continued existence of gender biases. These biases are present in just about every aspect of life, including but not limited to unequal salaries and the lack of representation of women in STEM-related fields. Understanding the origins of these biases and the differences between natural language from men and women may provide valuable insight into what steps should be taken in order to combat these biases, which is what many previous papers have explored and what this paper continues to build upon.

As the tools and available data for Natural Language Processing continue to grow, there has been a considerable amount of interest in studying gender biases through words, such as which adjectives are commonly used to describe each gender (Fast et al., 2016; Madaan et al., 2017). The majority of this work has been concentrated towards very specified domains that have little overlap with each other, and in doing this, high levels of accuracy have been achieved in these areas. Here, we extend this work to cross domain classification using the idea of transfer learning from neural networks (Mou et al., 2016). Given the specificity of the models previously trained, they are unable to generalize to other situations, which motivates us to look for methods of creating a general classifier. We approach this in a variety of ways: by comparing the features and performance of gender classification on separate datasets, by using one dataset in order to train our classifiers to predict the gender of lines spoken in another dataset, and by mixing datasets in order to train and test our classifiers.

2 Related Work

In recent years, the emergence of easily scrapeable domains has led to an advent of easily accessible natural language data. Examples of this include the datasets used in (Fast et al., 2016), (Madaan et al., 2017), and (Crowdflower, 2016), where the majority of the data came from information available on the Internet. This has led to an explosion of gender classification analysis, and here we discuss several of the most relevant works to our project.

While the Cornell Movie-Dialogs Corpus is one of the most popular datasets for analysis, being used in (Danescu-Niculescu-Mizil and Lee, 2011) and (Agrawal et al., 2016), there has been work done in other areas of cinema. In (Madaan et al., 2017), the researchers looked at Wikipedia summaries and movie scripts of movies in the Bollywood industry, and found several areas of considerable difference between men and women. They concluded that men tended to have a more prominent role in movies coupled with success and power while women tended to be pushed to the sidelines and marginalized across the 4,000 movies examined. This reinforces the hypothesis that gender biases do exist and are perhaps even magnified when portrayed on the screen.

Returning to the Cornell Dataset, (Schofield and Mehr, 2016) did a thorough analysis of male versus female classification and ultimately achieved an accuracy of 73.6%. In addition to exploring both a Naive Bayes and Logistic Regression approach, they also augmented their training with sentiment labels from VADER and information about arousal, valence, and dominance for a list of pre-labeled English words. One of the main findings was the prevalence of curse words for male lines, confirmed also in (Agrawal et al., 2016) and (Danescu-Niculescu-Mizil and Lee, 2011).

Outside the cinema industry, researchers have also taken advantage of the cornucopia of natural language online, generated daily by social media users and online writers (Fast et al., 2016) (Burger et al., 2011). One of the most interesting works in this area comes from an analysis of amateur fiction writers done in (Fast et al., 2016) that elucidate several intriguing trends in the works by male and female authors. The main goals of this analysis were to discover the portrayal of men and women in stories, the effect of stereotypes on story success, and how each gender wrote male and female characters in their stories. Ultimately, they did discover there to be a significant association of different adjectives with each gender. For example, their analysis found that men are more likely to be described as *strong, sexual, and arrogant* while women are associated with adjectives such as *weak, submissive, and childish*. Apart from this, they also found that no matter the gender of the author, there was a consistent trend of gender stereotypes in the stories, indicating that even when given the liberty of writing anything, people will subconsciously conform to traditional portrayals of gender roles.

On the topic of the liberty to write anything, the advent of social media in the past decade has introduced a signficiant amount of promising text information for many different types of Natural Language Processing tasks. One of the most popular social media platforms is Twitter, where users can informally type short thoughts or status updates, called *tweets*, and post them online. This form of text grants similar creative freedoms as those granted to the authors in (Fast et al., 2016), but are a representation of oneself rather than of a character. In (Burger et al., 2011), the researchers tackle the problem of binary classification of gender, much like (Agrawal et al., 2016), using an assortment of methods such as Naive Bayes, Support Vector Machines (SVM), and Balanced Winnow, a technique similar to Stochastic Gradient Descent, and ultimately reach a peak accuracy of 76%, which surpasses the performance of most humans performing the same task.

One of the most important aspects of our task is the application of a single trained model across multiple datasets, an idea drawn from transfer learning. Transfer learning is a widely popular idea in the field of deep learning, particularly Computer Vision, and is used to reduce the redundancy in training models from scratch by instead using pretrained weights. In (Mou et al., 2016), the researchers explored the effects of transfer learning and its applications to the task of binary classification and sentence pair classification. The paper ultimately finds that between two semantically different tasks, transfer learning is not successful, leading to little improvement, if any. For semantically similar tasks however, such as training a sentiment classifier on one dataset and then using that trained classifier on another dataset, they did find improvements. While the paper does not examine models other than neural networks, the same concept applies to any model that uses weights and features, such as Naive Bayes and Stochastic Gradient Descent.

3 Datasets

After conducting a review of previous work done in the field, we decided to use the four datasets described below for our problem. We concluded that in order to best develop a general gender classifier, the datasets chosen must be diverse enough to satisfy the large number of areas where natural language can be used for, while keeping in mind the most significant forms of communication.

3.1 Cornell Movie-Dialogs Corpus

The Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011) is a widely popular dataset that has seen an extensive amount of analysis due to the high amount of metadata information provided (Agrawal et al., 2016) (Schofield and Mehr, 2016). The corpus itself draws from 617 movie scripts from all over the cinema industry. While the dataset boasts over 304,000 lines among over 9,000 characters, we were only able to make use of a subset of the data due to the incomplete labeling. Out of the 9,035 characters, there are only 3,774 of them with a labeled gender, which led to a total of 66,228 female lines and 154,133 male lines. However, as in (Schofield and Mehr, 2016), we decided to use an equal number of male and female lines from each movie in order to remove sample size biases from the classifiers, which led to 19,165 lines from each gender.

We decided to use this dataset because there has been a significant amount of work regarding the biases of male/female portrayal in cinema (Madaan et al., 2017), and it would prove interesting to see if a model that has learned these biases would perform well on other areas of dialogue.

3.2 Switchboard Dataset

To capture the nuances of real life conversation, we turned to the 1992 Switchboard Dataset (Godfrey and Holliman, 2016), which features approximately 2,400 telephone conversations with speakers from the United States. The speakers did not know each other prior to their conversations and were given a topic to discuss from a set of 70 topics. The dataset included not just the words spoken by individuals but also linguistic notations, so we preprocessed the dataset by removing these notations to make sure they were not accidentally included as features during training. The dataset is slightly imbalanced in that there were 302 male speakers and 241 female ones, but after analyzing the dataset, we found that there were actually significantly more female lines spoken, 130,729 to 92,887. However, as in the movie corpus, we used an equal number of lines per gender from each topic of conversation, which led to 21,352 lines from each gender.

3.3 Twitter Dataset

One of the difficult aspects of language is the rate at which it evolves, especially amongst the younger population, and the advent of social media has certainly magnified this effect. As a result, we turned to Twitter as an area of natural language that captures the conversations and speech patterns of the younger generation today, which may differ drastically from that of 20-25 years ago, when the Switchboard Dataset was released. The dataset we found with the appropriate labeling and metadata (Crowdflower, 2016) is much smaller compared to the other datasets we are relying upon and thus, it is much more viable as a testing set rather than a source for training a model. Amongst the tweets labeled with a guaranteed gender, there are 4,653 male and 5,367 female tweets. Since these values are relatively close and the dataset is already small, we did not alter the number of tweets used per gender.

3.4 Blog Posts Dataset

The final dataset we used to help train and test our models was a blog posts dataset published on Kaggle and previously used in analysis for (Schler et al., 2006). This was by far the largest dataset with 681,284 unique blog posts coming from 392 male and 200 female authors. As with the movie and Switchboard datasets, we used an equal number of posts per gender from each topic, which led to 280,879 posts per gender. However, to more closely match the size of the other datasets, we usually analyzed a subset of 20,000 posts per gender. Given that the previous datasets chosen involved data that would be relatively short (e.g. single movie lines, short tweets), this dataset allows us to examine the ability to classify natural language given a more expansive set of writing.

4 Models and Methodology

4.1 Tasks

We conducted the following gender classification tasks with our datasets:

- Using one dataset to both train and test our gender classifiers.
- Using one dataset to train our classifiers, and another to test.
- Using a mix of datasets to both train and test our gender classifiers.

4.2 Models and Setup

For all individual dataset and mixed dataset tasks, we used 75% of the data for training, and the remaining 25% for testing. For the cross-dataset tasks, where one dataset is used to train and the other to test, we used 100% of each dataset for their respective role.

We begin with single datasets using each classifier to measure a baseline performance of how well the dataset is able to classify a randomly selected test set of itself, and we use this to compare cross-dataset performances. For our cross-dataset performances, the performances there are directly compared to see how well each feature weight set generalizes to a completely different domain.

For the majority of our models, we used the scikit-learn library, a machine learning library in Python with many classification models already implemented for easy use (Pedregosa et al., 2011). From this library, we utilized their Naive Bayes, Logisitic Regression, and Stochastic Gradient Descent models. These were chosen both because almost all of the above papers we reviewed (Agrawal et al., 2016; Fast et al., 2016; Burger et al., 2011; Boulis and Ostendorf, 2005; Schofield and Mehr, 2016) utilized these models, and they are all linear models, which enables us to better contextualize the feature weights in order to analyze the nuances of gendered speech.

4.3 Features

As said in the previous subsection, we wanted to keep our features easily interpretable, thus we used variations and combinations of two core features. First, we used two matrices which contained word counts for each line, one for unigrams and one for bigrams. Then, we used two approaches to TF-IDF, both at the word level and the n-gram level, where we included both bigrams and trigrams. In addition to training each model on each feature separately, we also trained each model on both count matrices as a feature, and a union of both count matrices and TF-IDF feature vectors.

5 Results

5.1 Metrics

In order to analyze the performance of each classifier, we used the classifier's precision, recall, and f-1 score for each gender, and then used the average of the scores for each gender to measure the overall performance of the classifier. For a given gender, *precision* is the percentage of correct predictions (true positives) of that gender out of all text predicted to have that gender (true positives and false positives), *recall* is the percentage of correct predictions of that gender out of all text that should have been predicted that gender (true positives and false negatives), and the *f1-score* is the weighted average of precision and recall. The three equations are:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

Method	Precision	Recall	F-1
Naive Bayes	0.60	0.59	0.59
Regression	0.59	0.59	0.59
SGD	0.80	0.53	0.60

Table 1: Switchboard Dialogues Performance

Method	Precision	Recall	F-1
Naive Bayes	0.61	0.60	0.60
Regression	0.60	0.59	0.60
SGD	0.87	0.52	0.62

Table 2: Cornell Movie-Dialogues Performance

5.2 Individual Dataset Performance

5.2.1 Telephone Conversations, Movies, Twitter

For the first three datasets we examined individually, we saw a very similar performance across each of our metrics. This mirrors the very similar average datapoint length of each of these datasets, which can be seen in Figure 1, suggesting that this is directly related to how well a model does in predicting gender. As shown in Tables 1, 2, and 3, the

Method	Precision	Recall	F-1
Naive Bayes	0.62	0.61	0.62
Regression	0.63	0.61	0.62
SGD	0.60	0.60	0.60

Table 3: Twitter Performance

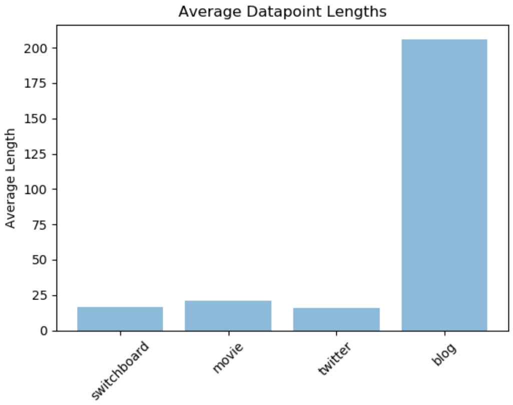

Figure 1: Average Lengths for each Dataset

Method	Precision	Recall	F-1
Naive Bayes	0.70	0.68	0.69
Regression	0.71	0.71	0.71
SGD	0.72	0.70	0.70

Table 4: Blog Posts Performance

F-1 score seems to reach a cap at around 0.60, indicating a noticeable improvement over the baseline accuracy but still with a ways to go. We also see that there is no single best classifier, with SGD edging out in Tables 1 and 2 and Naive Bayes and Linear Regression being comparable in Table 3.

5.2.2 Blogs

There is 0.10 jump in F-1 score across the board when training and testing on the blog post dataset, indicating that the size of this dataset and its average length of lines as seen in Figure 1 may have led to a better performance. However, as you can see in Table 4, there is still no classifier that is significantly better than the others. Looking at all the datasets, it is clear that the specific classifier used does not seem to make a drastic impact on the overall performance, and it is more important to have a good, robust dataset on which to train and test.

5.3 Cross-Datasets Performance

We next moved toward our goal of a general classifier by using two datasets at a time, one for training and the other for testing. We found that in all cases, the use of feature weights trained on a different dataset did not improve performance, and the majority of times, the performance was lower, which can be seen in Table 5, where each row is a training dataset and each column is a testing dataset. We did not train on the Twitter dataset for any test sets except itself due to its small size. From this table, we see that the combinations using 75% training and 25% testing from the same dataset consistently perform better than all cross-dataset trials in terms of the top F-1 score, with notable results in both the Twitter and Blogs columns; training on blog posts performs almost as well as training on tweets when testing on tweets, which is most likely due to the common style of online natural language throughout contexts, while testing on blogs does not perform well training under any other dataset, mostly likely due to the drastic difference in average datapoint length between the blog post dataset and all other datasets, as seen in Figure 1.

	Movie	Phone	Twitter	Blogs
Movie	0.62	0.55	0.56	0.56
Phone	0.54	0.60	0.56	0.59
Twitter	n/a	n/a	0.62	n/a
Blogs	0.60	0.55	0.60	0.71

Table 5: Top F-1 Score From Each Dataset Combination[1]

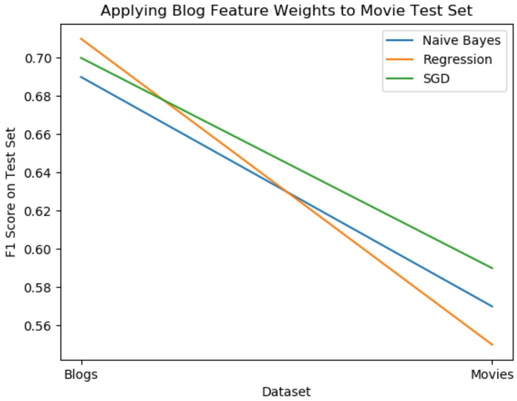

Figure 2: Blog Features Performance on Blog Test Set vs. Movie Test Set

Further, Figure 2 gives us a deeper look at how drastically performance declines when faced with an unfamiliar dataset. Using features and weights trained on the blog post dataset, there is about a 0.10 to 0.13 decrease in F-1 score when comparing the results on the blogging test set versus the movie test set. This is most likely the result of a large number of features that are important to the classification of blogs being useless when classi-

Method	Precision	Recall	F-1
Naive Bayes	0.55	0.54	0.54
Regression	0.58	0.54	0.55
SGD	0.58	0.54	0.55

Table 6: Mixed Datasets Performance

Word	Movie	Phone	Twitter	Blogs
he	1.19	1.06	1.05	1.05
me	1.13	0.97	1.40	1.13
you	1.05	0.96	1.14	1.07
people	1.05	1.06	1.04	1.02
home	0.99	1.05	1.03	0.98
husband	0.92	1.02	1	0.98
children	0.97	1.08	1	1.04
hate	0.97	0.97	1.19	1.03
fucking	0.93	n/a	1.1	1.04
hell	0.94	n/a	1.14	1.03

Table 7: Female : Male Word Frequencies

fying the movie set, possibly due to a different vocabulary set and differing contexts for common words. This implies that there aren't consistent enough features across datasets with drastically different contexts, such as movie scripts vs blog posts, to contribute to a general classifier, which we will discuss in more detail in sections 6.1 and 6.2.

5.4 Mixed Datasets Performance

The final combination of datasets we experimented on was a mixture of the four datasets we discussed earlier. Applying the same ratio of 75 to 25 train/test split, we shuffled all the datasets together and randomly split the resulting superset into training and testing. We ran the same analysis as before, and we found that this cross-domain feature learning led to decreased performance, which can be seen in Table 6. This decrease in performance likely resulted from the wide variance in usage of words across each dataset, and the confusion caused by shuffling them together led to much weaker feature weights that did not contribute to the classification as meaningfully. We discuss this further in the feature analysis section.

6 Discussion

6.1 Feature Analysis

For each single dataset task, we extracted the 1000 highest weighted features for each gender. Then, in order to compare the usage of each word between the genders more clearly, we calculated the ratio *female weight : male weight* for each word; in this analysis, we focus on the unigram count feature, meaning that this ratio is the *female word frequency : male word frequency*. In this section we analyze notable trends in these ratios among the datasets, especially in comparison with one another, and suggest sociolingual implications from these trends. All of the trends we discuss can be seen in Table 7.

First, we will discuss the first four rows of Table 7, which contain the words he, me, you, and people. In almost all datasets for all four words, the frequency for women is higher than it is for men, which implies that women generally talk about people more often than men do; the values are notably high for women talking about themselves on Twitter (me) and about men in movies (he). The former value might make up for the fact that in more restrictive conversations, such as in the phone conversations with strangers, women tend to talk less about themselves (me) and about the person they are talking to (you), which may be out of politeness conventions and social expectations of women to be more polite than men. The latter value most likely correlates with the fact that many movies unfortunately fail the Bechdel test [2].

Further, the next three rows focus on family words - home, husband, and children. We found it interesting that men appear to speak about all three topics more than women in movies, despite (Agrawal et al., 2016) stating that movies tend to show language through a more stereotypical lense, and previous works finding that women talk about these topics more often (Agrawal et al., 2016). Additionally, although women speak more about these topics with strangers on the phone and on Twitter, men speak about the home and husbands more often on blogs; the latter might either imply that men discuss wanting to be better husbands on their blogs, or a larger presence of the gay community in the blogging community. Interestingly, children being discussed in blog posts is the only

[2] The Bechdel Test is a commonly known test for movies which states that a movie should have at least two women in it who talk to each other about something besides a man.

per-dataset outlier in this section; despite women speaking less about the home and husbands, they do tend to talk about children slightly more, which aligns with previous findings of women talking about the family (Agrawal et al., 2016).

Finally, the last three rows of the table contain three strong or expressive words, hate, fucking, and hell. Although prior works have found that women curse less than men (Boulis and Ostendorf, 2005; Schofield and Mehr, 2016), we have found that this does not seem to be the case; this only seems to hold true in the case of movies, which (Agrawal et al., 2016) has noted tend to show a more stereotypical, extreme portrayal of gendered language. Of course, the Switchboard corpus did not contain curse words, as people tend not to curse the first time they meet someone. Meanwhile, on Twitter and in blogs, women consistently curse more and use strong words more often, such as hate. Similarly, in the movie and Switchboard context, women generally spoke less than men (most of the frequency ratios were less than 1) and consequently, with a smaller variety of words, while in the Twitter and blog context, women spoke about the same amount as and with similar variety to men. This may show a divide between restrictive contexts, movie scripts and phone conversations with strangers, where women are expected to act a certain way, and expressive contexts, Twitter and blogs, where women are less restricted by societal expectations. Alternatively, the expressive context could be seen as an online context; we would need to analyze an expressive, offline dataset to draw an accurate conclusion.

6.2 Further Error Analysis

Building off the feature analysis, we can now understand why a decrease in performance is plausible given the word frequency variance across the datasets. When the datasets are trained and tested separately, their individual nuances allow for some success in classification, but when brought together, these nuances may become mixed and distorted, resulting in unhelpful weights.

This problem only becomes amplified and compounded due to the varying sizes of each dataset, ranging from about 10,000 tweets to 600,000 blog posts. This leads to an uneven scaling of words from different contexts, and it may have had a drastic negative impact on performance.

In addition to the classifiers mentioned above, we also tested a LSTM Recurrent Neural Network model using pretrained GloVe vectors (Pennington et al., 2014) to measure the potential of deep learning for this task. We found this approach to be too computationally intense without any improvement over our more simple classifiers. A training period of 5 epochs over 12 hours led to about a 0.55 F-1 score on the individual dataset tasks, leading us to abandon this approach. We suspect that while GloVe vectors are an excellent starting point for NLP tasks such as question answering, it falters when dealing with different domains where words are contextualized very differently.

7 Conclusion and Future Work

Looking at the different combinations of datasets and models, we found that there was no single classifier out of the three (Naive Bayes, Regression, and SGD) that was overall better suited for binary classification in this task. Furthermore, we found that applying learned feature weights from one dataset to a new, foreign one did not lead to any improvement. In some cases, the performance was comparable, with a few points of difference, but in other cases, such as when testing on the blog dataset, it led to a significant decrease. In conclusion, we find that it is extremely difficult to build a general classifier that performs well on all areas of natural language (social media, blogging, scripts, etc.) due to the nuances in which words are used in each domain.

There are several areas of interest that we would like to explore in future work. When running our experiments, we also took advantage of pretrained GloVe vectors (Pennington et al., 2014) to try an experimental Recurrent Neural Network model, but we quickly abandoned this approach due to computational cost and lack of improvement compared to other classifiers. It would be interesting to try more complicated deep learning architectures such as bidirectional RNNs, GRUs, and multilayer networks to see if there is possible improvement. However, with this approach, there would be need for bigger datasets due to the data hungry nature of deep learning models.

Lastly, there are many more datasets that are worth exploring that we did not touch in this project. One potential dataset worth generating and exploring is a corpus of authentic conversations between friends and family. The Switchboard conversations' drawback is that all transcripts are between strangers, leading possibly to a deviation from actual real life conversations and natural gendered speech patterns. It would be enlightening to compare and contrast these two datasets and see how language changes given the two situations.

8 Code

All code used for this project can be found on our Github repository.

References

Monica Agrawal, Priya Ganesan, and Catherine Wong. 2016. You talkin to me? gender classification of unseen conversational partners.

Constantinos Boulis and Mari Ostendorf. 2005. A quantitative analysis of lexical differences between genders in telephone conversations. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 435–442. Association for Computational Linguistics.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.

Crowdflower. 2016. Twitter user gender classification.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Ethan Fast, Tina Vachovsky, and Michael S Bernstein. 2016. Shirtless and dangerous: Quantifying linguistic signals of gender bias in an online fiction writing community. In *ICWSM*, pages 112–120.

John Godfrey and Edward Holliman. 2016. Switchboard-1 release 2 ldc97s62.

Nishtha Madaan, Sameep Mehta, Taneea S Agrawaal, Vrinda Malhotra, Aditi Aggarwal, and Mayank Saxena. 2017. Analyzing gender stereotyping in bollywood movies. *arXiv preprint arXiv:1710.04117*.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in nlp applications? *arXiv preprint arXiv:1603.06111*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

J Schler, M Koppel, S Argamon, and J Pennebaker. 2006. Effects of age and gender on blogging in proceedings of 2006 aaai spring symposium on computational approaches for analyzing weblogs.

Alexandra Schofield and Leo Mehr. 2016. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39.