2-2021

When Misclassification Is Misgendering: Gender Prediction in the Context of Trans Identities

Sean Miller
The Graduate Center, City University of New York

When Misclassification Is Misgendering:

Gender Prediction in the Context of Trans Identities

by

Sean Miller

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2021

When Misclassification Is Misgendering:

Gender Prediction in the Context of Trans Identities

by

Sean Miller

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

_____	_____
Date	Kyle Gorman
	Thesis Advisor
_____	_____
Date	Gita Martohardjono
	Executive Officer

The City University of New York

ABSTRACT

WHEN MISCLASSIFICATION IS MISGENDERING:

GENDER PREDICTION IN THE CONTEXT OF TRANS IDENTITIES

by

SEAN MILLER

Advisor: Kyle Gorman

As a subdomain of author profiling, gender prediction (sometimes called gender inference) has received a substantial amount of attention—both as a task in itself, and for other downstream analyses. Throughout the existing literature various statistical and machine learning methods have been applied to extract features in order to either characterize and differentiate female and male writing styles, or simply to achieve maximum accuracy on gender prediction as a binary classification task. However, researchers often do not disclose how they conceptualize gender nor do they consider the implications that gender prediction has for non-binary and trans individuals. Along with an overview of the previous research, I apply pre-existing, well known statistical and machine learning methods to data from trans individuals in order to extract linguistic features and characterize their writing styles. I find that several of the features pattern with features found in previous research, but are in contradiction with the gender-marked writing styles they have been shown to characterize—suggesting that trans individuals are likely to be misclassified by standard state-of-the-art methods of gender prediction. Misclassification in gender prediction is indistinguishable from misgendering, and therefore has great capacity for harm to individuals of trans experience.

ACKNOWLEDGEMENTS

Thank you to my colleagues in the CUNY Linguistics program who sat with me, supported me, and worked through it all with me. I am also deeply grateful to Jason Kandybowicz, Sam Al Khatib, and Juliette Blevins for their knowledge, generosity and encouragement—I would have taken every course offering if I could have. Thank you to Thomas, Mikey, James and so many other friends and family for an abundance of moral support. Finally, I'd like to express my immense appreciation to my brilliant advisor, Kyle Gorman, an excellent instructor and unfailing resource who guided me in completing this work.

Contents

List of Tables

List of Figures

1 Introduction

Automated gender prediction has received substantial attention from the Natural Language Processing (NLP) community for decades, but what are its implications for trans individuals? Much of the existing literature focuses on gender prediction as a binary classification task and shows no consideration for non-heteronormative gender identities. In an effort to expand on gender prediction research, this study seeks to examine the manner and extent to which trans users on Reddit make use of language that is marked with their gender identity.

For the purposes of this investigation, *trans* is intended to encapsulate users who identify as transgender, transsexual, or any other related transmasculine or transfeminine identities present within the data collected from two Reddit subcommunities, which I describe in greater detail in Section 3. It is important to note that some limitations are already apparent in this approach. First, trans identities, just like cisgender identities (as noted by Bamman, Eisenstein, and Schnoebelen 2014 and Nguyen et al. 2014), are highly nuanced and such reductions could be an over-simplification failing to account for other influential and intersectional factors, e.g., sexuality, race, nationality/culture, family upbringing, and the like. Additionally, these Reddit users represent only a subset of the trans community, which also consists of identities that reject the gender binary entirely. In any case, it should not be misconstrued that this is a comprehensive NLP investigation of trans identities.

This research is largely inspired by the suggestions for ethical frameworks regarding gender as a variable in NLP from Larson (2017) and examines gender through the performative view. Under this conception individuals exhibit characteristics and behaviors which align to their gender identity that both contribute to and draw from social constructs. Automatic gender prediction for trans individuals presents a potentially harmful impact on these populations, who are subject to misgendering, pathologizing ideology, transphobia, erasure, and other forms of oppression based solely on their gender identity. The main point of this work is to examine whether the previous research that treats gender prediction as a binary

classification task can be recreated in this context and how the gender-marked features identified in previous research align with the features in the target dataset of this study.

2 Related Work

In this section I will begin with a brief overview of two earlier works related to gender and marked language. This is followed with a discussion of the evolving understanding of gender, including analyses from two authors whose work focuses primarily on gender, sex, and sexuality. Finally, I will look at more recent work specific to using social media data to predict gender.

2.1 Early Work

Early work from Robin Lakoff in 1973 provides an introspective account of the ways in which men and women use language differently. In this highly influential work for studying language in the context of gender Lakoff proposes a "Women's language" underpinned by the notion that the marginalization of women is reflected in their behavior and use of language. This "language" denies women the means to express themselves strongly and trivializes the subject-matter about which they do speak. Lakoff specifically points to examples such as the level of specificity in choosing color terms (e.g. "mauve" vs. "purple"), choice of swear-words and certain adjectives (e.g., "terrific" vs. "divine"). Lakoff also makes the claim that tag questions and rising intonation in declarative statements (as in yes/no questions) belong to women's language. Additionally, while in some cases it is socially acceptable for women to speak in a way that is more masculine, for a man to unsarcastically stray into "women's language" would cause him to be subjected to ridicule and/or questioning of his masculinity.

In a more data-driven approach, Argamon et al. (2003) used texts from the British National Corpus to identify features that are indicative of female or male authors. They use 604 documents in total from various genres of both fiction and non-fiction. Care was taken to make sure the corpus was balanced with documents evenly from 123 male authors and 123 female authors, but it is not made explicit how gender was either

determined or indicated in the corpus. The hand-crafted distributional features consist of 467 function words, and parts-of-speech n-grams including 500 of the most common trigrams, 100 of the most common bigrams, and all 76 singular tags.

The features were assigned weights representing their associations with either male of female gender using the exponentiated gradient algorithm (Kivinen and Warmuth 1995). At convergence, only 50 features were indicated as being useful for distinguishing the male-authored from the female-authored texts. For the males, certain "noun specifiers" such as determiners and quantifiers were highly weighted, as well as the part of speech tags for two types of determiners and cardinal numbers. On the other hand, pronouns were marked as indicative of female authorship. Meanwhile, the authors found no major differences in the frequencies of nominals between male and female authored documents. Given these features, the authors provide an analysis that the work is consistent with studies of epistolary writing from the 17th and 20th centuries (Biber, Conrad and Reppen 1998, Palander-Collin 1999), where a difference was found on the "involvement-informational" dimension (Biber 1995). According to their analyses, men talk more about objects or classes of things, and the features linked to male authorship are also features found more prevalently in non-fiction writing—features which were identified in the previous work as "informational". In contrast, female authors use features that were identified as "involved" such as pronouns which suggest that females write more about relationships. Other features found to be consistent with the "involved" dimension included analytic negation (primarily *not* statements), contractions and present-tense verbs. With their consistent findings over millions of words, the authors recognize that there is more work to be done to understand how writers develop a personal writing style that can be somehow reflected in a set of given linguistic features, how that is affected by genre differences, and how they can be recognizable as belonging to a speech community.

While detailed in their analyses, neither Lakoff nor Argamon make explicit their conceptions of gender, nor do they acknowledge that gender is a moving target (Larson 2017). For instance, Fausto-Sterling (2012) points out how the ideas of European masculine dress in the mid to late 1600's would be considered rather feminine by today's standards—hats heavily laden with ostrich feathers, frilled breeches and bibs, even

rouge-adorned cheeks. Twenge (1997) found that studies using the Bem Sex Role Inventory (BSRI, Bem 1974) showed steadily increasing masculinity scores for women over a period of 15 years, correlating with steadily decreasing masculinity scores for men in the same time frame. With these indications that gender expression changes over time, it seems likely that the linguistic features associated with gender would change as well.

In addition to broader changes in gender expression, phenomena of language change over the lifespan have also been observed in previous work. Many of these studies have focused on phonological/phonetic changes, such as the Queen's pronunciation of vowels in Christmas broadcasts (Harrington 2006) or the shift from apical /r/ to dorsal [ʀ] among Québécois French speakers (Sankoff and Blondeau 2007, Sankoff 2019). However Sankoff (2019) also observed evidence of morphosyntactic changes in Québécois French speaking adults, viz. a shift from using the inflected future (e.g. *demandera*) to the periphrastic future (*va demander*). In all of these studies, changes in the speakers' greater communities were considered motivators for various trajectories of language change. Another study (Pennebaker and Stone 2003) found word choice changes in correlation with increasing age, such as using fewer self-references, using more positive and fewer negative affect words, and using more future-tense and fewer past-tense verbs. So it is clear, individuals do in fact change in their language norms over time. Certainly there are such dynamics within the trans population, perhaps even on a magnified scale, influenced by the process of gender transitioning, the rejection of sex-assigned-at-birth norms, and changes in one's community related to coming out as trans. So it's possible that where an individual is in their transition has an impact on the language they use. This dimension is not within the scope of this work—the data collected (see Section 3) does not include information on where each individual is in the timeline of their transition, but it is worth consideration as a potential covariate in the experiments that follow in Sections 4 through 6, and a possible area for further investigation in future work.

2.2 What Is Gender?

There is extensive literature on gender prediction, the various ways to go about it, what methods achieve the best accuracy, and even how it can be useful for downstream tasks, but there is little consideration for one key thing of which there also happens to be a vast amount of literature—what is gender, exactly? Of course, as Larson (2017) points out, it should not be expected that every researcher provide an extensive account of their understanding of gender, but at very least, it would behoove them to state their general conception. Are gender and biological sex one in the same? Is gender socially constructed or is it intrinsic to nature? Or is it somehow both? Is it binary? Yes, one might be able to glean assumptions from the experiments, i.e., using gender as a binary variable, but given the evolving understandings of gender we have today this is not enough. Larson also advises that researchers should indicate whether gender categories have been self-identified by participants, ascribed by the researcher, or designated by a third-party (e.g., gender labels come with the dataset). The article details that if the researcher has done the labeling, it is necessary to explain the process for this, and in the case of the third-party designation, it is important to recognize this as a limitation. In many studies that follow, these considerations are often not made explicit.

In the expansive and rich literature on gender there is much discussion on what gender is, how it comes to be, and how it relates to sex, sexuality and other aspects of identity. For Butler (1993), gender is a form of *performativity* arising in "an unexamined framework of normative heterosexuality," (p. 97). The term *performativity* derives from the concept of a *performative* found in speech act theory, where it is defined as a "discursive practice that enacts or produces that which it names." In terms of gender, performativity differs from performance in that the former both draws from and contributes to the societal norms and expectations of the actor. The "I" as Butler puts it, is subjected to and subjectivated by gender, and thus neither precedes nor follows the process of "gendering" but emerges both within and as the matrix of gender relations themselves. Butler goes on to explain that the activity of "gendering" is not simply a human act or expression, willful appropriation, or putting on of a mask—it is a recursive conditioning which qualifies its

own cultural existence and emerges prior to the state of being human. This is supported with evidence from the medical practice of assigning sex to a baby.

> Consider the medical interpellation which…shifts an infant from an "it" to a "she" or a "he," and in that naming, the girl is "girled," brought into the domain of language and kinship through the interpellation of gender…that founding interpellation is reiterated by various authorities and throughout various intervals of time to reenforce or contest this naturalized effect. The naming is at once the setting of a boundary, and also the repeated inculcation of a norm. (Butler 1993, pp. 7-8)

In Butler's analysis, (binary) gender is imposed on the infant as sex, the materiality of which is often supported through the identification external genitalia. However, this practice contributes to power structures which delimit and define that which qualifies as human—a truth which is observed in the way a person's humanity is thrown into question when that individual appears to be improperly gendered. Therefore the demarcation of sex as a material truth produces a division between legitimate sex and delegitimated sex, wherein only the legitimated bodies qualify as bodies that matter and all others are denied the right of cultural articulation.

Taking a more biological approach, the intricacies of sex and gender are also explored by Fausto-Sterling (2012). Consider Figure 1, wherein there are different "layers" of sex presented, some of which feed into another, while others bifurcate to form a new chain of sex/gender "events". This chart originates from Money and Ehrhardt's 1950's study of intersex children and adults—individuals born with rare combinations of sex markers—and it effectively demonstrates that biological sex is not a simple two-way path. The process begins with *chromosomal sex* that occurs when sperm meets egg, then in the early stages of gestation all fetuses pass through a stage of undifferentiated sexual anatomy until they develop ovaries or testes. Notably, the chart makes clear that *genital sex, fetal internal reproductive sex,* and *brain sex* (though the concept of a "brain sex" is debated in the scientific community) all develop separately from one another. While *juvenile gender identity* is singularly based on genital sex and the socialization that follows from it, *adult gender identity* is informed by all three.

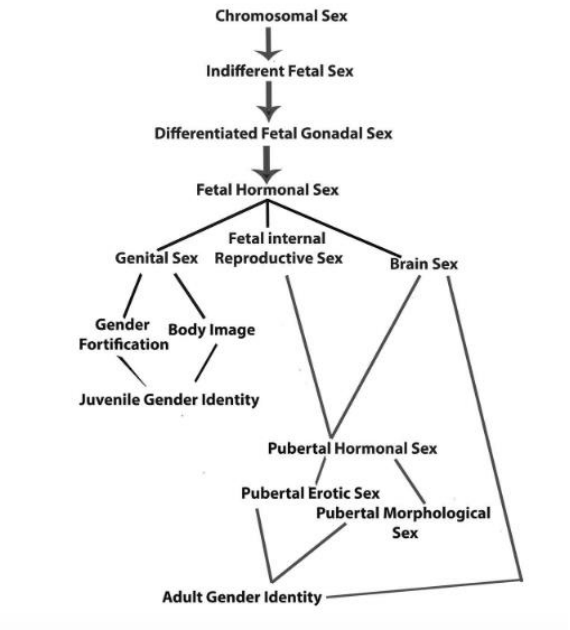

Figure 1: Fausto-Sterling's (2012) expanded version of Money and Ehrhardt's (1972) "levels of sex".

For a substantial portion of the United States population, gender is not as simple as the biological sex assigned to them at birth. According to a UCLA Williams Institute Study (Flores et al. 2016), an estimated 1.4 million adults identify as transgender. Notably, this was twice the amount estimated using data from roughly a decade prior. It is hard to know exactly how many trans people there truly are living in the US given that many do not openly identify as trans. Additionally, this statistic includes people who identify as non-binary or gender-nonconforming. How then, do researchers account for these marginalized persons when performing a gender study, if at all? Despite the relatively small likelihood of any such identities showing up in a dataset, is it unethical to risk ascribing the wrong gender label to someone who is trans? While the harm caused may not be direct, the practice of training algorithms to predict gender re-enforces the problematic practice of assuming gender labels based on perceived features rather than actual personal identity. Gender misclassification is simply another form of misgendering, and it could be particularly harmful in cases such as using gender prediction to target a particular audience for marketing of a gendered product. With the increased use of artificial intelligence for marketing and advertising the potential for psychological impact is becoming even more damaging. The analyses of both Butler and Fausto-Sterling

highlight that individuals are gendered from birth and their gender identities are heavily influenced by gender norms which are upheld by various "authorities". In the case of trans people, there is likely to be an incongruity between juvenile gender identity and adult gender identity, leaving open the possibility that influences from childhood remain as latent characteristics. These latent characteristics could include aspects of linguistic expression, making it possible that standard methods of gender prediction could misclassify and therefore misgender these individuals to potentially harmful effect.

2.3 Gender on Social Media

The task of gender prediction using text from social media platforms has received a large amount of attention from the NLP community. Though many researchers have reached some level of success, the task is far from simple and no single approach has mastered it. Among the many difficulties, different social media platforms impose a level of linguistic diversity that has proven tricky to navigate. Twitter appears most often in the studies, likely because of the amount of data available and how easy it is to obtain. However, it also poses a particular challenge because gender is not explicit in user metadata and posts are limited by character length. Facebook, on the other hand, provides data with users' self-identified gender, yet it requires some more work (or money) to obtain. In some more recent studies like Ljubešić, Fišer, and Erjavec 2017 and Goot et al. 2018, researchers have attempted to pare down their linguistic approach to be able to generalize across languages, but this has consistently come at a cost. Finally, there is the issue of gender itself and how it's considered in the literature. While most have treated the gender variable as binary and static (Sap et al. 2014, Ljubešić, Fišer, and Erjavec 2017, Goot et al. 2018), others have explored the value in questioning this treatment of gender and evaluating how and why some users stray from what's expected of their gender label, (Bamman, Eisenstein, and Schnoebelen 2014, Nguyen et al. 2014).

2.3.1 Motivations

Gender prediction falls under a larger category known as *author profiling* in the literature (Goot et al. 2018) where text is evaluated for distinguishing characteristics as indicators of latent demographic information.

As addressed by Sap et al. (2014), motivations for gender prediction can range from business and marketing to social science applications. For Bamman, Eisenstein, and Schnoebelen (2014), their model for gender prediction was used as a means for down-stream analyses such as the interaction of gender and other aspects of personal identity and interests. Both Nguyen et al. (2014) and Ljubešić, Fišer, and Erjavec (2017) mentioned the use of gender prediction on social media as a means to improve other predictive tasks such as sentiment classification (Volkova, Wilson, and Yarowsky 2013), and cyber-bullying detection (Dadvar et al. 2012).

For several authors, the main goal of their paper was to improve upon previous approaches. As their title suggests, Sap et al. (2014) focused on predictive lexica, using the bag-of-words approach, which seems to be the most useful for in-language prediction of gender. Alternatively, Ljubešić, Fišer, and Erjavec (2017) and Goot et al. (2018) focused on language-independent approaches which can predict gender over a range of languages.

2.3.2 Data

A majority of the literature available on gender prediction for social media makes use of Twitter data, which is true of every article considered in this review. Up until 2017, tweets were restricted to a 140 character count, but today users can make use of 280 characters. In either case, this poses obvious issues for generalization across other social media platforms because of the resulting noise from truncations, abbreviations and the like. Bamman, Eisenstein, and Schnoebelen (2014) demonstrated that Twitter offers a rich data source that is as diverse as it is easy to collect; notably, nearly 14-15% of all female and male internet users are on Twitter. Unfortunately, tweet metadata does not include an explicit field for gender, so user self-identification generally isn't an option when scraping from the API. This adds on the extra task of gender-labeling using human annotators or algorithms that use other information (such as names, handles or helpful links).

In their study, Bamman, Eisenstein, and Schnoebelen (2014) used the Twitter API to gather a corpus over a period of six months in 2011. The tweets were then filtered for several criteria. The first criteria was

to include only English speakers who used 50 of the 1,000 most common words in the US. Second, the authors targeted only mutual interactions (via @ mentions) to filter for social networks and rid the data of 'broadcasting' accounts, bots and spam. It should be noted that focusing on social networks within the data was novel approach that wasn't observed in the other studies. This allowed the authors to provide some compelling analyses later on when the authors incorporated social networks to re-examine the use of gender-marked language. Finally, the authors used the names of the Twitter users to identify their gender by matching with historical census information, i.e., a user was labeled with the gender associated with the majority count of the first name in the census data. Any names occurring less than 1,000 times were filtered out from the data, leaving them with over 14,000 users and 9 million tweets. The authors do recognize that this method is not fool-proof but should be effective in aggregate.

Nguyen et al. (2014) also collected tweets using Twitter's API, but used annotators to indicate the "biological sex" using all the information available from the tweets, users' profiles, as well as their Facebook and LinkedIn. All told, they were able to collect 20 to 40 tweets for 3,000 Dutch users. This is arguably a very limited amount of data per user, which is reflected later when discussing certain parts of their methodology.

The primary data for Sap et al. (2014) was from Facebook between 2009 and 2011, where they collected user status updates and self-indicated gender via their MyPersonality application on Facebook (Kosinski and Stillwell 2012). Two secondary sources were also used, since a goal of the study was to generalize across social media platforms. The authors made use of gender-annotated blogs from 2004 (Schler et al. 2006) and Twitter data (Volkova, Wilson, and Yarowsky 2013)—gender was self-reported in both cases either directly from the blogging site or indirectly from MySpace or Facebook. In all cases, the users in this study were offered a binary choice of male or female, and there was no consideration of users who might not feel their gender identity falls perfectly into either of these categories—which may be helpful in defining male- vs female-marked lexica, but seems neglectful to users who either reject the gender binary or have difficulty within its framework. Note that beginning in February of 2014, Facebook started offering a custom gender field allowing users to select from a wide range of genders (Oreskovic 2014).

Both Ljubešić, Fišer, and Erjavec (2017) and Goot et al. (2018) made use of the TwiSty Corpus (Verhoeven, Daelemans, and Plank 2016) manually annotated for gender and contains over 18,000 authors across six languages. Ljubešić, Fišer, and Erjavec (2017) discarded users with less than 100 tweets whereas Goot et al. used 200 tweets per user and included additional tweets from English-speaking users (Plank and Hovy 2015).

2.3.3 Methods

The bag-of-words approach is a common thread across many of the articles mentioned above and appears to be most consistent for language-specific gender prediction with reported accuracies as high as 91.9% (Sap et al. 2014) depending on the dataset. Put simply, the bag-of-words approach turns documents into some form of frequency distribution over tokens or some other rendering of a text (e.g. parts-of-speech tags, stemmed or lemmatized text, case-folded tokens). While it often performs well, it has the major shortcoming of losing all syntactic composition—with the exception of short-distance dependencies if it is a distribution of n-grams. For Bamman, Eisenstein, and Schnoebelen (2014), their approach included training a logistic regression classifier by partitioning their data for ten-fold cross-validation, training on 80%, tuning on 10% and testing on 10%, resulting in 88% overall accuracy on their testing data. They then used their classification results to identify categories of words most associated with either male or female gender and compared them with findings from previous literature. Among the female markers were pronouns, emotion terms, emoticons, most family-oriented kinship terms, abbreviations, ellipses, exclamation marks, question marks, backchannels, assent terms such as *okay* and *yes*, negation terms such as *no* and *cannot*, and expressive lengthening, e.g., *yessss*. As for the male markers, these included numbers, technology words, swear words, kinship terms such as *wife* and *bro*, and informal negation terms like *nah* or *ain't*. This was followed with further analysis involving word categories defined by the authors—named entities, taboo words, numbers, hashtags, punctuation, dictionary words, pronounceable non-dictionary words, and non-pronounceable non-dictionary words—finding that males were more likely to use named entities, and females were more likely to use emoticons and non-pronounceable abbreviations. Furthermore,

the authors used a clustering algorithm on the same features used for the classification task to cluster the authors. Some of clusters were heavily gender-oriented, suggesting multiple expressions of gender, and variation in word categories among the clusters indicated strong interactions with other aspects of identity. Additionally, their analysis found that users with social networks that were more skewed toward opposite gender were significantly rated lower in terms of gender classifier confidence.

Nguyen et al. (2014) made use of both crowd-sourcing via a gaming app and a bag-of-words model. The game was created as a web by the authors, where users were asked to guess a Twitter user's "biological sex" based only on a selection of their tweets. The human evaluations of user tweets yielded 84% accuracy just by linking majority votes with users' tweets. The bag-of-words model did not perform as well, with an accuracy of 69% which the authors attributed to insufficient data (20-40 tweets per user), so the majority of the discussion rested on the crowd-sourcing portion. The authors specifically focused on how disagreement among the crowd was a reflection of how users express their gender in varying degrees and different ways, with the conclusion that treating gender as a binary variable is too simplistic.

In the other approaches, there was less consideration for the variation within gender categories and more focus on constructing an accurate binary model. Sap et al. (2014) took the bag-of-words approach to generate a gender-specified lexicon which they provide publicly. The authors assigned unigram coefficients to tokens based on their probability distribution by using a support vector machine (SVM) classifier with L1 penalization to enforce sparsity on the lexicon. The top performing model was trained on data from Facebook, Twitter and blogging sites and achieved 91.9% accuracy when tested on random Facebook data.

The language-independent models from Ljubešić, Fišer, and Erjavec (2017) and Goot et al. (2018) made use of abstracted features from their Twitter data. The features used by Ljubešic et al. included the percentage of user tweets that satisfied a given condition (e.g., contains an emoji or link), means, medians and variance of items from tweet metadata (e.g., time of posting and tweet length) and variables from user metadata (e.g., tweet counts and favorite counts). Their models were trained on standardized (zero mean, unit variance) features using SVM with a radial bias function kernel and they optimized hyperparameters using five-fold cross-validation. While cross-linguistic accuracy improved over the bag-of-words approach,

their accuracy never exceeded 70%, but further discussion showed that certain features were skewed for each gender.

On the other hand, Goot et al. (2018) focused on "bleached text" or abstracted features particular to language, i.e., word frequencies, character length, abstract alphanumeric token representation, non-alphanumeric characters (emojis, emoticons and punctuation), token 'shapes' (upper- or lowercase, digits or other), and vowel-consonant structures. Their study used an implementation of a linear kernel SVM with L2 regularization, and they found that the bleached model provided most robust results when trained with all features combined. The n-gram size was tuned through in-language cross-validation, finding that five-grams performed best. The bleached model was shown to generally perform better than lexical and multi-lingual embedding models with accuracy as high as 69.2%, and therefore comparable to results based on user meta-data (like above). Their study also examined cross-linguistic human evaluation of gender and found that performance between humans and the bleached model matched, suggesting that humans may in fact be relying on more abstract linguistic cues for such tasks.

2.3.4 Discussion

In general, the studies had significant findings for gender prediction using NLP. Most notably Bamman, Eisenstein, and Schnoebelen (2014) and Nguyen et al. (2014) provide valuable insights that probe into the classification of gender as a binary variable. While framing the assumption of a person's gender as a game is highly insensitive to trans people, the observation of a *gender continuum* was surprisingly novel and stretches toward a more inclusive and thoughtful approach for gender using NLP. What separated these two studies from the others was their willingness to ask what is actually being observed when researchers attempt to look at language and gender. Additionally, using NLP to look at gender as a binary variable potentially overlooks intersectional factors that influence the way in which individuals use language to construct identity and position themselves within their social network. Especially when considering Twitter, where language (as opposed to photos or videos) is the primary means to project all aspects of identity or affiliation, to attempt to extricate just one part seems an oversimplification.

As for the studies focused on gender predictive models, it is unclear whether such questions are being asked. Sap et al. (2014) seem to have developed a model for general detection of rhetoricity and/or topicality among male and female Facebook users, which receives little discussion beyond its efficacy to do so. On the other hand, the results from Ljubešić, Fišer, and Erjavec (2017) do not seem to justify the use of social networking activity as a means to discriminate gender in light of the "bleached text" from Goot et al. But abstraction from language for cross-linguistic gains consistently comes at an overall cost in accuracy. Interestingly, Goot et al. do not mention cultural differences as a limitation in their study, which is likely to play a role whenever cross-linguistic differences are present. However, they offered interesting insights about the relationship between the bleached model and human evaluation of gender.

Overall, gender prediction for social media heavily leans toward bag-of-words models that identify lexica observed in correlation with self-reported or ascribed male or female gender. Some approaches have attempted to circumvent the bag-of-words model in order to generalize beyond a specific language, but there has consistently been a significant trade-off. Additionally, gender is most prominently featured as a binary and static variable which is limited at best, and unethical at worst. Considerations for intersectionality, non-binary and trans identities are largely not observed in many NLP studies that incorporate gender prediction, which may pose some ethical concerns (Larson 2017). This is especially notable given that gender prediction has been proposed to increase efficacy in other tasks (Dadvar et al. 2012, Volkova, Wilson, and Yarowsky 2013) where only male and female genders are considered. While the work does present some interesting findings on how both humans and machines can recognize gender in natural language, there are equally rich results when the expression of identity through language is not so heavily focused on static, binary gender and instead takes a more probing look into how gender is socially constructed and expressed across different contexts.

3 Materials

For my study, I focus directly on the population that seems to go overlooked in the previous literature. Trans people have the most at stake when it comes to gender prediction and being incorrectly classified could have distressing consequences for the wellbeing of these individuals. Since I know of no free or available datasets for trans people I have built my own corpus by scraping Reddit. In light of the fact that this study focuses on a marginalized population, the corpus will not be published or posted publicly and no names or texts will be produced in full to protect the privacy of the individuals.

The data was gathered by identifying the users who contributed to the subreddits r/ftm[1] (female-to-male) and r/MtF[2] (male-to-female), and accepting *Gender M* as the label for users from r/ftm and *Gender F* as the label for r/MtF. There are several things to note about this approach, beginning with the possibility of users whose gender identity doesn't necessarily fall within the range of identities of the target community, e.g., a cisgender ally looking to be supportive of a trans friend. Some due diligence was done to remove these through manual search, finding only few of such instances. The description for r/ftm reads, "support-based discussion place focused on trans men, trans-masc individuals, and other people assigned female at birth who are trans." The r/MtF description is, "a subreddit devoted to transgender issues pertaining to male-to-female or MAAB [male-assigned at birth] people." From both these descriptions, it is clear that many possible gender identities are being oversimplified in order to run these experiments. At the time of data collection both subreddits had roughly 95,000 subscribers and I assume that—in aggregate—those receiving the Gender F label belong to the community of trans people who were assigned male at birth, while those who received the Gender M label belong to the community of trans people who were assigned female at birth.

The members of these communities were then tracked across other subreddits to capture a natural variety of topics and interests. This resulted in collecting nearly 160,000 total combined submissions from

[1] https://www.reddit.com/r/ftm/
[2] https://www.reddit.com/r/MtF/

1,873 Gender M users and 1,866 Gender F users. This data was then filtered for text written in English using language detection from spaCy, a free and open-source library for natural language processing (Honnibal et al. 2020). Compared to human annotators with a Cohen's κ agreement of .96, the language detection achieved an accuracy score of .92, with a precision score of .98 and .89 for recall. After language detection was performed, the remaining submissions were tokenized, lemmatized and tagged for parts-of-speech also using spaCy's largest model.[3] The resulting dataset is presented in Table 1.

	Users	Submissions
Gender M	1,372	63,600
Gender F	1,253	73,143
Total	2,625	136,743

Table 1: The total users and submissions remaining after filtering the data.

4 Experiment 1: Log-odds Ratios, Informative Dirichlet Prior Method

Much of the more recent literature has been focused on using modeling approaches to achieve the best accuracy possible without providing much in the way of feature analysis that might explain the differences between gender classes. In contrast, the earlier work from Argamon et al. (2003) combined modeling and statistical methods to provide more thorough analyses of features. In this vein, I will also provide some simple statistical analyses to explore the data and find what features are statistically linked to each class.

One of the approaches to comparing corpora, as explained by Monroe, Colaresi, and Quinn (2009), is to use the log-odds ratios with an informative Dirichlet prior. To explain, it first begins with the plain log-odds ratio, which can be simplified to the equation below,

$$lor(w) = log\left(\frac{f^i(w)}{n^i - f^i(w)}\right) - log\left(\frac{f^j(w)}{n^j - f^j(w)}\right)$$

[3] https://github.com/explosion/spacy-models/releases//tag/en_core_web_lg-2.3.1

The above does a simple estimate of whether the word, w, has higher odds in corpus i or corpus j through taking the frequency count of w in corpus i, represented as $f^i(w)$, dividing it by the result of the frequency subtracted from total number of words in the corpus, n^i, and the ratio is completed by subtracting the same operations for corpus j. The resulting ratio can be any real number, where positive ratios indicate higher odds in corpus i, and negative ratios in corpus j. Using the informative Dirichlet prior method, we can get more corpus-focused results by including a background corpus. The modification is as follows,

$$log\left(\frac{f^i(w) + f^k(w)}{n^i + n^k - (f^i(w) + f^k(w))}\right) - log\left(\frac{f^j(w) + f^k(w)}{n^j + n^k - (f^j(w) + f^k(w))}\right)$$

where $f^k(w)$ represents the count of word w in background corpus k, and n^k represents the size of the background corpus. This background corpus serves to inform the log-odds ratios through including information about the prior distribution of a given word, so words that are high frequency in the background corpus will work to cancel themselves out in the target corpora, allowing words that are more unique to the target corpora to be more heavily weighted in their log-odds ratios.

For the purposes of this portion of the study, I gathered a corpus of submissions over the period of July 2018 through July 2020 from across all of Reddit to act as a background corpus, making sure to remove any submissions from the authors in the gender dataset. Since the original corpus was pulled entirely from Reddit, Reddit seemed the most appropriate source for a background corpus. The background corpus received the same aforementioned treatment of language detection, tokenization, lemmatization and parts-of-speech tagging using the spaCy model, ending up with a total of 184,225 submissions from 149,165 authors.

Using target dataset along with the background corpus, the log-odds ratios were applied in several iterations to gather insights in a variety of ways. First, I use the plain tokenized text with case-folding applied, selecting only the top one thousand most frequent tokens (excluding punctuation, but maintaining emoticons and emojis) to capture the most useful function words and content words for analysis. Next, I used the same approach applied to the lemmatized tokens in both case-folding and limiting the number of

tokens. Finally, I applied the log-odds ratios to parts-of-speech tags to get a more generalized syntactic look at the features.

4.1 Raw Text

Beginning with the raw text, Figure 2 shows a blue color to indicate Gender M favoring ratios, while the pink color indicates Gender F. Right away there are some clear categorical associations forming. Almost all of the contractions are associated with Gender M, including *'m, n't, 've, 'll, 'd, 're, y',* even *gonna* and *wanna*. In previous literature (Argamon et al. 2003), these were said to fall among the more "involved" features and therefore linked to female gender. Also, almost all personal pronouns, *he, him, his, she, her, it*

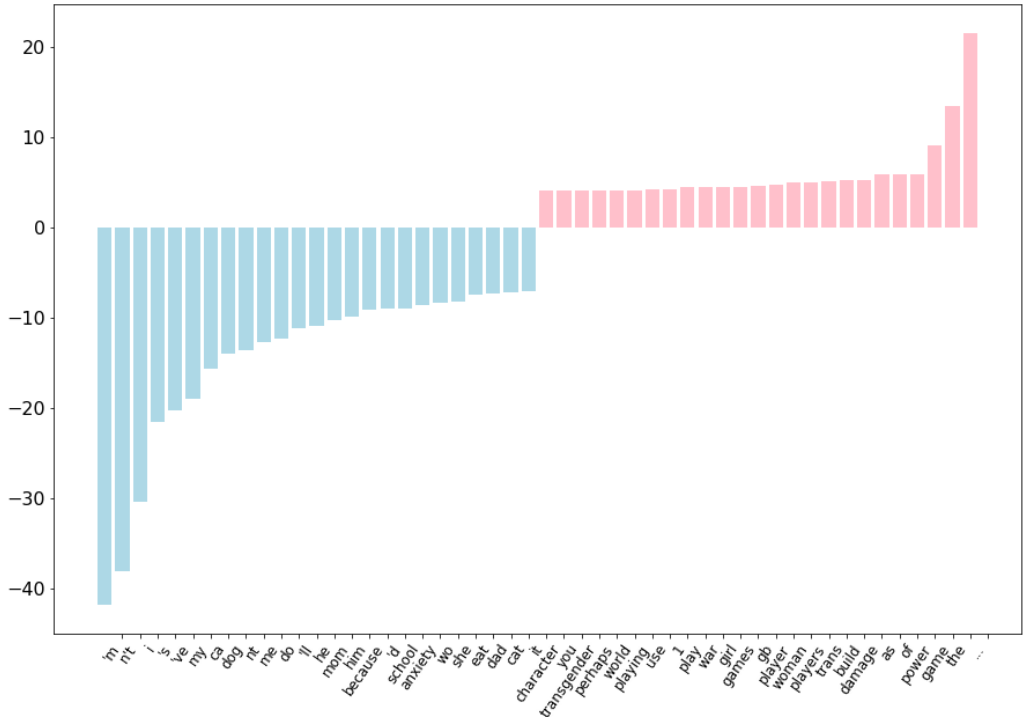

Figure 2: The top 25 log-odds ratios for each class using case-folded text.

and especially the first person *I, me*, and *my*, show associations with Gender M, where previous work has shown this to also be linked to female gender (Argamon et al. 2003, Schler et al. 2006, Bamman, Eisenstein,

and Schnoebelen 2014). However, certain pronouns such as the second person *you*, possessives *your*, *its*, *their* and reflexives *itself* and *themselves* all showed association with Gender F. This could indicate that Gender F users are directly engaging with their community more frequently, while Gender M users are expressing themselves and talking more about relationships with others. There are also indications that Gender M users talk more about their home and family life with words such as *mom, dad, family, brother, home*, most of which have previously been linked to female gender (Schler et al. 2006, Bamman, Eisenstein, and Schnoebelen 2014). Gender M users also tend to use words related to emotional states such as *anxiety, feel, depression, depressed, pain, scared, upset, hurt, sad, angry.* Much like the aforementioned categories, previous work has shown that emotion words are linked to female gender (Schler et al. 2006, Parkins 2012, Bamman, Eisenstein, and Schnoebelen 2014). In a slight departure, the category of swear words seems to be associated with the Gender M users, e.g., *shit, fuck, fucking, damn, hell*, aligning with previous work where this has been associated with male authors (Lakoff 1973, Bamman, Eisenstein, and Schnoebelen 2014).

Moving to the other side, there seem to be associations with gaming (*game, players, play, playing war, level, attack*) and technology (*gb, pc, system, windows, server, screen, code, video, computer)* for the Gender F users, both of which have been associated with male gender (Schler et al. 2006, Bamman, Eisenstein, and Schnoebelen 2014). These categories also contribute to an overall sense of talking about personal interests, rather than relationships and feelings being the topic of discussion among Gender F submissions. Numbers are also prevalent among the log-odds ratios for Gender F users, which has also been linked to male gender in previous work (Argamon et al. 2003, Bamman, Eisenstein, and Schnoebelen 2014). Similarly, certain "noun specifiers" (as Argamon et al. refer to them) such as *the, these, those, this,* and *that* all showed potential association with Gender F users, despite being historically linked to male gender. Interestingly, words referring to trans experience such as *trans*, *transgender*, *transition,* seem to be slightly more associated with Gender F despite them being applicable to both classes.

4.2 Lemmatized Text

Moving onto the lemmatized tokens, Figure 3 paints a similar picture. Replacing personal pronouns with the tag *-PRON-* places it squarely on the Gender M side, which is to be expected, however *i* stands separately, which is likely due to non-capitalization causing an error. Nearly all of the contractions also maintained their standing, but the possessive *'s* has made its way into the top 25 log-odds ratios for Gender F. Figure 3 also suggests that Gender M users use more common transitive verbs, such as *feel*, *tell*, and *know*, and further probing into the data confirms this—*eat*, *have*, *get*, *want*, *try* are all somewhat more strongly associated with Gender M. A deeper look into the data also reveals associations between Gender M and intensifying adverbs *so*, *very*, *always*, *constantly*, *especially*, *extremely*, *definitely*, *absolutely*, *completely*, and *literally* which could be considered more expressive, and therefore linked to female gender (Palander-Collin 1999, Parkins 2012). Notably, both *boyfriend* and *girlfriend* (as well as plain-old *friend*) show possible associations with Gender M, which adds more evidence that Gender M talks more about

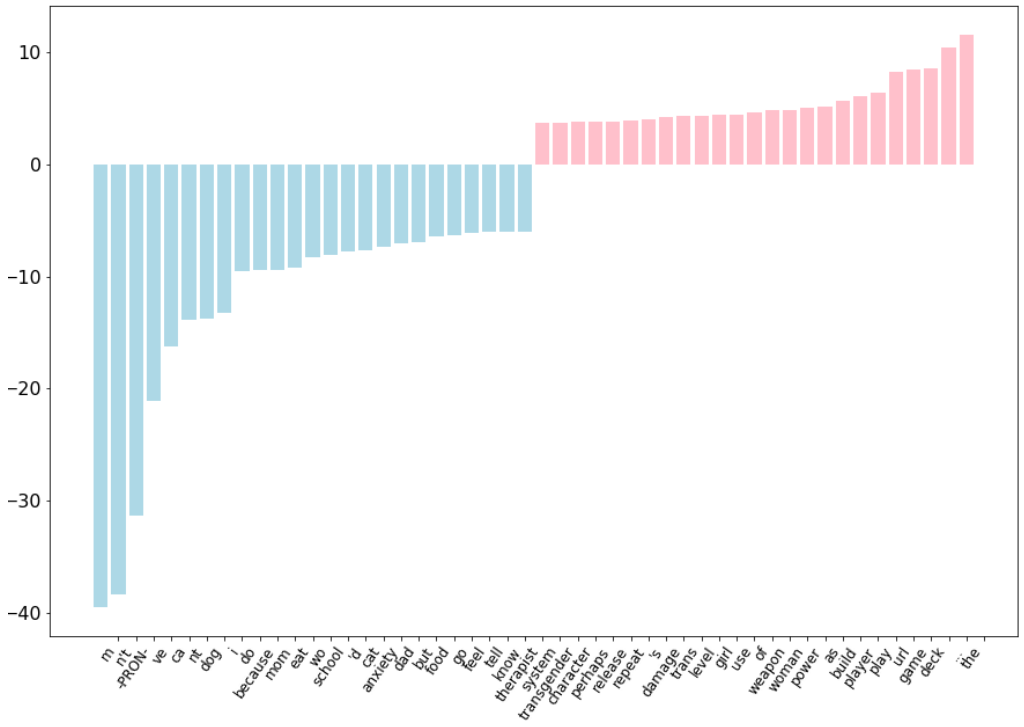

Figure 3: The top 25 log-odds ratios for each class using lemmatized tokens.

relationships. However *wife* shows possible associations with Gender F, which is a kinship term that has previously been associated with male gender (Bamman, Eisenstein, and Schnoebelen 2014). Other seemingly marked language for Gender M include time-associated words like, *day, time, week, month, hour, tomorrow, yesterday, weekend,* and *year*, as well as words in the medical domain, e.g., *therapist, med, medication, therapy, surgery, diagnose, doctor, symptom*. Neither of these categories necessarily align with anything found in the previous literature, but perhaps show a higher tendency for these authors to share their personal stories and experiences, compared to the more personal interest focus suggested by the marked language for Gender F.

The Gender F associations are also mostly consistent with the previous results, especially with technology and gaming words. Words associated with violence (stereotypically a masculine trait) are slightly more prevalent, *weapon, damage, attack, fight, shoot, death, hit, beat,* but given the context, these are also most likely used in discussions about gaming. A couple assent terms, *yeah* and *yes*, appear to be associated with Gender F, as well as the emoticon *:)*, both of which have previously been associated with female gender (Schler et al. 2006, Bamman, Eisenstein, and Schnoebelen 2014). Overall, both the raw text and lemmatized text results suggest that Gender F discussions revolve around building community through shared interests, while Gender M discussions are more focused on personal experiences and relationships.

4.3 Parts-of-Speech Tags

Moving beyond the vocabulary and into a more syntactic view, there are several things of note which are demonstrated in Figure 4. First, unsurprisingly, the tag for personal pronouns (`PRP`) has taken the top spot for association with Gender M followed not to distantly by the tag for possessive personal pronouns (`PRP$`), both still indicating a contradiction with the previous literature that pronouns are associated with female authors. Both tags indicating present-tense verbs (`VBP` and `VBZ`) are also showing possible association with the Gender M authors. According to Argamon et al. (2003), present-tense verbs belong to the involved dimension which has been associated with female writing styles. Bamman, Eisenstein, and Schnoebelen (2014) also discuss that previous literature found conjunctions to be associated with female writing styles,

but here the tag for coordinating conjunctions (CC) shows association with Gender M writing styles. Consistent with some of what was seen in the lemmatized text analysis, adverbs are spread between classes with adverbs and adverb particles, RB and RP respectively, showing association with Gender M, while superlatives (RBS) and comparatives (RBR) showing slight association with Gender F. Another difference between the classes seems to be demonstrated in the non-lexical tags for punctuations and spaces, which are mostly associated with Gender M. Also, while most of the verb tags show association with Gender M, all of the noun tags as well as the tags for determiners and numbers show possible association with Gender F, which are also consistent with earlier results.

Figure 4: The log-odds ratios for singular POS tags.

5 Experiment 2: Logistic Regression Classification

Machine learning is by far the most often used method for text-based gender prediction. As was reviewed previously, there are plenty of algorithms to choose from. In many cases, machine learning methods are used with bag-of-words features (i.e., some form of frequency distribution over words) to assign feature weights to tokens that the algorithm finds most useful for determining gender. Such an approach then allows for a post-hoc analysis of these features to postulate why a given feature or set of features might be associated with a given class. In addition to the feature analysis, logistic regression outputs probability measures on feature sets which correspond to each class, and these probabilities can be understood as confidence measures which can be used for further analyses—these will come in handy for the next section where I use as clustering analysis to group authors into speech communities to gain insight into why certain authors are misclassified.

For my implementation of the logistic regression classifier, I treat the author gender label as the dependent variable, and all tokens used by at least 20 authors in the corpus as the independent variables. The tokens included case-folded lexical items, punctuations, symbols, emojis and emoticons. Different options were explored for features in the data, including using the raw text, lemmatized text, parts-of-speech tags and iterations using variations on n-gram counts from unigram to trigram, as well as switching from raw frequencies to term frequency–inverse document frequency measures (TF-IDF). For the issue of overfitting to the training set, I explored options for the regularization parameter using L1, L2 and elastic-net. The L1 regularization or "lasso regression" (Tibshirani 1996) seeks to induce model sparsity by heavily penalizing features that are not useful for classification, setting as many as close to zero as possible—making it an ideal choice for feature selection among a large set of features. L2 regularization, or "ridge regression" (Hoerl and Kennard 1970) maintains a larger set of features but penalizes extremes to smooth out feature weights, and elastic-net offers a combination approach using both with the option of adjusting the ratio of L1:L2. To tune the model parameters, I used 10-fold cross validation on 80% training data and 10% development data, with the remaining 10% held out as a test set.

	Train	Dev	Test
Unigram (TF-IDF)	.9052	**.**7714	**.7421**
Bigram (TF-IDF)	.8224	.7378	.7200
Unigram (freq)	.9224	.7181	.7105

Table 2: Model accuracy by feature set.

	P (F)	P (M)	R (F)	R (M)	F1 (F)	F1 (M)
Unigram (TF-IDF)	.7130	.7701	.7364	.7485	.7245	.7591
Bigram (TF-IDF)	.6988	.7406	.7240	.7164	.7112	.7283
Unigram (freq)	.6988	.7218	.7098	.7111	.7043	.7164

Table 3: Precision, recall and F1-Score for each test set.

For the remainder of this section I will be focusing on the three top performing models from my experimentation. All three were trained using L1 penalization, and their results can be found in Tables 2 and 3. Lemmatization had no substantial effects on model performance, and parts-of-speech features did not achieve high enough accuracies to indicate that they were useful, so all models shown were trained on the unlemmatized text features. The model trained on unigram raw frequencies achieved a the highest training accuracy, but performed the poorest on the held-out test set, while the model using TF-IDF unigram features performed the best on the development and test sets. These accuracies are low for the task when compared to the previous work, suggesting limitations with the dataset or that more data per user may have been required to achieve better results; however previous work has never specifically looked at trans individuals, so it cannot be ruled out as an indication that classification for these individuals is more difficult. Still, these accuracies are more at least 20% greater than chance, suggesting that the model did learn some useful features for classification. In all cases, precision was higher for Gender M, with only slight differences in recall showing that the models we generally better at identifying the majority class.

Figure 5: Unigram (TF-IDF) learned using logistic regression with L1 penalization.

The results show some differences from the log-odds ratios of the previous section. Most noticeably, gender terms are more heavily weighted for each class—*girl, girls, MTF, woman, women, female, feminine*, and *AMAB* (assigned male at birth) are all markers for Gender F as are gender-neutral terms *trans* and *transgender*, while *guy, guys, man, men, male, boy*, and *FTM* are markers for Gender M. This suggests that, in some cases, the authors are likely self-identifying their gender and the classifier is able to pick up on it. Words referring to hormones are also distributed in ways to be expected—*hrt* (hormone replacement therapy), *spiro* (spironolactone, a testosterone blocker) and *estradiol* (an estrogen steroid hormone) are Gender F markers, and *testosterone* is a Gender M marker.

Certain themes also seem to carry over from the log-odds ratios—technology words (e.g., *data, ios, java, usb, gb, pc, macbook, hdd, vr, console, software,* etc.) are still strongly associated with Gender F, as are most words related to violence (*weapons, damage, shoot, combat, kill, attack*) and gaming (*rpg, mario, game, players, gameplay*) however, *sims* and *pokemon* both show up as weighted toward Gender M. The models also provide further evidence that determiners and now quantifiers (e.g., *some* and *any*) are markers for Gender F. All three models also showed that personal pronouns were still more strongly associated with Gender M. Most family and kinship terms are also still linked to Gender M, including *dude* and *husband*, but *ex* and *wife* are still markers for Gender F and, in a slight departure, two of the models weighted family terms *brother* and *son* toward Gender F as well. Many emotion words (*stressed, worried, terrified, pissed, mad, upset, happy, proud* and *excited*) still show up as markers for Gender M. The terms of assent, *ok, yes, yeah* as well as dissent term *nope* are all weighted toward Gender F, but *ya* is weighted toward Gender M. The conjunctions *because, if, but* all show stronger associations with Gender M. As previously mentioned, Bamman, Eisenstein, and Schnoebelen (2014) report that conjunctions have been said to be markers of female gender. Certain greetings such as *hi* and *yo* seem to be markers for Gender F, but *hey*, *hello* and *howdy*, are all markers for Gender M. Abbreviations were mixed between classes, for example, *lol, lmao, ppl* were indicators for Gender M, but *irl* and *pls* were markers for Gender F. Such abbreviations were linked to female gender by Bamman, Eisenstein, and Schnoebelen (2014). Unlike the log-odds ratios, the

models showed no significant differences regarding swear words, emoticons or most contractions—though *'ve* and *wanna* do show up as weighted toward Gender M.

Figure 6: Unigram & bigram (TF-IDF) coefficients learned using logistic regression with L1 penalization.

Looking over the results from the models gives a fairly clear idea about where personal interests and topics of discussion overlap and where they differ. Quite obviously, many of the authors on both sides are discussing their experiences of being trans as evidenced by the discussion of hormones, trans identities and other words that are potentially relevant to this discussion such as *makeup, feminine, transition, breast,* and *'to female'* for Gender F, and *'top surgery'* and *scars* for Gender M, as well as the other medical or health-related terms dispersed throughout each. While both sides seem to be discussing the physical aspects of their transition regarding their appearance and hormonal changes, the abundance of emotion and mental health words observed to be weighted more heavily toward Gender M suggests that perhaps these authors are more likely to discuss their feelings around it, and the significant differences in pronouns and kinship words suggest that they may also be more likely to discuss how it is affecting their relationships with others.

As for the differences in interests, words like *vegan*, *books*, *flowers*, *school*, *art, dog,* and *cat* seem to give some insight into the personal interests being discussed among the authors of the Gender M class. On the other hand, there is a rather clear sense that the Gender F class has more specific interests related to gaming, technology and anime—interests which involve communities that are stereotypically male. However, these communities are also largely internet-based, allowing their members to interact and feel connected with each other from the privacy of their own homes. This could possibly be related to the disproportionate amount of violence faced by trans women. For example, of the 44 murders of trans and gender nonconforming people in the United States reported on the Human Rights Campaign's website in 2020, 37 (84%) were trans women (HRC 2020). In light of this, it is possible that many trans women opt to seek human connection and build their communities from the safety of their own home. Additionally, sharing their experiences of violence with their online communities also offers another possible explanation as to why terms related to violence are weighted more heavily toward Gender F. So, while at first glance some of these results may seem surprising, they have possible explanations from the differing experiences within the trans community.

Figure 7: Unigram (frequency) coefficients learned using logistic regression with L1 penalization.

6 Experiment 3: Non-Negative Matrix Factorization

For the final experiment, I use non-negative matrix factorization (NMF; Xu, Liu, and Gong 2003) to cluster the authors—essentially using an unsupervised method to identify possible topics and/or speech communities within the data. In basic summary, the NMF algorithm is able to approximate a larger matrix via estimation of a decomposition into two matrices of reduced dimensions—the basis matrix that corresponds to the original matrix's basis elements (recurring elements within the original matrix) and a reconstruction matrix that provides a mapping to those elements in an approximation of the original matrix. In text mining, the original matrix is then composed of the bag of words token distributions (frequencies or TF-IDF), where each row corresponds to a word, and each column corresponds to a document—or, for my purposes, each column represents an author. In this representation, the basis elements then correspond to the recurring tokens that characterize the topics extracted from the data. Similar to other topic modeling algorithms, NMF distributes these among a predetermined number of topics using a specified number of features.

After several rounds of experimentation with the NMF algorithm, I was able to produce coherent topics among the 2,625 authors using TF-IDF features from the lemmatized unigrams, setting the number of features to 1,500, and the number of topics to 20. Table 4 provides an overview of the topics with the key terms, as well as the distributions for each gender label. As expected, many of the topics were imbalanced for class suggesting association with a gender label. The shading in the table highlights those topics where the ratio of the majority class to the minority class is approximately equal to 2:1 or greater. Many of these topics are consistent with findings in the previous two sections—topics where the Gender M is the dominant class are characterized by terms largely relating to family (#6), job/career (#7), health and diet (#9), pets (#8, #12), education (#15), and feelings and relationships (#20); topics dominated by the Gender F class are characterized by technology terms (#2), financial terms (#3), transgender identity and experience (#10, #16), and games and gaming (#11, #18). Topics that were more balanced included more general terms (#1),

#	gender-m	gender-f	top terms
1	534	493	like, just, know, make, time, think, use, look, ve, thing, try, really, people, good, want
2	38	***99***	gb, use, cpu, build, pc, laptop, windows, monitor, gaming, computer, keyboard, screen, ram, product, drive
3	15	***26***	card, pay, money, credit, account, rent, gift, bank, buy, send, loan, month, mail, list
4	47	31	nt, wanna, ca, ill, lol, idk, bc, fucking, fuck, na, kinda, gon, n't, sorry, buy
5	27	33	url, link, product, video, auto, 99, image, post, look, channel, art, pic, list, hey, picture
6	***131***	61	tell, mom, say, dad, friend, want, mother, know, parent, family, talk, sister, come, kid, make
7	***71***	37	work, job, pay, month, week, day, hour, time, manager, money, year, company, need, rent, car
8	***23***	13	cat, vet, food, pet, med, old, day, apartment, bathroom, room, lady, feed, night, skin, concerned
9	***45***	20	eat, weight, calorie, food, lose, fat, diet, day, pound, gain, meal, healthy, lb, vegan, week
10	40	***69***	woman, trans, man, gender, male, girl, transition, transgender, female, people, dysphoria, sex, feminine, dress, identify
11	19	***83***	game, play, player, team, pc, steam, fun, server, kill, mod, enemy, new, win, quest, build
12	***37***	7	dog, training, train, owner, pet, house, animal, service, walk, home, vet, apartment, park, plan, food
13	14	20	just, like, fucking, fuck, ca, shit, hate, really, say, gon, na, wanna, want, idk, kinda
14	36	35	thank, hi, help, look, advance, hey, appreciate, know, want, hello, good, new, advice, wonder, suggestion
15	***63***	22	school, college, class, year, student, program, graduate, apply, high, grade, semester, degree, university, study, want
16	19	***37***	mg, hrt, month, dose, testosterone, day, doctor, pill, breast, dr, start, result, level, 50, 100
17	19	27	hair, wash, use, skin, grow, dry, cut, product, oil, long, feminine, look, face, girl, short
18	32	***51***	character, player, party, dm, campaign, spell, level, dragon, world, magic, roll, story, weapon, make, group
19	13	9	ve, year, month, recently, try, ago, week, day, start, hey, advice, notice, past, love, couple
20	**149**	80	feel, like, really, know, want, life, think, just, time, people, friend, relationship, thing, feeling, depression

Table 4: Topics extracted using NMF on the top 1,500 TF-IDF features, with the number of topics set to 20.

terms indicating frustration and tentativeness (#4, #13), link and media shares (#5), suggestions and advice (#14), and physical appearance (#17).

Overall these distributions affirm much of what was seen in the previous experiments. The family and relationship clusters are both heavily skewed toward Gender M individuals, showing that they are roughly twice as likely to discuss these topics compared to their Gender F counterparts. On the other hand, the technology cluster shows that Gender F individuals are almost three times more likely to participate in these discussions. This is slightly troubling considering that in both cases, these individuals would likely be misclassified by standard gender prediction methods, where it has been consistently found that technology terms are linked with male gender and female gender has been associated with the discussion of relationships.

The distributions of the classes within each topic also provide further insight into the results from the log-odds ratios and coefficients from the classification experiments. Topics #11 and #18 suggests the link between terms relating to games and violence with the inclusion of the words like *kill* and *weapon*. Also, while topic #18 is slightly more balanced, Topic #11 demonstrates one of the greatest imbalances of the topics at a Gender F to Gender M ratio of over 4:1. Judging from the top terms for each, it's likely that

Topic #11 consists of mentions about video games and Topic #18 is more specific to Dungeons & Dragons—a tabletop fantasy role-playing game. Given that technology and video games go hand in hand, it's not surprising that both these topics are dominant within the same class, while the non-technological game shows slightly more balance. Topics #10, #12, and #17 all appear to involve discussion of gender identity and appearance, each of which has a larger number of Gender F participants. One might think that these topics would show more balance given that they relate to the trans experience on a more general level, but the preoccupation with appearance and perception of others—suggested by words like *people, dress, breast, look* and *hair*—could be influenced by the violence against trans women (as mentioned in the previous section). In other words, perhaps discussions about appearance and passing are more prevalent among trans women as a way of trying to avoid harassment and violence.

In a more quantitative analysis, Table 5 provides precision, recall and average confidence scores broken down by gender class and topic. Precision scores are generally higher for Gender F, indicating that the model is more conservative with predictions for Gender F, likely because there are slightly fewer training samples for this class. On the other hand, recall scores show some of the most drastic differences among the clusters. The Gender M individuals in topics #2 and #11, the technology and gaming clusters, had extremely low recall scores with only 52% and 65% of individuals being classified correctly. Gender F users in almost all of the Gender M dominated clusters fared similarly, with recall scores ranging from 43% to 68%. Notably, the recall scores for the trans identity and experience clusters (#10 and #16) ranged from 88-95% despite being dominated by the Gender F users. As discussed in the previous section, the authors in these clusters were likely explicitly indicating their gender identity, providing the classifier with useful features to distinguish the classes.

Moving on to the average confidence scores, there are some more clear patterns that arise. For each class, the average confidence score only exceeds 80% for clusters where that class is dominant, thus making their language highly marked. In many of these clusters there is greater than a 10% gap between confidence scores for each class, suggesting that individuals in the minority class are using highly marked language

that is at odds with what the classifier has determined for their gender. These metrics demonstrate one of the biggest weaknesses of using the bag-of-words approach for gender prediction. The issue with using

Topic	lean	prec-m	prec-f	recall-m	recall-f	avg-conf-m	avg-conf-f	conf_diff
1	ns	.8526	.8836	.8989	.8316	.7632	.7216	(.0416)
2	f	.7813	.8762	.6579	.9293	.6957	.7748	.0791
3	f	.8750	.9600	.9333	.9231	.7646	.8521	.0875
4	ns	.8627	.8889	.9362	.7742	.7798	.7008	(.0790)
5	ns	.8276	.9032	.8889	.8485	.6893	.7332	.0439
6	m	.8514	.8864	.9618	.6393	.8098	.6902	(.1197)
7	m	.8519	.9259	.9718	.6757	.7891	.7250	(.0641)
8	m	.9167	.9167	.9565	.8462	.8356	.6924	(.1431)
9	m	.8491	1.0000	1.0000	.6000	.8153	.7522	(.0631)
10	f	.8537	.9265	.8750	.9130	.7281	.8468	.1187
11	f	.8333	.9000	.5263	.9759	.6639	.8045	.1405
12	m	.9024	1.0000	1.0000	.4286	.8996	.7155	(.1841)
13	ns	.8235	1.0000	1.0000	.8500	.7428	.7526	.0098
14	ns	.8056	.8000	.8056	.8000	.7695	.7141	(.0553)
15	m	.8493	.9167	.9841	.5000	.8190	.6010	(.2180)
16	f	.8571	.9714	.9474	.9189	.7687	.8565	.0878
17	ns	.8095	.9200	.8947	.8519	.7312	.7553	.0241
18	f	.9000	.9057	.8438	.9412	.7134	.8013	.0879
19	ns	.8571	.8750	.9231	.7778	.7597	.6632	(.0966)
20	m	.7989	.8800	.9597	.5500	.8037	.7209	(.0829)

Table 5: Precision, recall and average confidence measures separated by class and topic.

language to predict gender is that gender is only one of many aspects of identity that are revealed through language. As Bamman, Eisenstein, and Schnoebelen (2014) state, “to the extent that a linguistic resource indexes gender, it is pointing to (and creating) the habitual, repeated, multifaceted positionings inherent in every situated use of language.” Social media and internet platforms like Reddit provide a space for community building through shared interest, and within these communities there can be a diverse array of personalities, cultures and backgrounds. Fitting these multidimensional aspects into a binary model of gender prediction can overlook these other influences and rather than examining how these influences interact within a community, the predictive model can only see them as outliers.

7 Conclusion

Basing my approach in a way that can be compared to the previous literature on gender prediction I began with two standard statistical methods for comparing corpora—token log-odds ratios, and logistic regression classification. Using these methods, I provided findings for linguistic features that appeared to be highly marked for each gender class. When comparing these findings to the previous literature, I found discrepancies between the features marked for cisgender identities and those of the transgender classes. In the case of Gender M, the transmasculine class, several features pattern with those previously identified as being markers of female gender, particularly the heavier usage of pronouns—which have been the most consistent female markers in previous literature—as well as emotion words, coordinating conjunctions, and other features associated with the *involvement* dimension. Features that were previously found to be markers of male gender were also prominently featured among the results for the Gender F class, such as the use of determiners, numbers, quantifiers and technology words. Further exploration into the data with the use of author clustering provided a topic-level analysis that was consistent with these findings and showed further evidence of variation from previously established gender norms—namely that Gender M users have a greater tendency toward discussing relationships, while Gender F users write more about personal interests such as technology and gaming. In addition to providing deeper insight into the individual features, topic clustering also demonstrated some of the shortcomings of treating gender as a binary classification task in its inability to handle outliers and extricate gender identity from other factors which influence linguistic expression.

There is no denying the reality of the heteronormative categories of “male” and “female” and their influence on individual expressions of gender identity, but simply examining gender identity in this framework could prove harmful to transgender individuals. There are a number of possible reasons why some of the marked features from this dataset do not align with previously established gender norms. Perhaps underlying influences from the socialization and indoctrination under a certain imposed juvenile gender identity remain despite the agency in establishment of a disparate adult gender identity. Perhaps the

transgender experience has a profound effect on the ways in which individuals interact with the world and express themselves, whether it be in reaction to marginalization or violence, or simply finding and positioning oneself in a community in which one feels safe and affirmed. In any case, the standard state-of the-art methods which use linguistic features to predict binary gender are not equipped to handle variation beyond the norms that are established within a given dataset, nor can they differentiate features which may or may not be related to any number of factors, demographic or otherwise. In light of what has been observed, these gender prediction methods are implicitly harmful in their capacity to incorrectly ascribe gender labels to trans people. In an age where people are becoming increasingly aware of the use of artificial intelligence in our everyday lives, the potential to cause harm through algorithmic misgendering, directly or indirectly, should be recognized. Approaches to gender prediction are, more often than not, based on either unclear or problematic notions of gender (e.g., gender = sex assigned at birth). As a technology, it has enormous potential for harm to people of trans experience—whether they find out about their result directly, or receive a particular treatment because of the result. If it is necessary to know the gender of a subject, that information should be willingly provided by the subject. Studies involving gender should show consideration for individuals that do not fit perfectly within the conception of gender as a binary and static attribute. In this way, research involving gender can be formulated ethically and reduce the risk of causing harm to people who are marginalized because of their gender identity.

8 Future Work

The work in this paper provides a starting point for further investigation into various manifestations of gender through language or other means. Other studies could involve investigations into the ways in which gender norms and marked language might change over time in formal or informal texts—this could involve studying the same individuals over their lifetime, or generational changes across general populations. Furthermore, I would be interested to see if there is covariation between perceptions of masculinity/femininity and attitudes toward individuals, such as in online dating profiles, college professor

ratings, or public figures, and how that interacts with other demographic factors. Gender variation could also be of particular interest in languages where speech and language patterns are culturally marked for gender, such as Japanese. Gender stereotypes are also often reflected in language, so it is potentially possible to quantify them and compare them across various media and genres (e.g., news sources, movie/television scripts, books, etc.). How do gender associations differ between the New York Times and the Wall St. Journal? Are they reflected differently in movie genres such as horror as opposed drama? What about children's books and young adult books? Gender and parenting could also shed some light on early gender fortification—do parents speak differently to a child that has been assigned male at birth vs female? How and in what ways? The answers to these questions can help illuminate the ways in which gender is constructed collectively, and how individuals are influenced to think about gender differently based on certain affiliations, personal preferences, or even at different stages of life.

9 References

Argamon, S., Koppel, M., Fine, J., and Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text & Talk*, 23(3), pages 321–346.

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

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, 42(2), pages 155–162.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.

Butler, J. (1993). *Bodies That Matter: On the Discursive Limits of Sex*. Taylor & Francis.

Dadvar, M., de Jong, F., Ordelman, R., and Trieschnigg, D. (2012). Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop* (DIR 2012), pages 23–25.

Fausto-Sterling, A. (2012). *Sex/gender: Biology in a Social World*. Routledge.

Flores, A. R., Herman, J. L., Gates, G. J., and Brown, T. N. T. (2017). How Many Adults Identify as Transgender in the United States? The Williams Institute 2016.

van der Goot, R., Ljubešić, N., Matroos, I., Nissim, M., and Plank, B. (2018). Bleaching Text: Abstract Features for Cross-lingual Gender Prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (Volume 2: Short Papers), volume 2, pages 383–389.

HRC. (2020). Fatal Violence Against the Transgender and Gender Non-Conforming Community in 2020. https://www.hrc.org/resources/violence-against-the-trans-and-gender-non-conforming-community-in-2020. Retrieved Jan 20, 2021.

Harrington, J. (2006). An acoustic analysis of ‘happy-tensing’ in the Queen's Christmas broadcasts. *Journal of Phonetics*, 34(4), pages 439–457.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1): pages 55–67.

Honnibal, M., Montani, I., van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo. https://doi.org/10.5281/zenodo.1212303

Kosinski, M. and Stillwell, D. J. (2012). mypersonality project. http://www.mypersonality.org/wiki/.

Kivinen, J. and Warmuth, M. K. (1995). Additive versus exponentiated gradient updates for linear prediction. In *Proceedings of the 27th Annual ACM Symposium on Theory of Computing*, pages 209–218.

Lakoff, R. (1973). Language and woman's place. *Language in Society*, 2(1), pages 45–79.

Larson, B. (2017). Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 30–40. Association for Computational Linguistics.

Ljubešić, N., Fišer, D., and Erjavec, T. (2017). Language-independent gender prediction on Twitter. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 1–6. Association for Computational Linguistics.

Money, J. and Ehrhardt, A. A. (1972). *Man and woman, boy and girl: Differentiation and dimorphism of gender identity from conception to maturity*. Johns Hopkins University Press.

Monroe, B. L., Colaresi, M. P., and Quinn, K. M. (2009). Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4), pages 372–403.

Nguyen, D., Trieschnigg, D., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., and de Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowd-sourcing experiment. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1950–1961.

Oreskovic, A. (2014). In new profile feature, Facebook offers choices for gender identity. *Reuters*. Thomson Reuters. https://www.reuters.com/article/us-facebook-gender-idUSBREA1C1RU20140214 Retrieved Jan 20, 2021.

Palander-Collin, M. (1999). Male and female styles in 17th century correspondence. *Language Variation and Change*, 11(2), pages 123–141.

Parkins, R. (2012). Gender and emotional expressiveness: An analysis of prosodic features in emotional expression. *Griffith Working Papers in Pragmatics and Intercultural Communication*, 5(1), pages 46–54. Griffith University.

Pennebaker, J. W. and Stone, L. D. (2003). Words of wisdom: language use over the life span. *Journal of Personality and Social Psychology*, 85(2), pages 291–301.

Plank, B. and Hovy, D.. (2015). Personality traits on twitter—or—how to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–9. Association for Computational Linguistics.

Sankoff, G. and Blondeau H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, 83(3), pages 560–588.

Sankoff, G. (2019). Language change across the lifespan: Three trajectory types. *Language*, 95(2), pages 197–229.

Sap, M., Park, G., Eichstaedt, J. C., Kern, M. L., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Schwartz, H. A. (2014). Developing age and gender predictive lexica over social media. In *Proceedings of EMNLP*, pages 1146–1151.

Schler, J., Koppel, M., Argamon S., and Pennebaker, J. W. (2006). Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Volume 6, pages 199–205.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*. Series B (Methodological), pages 267–288.

Verhoeven, B., Daelemans, W., and Plank, B. (2016). Twisty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), pages 1632–1636. European Language Resources Association (ELRA).

Volkova, S., Wilson T., and Yarowsky, D. (2013). Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing*, pages 1815–1827.

Twenge, J. M. (1997). Changes in masculine and feminine traits over time: A meta-analysis. *Sex Roles*, 36(5-6), pages 305–325.

Xu, W., Liu, X., and Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 267–273.