Tweet Like a Girl: A Corpus Analysis of Gendered Language in Social Media

Margaret Ott

Advisor: Bob Frank

Submitted to the faculty of the Department of Linguistics
in partial fulfillment of the requirements for the degree of
Bachelor of Arts

Yale University
20 April 2016

Abstract

Thanks to social media platforms such as Twitter, there is a plethora of written language available and accessible online. Previous work has been done to use large sets of written language such as these to make interesting inferences. Sometimes it is used to extract information, such as sentiment analysis (Kouloumpis et al., 2011), and sometimes it is used to build classifiers, for categories such as gender and age group (Koppel et al., 2002). The goals of analyses such as these are usually utilitarian: companies want to extract the sentiments about their products from social media posts, or people want to be able to automatically and accurately predict the gender of an author of a piece of writing without being told.

This project is an investigation into the existence and nature of gendered language use on Twitter. The motivation is academic and not practical. Some other researchers interested in writing style have investigated gender stereotypes through an academic lens, but often using human readers for data collection rather than computational methods. The project formalizes previous ideas from both sociolinguists and the general population about gender and language and applies them to a relatively new language medium (Twitter). It uses computational techniques to test the validity of some of these previous answers to the question of how language is used differently by the genders.

The project made use of a large corpus of tweets, for which a binary classifier for gender was built, using the Naive Bayes algorithm to train the classifier using data marked for author gender. The features included in the classifier were words, parts of speech, n-grams of both words and parts of speech, as well as pairs of syntactically dependent words. The project also included a binomial logistic regression using lexical and semantic word classes of the LIWC.

These statistical models were able to classify tweets by author gender with some success. They showed that men and women do tweet in some statistically differing ways. Women's tweets are shorter and more often about their home lives, while men's tweets are longer and more focused on the outside world. There were also many ways in which the models failed to show evidence of some widely held beliefs about gender and language.

Acknowledgments

First and foremost, I would like to thank my advisor, Bob Frank, for his guidance throughout this project. Many thanks as well to Raffaella Zanuttini and my fellow linguistics seniors - Grace, Ethan, Alexa, Ralph, Kyle, Rose, and Alexandra - for all of their feedback and support during the process. Thanks to Kieran Snyder for inspiring the project through her work on gender in the workplace. Huge thanks to my family and friends for their support and encouragement. And to my girlfriend, Katharine Walls, for being by my side the whole time.

Contents

1 | Introduction: Gender as Difference

1.1 Motivation and Objectives

In the summer of 2015, Kieran Snyder published an article investigating the differences between the way women and men write resumes (Snyder, 2015). She analyzed 1,100 resumes and found what she called "stark" differences between the way women and men present themselves to potential employers. She wrote that women's resumes were "longer, but shorter on the details."

Evidence given included that the women's resumes averaged 80% more words than the men's, but men presented far more specific content. In addition, 91% of men included bulleted verb statements, while only 36% of women do. But, Snyder says, despite these stylistic differences, the women and men in her study had similar backgrounds, in terms of experience and credentials.

The article's closing implored hiring managers to "recognize different resume communication styles and the skill sets behind them." The conclusion was not that women and men must be inherently different, which is an argument that can easily be used against women's rights. Rather, her conclusion was that women should not be penalized for the way they use language, when it can be shown to be a pattern found within their gender.

Similarly, the motivation of this project is not to shed light upon the differences between women and men for their own sake. No claims will be made about how or why women and men are inherently different. The objective is to investigate how they use language differently when they present themselves in certain settings on the internet, namely, on Twitter.

Furthermore, the focus will be on language use, and not on the way women and men are talked *about* differently. There are interesting things to be said on that topic, but they will not be addressed here.

1.2 A Note on Gender and Identity

It must be noted that there are more than two genders. Not everyone can be categorized into “female” and “male,” and everyone who fits into either of those categories does not use language the same way as everyone else in that category. As explained later (see 3.1), this project only used data that included self-reported gender labels.

There are also countless other factors and identity categories that have been shown to influence language use in similar ways to gender. These include sexuality, region, and many others. This project is not meant to imply the unimportance of any other factor when considering varied language use across identity categories.

2 | Background: Gender and Language

2.1 The Sociolinguistics Approach

2.1.1 Origins: Women's Liberation and Male Dominance

The study of Gender and Language is a relatively young one. It began in the early 1970s and was arguably established by Robin Lakoff with her flagship book Language and Woman's Place, published in 1975. The book introduced ideas about gendered language (mostly women's language) that established foundational ideas from which further study would stem. She outlined specific tendencies she noticed that were unique or more prevalent in women's language, including hedges, apologies, tag questions, as well as vague notions such as "avoidance of coarse language" and "hyper correct language and punctuation."

But as is true of any groundbreaking work, some of her methods were questionable. She used mostly native speaker intuition, or "introspection," as her investigation tool. She was not able to present any statistical evidence of her claims, and attributed them to her own intuition. Of tag questions (*The way prices are rising is horrendous, isn't it?* (p. 49)), she wrote, "It is my impression, though I do not have precise statistical evidence, that this sort of tag question is much more apt to be used by women than men" (Lakoff, 1975, 49). In defense of this introspective method of investigation, she wrote towards the beginning of her book:

> "Any procedure is at some point introspective: the gatherer must analyze his data, after all. Then, one necessarily selects a subgroup of the population to work with: is the educated, white, middle class group that the writer of the book identifies with less worthy of study than any other? And finally, there is the purely pragmatic issue: random conversation must go on for quite some time, and the recorder must be exceedingly lucky anyway, in order to produce evidence of any particular hypothesis, for example, that there is sexism in language." (Lakoff, 1975, 4-5)

Here Lakoff defends her methodology of introspection (ironically, using a male pronoun to refer to the unknown researcher). There are criticisms to be made of the idea that all investigation is at some point introspective. It is always subjective, perhaps, at a certain point, but research can certainly be improved by empirical techniques she was not employing. Lakoff's use of introspection as methodology allows for no control of how much of her own prejudices entered into her findings.

Furthermore, she criticizes the potentially limited demographic reach of empirical studies. While this is definitely a downside to many research endeavors, it is certainly possible for someone to investigate the speech of white, middle class women as long as they don't generalize their findings to a wider group. Finally, she claimed that sexism is hard to find in a given random conversation. While sexism, or the way in which people of some gender are talked about, may not be present in a given stretch of language, this is less true of gendered tendencies in speakers, or the way in which people of some gender speak, which is what Lakoff was making claims about.

While there are many criticisms that have been made of Lakoff's book, she did succeed in challenging other researchers to investigate the issues and phenomena she wrote about, thereby beginning a period of important groundbreaking research in gender and language. In her own words, "But granting that this study does in itself represent the speech of only a small subpart of the community, it is still of use for indicating directions for further research in this area, ...a means of discovering what is universal in the data and what is not, and why" (Lakoff, 1975, 40). She put forth her ideas as suggestions for future investigation, and in doing so, inspired much of the research that has been done in the field to this day.

Lakoff's work framed the differences between the language of men and women around the concept of male dominance. She looked for and reported on ways in which women's language reflected their subordination to men. She saw language as a process that kept women "in their place" (hence the title of the book). They are taught to speak in "lady-like" ways as girls, and then those tendencies are what keep them in a demeaning position later in life, as people take them less seriously because of the way they talk.

Lakoff's work was followed closely by Fishman (1977) who famously discussed women engaging in what she called "interactional shitwork," which among other things involved using questions and hedges to force responses from men in order to facilitate conversation. This claim is based on the same idea of male dominance that Lakoff was using, but Fishman turned the focus onto the ways in which women must accommodate for their language not being taken seriously in male-dominated society.

This period of the study of Gender and Language was strongly informed by the women's liberation movement of the 1970s. In the same way the movement focused on the ways society disadvantaged women, the period of research focused on the way language also revealed women's disadvantaged position. Feminist-inspired studies investigated issues such as interruption (West & Zimmerman, 1983), verbosity, as well as questions and back channeling (Fishman, 1983). The methods of investigation in these studies were most frequently conversational analysis. Conversations between small groups of people were recorded and carefully annotated by humans looking for phenomena such as interruption. Researchers focused more on the *who* and *how* of communication, looking closely at spoken conversations in small groups, and less on what was said, public talk, or written texts.

2.1.2 The 1990s: Difference Displaces Dominance

By the 1990s, "dominance" had taken a back seat to the "(cultural) difference" approach to the investigation of gender an language, and the field became much less critical and much more objective. The underlying framework changed from being based upon male dominance to being inspired by cultural differences. Deborah Cameron wrote about this paradigm shift in the following way:

> "The gradual ascendancy of difference over dominance was almost inevitable given the ideology of twentieth-century linguistics, especially its anthropological and sociological variants. Difference, and not inequality, is what the framework of structural linguistics is designed to deal with." (Cameron, 1995, 35)

An example that illustrates this "difference" framework well is a book by Deborah Tannen. She published a New York Times bestseller written for the public entitled *You Just Don't Understand: Women and Men in Conversation* (Tannen, 1990). In it, she explains that boys and girls spend their childhoods learning different ways of communicating, which she calls "genderlects." She claims that women learn "rapport-talk," language that is meant to promote social affiliation and emotional connection, while men learn "report-talk," language that is meant to convey information with no emotional implications. She says that these differences can lead to misunderstandings later in life, and there were even self-help books written based on this notion intended to help couples discover the cause of their misunderstandings.

2.1.3 And Now: Discourse

Most recently, Gender and Language research has become focused on an approach known as Critical Discourse Analysis (CDA), which views language as a form of social practice. The prevailing theory of gendered language is that gender is formed and reproduced through discourse, and it interacts with speakers' other identity categories to produce their language. There is no explicit methodology required for CDA, as long as it provides insight into the way discourse reproduces or resists societal power and inequality.

Research is often concerned with "Communities of Practice," or groups of people "who come together around natural engagement in an endeavor" (Eckert & McConnell-

Ginet, 1992, 464). CDA concerns itself with questions of how societal power relations are established and reinforced through language. These studies still tend to involve close, qualitative analysis of a small number of short texts, taking into account the nature of the creation, distribution, and reception of those texts.

Even with the advancement of technology and tools for analyzing language data, corpus analyses in the field of Gender and Language within linguistics are rare. There have been a few over the past decade, and they usually focus on investigating a particular feature of language within a certain context, such as the presence of emotional talk in men's speech about their experiences of illness (Charteris-Black & Seale, 2009). This project brings some of the conceptual aspects of Gender and Language research together with the power of corpus analysis to show to what extent gendered language exists and persists online.

2.2 The Text Classification Approach

As sentiment analysis and text classification have become hot topics in modern NLP, some computer scientists have taken up the task of building accurate text classifiers of corpora by gender. In some cases, they are able to be relatively successful, reporting accuracies as high as 80% (Koppel et al., 2002). Burger et al. (2011) reported over 90% accuracy on the task of classifying tweets by author gender.

In cases like these, where the goal is simply an accurate classifier, the features that are taken into account are mostly ones that can be easily identified and counted en masse by computers. In Koppel et al. (2002), the only features used by the classifier were function words and n-grams of parts of speech. From only that information and what they call "a variant of the Exponential Gradient algorithm," they were able to classify texts marked for author gender in the British National Corpus with about 80% accuracy.

In the case of Burger et al. (2011), it seems as if their method was to simply use as many features as they could get their hands on or create to build huge classifiers. They write in the "Experiments" section of their paper that there were "feature vectors for some experiments requiring over 20 gigabytes of storage" (Burger et al., 2011, 1303). They used character n-grams (up to n=5) as well as word n-grams (up to n=2), for each tweets's username, full name, description, and tweet text. Not only did they use a huge number of features (numbering in the millions), but in the cases when they reached 90% accuracy, they were taking into account tweet metadata including the user's screen name, full name, and description. So, for example, if a user's description included the word *mother*, or if their full name included the character trigram *Joh*, the classifier was able to take that into account.

It should also be noted that in the case where they achieved over 90% accuracy, it was for determining the gender of a user given the text of all of their tweets and their associated metadata, and they had an average of 18-24 tweets per user. Given only the text of one tweet, their classifier was 67.8% accurate on their development set and 66.5% accurate on their test set. Their abstract begins by explaining that automated demographic prediction is valuable for marketing, personalization, and legal investigation. For these purposes, it appears they have been successful, since they report that their classifier outperforms even humans.

2.3 The Current Approach

This project bridges the gap between NLP and sociolinguistics in a way it hasn't been before, by using statistical and machine learning tools to carefully investigate patterns of language use online. Learning algorithms common in the modern field of statistics and NLP text classification are used to identify where aspects of language are strongly gendered in their use in the context of Twitter.

3 | Data and Analysis

3.1 The Corpus

3.1.1 Origin of the Dataset

The explosion of social media and online writing over the recent years has led to a huge increase in user-generated text accessible publicly on the Internet. This is exemplified notably on Twitter, where users are constantly publishing short strings of characters for the world to see. It is estimated that Twitter users produced about 500 million tweets per day (Kriorian, 2016). Twitter even allows this never-ending stream of language to be easily accessed and used using their API, where anyone who knows a little about programming can download many public tweets, along with their accompanying metadata (username, "favorite" counts, etc).

Twitter profiles, however, do not have a field where users can report gender, so gender metadata is harder to come by. This project began with a corpus originally collected by Burger et al. (2011) of about 1 million tweets, with gender labels that were extracted from the Twitter users' self-reported gender on Facebook, Myspace, or other blogging profiles linked to their Twitter profiles.

The investigators in Burger et al. (2011) did a small-scale quality assurance exercise in which they took a random sampling of 1000 twitter users from their corpus and manually examined their description fields. They determined that all of the users with gender cues in their description (i.e. *mother of 3 boys* or *just a dude*) were consistent with the gender they reported in their linked blogging profiles. They also purport that by limiting their dataset to users with blogs or other social media linked in their profile, they largely eliminated spam users, which are commonplace on Twitter.

3.1.2 Manipulation and Use

Twitter's data sharing policy restricts the sharing of actual tweets, so Twitter datasets are in the form of pairs of tweet ids and user ids. The current work makes use of the set of gender-labeled tweets collected by Burger et al. (2011) and distributed by Svitlana Volkova (http://www.cs.jhu.edu/ svitlana/), who used them as the basis of work reported in (Volkova et al., 2013).

The goal of Volkova et al. (2013) was to use gendered subjective language use to incorporate gender into sentiment analysis models. They found, for example, that *perfecting* was used with positive polarity for women and negative polarity for men, while *dogfighting* was used with negative polarity for women and positive polarity for men.

The dataset of tweetids, userids, and gender labels was used in conjunction with Twitter's API to gather tweet text and associate it with a gender label. Due to a combination of time, API rate limit, and memory constraints, this project's dataset came from the first 100,000 datapoints in the Volkova dataset. After having to eliminate some due mostly to a change in a Twitter user's preferences (their profile having become private since Burger et al. (2011) collected their data), the final dataset was 90,008 tweets, with more female-authored than male-authored (which was also true in the original Burger et al. (2011) dataset). 52,851 of the tweets were female-authored, while 37,157 were male-authored.

From there, the tweets were part-of-speech tagged using the tagger described in Gimpel et al. (2011) and Owoputi et al. (2012). The parts of speech used by this parser are not in the traditional Penn Treebank style, but are rather one-letter labels that are more relevant to what occurs in tweets. There are categories for URLs (U), for emoticons

(E), and for hashtags (#). As is reported in Owoputi et al. (2012), the tagger has been shown to be very reliable.

3.2 Text Classification and Features

3.2.1 The Learning Method

In order to take advantage of modern computational speed and accuracy as well as a large dataset, this project began with the use of text classification tools to investigate quantifiable differences in gendered language. It should be noted, though, that the motivation is not to build a better-than-state-of-the-art classifier by author gender. The classifiers were used as a window into the most strongly gendered aspects of language on Twitter.

Preliminary tests were performed using both a Naive Bayes classifier, as well as a MaxEnt classifier, using NLTK implementations (Bird et al., 2009). Their accuracies were comparable, and they tended to agree on which features were most informative of author gender. The results below were obtained using a Naive Bayes classifier. In each case, the classifier was trained on 90% of the (around 90,000) tweets and tested on the remaining 10%. A Naive Bayes classifier can be defined as a function that assigns a class label $y = C_k$ as defined below, drawn from a finite number (K) of classes (in this case only F or M) to a problem instance (in this case a tweet), represented as a feature vector $\mathbf{x} = (x_1 ... x_n)$, using conditional probabilities and "naive" conditional independence assumptions.

$$y = \arg\max_{k \in \{1,...,K\}} p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

3.2.2 Word Features

When trained on simply the presence of individual tokens, the Naive Bayes classifier was 64.7% accurate on the test set. One could also call this category "unigrams of words." In this case, all words that were present in all tweets were considered (for a total of 118,313 tokens), and all the characters were made lowercase first. The following features were the most informative:

Female Token	**F:M Probability**	**Male Token**	**M:F Probability**
giveaway	24.1 : 1.0	#news	20.3 : 1.0
etsy	17.6 : 1.0	inflation	14.6 : 1.0
hubby	16.7 : 1.0	sci	13.7 : 1.0
baking	11.5 : 1.0	muslims	12.8 : 1.0
hubs	8.8 : 1.0	afghanistan	11.8 : 1.0
knitting	8.6 : 1.0	forces	11.8 : 1.0
makeup	8.3 : 1.0	gospel	11.8 : 1.0
edward	8.2 : 1.0	world's	11.1 : 1.0
jewelry	8.0 : 1.0	stocks	10.9 : 1.0

A classifier that was trained entirely on the presence of bigrams of tokens (i.e. without the use of unigrams) was 62.2% accurate, and the following features were the most informative:

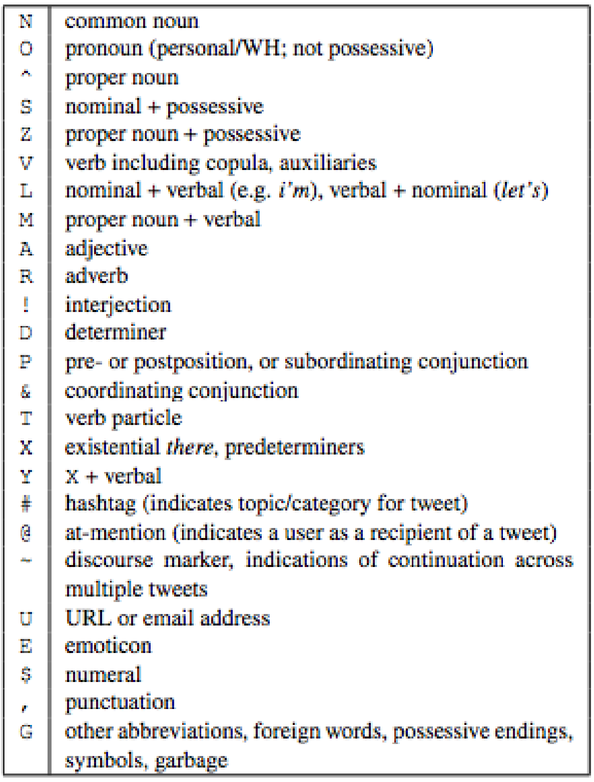
Female Bigram	**F:M Probability**	**Male Bigram**	**M:F Probability**
my cat	12.0 : 1.0	bbc :	32.6 : 1.0
giveaway !	17.6 : 1.0	:d xxx	14.6 : 1.0
love ya	9.6 : 1.0	! cheers	13.7 : 1.0
us ...	9.2 : 1.0	protect your	12.8 : 1.0
gossip girl	8.7 : 1.0	the world's	11.8 : 1.0
too funny	8.7 : 1.0	inflation is	11.8 : 1.0
!! so	8.7 : 1.0	mail -	11.8 : 1.0
hee hee	8.6 : 1.0	of gold	11.6 : 1.0
like '	8.2 : 1.0	songs .	10.9 : 1.0
feel about	8.2 : 1.0	app store	10.9 : 1.0
wink *	8.2 : 1.0	on u	10.9 : 1.0

3.2.3 Part-Of-Speech Features

Similarly to the way the token unigrams and bigrams were isolated respectively in order to find detailed results about the most informative features in each category, classifiers were also trained on only part-of-speech n-grams. The following results are for a classifier trained on the 500 most common part-of-speech bigrams, and showed 60.7% accuracy. Neither unigrams and trigrams produced as accurate a classifier as bigrams.

Female Bigram	F:M Probability	Male Bigram	M:F Probability
E L	3.7 : 1.0	# ˆ	5.3 : 1.0
L S	3.5 : 1.0	G Z	3.3 : 1.0
O ˜	3.2 : 1.0	L U	3.3 : 1.0
L E	3.1 : 1.0	T X	3.1 : 1.0
! Z	3.1 : 1.0	S #	3.1 : 1.0

Below (Figure 3.1) is the part-of-speech tagset being used, from Owoputi et al. (2012):



N	common noun
O	pronoun (personal/WH; not possessive)
^	proper noun
S	nominal + possessive
Z	proper noun + possessive
V	verb including copula, auxiliaries
L	nominal + verbal (e.g. *i'm*), verbal + nominal (*let's*)
M	proper noun + verbal
A	adjective
R	adverb
!	interjection
D	determiner
P	pre- or postposition, or subordinating conjunction
&	coordinating conjunction
T	verb particle
X	existential *there*, predeterminers
Y	X + verbal
#	hashtag (indicates topic/category for tweet)
@	at-mention (indicates a user as a recipient of a tweet)
~	discourse marker, indications of continuation across multiple tweets
U	URL or email address
E	emoticon
$	numeral
,	punctuation
G	other abbreviations, foreign words, possessive endings, symbols, garbage

Figure 3.1: Owoputi et al. POS Tags

3.2.4 Dependency Features

The tweets were parsed using a dependency parser designed for tweets described in Kong et al. (2014). The parser, known as "TweeboParser," returns CONLL style dependency trees of each tweet, given a set of tweets. The parser does not consistently provide dependency relation labels, but pairs of dependent words were successfully incorporated into the classifier without regard for what particular relation they were in. The most informative features for a classifier trained only on dependency pairs (with 59.7% accuracy) are listed below. In these pairs, the second word is dependent upon the first.

Female Pair	**F:M Probability**	**Male Pair**	**M:F Probability**
husband, my	35.6 : 1.0	is, inflation	11.8 : 1.0
cat, my	15.3 : 1.0	making, money	9.9 : 1.0
am, tired	10.1 : 1.0	of, duty	9.9 : 1.0
funny, too	9.2 : 1.0	involved, in	9.9 : 1.0
love, ya	8.7 : 1.0	for, users	9.0 : 1.0
will, too	7.8 : 1.0	holla, at	9.0 : 1.0
recipe, the	7.8 : 1.0	was, playing	9.0 : 1.0
entered, contest	7.8 : 1.0	why, or	9.0 : 1.0

3.2.5 Performance of Classifiers

Below is a chart with metrics of the classifiers' performance when considering various combinations of features. The classifier reached its maximum accuracy when considering unigrams and bigrams of words, bigrams of parts-of-speech, and dependency pairs. It consistently performed better on female tweets than on male tweets.

Features included	**Accuracy**	**Female F-score**	**Male F-score**
Baseline (all F)	58%	NA	NA
Word bigrams and POS bigrams	63.8%	72.8	46.0
Word unigrams & bigrams and POS bigrams	65.6%	72.8	46.9
Word 1 & 2grams and POS 2 & 3grams	65.3%	74.3	46.7
Word 1 & 2grams and POS 2grams & Deps	66.2%	74.3	50.6

3.3 KL Divergence of Verbs

In addition to the text classification analyses, KL (Kullback Leibler) divergence was used as a statistical indicator of differences between the language of men and women on Twitter. KL divergence is a calculation of the difference between two probability distributions. It is often used to compare an ideal distribution of data P to another distribution Q - be it a theory, model, or approximation of the first distribution, P. For discrete probability distributions P and Q, the symmetric KL divergence of Q from P is defined to be

$$D_{KL}(P||Q) = \sum_i P(i)\log(\frac{P_i}{Q_i}) + \sum_i Q(i)\log(\frac{Q_i}{P_i})$$

In this case, the KL divergence was between two distributions P and Q - one representing a distribution for women and one for men, of all the words that appeared in dependency pairs in the dataset. Probability distributions for all verbs were calculated for both women and men, after having been stemmed by NLTK's Porter Stemmer. The distributions were the probabilities assigned to words that were dependent upon the verbs in question. So, for example, the KL divergence of the word *love* would be a calculation over the probability distributions of all words that were dependent upon *love* for both women and men. It would be roughly a calculation of how differently women and men use the word *love*, with respect to the words they use in dependency relations with it. The implementation of KL divergence used here employed add-λ smoothing for missing counts, where $\lambda = 0.1$.

The verbs that were found to have the largest KL divergences were *got, be, are, make, get, had, see, is, know, go, been, made, have, am, do, can, need, should,*, and *love.* In many cases, upon closer inspection of the probability distributions, the words used most commonly as dependent upon these verbs were similar, but ordered differently and with different probabilities, so specific conclusions couldn't easily be drawn about the way these verbs were used differently. For example, for *got*, the verb with the highest KL divergence, the highest values in the probability distributions for women and men are below:

Word	**P (Women, got)**
home	0.062
back	0.041
done	0.025
email	0.020
like	0.015
phone	0.014
time	0.013
so	0.010
lot	0.010
lost	0.008

Word	**P (Men, got)**
back	0.058
home	0.030
done	0.013
like	0.013
text	0.011
phone	0.009
game	0.007
so	0.007
shit	0.006
email	0.006

3.4 Binomial Logistic Regression

The features that the Naive Bayes classifier showed were most informative appeared to fall into some broad trends and categories, as *my cat* and *my husband* were consistently strong indicators of female tweets, while *bbc* and *#news* were consistently strong for male tweets. The next step was a Binomial Logistic Regression using word categorization by LIWC (Linguistics Inquiry and Word Count) (Tausczik & Pennebaker, 2010), using SPSS software. This was performed in order to ascertain the statistical effects of certain word categorization on the likelihood that a tweet was authored my someone of a certain gender. The regression was run on 24,199 tweets, limited to one per user, to eliminate unequal effect sizes from different users.

The LIWC provides over 80 classes of words. For each tweet, or data point, a vector was calculated that contained 92 values. For the values associated with word classes, the value was the percentage of the words in the tweet that fell into a given LIWC word class. There were some other variables, such as word count, that were not representations of word classes, and were therefore not percentages.

A binomial logistic regression was performed, with the tweet's author's gender as the dependent variable (F = 1.0, M = 0.0), and 92 independent variables. The final model categorized 74.3% of the female-authored tweets correctly and 43.0% of the male tweets. Omnibus Tests of the model coefficients tested whether the explained variance of data was significantly greater than the unexplained variance. They revealed that the logistic regression model was statistically significant ($\chi^2(93) = 1199.003, p < .001,$) and explained 6.5% of the variance (Nagelkerke R^2). While it is clear the model did not fit the data perfectly, many significant variables were revealed.

Below are the variables that showed statistical significance and a positive coefficient (meaning they were indicators of female authorship). Also listed are their coefficients, standard errors, Wald test values, degrees of freedom, and p values.

Some of the word classes listed are self explanatory, but here is a short summary of the less obvious variables. *Dic* was a percentage of the words that were in LIWC dictionary. The significance of this variable for women could explain why the model performed so much better at classifying women's tweets than men's. *Function* is function words, and *verb* is verbs. *Female* includes female references that are pronouns (*she*, *her*) and nouns (*women*, *girl*). *Bio* is "biological processes." *Netspeak* includes abbreviations like *btw*, *lol*, and *thnx*. *Nonflu* is nonfluencies like *uh* and *rr**.

Class	**Coefficient**	**S.E.**	**Wald**	**df**	**Sig.**
Dic	0.006	0.002	10.065	1	0.002
function	0.011	0.004	7.311	1	0.007
verb	0.006	0.003	4.315	1	0.038
family	0.021	0.008	6.668	1	0.010
female	0.033	0.008	17.882	1	0.000
bio	0.023	0.008	9.117	1	0.003
home	0.035	0.006	30.415	1	0.000
netspeak	0.022	0.007	10.037	1	0.000
nonflu	0.036	0.009	17.599	1	0.000

And below, there are variables that showed statistical significance and a negative coefficient (meaning they were indicators of male authorship). *Intercept* was simply the intercept of the logistic model that was built, not an independent variable.

Class	**Coefficient**	**S.E.**	**Wald**	**df**	**Sig.**
Intercept	-0.643	.148	18.839	1	0.000
WPS	-0.026	0.003	60.989	1	0.000
article	-0.015	0.005	10.709	1	0.001
auxverb	-0.010	0.004	5.312	1	0.021
anger	-0.016	0.008	4.675	1	0.031
male	-0.024	0.007	11.029	1	0.001
cause	-0.026	0.007	12.668	1	0.000
tentat	-0.018	0.005	11.173	1	0.001
health	-0.018	0.009	3.966	1	0.046
sexual	-0.053	0.012	18.922	1	0.000
drives	-0.014	0.006	5.620	1	0.018
work	-0.010	0.003	9.291	1	0.002
religion	-0/019	0.007	7.478	1	0.006
death	-0.023	0.010	5.287	1	0.021

WPS was words per segment (or words per tweet). *Cause* is a category of causation words, like *because*, *effect*, and *hence*. *Tentat* are tentative words, like *maybe*, *perhaps*, and *guess*. *Drives* is an overarching dimension that captures affiliation (references to others), achievement, power, reward focus, and risk focus.

4 | Conclusion

4.1 Summary of Findings

4.1.1 So, How Do Women and Men Tweet Differently?

The work done here shows that overall, women and men do show some notable differences when it comes to the way they tweet. Women are generally more likely to tweet about their home lives and personal concerns. According to the Naive Bayes model, the tokens *hubby* and *hubs* both appeared in the top 5 most indicative features for women, while no such token indicating a female spouse appeared even in the top 40 most indicative features for men. Women were 12 times more likely to include the bigram *my cat* (and 15.3 times if it was a dependency pair).

These generalizations were confirmed by the logistic model, which showed the *home* and *family* word classes to be statistically significant predictors of female authorship. Women's tweets are shorter, and they are more likely to use abbreviations and nonfluencies. Women appear to be more likely to use Twitter to send out short, abbreviated updates on their day-to-day home and family life.

Men, on the other hand, tend to tweet more about the "outside" world, through topics like the news and technology. According to the Naive Bayes model, men were 20 times more likely to include *#news* in their tweets, 32.6 times more likely to use the bigram *bbc :*, and 10.9 times more likely to mention the *app store*. The logistic regression showed that *work* was a statistically significant predictor of male authorship. The model also showed some other stereotypically male concerns born out in the data, including some that fit well into a “male dominance” framework of language use. Men were more likely to tweet sexual words, to tweet about anger, and to demonstrate the *drives* metric, which includes affiliation, achievement, power, reward focus, and risk focus. At the same time, though the model showed that men used more *tentative* words, which is a metric one would expect to be associated with women in such a framework.

Overall, male’s tweets are longer, and they use more articles and auxiliary verbs. Men appear to be more likely to use Twitter to send out more polished messages about the world outside of their home, whether it is work or current events.

4.1.2 How Don’t They?

Most of the variables in the logistic regression were not statistically significant predictors of author gender of tweets. Some of those are worth noting here. No indicators of questions were predictors of male or female authorship, including interrogatives and question marks. There were many different word classes representing different emotions, including positive emotion, negative emotion, sadness, and anxiety, none of which proved to be significant. Anger was the only significant emotion variable, and it was a predictor of male authorship. The word class of swear words was not only insignificant, but had one of the closest coefficients to 0 of any variable.

While the models described here were able to classify tweets into classes of female and male authorship better than chance, they weren’t performing outstandingly well. The ways in which women and men tweet differently are mostly related to what they tweet about. In addition to these semantic differences, their styles seem to differ as well, with men tweeting closer to complete sentences and women abbreviating more.

There is not much evidence here that supports either the “male dominance” or “rapport-report” frameworks. While the *drives* word class was a statistically significant indicator of male authorship, as an overarching category, the *achieve* and *power* subcategories were not significant. As mentioned above, *tentative* words were in fact significant variables for the male side. Some of Lakoff's specific claims, including that women avoid "coarse language" and use "super-polite forms" were not found to be true in this case.

Furthermore, there was not much evidence to support the "rapport-report" dichotomy. Women were not found to express their emotions more on Twitter. One could argue that men's tendency to talk about current events qualifies as "report talk," but there is no evidence to suggest that they do not tweet about their emotions (in fact, they tweet about their anger more than women do). Additionally, women tweeting about their cats should not be categorized as less of a "report" than men tweeting about their jobs. Overall, Twitter does not seem to be a place where previous ideas about gender and language play out, really.

4.1.3 Limitations

First, it should be noted that Twitter is a linguistically unique platform, since a tweet is limited to 140 characters. It is a platform on which posts are very short, but also very frequent. It is common practice to send frequent (obviously short) thoughts and updates to Twitter, in a way that is not usually found on social media platforms in the style of Facebook, and certainly not on blogs. This creates a unique, but linguistically interesting, environment. Tweets tend to be brief updates on often mundane aspects of one's everyday life, visceral reactions to current events, or similarly short bursts of thought users have.

These findings cannot be generalized to make conclusions about men and women in general, or even how they may use language. Like other social media, its language is not usually a conversation, but is often language in isolation. Twitter is a platform on which people send their ideas to no one in particular, to the ether. It is a place where people often get things "off their chest," so to speak. It is even possible that perhaps Twitter is also the place where people go to use language in ways they feel they can't in the real world, if only subconsciously - where women are free to be sloppy with their words and men are free to be tentative. That is not what is being addressed here, but is just to say that this project is not a claim about gendered language in the world at large. As Lakoff so carefully pointed out (and is paraphrased here), generalizing past a data set is a dangerous game.

4.2 Future Work

Further work on this topic could try to formalize more of the claims of earlier gender and language researchers, like interrupting, back-channeling, and tag questions, that were not directly addressed here. Further investigation could be done to test the idea that men's tweets are more polished and come closer to complete sentences. This kind of work could also easily be extended to other kinds of social media and other kinds of language.

Bibliography

Bird, Steven, Ewan Klein & Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Media.

Burger, John D., John Henderson, George Kim & Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing* EMNLP '11, 1301–1309. Stroudsburg, PA, USA: Association for Computational Linguistics.

Cameron, Deborah. 1995. *Verbal hygiene.* London: Routledge.

Charteris-Black, Jonathan & Clive Seale. 2009. Men and emotion talk: Evidence from the experience of illness. *Gender and Language* 3(1). 81–113.

Eckert, Penelope & Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual Review of Anthropology* 21. 461–490.

Fishman, Pamela. 1977. Interactional shitwork. *Heresies* 1(2). 99–101.

Fishman, Pamela. 1983. Interaction: The work women do. In B. Thorne, C. Kramarae & N. Henley (eds.), *Language, gender, and society*, Rowley, MA: Newbury House.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan & Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies: Short papers - volume 2* HLT '11, 42–47. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kong, Lingpeng, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer & Noah A Smith. 2014. A dependency parser for tweets .

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

Kouloumpis, Efthymios, Theresa Wilson & Johanna D. Moore. 2011. *Twitter sentiment analysis: The good the bad and the omg!* 538–541. AAAI Press.

Kriorian, Raffi. 2016. http://www.internetlivestats.com/twitter-statistics/.

Lakoff, Robin. 1975. *Language and woman's place.* New York: Harper & Row.

Owoputi, Olutobi, Chris Dyer, Kevin Gimpel & Nathan Schneider. 2012. Part-of-speech tagging for twitter: Word clusters and other advances. Tech. rep.

Snyder, Kieran. 2015. The resume gap: Are different gender styles contributing to tech's dismal diversity? http://fortune.com/2015/03/26/the-resume-gap-women-tell-stories-men-stick-to-facts-and-get-the-advantage/.

Tannen, Deborah. 1990. *You just don't understand: Women and men in conversation.* New York: William Morrow & Co.

Tausczik, Yla R. & James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods http://homepage.psy.utexas.edu/homepage/students/Tausczik/Yla/index.html.

Volkova, Svitlana, Theresa Wilson & David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 conference on empirical methods in natural language processing (emnlp)*, .

West, Candace & Don H. Zimmerman. 1983. Small insults: A study of interruptions in cross-sex conversations between unacquainted persons. In B. Thorne, C. Kramarae & N. Henley (eds.), *Language, gender, and society*, 102–117. Rowley, MA: Newbury House.