Lexical Similarity Analysis between Transgender groups and Cisgender groups on Reddit

Assessing the lexical variation and lexical distance between transgendered and cisgendered individuals on the website Reddit.

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

ABSTRACT

Lexical differences exist between demographics, as different groups have different preferences and interests which reflect in word choice and word frequency. I examine the lexical differences, in the form of word usage and frequency, within and between two transgendered groups: assigned male at birth individuals (AMAB) who choose to identify as females, commonly referred to as trans women or transgender females, and assigned female at birth (AFAB) individuals who identify as males, commonly known as transgender males or trans men.

Over two million words are mined, filtered and transformed into 4 classes from subforums relating to transgender females, transgender males, cisgender males and cisgender females from the popular website "Reddit".

Three questions of note regarding gender identity: is there more lexical distance between transgender groups and their gender assigned at birth or is there more lexical distance between transgendered persons and their chosen gender? How much distance is there between all four different groups? Can modern machine learning techniques identify the gender identity of a random reddit user with a high degree of accuracy? I use WordScore, Jensen-Shannon and RandomForest to answer these questions respectively. It is found that transgender persons on Reddit have a written lexicon which is closer to their assigned gender at birth. Out of the 4 different subgroups; transgender males and cisgender females have the least lexical distance from each other, while transgender females and cisgender females have the most lexical distance from one another. RandomForest predicted the gender class of our test data with 69% accuracy when trained on a dataset of 464 users and tested on a dataset of 125 users.

KEYWORDS

Transgender, LGBT, Lexical, Natural Language Processing, WordScore, Random Forest, Machine Learning, Jensen-Shannon

1 Introduction

There has been remarkable growth in two trends in the past decade: the widespread usage of internet forums such as Reddit, Twitter, 4Chan 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 many widespread trends relating to gender and gender identity as non-binary identities as come to the forefront of our public consciousness. Transgender identity has become a prominent topic in public discourse with issues such as restroom access and legal identification frequently making headlines and influencing legislative debates. Moreover, public acceptance of transgender issues and willingness to accommodate transgender identities have been ingrained into our culture exemplified by the increasing use gender pronoun labelling (he/him) before titles. 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, which had historically been understudied, often compelled to conceal their identities has been illuminated by new forms of anonymous digital communication, such Reddit, which opens many avenues of deep research into uncharted territories as questions pertaining to gender identity, lexicon of subpopulations, can be analyzed under new lenses as researchers can now apply new computational techniques to formerly obscure and marginalized demographics.

I attempt to attempt to answer three questions pertaining to this once obscure demographic: whether the transgender populations have a written lexicon more closely resembling their birth gender or to their preferred gender? How much distance is there between all the groups? Can individuals be classified into one of the 4 classes (cisgender female, cisgender male, transgender male, transgender female) from their comments alone with any semblance of accuracy?

2 Related Literature

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. That research generally focuses on the impact of nurture and gender reassignment surgery on the subject's gender identity. Specifically, whether intersex children who have had gender reassignment surgery identified as their birth gender or their present gender. Many of these studies specifically investigated the nature vs nurture aspect of gender reassignment for intersex children at birth. The literature concluded that in 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, wherein the child was generally born with an intersex condition or a genital deformity. The surgeries were done without the consent of the child, and in many cases, the child was not even aware of this situation until much later in life.

Furthermore, most of these studies are aged as they were either produced decades ago and these studies 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 securely determined well into adulthood. Furthermore, many of these studies were produced in a less accepting era, and the subjects likely felt more need to feel the need to conform with gender norms. In addition, these studies were conducted as longitudinal behavioral interviews, rather than digitalized big data analysis.

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

Other studies focusing transgender people on online forums attempted to study the macroscopic behavior of the communities at large. These studies attempted to analyze the sentiment of communities on popular subreddits on Reddit, to determine whether posts within transgender communities were positive, negative or neutral (Li, Wang, Zhao, Li 2020). Other studies 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, and much reddit's community is global (Triggs 2019). There were only a few that researched lexical choices.

However, these studies only had a few distinguishments, between different transgender identities, as studies simply pooled any non-binary identity together instead of differentiating between different aspects of gender identity. Studies which utilized data from Twitter to whether classify posters were LGBT or not, did not differentiate between transgender populations and other LGBTQ populations (Karami 2021). Much of the literature generally lumps transgendered populations into the wide basket of LGBTQ, never making the distinction between "transgender males" and "transgender females". To the best of my knowledge, there was absolutely no analysis that I could find on the lexical variation between the transgendered compared to the cisgendered on internet forums.

3 The Corpus

A large corpus was be created and processed to analyze cisgender and transgender posts. I mined posts from Reddit.com using PRAW or the Python Reddit API Wrapper (Boe B.) A total of 396,031 raw comments were scraped and compiled.

Reddit is structured into 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" (colloquial acronym for female-to-male) relates to issues that transgender males might find important and the subreddit "MTF" (colloquial acronym for female-to-male) would contain posts relevant to transgender females. Each subreddit contains posts which are submitted by an author. Therefore, although impossible to fully verify, it is assumed that most of the posts in a specific forum were posted by the gender that the forum is designed for. It is unlikely for an individual to post in a forum outside their gender identity, however flairs help to further filter our dataset. We used the subreddit, "TwoXChromesomes" to select posters to build our corpus for cisgendered females, the subreddit "mtf" to select posters to build our corpus for transgender females, the subreddit "ftm" to select posters for transgender males and the subreddit, "AskMen" to select posters build a male corpus.

The dataset is generated by finding the authors of the most recent posts from the subreddits: "mtf", "ftm", "TwoXChromosomes" and extracting their comments outside of these subreddits. It is equally beneficial as many of these subreddits have specific flairs for authors to tag their identity. Within the subreddit "mtf", only authors which had a tag including "trans" or "mtf" will be selected. The subreddit "ftm" will exclude all authors with "GuestPost" in their flair as the subreddit rules mandate all guests should use the flair "GuestPost" in their post. The subreddit "TwoXChromosomes" does not have any flairs, however, the subreddit is intended only for women as the subreddit rules state, "We ask that you keep this community awesome by submitting content that is relevant to our experiences as women, for women, or about women".

Code would scrape to extract the 300 most recent unique authors of the posts on their corresponding subreddits (500 for "FtM" and "MtF"), that fall within the flair rules, and then scrape for the 1000 most recent comments from the author with all subreddits posted.

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 for posters to receive answers from males, therefore it is almost certain that only men would respond to posts on this subreddit. In addition, there are "Female", "Non-binary" or "Agender" flairs that that non-

males could use when posting. Unique authors, who did not have any non-male flair who posted top-level comments from the most recent posts within the Askmen forum were extracted. Post history from these authors, in the form of the 1000 most recent posts were scraped for and compiled into a dataset.

The terminology and the jargon of "ftm" and "mtf" subreddits are obscure that it is unlikely that anyone who was not "mtf" or "ftm" would post there. In addition, these subreddits are heavily moderated and malicious actors are likely to have their posts removed. Furthermore, the subreddit "TwoXChromosomes" is also 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. It is more likely that non-males might be first level commentors to submissions on the subreddit "AskMen" as there are no rules pertaining to Flair usage.

It is acknowledged that the populations maybe not be perfectly representative, as we are taking posters from three forums as our post authors. While we are identifying commenters as authors to scrape for from the male corpus.

This process generated a total of 396031 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 transgender people is distinctive from cisgender people and whether transgender lexicon share more similarities to their assigned gender at birth or their preferred gender. However, one issue with scraping reddit is that transgender populations post in locations designed solely for their gender identity. As an example, subjects who post on the subreddit, "ftm", would be assumed to be female-to-male. Posts in these specific subreddits would have specific jargon that would be far too indicative of one's gender. Therefore, all posts made within gender specific subreddits must be removed.

Words that define one's gender identity, such as "ftm", "trans", "non-binary" should likewise be removed from the corpus. The intent is this study is to analyze whether techniques can discern between different classes genders without self-labeling by the authors. Likewise, any posts made by females and males in locations which implicitly identified them or in posts which they explicitly self-identified their genders must be excluded from our corpus. In addition, gender identifiers for the opposite gender were also removed, as the removal of a gender identifier for one's own gender without the removal of the opposite gender would cause the existence of a word relating to the opposite gender to be a predictor. For example, the lack of "man" in the male corpus would simply cause "man" to be predictive of a female dataset.

All posts from these following subreddits would be removed and 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	Women	Male
cis	cis	Woman	Men
Women	Women	Girl	Man
Men	Men	Female	Boy
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 195,282, with transgender females having 53,875 comments, transgender males having 55,156 comments, males having 44,175 comments and females having 42,076 comments.

5 WordScore for Lexical Similarity Comparison

Answering the question: "Do transgender people communicate in a manner closer to their birth gender or their assigned gender?" is deceptively more difficult than it appears. One method for assessing whether a corpus would be predominantly male, or female is the scaling model WordScore.

WordScore operates through first creating a corpus reference set from two diametrically opposite corpuses, which are defined as opposites. These anchor corpuses are defined as -1 and 1. In this study, cisgendered males are -1 and cisgendered females is defined are 1.

The corpuses for cisgendered males and females are transformed into a document frequency matrix (DFM), with the rows being the texts and the columns being the words used and the intersection being frequency of each word occurs in each text.

```
Document-feature matrix of: 2 documents, 43,620 features (35.23% sparse) and 0 docvars.
          features
docs         w   r  v  g   u oppos partial  one anyth els
  cisMale   102 357 22 21 269    26       7 2807   723 469
  cisFemale  47 360 30 28 174    26      17 3090   685 424
[ reached max_nfeat ... 43,610 more features ]
```

The DFM is then put through a WordScore weighing wherein every word within both corpuses is assigned a value

according to the frequency of that word according to this formula:

$$w_j = \frac{\sum_{i=1}^{N} p_{ij} S_i}{\sum_{i=1}^{N} p_{ij}}$$

The WordScore score for each word (W_j) is computed by multiplying the number of occurrences of the word in each corpus multiplied by the reference score of each corpus. The sum would then be divided by the number of times that the word occurs.

For instance, if the word "fight" occurred 3 times in a male corpus of 10 words and the word "fight" occurred once in a female corpus with 20 words. The value of "fight" would lean male from the occurrence.

$$w_{fight} = \frac{(3/10)(-1)+(1/20)*1}{3/10 + 1/20} = -5/7$$

The word "fight" is declared male with a value of -5/7. This applied to every column or word within the DFM.

After the reference or anchor dictionaries are created, the anchors are then applied to virgin texts, or the two transgender corpuses. The equation to weigh the virgin texts is this:

$$Corpus\ Position = \frac{\sum_{j=1}^{N} f_j \cdot w_j}{\sum_{j=1}^{N} f_j}$$

This equation is applied to every word of the corpus, as f_j represents the raw count of how many times the j^{th} word occurs within the corpus and w_j represents the score of the word as defined by the anchor documents. The denominator is the total word count of the corpus. The is applied to all the words within both the transgender male and the transgender female corpus to establish a numerical position.

WordScore was specifically chosen due to the sheer size of the dataset, as WordScore is semi-supervised and does not need any significant labelling other than having an anchor document. Other methods such as naïve bayes require the labeling of every word within the dataset, which is relatively difficult with 43,620 features. In addition, methods such as Naïve Bayes can only assign words fully categorically as words must be defined as fully male or fully female.

Off the shelf dictionaries could have been another option. However, that was decided against as the paper wanted to fully capture the nuances of modern interest slang. The lexical ecosystem of Reddit is quickly evolving with far too much modern slang, jargon, emojis for an established dictionary to capture in time.

5.1 Data Preprocessing and Model Training

Preprocessing is required before implementation of a WordScore model to reduce computational complexity, especially with the current size of the dataset. Specifically, stemming is required to improve the level of predictive accuracy and reduce the level of superfluous noise. For example, if the word "fight" was declared a male word within the anchor text, but the word fighting appears within the virgin text then the word "fighting" would not be declared male, which would hurt the predictive accuracy of the WordScore model towards the virgin texts. However, it should be noted that there are notable differences between the tense choice of males and females. One example is that females choose to overregularize compared to males, as females would use words like "holded" rather than "held" (Kidd and Lum 2008). This is ultimately a trade off as keeping more tenses would increase the features at the cost of sparsity, potentially causing a situation where rarer tenses could not be labeled from the anchor texts, "fight" might be declared male but "fighting" might be declared.

The corpus ignored cases, removed punctuation, removed stop words from many languages and removed numbers. These choices were applied to every DFM for all four gender identities. Capitalization was ignored 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 adds to the noise and the computational complexity 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 total size of DFM, corpus word count, was 413468 words for the male corpus and 434840 for the female corpus. That DFM was used to as the anchor to classify the transgender corpus as male presenting or female present.

5.2: Most Strongly Aligned Words

WordScores created a set of the most strongly associated words (over -0.9 or 0.9) for males and females. The top 10 most strongly aligned terms are listed.

Male words	Female Words
▒	♡
┼	💜
⣿	و
të	xxxx
lebron	xoxo
jule	💔
🤌	damon
në	trimest
për	exmo
jokic	itchi

One issue surfaced after the dataset emerged. Many less spoken languages have stopwords and are such languages spoken

by some users. These words were “për” and “në” which are Albanian. Removing some of the stopwords from more obscure languages is difficult as these languages often do not have a dictionary of stop words. These are unique outliers that arise from scraping which are distinctive of a way that a certain individual communicates, such as the inclusion of their username at the bottom or common stopwords in obscure languages. These words wo;; be deemed as nearly entirely male or entirely female, and will present a problem in the rare instance of that specific type of word being presented in a transgender corpus; For example, if a female posts Albanian stopwords often in their reddit history or posted an MBTI at the end of every post such as “ISTJ” and a transgender male also posts that same MBTI signature or also posts in Albanian, then the transgender male corpus may skew closer to the female corpus than had the outliers been removed. It should be noted that these outliers are still somewhat contained as a large sample size was used. Each corpus uses the posts of at least 300 different users.

5.3: Testing

WordScore was trained on the entirety of the cisgender female and cisgender male DFM, with the value -1 being defined as female and 1 as male.

The testing data is compiled from aggregating comments. Three different tests were conducted, the first was testing 235 batches of 100 transgender male comments and testing 235 batches of 100 transgender female comments. WordScore attempted to estimate the gender identity of those 470 comments. The second test classified 150 batches of 200 transgender male comments and 150 batches of 200 transgender female comments. The last test created 1 batch of 20,000 transgender male comments and second batch of 20,000 transgender female comments. All three tests applied WordScore to their subsequent batches to generate a prediction regarding the lexical distance between all 4 demographics.

One final feature that must be noted it is very common for WordScore to produce estimates that are very close to zero, as the values of each specific WordScore must be between -1 and 1 and once divided by the total number of words used in the corpus, the WordScore measures often tend to be rather small. Therefore I rounded the WordScore to the closest of the two classifications

5.3: Results and Conclusion

The results are as follows:

Batches of 100 Comments	WordScore Classifies Batch as Female	WordScore Classifies Batch as Male	
Transgender Male	232	3	235
Transgender Female	50	185	235
	282	188	470

Batches of 200 Comments	Wordscore Assigns as Female	Wordscore Assigns as Male	
Transgender Male	146	4	150
Transgender Female	22	128	150
	168	132	300

The result for the batch of 20,000 comments is as follows: the transgender male corpus was given a score of -0.0266 which means that it leans female, and the transgender female corpus was given a score of 0.0157, which means that it leans male.

WordScore produced conclusive results, as both transgender groups use words that are closer to their assigned gender at birth rather than the one of their current gender identity. In addition, that transgender males use a vocabulary that is more closely shared with cisgendered females while transgender females have a vocabulary that is more distant to their closest relative, cisgendered males, as the magnitude of the WordScore of transgendered males is about 70% larger than the magnitude of the WordScore of cisgender females.

WordScore provided an answer to one of the questions stipulated, whether transgender individuals have more lexical similarities to their assigned gender at birth or their chosen gender identity.

6: Measures of Distance: Jensen-Shannon Divergence

WordScore is an excellent measure to quantify the difference between a virgin corpus towards an anchor. However, it fails to classify the distance between the virgin corpuses with each other and the distance between the anchors themselves. Jensen-Shannon divergence is a solution as this method that bypasses these shortfalls and allows for the quantification of the similarity or difference between any two distributions.

6.1 Data Preprocessing

Jensen-Shannon can be applied to any probability distribution, as the method works by taking the average of two distributions and calculating the distances between each distribution to the mean of these distributions. This is done at a high level by using the R library philentropy (HG D 2018). Therefore, only two steps are necessary; ensuring that the size of the corpus for each class is about the same, so that the model is not biased towards the majority class. Secondly, every row (class) needs to be converted into a probability distribution.

I compile the comments of all authors who have written more than 100 comments and then merge them into a single corpus

corresponding to their corresponding gender class. Then we divide the frequency of each instance of the word occurring by the total size of the corpus for each class. All of the class corpuses have total word counts between 341,057 and 391280 so they are balanced enough for our purposes. Afterwards Jensen-Shannon is applied.

6.2 Results and Conclusions

	CisFemale	CisMale	TransFemale	TransMale
CisFemale	0.00000000	0.05371910	0.06400633	0.04227861
CisMale	0.05371910	0.00000000	0.05046510	0.05993018
TransFemale	0.06400633	0.05046510	0.00000000	0.05037442
TransMale	0.04227861	0.05993018	0.05037442	0.00000000

The results of Jensen-Shannon confirm the values that were generated from WordScore: transgender females have closer lexical similarity to cisgender males then cisgender females. Transgender males share more lexical similarity to cisgender females than cisgender males. In addition, this confirms the previous assessment that transgender males are closer to cisgender females than transgender females are to cisgender males in lexical terms.

In addition, it is found that Transgender females and transgender males have close lexical similarity to each other.

7 Classification with Random Forest

A question that is brought up often in transgendered communities is the question of "passing" as their preferred gender. In the physical world, speech patterns, vocal pitch and physical appearance are instrumental in whether transgendered individuals pass as their preferred gender. In online communities, lexical choices and mannerisms are sometimes 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.

Therefore, I attempted to discern whether the 4 different classes could be predicted from training a model from the dataset and whether prior conclusions of distance and distinctiveness of the 4 classes would be reflected within a random forest model.

The questions of "passing" and whether individuals "pass" can be asked to a Random Forest classifier. WordScore and Jensen-Shannon have concluded that transgendered lexicons are closer to the lexicon of their biological gender. However, an extended question is whether independent Reddit users are unique enough to be correctly identified? Will the misclassifications of the Random Forest model reflect the distances found by Jensen-Shannon Divergence? Are Random Forest errors a suitable proxy for lexical distance?

Many options could have been used for multiclass classification such as naïve bayes, 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 ease of use as it is a semi-supervised method. K-Nearest neighbors was not used simply due to the size of the corpus and dimensionality issues.

7.1 Data Preprocessing

Data Preprocessing is done in the same way as Jensen-Shannon, with capitalization removed, words stemmed, stopwords and punctuation and numbers removed too. However, comments are not randomly sampled, and are instead lumped by author, as the identity of the authors must be preserved for classification purposes. The entire library of all the comments written by all authors with over 100 comments compiled, and given a class of "Transgender Male", "Cisgender Male", "Cisgender Female", "Transgender Female". 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 available for model construction. The study decided to use the conventional test-train of 80-20. The DFM created for this test had a total of 11384 features. The model's implemented hyperparameter tuning on these features: number of trees, the number of features considered while splitting each node (mtry), the minimum size of terminal nodes, and the maximum number of terminal nodes. Hyperparameter tuning was done with 5-fold cross-validation on the training data alone. The parameters with the best values was setting mtry at 3 times the square root of the number of features or 320, 1400 trees, having no limitations on the minimum size of the terminal nodes and having no limitations on the maximum number of terminal nodes.

7.3 Results

```
Confusion Matrix and Statistics

             Reference
Prediction    CisFemale CisMale TransFemale TransMale
  CisFemale          23       6           2         3
  CisMale             4      19           1         2
  TransFemale         3       6          21         5
  TransMale           0       0           4        16

Overall Statistics

               Accuracy : 0.687
                 95% CI : (0.5938, 0.7702)
    No Information Rate : 0.2696
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5818

 Mcnemar's Test P-Value : 0.1583
```

The metric the study used was balanced accuracy as the classes are nearly evenly proportioned (30, 31, 28, 26) and thus there is no strong reason to use F1-accuracy. The model was then applied to the test split and Random Forests had an impressive level of accuracy, as 69% of all classes were predicted accurately. The baseline accuracy derived from estimating the most popular class

repeatedly would be 26.95%, so the Random Forest algorithm exceeded that significantly and there is some predictive value from a Random Forest trained on labelled Reddit user data.

It is a stretch to infer results from Random Forest errors compared to other more commonly used methodologies, especially with such a small result set. It does appear that transgender females and transgender males are often misclassified with one another, which concurs with Jensen-Shannon as they share the closest lexical distance. Transgender males do not get misclassified as cisgender females which was something that should have occurred from the results of Jensen-Shannon. However, Cisgender males are accidentally classified as their two closest neighbors: transgender females and cisgender females. Cisgender males are never classified as their furthest neighbor, transgender males. However, classifications from Random Forests are not a direct measure of distance and should not be given too much credence.

8 Conclusion

Three questions were asked: whether transgendered persons have a written lexicon more closely resembling their birth gender or to their chosen gender. What is the distance between the four gender identities (transgender female, transgender male, cisgender male, cisgender female). Can these gender identities be predicted with a fair amount of accuracy? Although confined to a Reddit Dataset, it is concluded the lexicon of transgender males is most like their birth gender (cisgender female) and the lexicon of transgendered females is most similar to their birth gender (cisgender male). Secondly, there are vary degrees of lexical similarity but the two groups with the most similarity are transgender males with cisgender females and the pair with the least similarity are transgender females and cisgender females. It should also be noted that both transgender groups share a large degree of lexical similarity with each other. Lastly, it is possible to predict gender identity using RandomForest trained on a relatively small Reddit dataset.

REFERENCES

Boe B. PRAW: The Python Reddit API Wrapper. 2012-, https://github.com/praw-dev/praw/ [Online; accessed 2022-04-22].

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.

HG D (2018). “Philentropy: Information Theory and Distance Quantification with R.” *Journal of Open Source Software*, **3**(26), 765. https://joss.theoj.org/papers/10.21105/joss.00765.

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.