Lexicon-based Similarity Analysis between Transgendered groups and Cisgendered groups on Reddit

Assessing the lexical variation between transgendered and cisgendered individuals and whether transgendered lexicon on the website Reddit shares more similarities with their birth gender or their identified gender.

Howell Lu
Center for Data Science
New York University
New York, New York, United
States
hl4631@nyu.edu

ABSTRACT

This paper uses machine learning methods to explore and answer important questions regarding gender identity within and between two important transgendered groups: biological males who identify as females, commonly referred to as male-to-females or transgendered females or "MtF", and biological females who identify as males, commonly known as female-to-males or transgendered males or "FtM". We will be using all terms synonymously.

This paper attempts to answer two of the most pressing questions regarding gender identity: is the there more lexical variation between transgendered peoples and their birth gender or more lexical variation between transgendered persons and their preferred gender. Furthermore, this paper attempts to answer which of the four gender identities (male-to-female, female-to-male, male, female) are most similar to one another and which ones are most distinct.

This paper analyzes the transgendered community on the popular website "Reddit" which hosts many subcommunities which focus on gender and gender identity. Data mining has allowed the creation of a sufficiently large datasets for posts from MtF's, FtM's, biological males and biological males.

The dataset used attempts to answer questions regarding lexical similarity and lexical variation by utilizing a WordScore model trained on a male to female dimension to assess the femininity and masculinity of female-to-male lexicon and male-to-female lexicon. Further Machine Learning processes are applied to categorize transgendered (MtF and FtM) and cisgendered users (Male and Female) to assess which gender identities are most similar and which identities are commonly conflated together, to identify which gender identities are the most distinct from one another.

KEYWORDS

Transgendered, LGBT, Lexical, Natural Language Processing, WordScores, Random Forest, Machine Learning

1 Introduction

There has been a remarkable growth in two trends in the past decade: the widespread usage of internet forums such as Reddit, Twitter, 4Chan and many others which have changed the nature of communication as the world has becoming increasingly digitalized. Reddit has over 430 million users by the end of 2019 (Roettgers 2019) and has only grown since.

The past decade has also seen the proliferation of many trends relating to gender and gender identity as non-binary identities have increasingly been accepted by the public. The ability and the freedom to identity as more than just male or female have proliferated all aspects of western society, with companies allowing prospective employees to tick they/them or genderqueer on their job applications and an ever-growing population which has begun to list their pronouns on social media. Currently, over 1.2 million Americans now identify as non-binary and this number is only expected to grow (Williams Institute 2021).

The growth of this nascent unstudied demographic intersected with the novelty of communication via a new digital mediums such Reddit rather than traditional methods such in-person communication opens many avenues of deep research into questions pertaining to gender identity as researchers can now apply new computational techniques to extremely large datasets.

This study attempts to attempt to answer two of these questions today: whether the transgendered have a written lexicon more closely resembling their birth gender or to their chosen gender and whether some gender identities' written lexica are distinct than others and which of the four gender identities (male-to-female, female-to-male, male, female) are most similar to one another.

2 Related Literature

Much academic literature has been produced on the topics of vocabulary and lexical variation between cisgendered and transgendered individuals. However, most of the literature regarding gender identity in transgendered persons is sociological research which focuses on the impact of nurture and gender reassignment surgery on the subject's gender identity. Specifically, such research studies whether children who had gender reassignment surgery identified as their birth gender or their present gender. The literature found that cases of early gender reassignment surgery, the subjects chose the gender identity of the reassigned gender rather than their birth gender (Bradley, Oliver, Chernick, Zucker 1998). However, the vast majority of these cases occurred in infancy, without the consent of the child and the subject was usually intersex or had genital issues. The majority of the cases in present day occur much later in life and most gender assignment occurs with the consent of the subject. Furthermore, most of these studies are aged as they were either produced decades ago and the most recent ones have a large delay between the subject's gender reassignment surgery and the determination of what gender the subject identifies with as the judgement of the child's gender identity can only be truly and stably determined well into adulthood. Furthermore, many of these studies were produced in a less accepting era, and the subjects likely felt more need to conform with gender norms.

There has been research into the lexicon of the transgendered community, however, most the research was not done on digitalized media and focused more spoken word rather than written word. These studies focused on aspects vernacular speech with an emphasis on articulation and the lexical variables such as intensifiers and litotes (Hazenberg 2012) (Lal 2018).

Most of the studies regarding transgendered people on online forums attempted to study the macroscopic behavior of the communities at large. Many of these studies attempted to analyze the sentiment of communities on popular subreddits on Reddit (Li, Wang, Zhao, Li 2020). Another attempted to study how transgendered users on Reddit took advantage of technical anonymity to manage an anonymous identity as there is still great stigma in being LGBTQ in many parts of the world (Triggs 2019). There were few, if any that researched into lexical choices. Lastly, few distinguishments, if any, were made between most of the past literature, as studies simply pooled any non-binary identity together instead of differentiating between different aspects of gender as some studies which utilized data from Twitter to whether classify posters were LGBT or not (Karami 2021). There is very little significant literature when it came to transgendered users of internet forums as they were generally lumped into the wide basket of LGBTQ, furthermore there was even fewer studies that made the distinction between "Male to Female" and "Female to Male" transgendered. To the best of my knowledge, there was absolutely no analysis on the lexical variation between the transgendered compared to the cisgendered on internet forums.

3 The Corpus

Before analyzing transgendered posts and cisgendered posts, a large corpus must be created and processed. The method this article used to mine posts from Reddit.com was through the PRAW (Python Reddit API Wrapper).

Reddit is structured with popular subforums in the shape of "subreddits". A subreddit would contain posts relating to a certain issue. For example, the subreddit "TwoXChromosomes" would be a subreddit which relates to female issues, the subreddit "FTM" relates to issues that female-to-male transgendered persons might find important and the subreddit "MTF" would contain posts relevant to male-to-female transgendered persons. Each subreddit has posts which are submitted by an author. Therefore, it is fair to assume that almost, if not all the posts in a specific forum were posted by gender of which the forum is designed for as it makes no reason for a cisgendered male to post in a transgendered forum.

The method of which this paper derives its dataset from is by finding the authors of the newest posts from the subreddits: "mtf", "ftm", "TwoXChromosomes". Code would scrape for the authors of 300 most recent posts on these subreddits (500 for "FtM" and "MtF", convert it into a set to ensure that every author is unique and then scrape for the 500 most recent comments from the author (1000 for "MtF" and 1000 for "FtM"). These comments would come from not only the subreddits "mtf", "ftm", "TwoXChromosomes", but also non-gender orientated subreddits such as "Pokemon", "Socialism" and others.

The corpus for males was derived in a different manner as there was no specific subreddit designed for males. Therefore, I scraped for all the top-level comments for author's names in a subreddit called "AskMen". The purpose of this subreddit is posters to receive answers from males, therefore it is almost certain that only men would respond to posts on this subreddit. Therefore, the authors of all the top-level comments from the most recent three posts were scraped for. Then post history in the form of the 1000 most recent posts from 500 authors which commented on the three most recent submissions would scrape for and compiled into a dataset.

The terminology and the jargon of "ftm" and "mtf" subreddits were so obscure that it is unlikely that anyone who was not "mtf" or "ftm" would post there. Furthermore, the subreddit "TwoXChromosomes" is heavily moderated and as there are over 12 million subscribers which mainly discuss female topics, it is likely that any posts that are not females would be a rare outlier. However, it was unlikely but there was still a non-zero possibility that females might be first level commentors to submissions on the subreddit "AskMen".

This process generated a total of 396041 comments, 137035 were from the male-to-female corpus, 112579 was from the female-to-male corpus, 75442 was from the male corpus and 70985 was from the female corpus.

4 Removing Explicit Gender Identifiers

The purpose of this study is to discern whether the lexicon of transgendered persons can be discerned from the cisgendered and whether transgendered lexicon is closer to their birth gender or their assigned gender transgendered lexicon from the other group. However, one issue with scraping reddit is that transgendered males and transgendered females would generally post in locations in which their genders would be assumed. For example, if subjects post on the subreddit, “ftm”, it would be assumed that they are female-to-male. Posts in these specific subreddits would have specific jargon that would be far too indicative of one’s gender. Furthermore, words such as “ftm”, “trans”, “non-binary” should not be included in the corpus as the intent is this article is not to see whether modern NLP techniques can identify transgendered males and females in posts where their genders are assumed but rather if modern NLP techniques can discern the genders without any self-labeling by the subjects. Likewise, any posts made by females and males in locations which implicitly identified them or in posts which they explicitly self-identified their genders would be excluded from our corpus.

All posts from these subreddits would be removed. Posts containing these strings will also be removed(case insensitive):

“FtM” Corpus	“MtF” Corpus	Male Corpus	Female Corpus
Nonbinary	Nonbinary	Male	Female
Trans	Trans	Men	Women
Egg	Egg	Man	Woman
Gay	Gay	Gender	Gender
Non-binary	Non-binary	Boy	Girl
Female	Female	Askmen	TwoXChromosomes
male	male		
cis	cis		
Women	Women		
Men	Men		
Woman	Woman		
ennnnnnnnnnnnbbbbbby	ennnnnnnnnnnnbbbbbby		
man	man		
gender	gender		
ftm	mtf		
queer	queer		
traaaaaaannnnnnnnnns	traaaaaaannnnnnnnnns		
Boys	Boys		
Girls	Girls		
lgbt	lgbt		
enby	enby		
lesbian	lesbian		
.com	.com	.com	.com

After post-processing, the number of comments was heavily reduced to only 224124, with male-to-females having 58093 comments, females-to-males having 59444 comments, males having 48746 comments and females having 57841 comments.

6 WordScore for Lexical Similarity Comparison

The answer to the question: “Do transgendered persons communicate more like their birth gender or their assigned gender?” is deceptively more difficult than it appears. The traditional method for assessing whether a corpus would be predominantly masculine or feminine would be the dictionary approach and utilization of an arbitrary cutoff for masculine and feminine to categorize the male-to-female corpus and the female-to-male corpus as either masculine or feminine. The standard dictionary approach would not work as it is difficult to find a find a dictionary which is applicable to a modern internet forum. The lexical ecosystem in an environment such as Reddit would be filled with modern slang, jargon, emojis and thereby significantly different than any established dictionary. Methods such as the naïve bayes classifier would not work either as the purpose of this study is to see where the transgendered lexicon lay between two texts rather than creating a predictive model.

However, this question can be answered with the Wordscores technique. Wordscores uses reference texts with known positions and then compares novel texts to these reference texts. Wordscores will be applied by using two corpuses of 40,000 randomly selected comments from males and females and using these as our anchor documents. Transgendered documents with words that are commonly seen in the male corpus but not words in the female corpus will be classified male and documents with words seen in the female corpus but not words seen in the male corpus would be classified female. This methodology is essentially creating a dictionary from anchor documents to assess the similarities between the reference test and the novel text. The female corpus is assigned a score of -1 and the male corpus will be assigned the label of 1. The anchor corpus is composed of 40,000 random comments from the biological female corpus and 40,000 random comments from the biological male corpus. Documents from the transgendered corpus will be rounded to a value of -1 or 1 and that is how classification is done.

6.1 Data Preprocessing

Before implementing this model, preprocessing is required as creating a wordscore model is extremely computationally intensive, especially with the current size of the dataset and processing a DFM with superfluous information hurts the accuracy of the model. Therefore, decisions needed to made to reduce the number of types. The decision was made to ignore case, to remove punctuation, remove stop words from many languages and to remove numbers. These choices were applied to every DFM for all four gender identities. Capitalization often does not make a significant impact on the model as words that are capitalized due to the sentence structure are not inherently different than uncapitalized words, nor does capitalization sufficiently change the meaning of the word. Punctuation and stopwords are used by all genders and therefore these cues do not add significant information but add to the noise and the computational intensity of the model. Numbers needed to be removed solely as some users posted their phone numbers at the end of every post and a specific individual’s phone number is not indicative of the tastes and the preferences of an entire gender identity. Stop words from other languages were removed as many users are multilingual and often post in different languages.

The unorthodox decision not to STEM words was due to previous research showing that there is a sex difference in tense choice (Kidd and Lum 2008) . Furthermore, testing showed that results were much more conclusive when words were not stemmed so the decision was made not to stem words.

6.2: Most Common Words

WordScores created a set of the most common words for males, females, transgendered males and transgendered females. The top 10 most common terms (excluding some emojis which could not be parsed) are listed.

Male words	Female Words	Male to Female Words	Female to Male words
░	many	💃	gundham
┼	Human	♏	imeatingdirt
█	men	■	https://yubo.live/en...
girl	managing	□	dmd
⣿	mental	♋	dodo
të	mention	esteele22	istj
lebron	moment	❒	6700
girlfriend	boy	◆	6260
🤙	♡	gardian	2068
dhe	boyfriend	♑	accutane

Two issues occurred with the way the dataset emerged. Firstly, I could not filter out comments which often had signatures and thereby many names such as “imeatingdirt” became prevalent in the dictionary. Secondly, many obscure languages have stopwords and are such languages spoken by some users. Removing some of the stopwords from more obscure languages is difficult as these languages often did not have a dictionary of stop words.

6.3: Results and Conclusion

One interesting question of note is whether WordScore would classify comments from transgendered males and transgendered females as text from their it’s biological gender or as it’s assigned gender. WordScore solved this question evidently. The male-to-female corpus was closer to the male corpus and the female-to-male corpus was closer to the female corpus.

Wordscore was used to estimate different subsets of the Transgender corpus. The transgendered corpus was subsectioned thrice. The first subset took 20,000 random MtF comments and 20,000 random FtM comments to see where their wordscore estimates. Wordscore used the “LBG” correction too. In this subsection, the FTM corpus has a score of -1.009 and the MTF corpus has a score 0.99. The male corpus had a score of 1 and the female had a score of -1.

The second subset took 150 samples of MtF text comprised of 200 comments each and 150 samples of FtM text comprised of 200 comments each. The estimates for samples from this subset of the transgendered corpus will be rounded to a value of -1 or 1 for classification. WordScore labelled 99.7% of these samples as their biological gender with 299 of 300 samples being labeled as biological gender. The last subset approached classification in the same manner and took 235 samples of MtF text comprised of 100 comments each and 235 samples of FtM text comprised of 100 comments each, 451 of the 470 (96%) samples were labelled as their biological gender.

Results of WordScore Estimation for Samples of 100 comments.

	Wordscore Female Estimated	Wordscore Male Estimated	
Female to Male	229	6	235
Male to Female	13	222	235
	242	228	470

The conclusion to be made from the study of this corpus is that WordScores estimated that transgendered individuals have a lexicon much closer to their birth gender than their biological gender.

7 Classification with Random Forest

A question that is brought up often in transgendered communities is the question of “passing”. In the physical world the speech patterns, vocal pitch and physical appearance is instrumental in whether transgendered individuals pass. In online communities, lexical choices and mannerisms are specifically curated as to seem as close to one’s chosen gender as possible. There is still often a stigma against communicating too much like one’s birth gender.

The questions of “passing” can be asked to Machine Learning and whether individuals “pass” according to a RandomForest classifier. WordScores have concluded that transgendered lexicons are closer to the lexicon of their biological gender. However, a further question is whether the transgendered corpuses are unique enough to be correctly identified? What percentage will be misclassified as their birth gender, what percentage will be misclassified as their assigned gender, will transgendered males be misclassified as transgendered females? This can also be considered a proxy for lexical distance.

Many options could have been used for multiclass classification such as naïve bayes, decision trees, and k-Nearest neighbors. The decision to use Random Forest instead of other methods is due to the flexibility of the model, robustness against overfitting and the ability of the model to account for non-independence between variables. K-Nearest neighbors was not used as there was no guarantee whether gender identity would be the most discriminant latent variable.

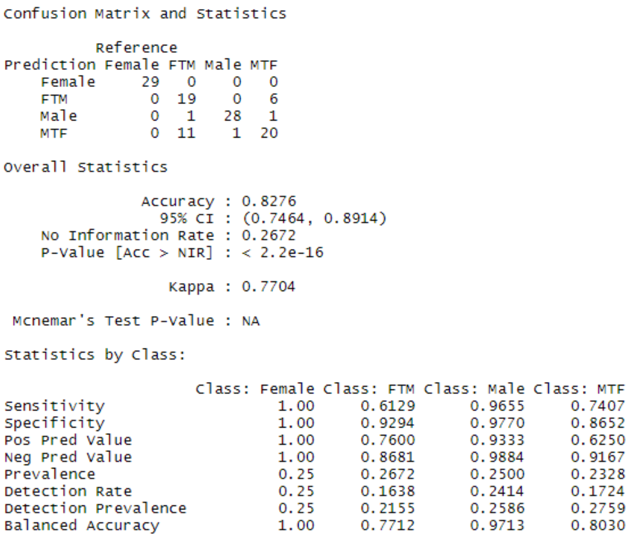
7.1 Data Preprocessing

Data Preprocessing is done in a similar way to the data preprocessing was done for WordScores, with capitalization removed, stopwords removed, punctuation and numbers removed. However, comments are not aggregated, as the identity of the authors must be preserved for classification purposes as to ascertain whether individuals "pass" according to computers. Therefore, all comments from an author are aggregated into a document, and all such documents are given a class of "FTM", "Male", "Female", "MTF". Then the data is split up into an 80:20 train-test split and the model is trained to be used for classification.

7.2 Model Construction

There are a multitude of settings that are available for model construction. Questions regarding the number of trees generated and the number of variables sampled at each split and test-train split was also an important question. The study decided to keep test-train split at 80-20 due to convention. The DFM created for this test had a total of 12416 features, therefore this study used 3(sqrt(features)) as the number of features to be mtry, or the number of features sampled at each iteration. I decided to also generate 1400 trees for ntree as the number of trees generated. These values were fine-tuned from testing.

7.3 Results and Conclusions



```
Confusion Matrix and Statistics

          Reference
Prediction Female FTM Male MTF
    Female     29   0    0   0
    FTM         0  19    0   6
    Male        0   1   28   1
    MTF         0  11    1  20

Overall Statistics

               Accuracy : 0.8276
                 95% CI : (0.7464, 0.8914)
    No Information Rate : 0.2672
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7704

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: Female Class: FTM Class: Male Class: MTF
Sensitivity                   1.00     0.6129      0.9655     0.7407
Specificity                   1.00     0.9294      0.9770     0.8652
Pos Pred Value                1.00     0.7600      0.9333     0.6250
Neg Pred Value                1.00     0.8681      0.9884     0.9167
Prevalence                    0.25     0.2672      0.2500     0.2328
Detection Rate                0.25     0.1638      0.2414     0.1724
Detection Prevalence          0.25     0.2155      0.2586     0.2759
Balanced Accuracy             1.00     0.7712      0.9713     0.8030
```

Random Forests had an impressive level of accuracy, as 83% of all classes were predicted accurately and some classes were nearly perfectly predicted such as the male class and the female class. The baseline accuracy derived from estimating the most popular class repeatedly would be 26.72%, so the Random Forest algorithm exceeded that significantly.

The metric the study will be using for accuracy per class will be the balanced accuracy as the classes are nearly evenly proportioned (29, 31, 29, 27) and thus there is no strong reason to use F1-accuracy.

What's found from the Random Forest algorithm is that Females and males are very distinctive and rarely ever get confused with each other or any transgendered gender identity. However, MtF's and FtM's are commonly confused with one another. 6 "MTFs" were wrongly assigned FTM labels and 11 "MtF's" were wrongly assigned the "FTM" label. Males and Females had near perfect sensitivity and specificity scores, whilst transgendered males and females had balanced scores around the 80% range, but almost all the mispredictions were within other transgendered groups. This implies that there is more lexical similarity between transgendered males and transgendered females than there is between the transgendered and any cisgender group. This heavily implies that transgendered lexicon is unique and does not mirror cisgender lexicon.

8 Conclusion

This paper initially set out to answer two questions: whether transgendered persons have a written lexicon more closely resembling their birth gender or to their chosen gender and which of the four gender identities (male-to-female, female-to-male, male, female) are most similar to one another. These questions have been solved, the lexicon of transgender males is most like their birth gender (female) and the lexicon of transgendered females is most similar to their birth gender (male). Secondly, there is more lexical similarity between transgendered males and transgendered females is greater than the lexical similarity between any transgendered group and any cisgender group. The lexicons of males and females are distinct from any of the other gender identities.

REFERENCES

Hazenberg, Evan Nicholas Leo. "Language and Identity Practice : A Sociolinguistic Study of Gender in Ottawa, Ontario." *Memorial University Research Repository*, Memorial University of Newfoundland, Sept. 2012, https://research.library.mun.ca/2346/.

Karami, Amir, et al. "Automatic Categorization of LGBT User Profiles on Twitter with Machine Learning." *MDPI*, Multidisciplinary Digital Publishing Institute, 29 July 2021, https://www.mdpi.com/2079-9292/10/15/1822/htm.

Kidd, Evan, and Jarrad A.G Lum. "Sex Differences in Past Tense Overregularization." *Developmental Science*, U.S. National Library of Medicine, Nov. 2008, https://pubmed.ncbi.nlm.nih.gov/19046157/.

KJ;, Bradley SJ;Oliver GD;Chernick AB;Zucker. "Experiment of Nurture: Ablatio Penis at 2 Months, Sex Reassignment at 7 Months, and a Psychosexual Follow-up in Young Adulthood." *Pediatrics*, U.S. National Library of Medicine, July 1998, https://pubmed.ncbi.nlm.nih.gov/9651461/.

Li, Mengzhe, et al. "Transgender Community Sentiment Analysis from Social Media Data: A Natural Language Processing Approach." *ArXiv.org*, 25 Oct. 2020, https://arxiv.org/abs/2010.13062.

Roettgers, Janko. "Reddit Ends 2019 with 430 Million Monthly Active Users." *Variety*, Variety, 4 Dec. 2019, https://variety.com/2019/digital/news/reddit-430-million-mau-1203423360/.

Triggs, Anthony Henry, et al. "Context Collapse and Anonymity among Queer Reddit Users." *New Media & Society*, vol. 23, no. 1, 2019, pp. 5–21., https://doi.org/10.1177/1461444819890353.

Triggs, Anthony Henry, et al. "Context Collapse and Anonymity among Queer Reddit Users." *New Media & Society*, vol. 23, no. 1, 2019, pp. 5–21., https://doi.org/10.1177/1461444819890353.

Williams Institute. "1.2 Million LGBTQ Adults in the US Identify as Nonbinary." *Williams Institute*, UCLA, 22 June 2021, https://williamsinstitute.law.ucla.edu/press/lgbtq-nonbinary-press-release/.

Zimman, Lal. "Transgender Language, Transgender Moment: Toward a Trans Linguistics." *Oxford Handbooks Online*, 10 July 2018, https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780190212926.001.0001/oxfordhb-9780190212926-e-45.