Andrea Quevedo
PPOL-628

Key Concerns in the Transgender Community: Before and After Obergefell v. Hodges

Introduction

The Supreme Court ruling (Obergefell v. Hodges) indicating that the fundamental right to marry is guaranteed to same-sex couples was a pivotal moment for the LGBTQ community and one that changed the discourse around queer topics. While not being explicitly directed towards transgender individuals, the ruling eliminated the mixed-sex requirement for civil marriage, making people's gender irrelevant to the right to marry and thus making it easier for transgender individuals to marry the person of their choice, irrespective of their sex. Additionally, the Obergefell v. Hodges marriage equality decision further "offer[ed] key support for the propositions that the constitution protects peoples' ability to define and express their gender identities, to shape their own destinies, and that courts can hold the Constitution to protect this gender autonomy", further empowering transgender individuals (Cruz, 2015). Using a corpus from the r/asktransgender subreddit including posts from 2009 to 2020, this project aims to examine whether the key concerns of individuals in the transgender community changed after June 26th 2015 (the day the Supreme Court ruling passed) when compared to prior to Obergefell through the use of topic models.

Corpus Overview

Data relevant to transgender issues, particularly open data, tends to be quite sparse or difficult to find. Yet, the transgender community faces an insurmountable amount of challenges and discrimination in terms of health/health access, housing, and unemployment, among others. With online platforms and communities emerging as sources of networking and support for sexual and gender minorities (SGM), and given the increase of online activity in the past couple of decades, online data from these platforms could provide unparalleled insight into the issues faced by the community as well as their concerns. One such platform is Reddit, which has become one of the primary forums for online discussion in numerous spaces and contexts, as it provides users with a relatively private and safe space in which they can express their opinions and engage in a manner divorced from personal identity. Additionally, having a 300 character limit for their title and a 40,000 character limit for their main body, Reddit posts tend to be longer and potentially more apt for NLP techniques when compared to microblog content originating from sites such as Twitter or Facebook. For this reason, I decided to scrape the subreddit /r/asktransgender, which intends to encourage discussion about transgender issues in mostly a question and answer format, to serve as the main corpus for this project.

The corpus was extracted using the Python Reddit API Wrapper (PRAW) and the Python Pushshift.io API Wrapper (PSAW), which allowed me to scrape the entirety of the subreddit and transform it into a data frame. The resulting csv file includes 258,634 entries (i.e. unique posts) and the following 9 features:

- **ID** : unique alphanumeric ID per post
- **Subreddit**: subreddit each post belongs to (asktransgender in this case)
- **Title**: Post topic
- **Body**: Post content
- **Score** : Post popularity (net up-votes). Number of people who up-voted the content - Number of people who down-voted the content.
- **Number of Comments** : Number of comments for each post
- **Date Created**: UNIX timestamp (ex. 1568978501)
- **Timestamp**: datetime object (ex. 2019-09-20 07:21:41) generated from the Date Created feature

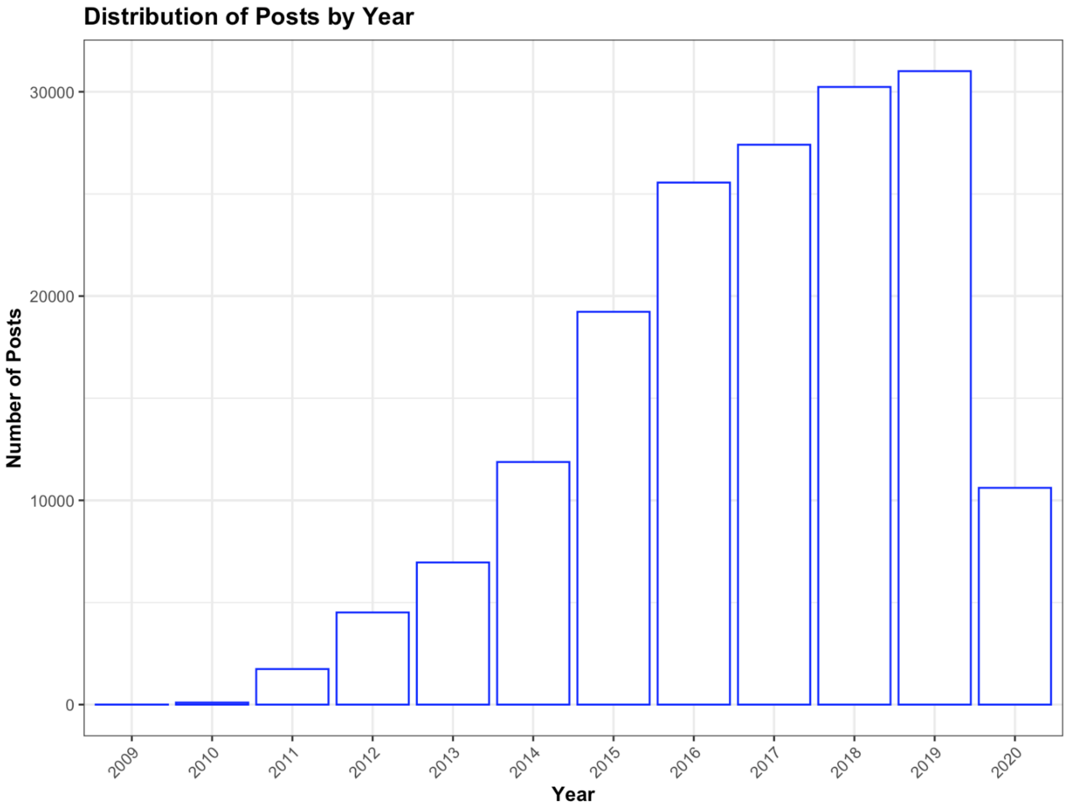
The entries in the data frame range from August 8th, 2009 (the date the sub-reddit was created) until May 6th 2020 (the date the subreddit was last scraped) and are sorted by score in descending order (i.e. starting with the post with the highest score).

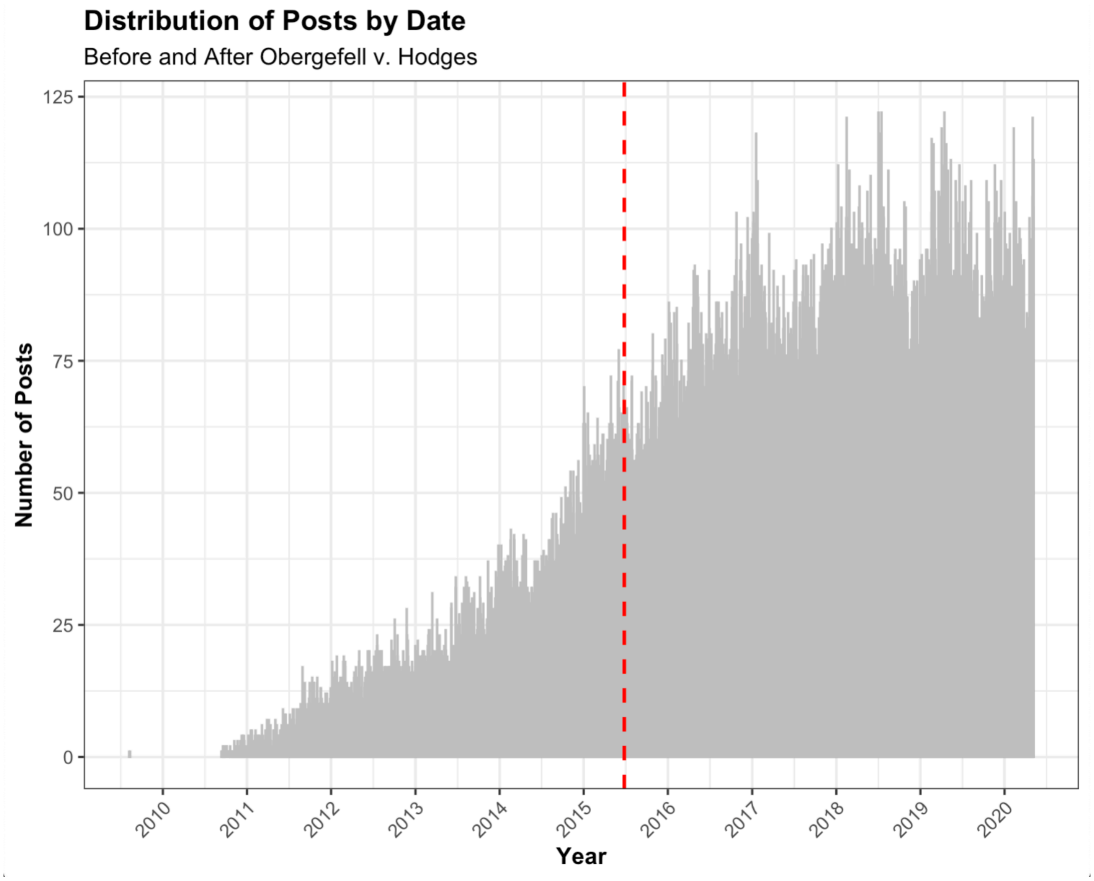
Data Cleaning and Covariate Creation

After filtering out incomplete entries (i.e. posts with NAs in the `body` field, as this was the field with the most missing values), I used regular expressions to remove text not pertinent to users' written speech in both the `title` and `body` columns. This included removing entries whose bodies were composed of only the words "[deleted]" or "[removed]", filtering out rows whose bodies had no blank spaces (i.e. bodies consisting of only emojis, hyperlinks, one word texts, or only blank spaces), and removing hyperlinks, mentions of subreddits and usernames, hashtags, new-line characters, non-ASCII characters (emojis, in this case), and non-alphanumeric characters with the exception of periods and commas. Before performing the latter, I used the textclean package to replace contractions with their long form (e.g. isn't to is not and I'd to I would) in order to increase the interpretability of words in the clean data frame. Without performing this step, when exploring the most frequent words in the corpus as part of data exploration, a lot of the resulting words were the endings of contractions (e.g. n't, 'm, 's, 've, 'd, 'll, 're, etc). Lastly, I trimmed whitespaces in both the `title` and `body` to ensure uniformity in spacing across posts.

Before converting the data frame into a corpus object, I generated the following features to aid in the subsequent analysis of the texts:

- **All_Text**: feature combining the text in `title` and `body` columns. Given that the title has a 300 character and the body has a 40,000 character limit, titles can be informative in their own right. Hence, combining them made more sense for analysis.
- **Char_Length**: Length of post in characters
- **Gay_Marriage**: Binary categorical variable with posts with time stamps before June 26th 2015 (the day the Supreme Court ruled that the fundamental right to marry is guaranteed to same-sex couples) being labeled `Before Approval` and the ones on that date and after labeled `After Approval`.
- **Year**: Year posted (ranges from 2009-2020)
- **Month:** Month posted
- **Weekday:** Day posted (Sunday-Saturday)





The resulting data frame is composed of 169,251 rows and 11 columns. As seen in the figure below, the number of posts seems to be increasing over time, with 2020 having less entries given that it is still ongoing. Additionally, when looking at the number of posts by date, we can see that the numbers also increase over time, with more posts being written daily after Obergefell v. Hodges compared to the years before. This means that overall, there seems to be an imbalance in the number of posts before and after the Supreme Court ruling; it is important to keep this in mind for the analysis going forward.

Preprocessing

After converting the cleaned data frame to a quanteda corpus object, I proceeded to preprocess it. Preprocessing is an important step when extracting the "bag of words" representation of text. In this particular corpus, each post can be thought of as a multi-set of words disregarding sequence and grammar; this allows us to count the number of times each unique word appears in a text and display these counts in a matrix to then perform both supervised and unsupervised techniques on. Preprocessing choices, such as removing stopwords, numbers and punctuation, lowercasing all words, and stemming, allow us to reduce feature size and word irrelevance (removing words that do not allow us to derive meaning from the text). For instance, stopwords including "is", "am", and "the", among others, are words that serve mostly a grammatical purpose and do not convey meaning. By removing them, along with removing word order, we can focus on words that speak to the content of the text itself—we remove the noise in order to derive meaningful insights from the corpus.

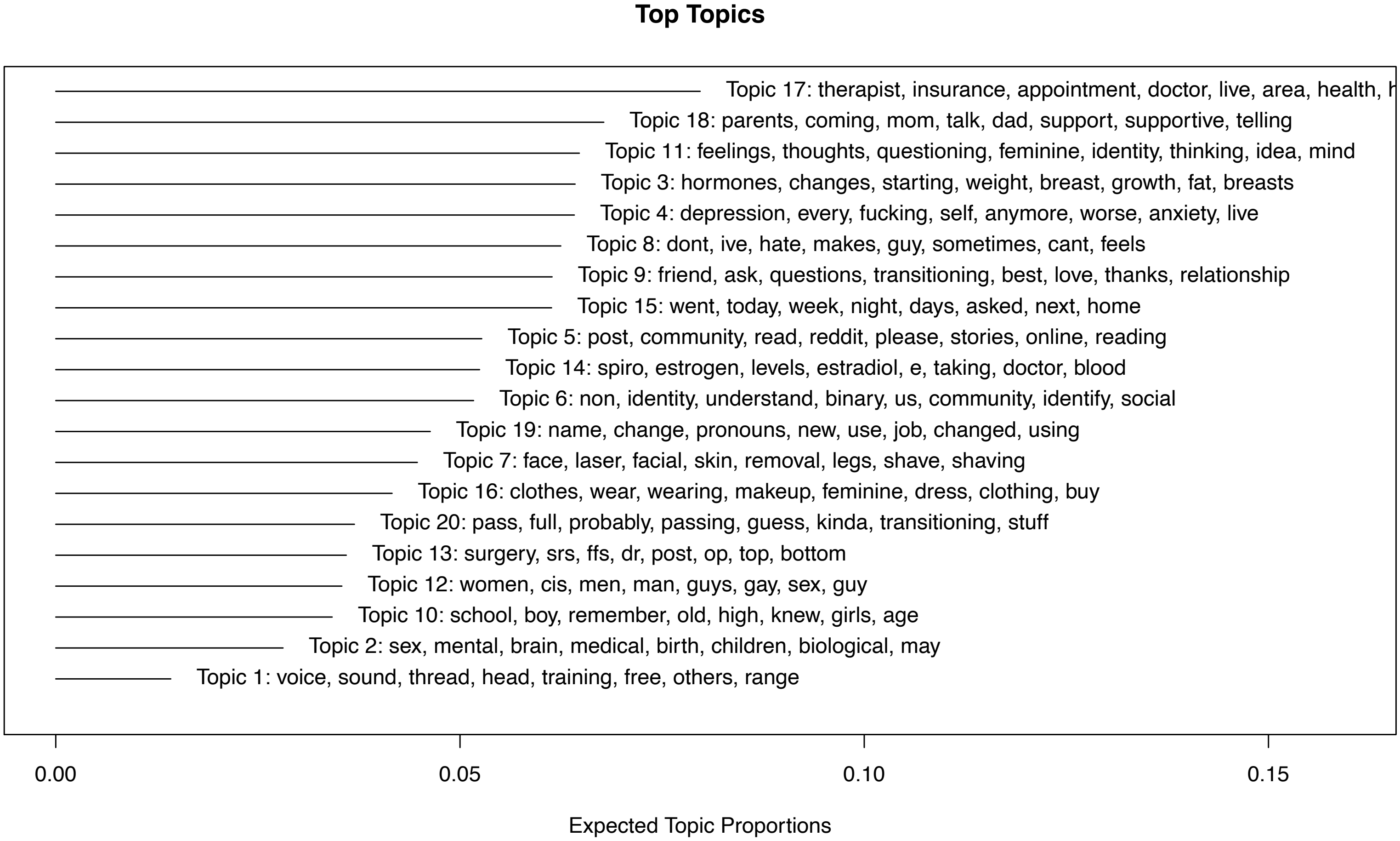
For this particular use-case, I implemented the following preprocessing steps:

- **Removing stopwords, numbers, and punctuation**, as they do not aid in the classification of topics.
- **Lowercasing**. Given that the "bag of words" representation of text is case sensitive, it is common practice to lowercase all words in English language texts in order for words like 'transgender' and 'Transgender' to be considered as equivalent. By lowercasing all words in my documents, I managed to significantly reduce the vocabulary size of the corpus.

- **DTM Trimming.** I removed terms that occur in more than 70% of the documents given that they are not likely to provide meaningful insights about the corpus. For instance, when looking at the most frequent terms in the corpus or when performing topic models, having general domain terms such as "gender", "transgender", and "like" in the output does not add much to the analysis and it might make it harder to decipher interesting topics and insights. I also removed highly infrequent terms (words appearing in less than 20 documents) in order to make the analysis more manageable by reducing the size of the dtm. The resulting DTM included 18,552 features, a significant reduction from the original 188,122.

Methods- Latent Dirichlet Allocation (LDA) and Structural Topic Models (STM)

In contrast to supervised machine learning methods, where we know what the output values should look like, unsupervised learning methods do not use labeled outputs in an attempt to "infer the natural structure present within a set of data points" (Soni, 2018). Hence, given that the aim of this analysis is to decipher the changes in key concerns in the transgender community following the Supreme Court ruling on marriage equality, I decided to implement two unsupervised classification approaches on my preprocessed DTM— Latent Dirichlet Allocation (LDA) and a Structural Topic Model (STM). Both these topic model variations have two general goals, deciphering recurring topics in a corpus along with their proportion in each document, and identifying the words that best represent each topic. Starting from the assumption that each document is a "bag of words" and that we know how many topics we are looking to decipher, the Latent Dirichlet Allocation (LDA) algorithm assigns words to each topic by calculating the probability that they belong to each one. Built on top of LDA, the Structural Topic Model (STM) allows us to incorporate document covariates into the the standard topic model, which in turn produces more accurate estimations and better qualitative interpretability (Lecture 11).



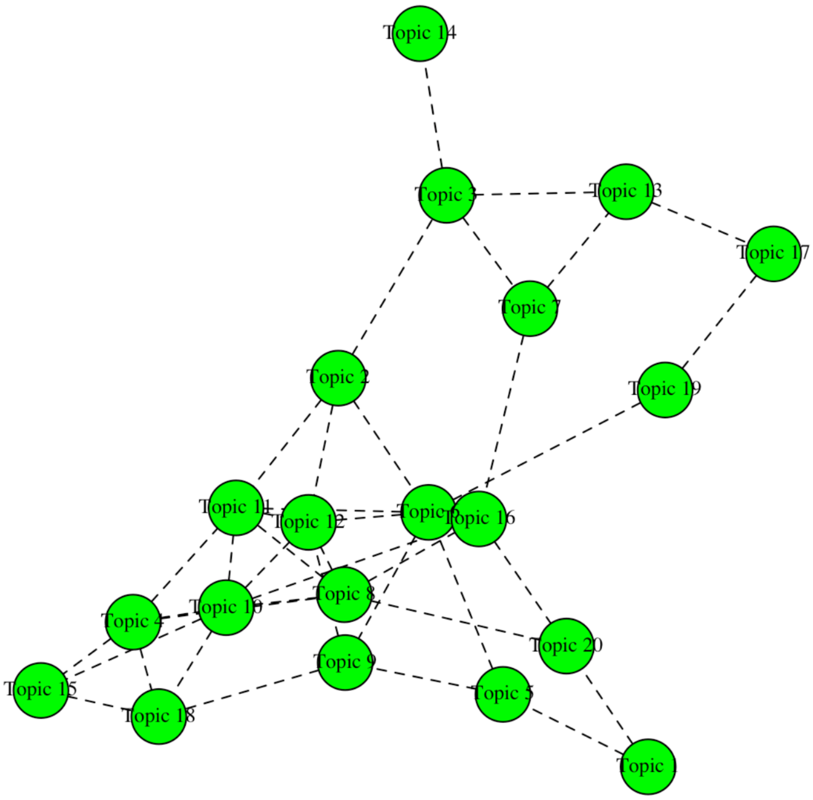
Latent Dirichlet Allocation (LDA) Results

After fitting a simple LDA model with 20 topics, the top 8 terms per topic along with each topic's overall proportion in the corpus is depicted in the following graph:
We can see that topics range from support systems (Topic 9 focuses on friends and relationships, Topic 18 on family, and Topic 5 on online communities), to mental health issues (Topic 4 alludes to issues potentially related to depression and anxiety), legal issues (Topic 19 deals with issues surrounding name change and use of correct pronouns), medical/surgical advice (Topic 13 relates to sex reassignment surgery (sms) and Topic 17 on health insurance and medical appointments), transitioning (Topics 1, 3, 7, 16, 20 focus on different aspects of transitioning), societal perceptions (Topic 6), and gender identity (Topic 11 and 12 focus on questioning one's sexual identity). The model has also produced less interpretable topics such as Topic 2, which seems to list terms pertaining to different topics. The following table depicts the top 20 terms in this topic:

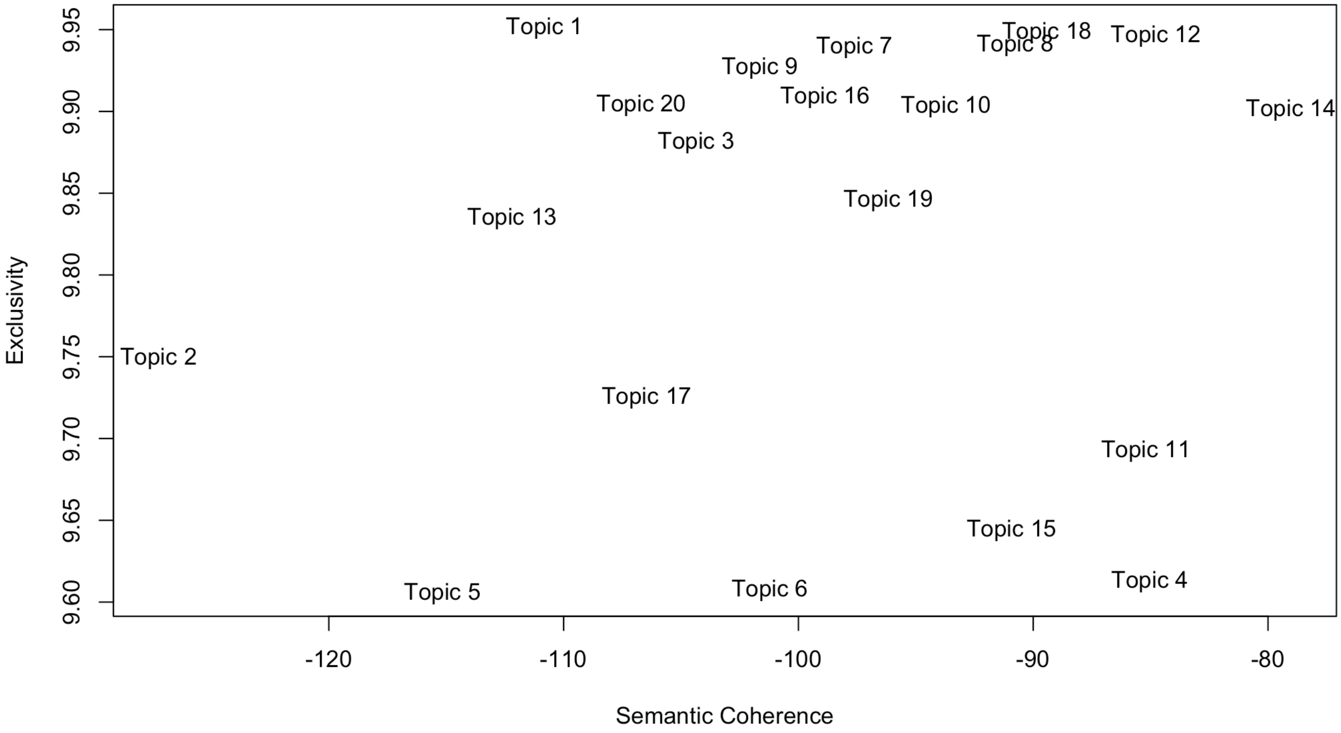
Topic 2
sex, mental, brain, medical, birth, children, biological, may, assigned, hormone, sperm, born, physical, treatment, disorder, different, possible, therapy, research, health

Overall, looking at the distribution of topics across the corpus, I was pleasantly surprised to see that the two most popular topics were related to family and medical support.

Looking at topic clustering and topic correlation within documents, we can see that the topics are very inter-connected, with only Topic 14 being connected to one other topic. When examining the clustering depicted below, for instance, seeing that Topics 11 and 12 co-exist within documents comes to no surprise, as these topics are related to sexual desires and gender identity and they delve into feelings of attraction towards individuals of specific genders and questioning of their own sexuality respectively. It was particularly interesting to see how Topics 10 and 18 co-exist within documents, as the former relates to mental health and general feelings of helplessness while the latter relates to supportive family members and the fear of "coming out" to them.



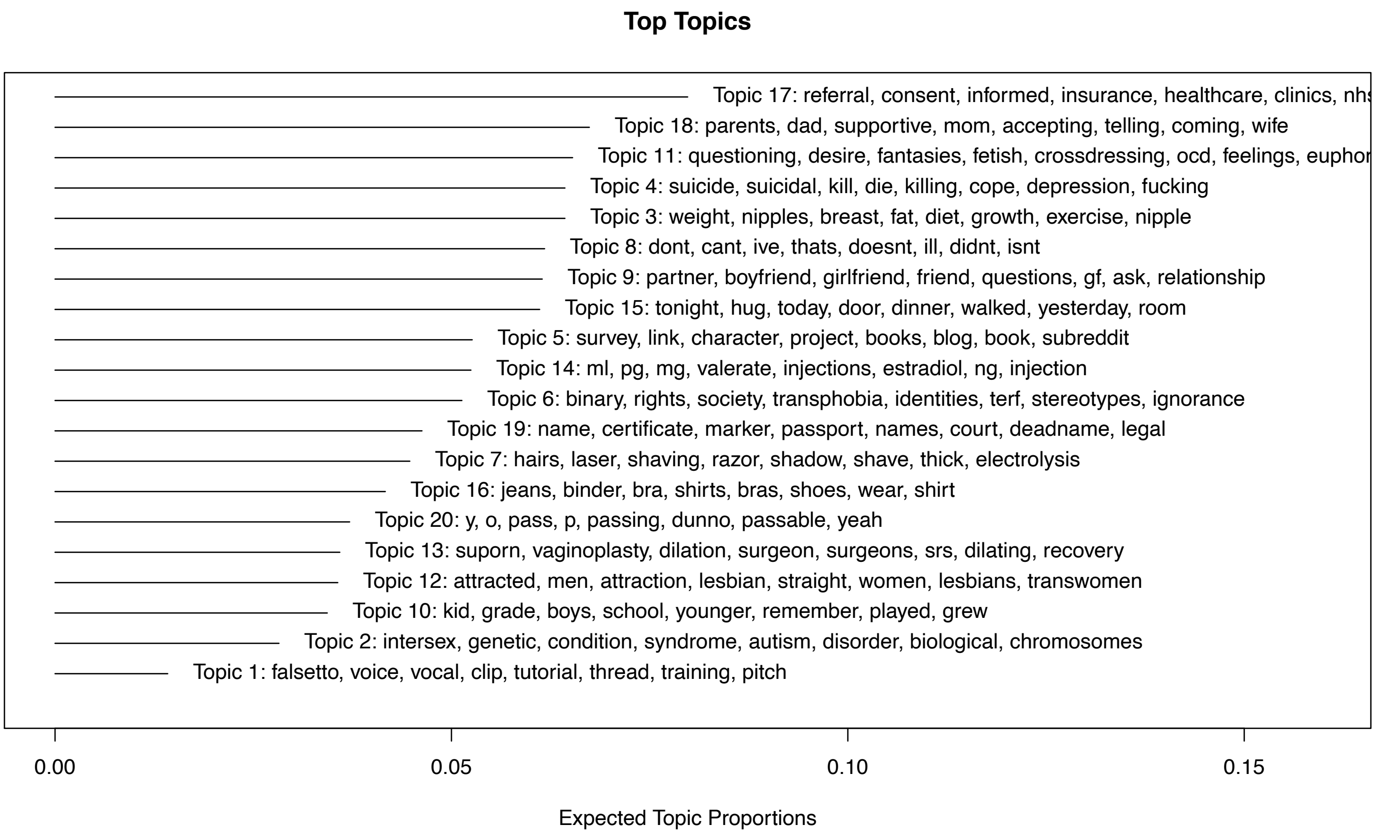
To evaluate the results of the LDA model, I computed both topic coherence, a "measure of the degree to which top words in topic co-occur in documents" (if top words co-occur a lot, this means that the topic is well defined across the corpus) and exclusivity, which increases as words are "both very common in a topic and very rare in other topics" and depicted them in the graph below for each of the topics; the closer to 10, the more exclusive the words in each topic, and the least negative, the more coherent/defined each topic is (Lecture 11). As we can see, Topics 14, 12, 18, 8, and 10 have very high exclusivity scores and have among the best coherence scores. Additionally, Topics 4, 11, and 15 have similar coherence scores to these, but lower exclusivity scores. Overall, we can see that more than half of the topics have very high exclusivity and coherence scores, meaning that they can be considered to be reliable topics. It is not surprising that Topic 2 had the lowest coherence score, as the words included within seem not to be very correlated (as shown above).



Structural Topic Model (STM) Results

While the simple LDA model allows us to derive meaningful insights on the corpus' structure and its recurring topics, it does not enable us to see how these topics change with regards to covariates of interest. Being built on LDA, STM allows two ways to include contextual information to the model, topic prevalence and content; prevalence allows us to see the relationship between a covariate and the occurrence/frequency of a topic (e.g. Democrats talk more about immigration than Republicans) while content allows us to examine different word usage from different groups in a given topics (e.g. Democrats and Republicans probably use different words when referring to immigration) (Lecture 11). For the current use case, I fitted an STM model with \`score\`, \`num_comms\`, and \`Gay_Marriage\` as prevalence covariates in order to examine the differences in topic occurrence before and after Obergefell v. Hodges and whether the proportion of topics within documents vary with increasing number of comments and higher scores.

In terms of topics produced and their overall proportion in the corpus, the STM model produced essentially the same topics as the LDA model with the exception that Topics 3 and 4 have swapped positions in terms of proportion (i.e. Topic 3 is now 5th while Topic 4 is now 4th). Different from the previous graph depicting the top 8 highly occurring terms by frequency of occurrence, the following graph depicts the top words based on FREX (frequency + exclusivity) scores (i.e. the terms with highest frequency per topic weighted by how rare they are within other topics). It is interesting to see how the FREX terms allow us to intuit the same topic categories as the ones derived previously.

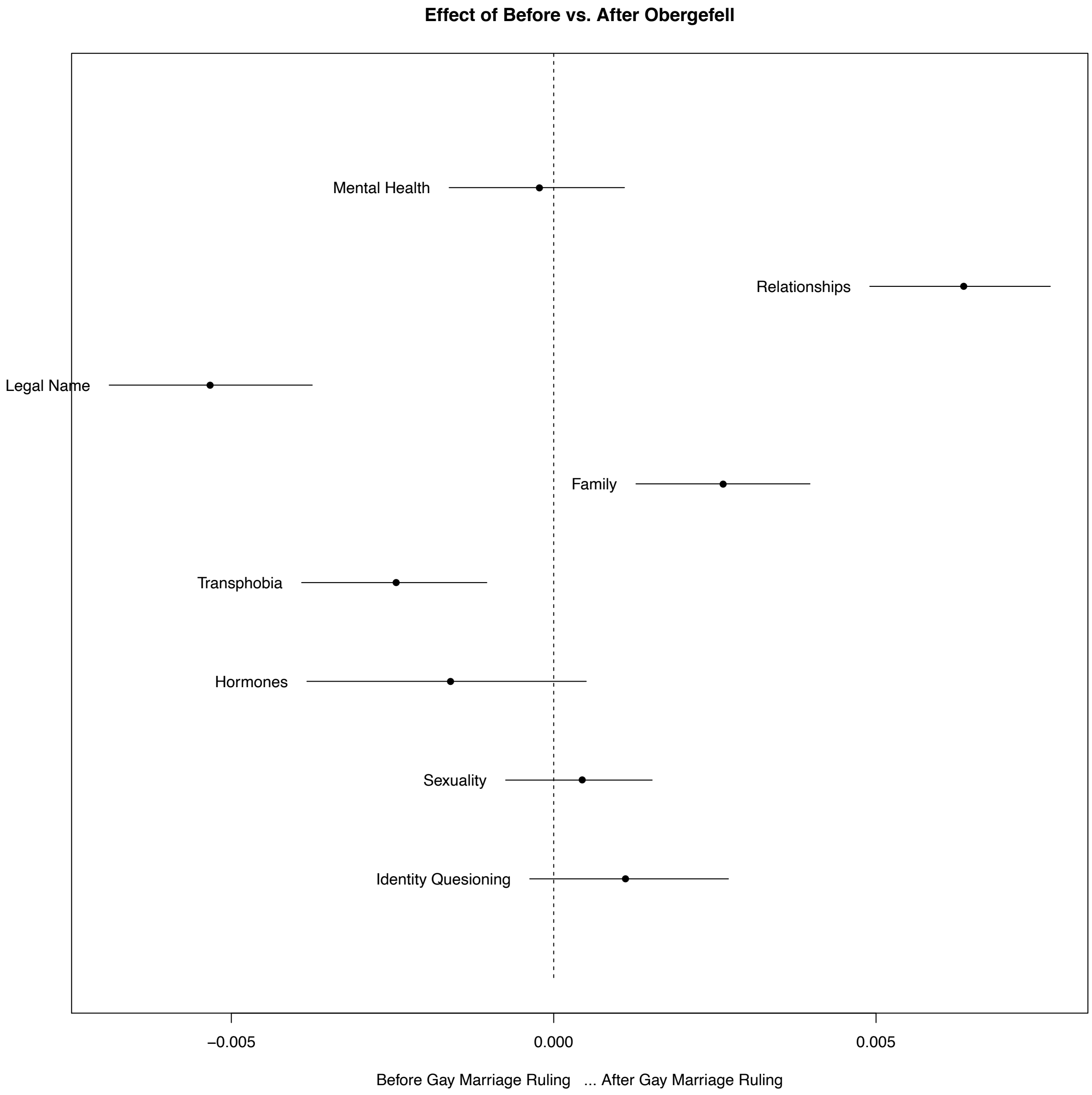


In order to examine the effect of `score`, `num_comms`, and `Gay_Marriage` as prevalence covariates, I decided to focus on the 8 most interpretable topics produced by the model:

Mental Health (Topic 4)
depression, every, fucking, self, anymore, worse, anxiety, live, nothing, bad, stop, away, end, depressed, shit, keep, job, enough, alone, living
Relationships (Topic 9)
friend, ask, questions, transitioning, best, love, thanks, relationship, ftm, asking, recently, partner, thank, girlfriend, pre, comfortable, answer, hi, sex, situation
Legal Name (Topic 19)
name, change, pronouns, new, use, job, changed, using, birth, legal, changing, call, names, preferred, old, company, correct, called, legally, state
Family (Topic 18)
parents, coming, mom, talk, dad, support, supportive, telling, mother, therapist, scared, wife, accepting, accept, understand, afraid, love, sister, live, brother

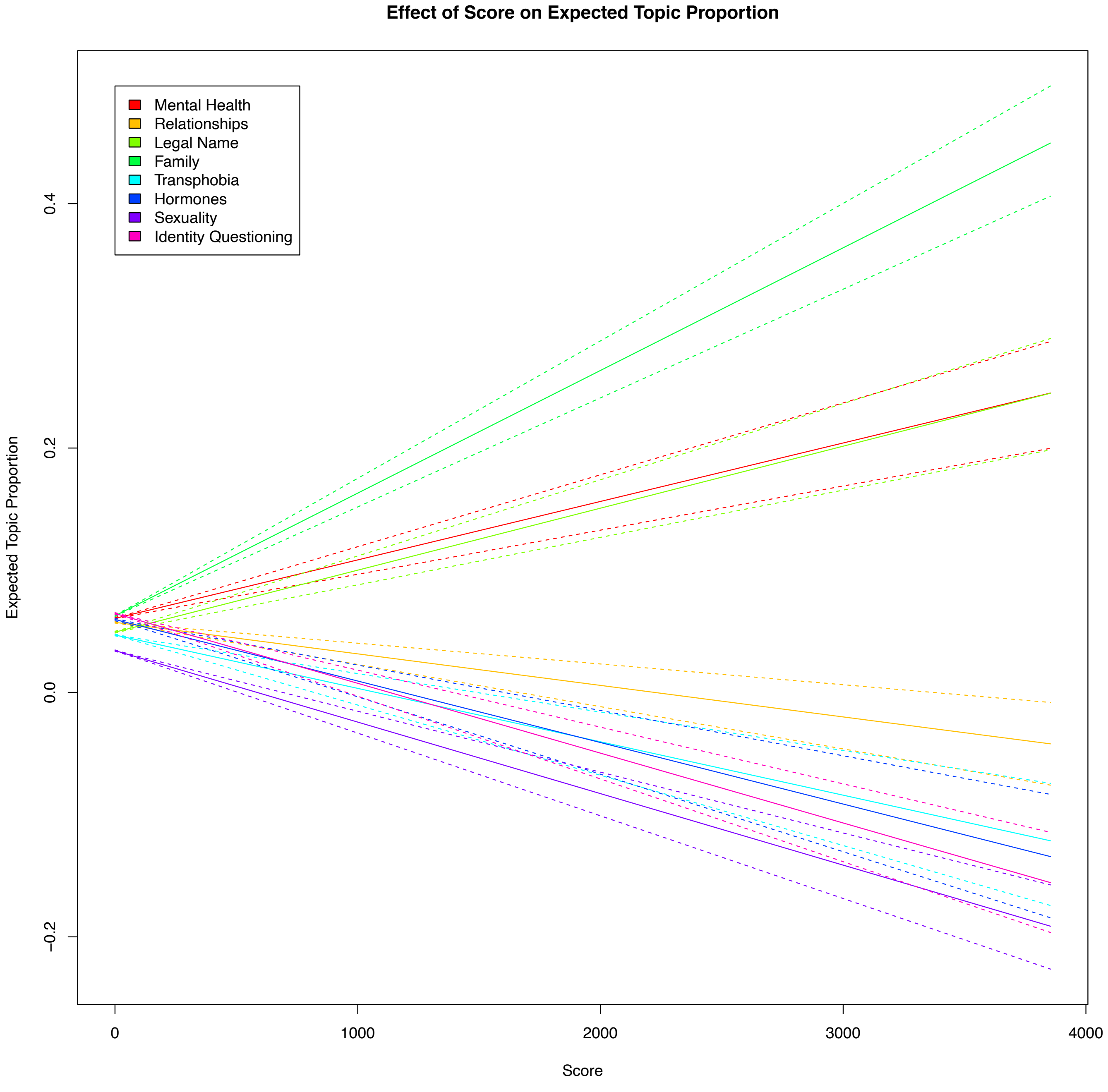
Transphobia (Topic 6)
non, identity, understand, binary, us, community, identify, social, society, mean, transphobic, many, word, seems, different, wrong, others, saying, believe, transphobia
Hormones (Topic 14)
spiro, estrogen, levels, estradiol, e, taking, doctor, blood, testosterone, dose, low, pills, test, mg, week, injections, progesterone, high, results, month
Sexuality (Topic 12)
women, cis, men, man, guys, gay, sex, guy, straight, attracted, girls, sexual, lesbian, sexuality, penis, attractive, bi, dating, sexually, attraction
Identity Questioning (Topic 11)
feelings, thoughts, questioning, feminine, identity, thinking, idea, mind, desire, past, sense, sometimes, sexual, whether, part, sort, masculine, self, confused, feels

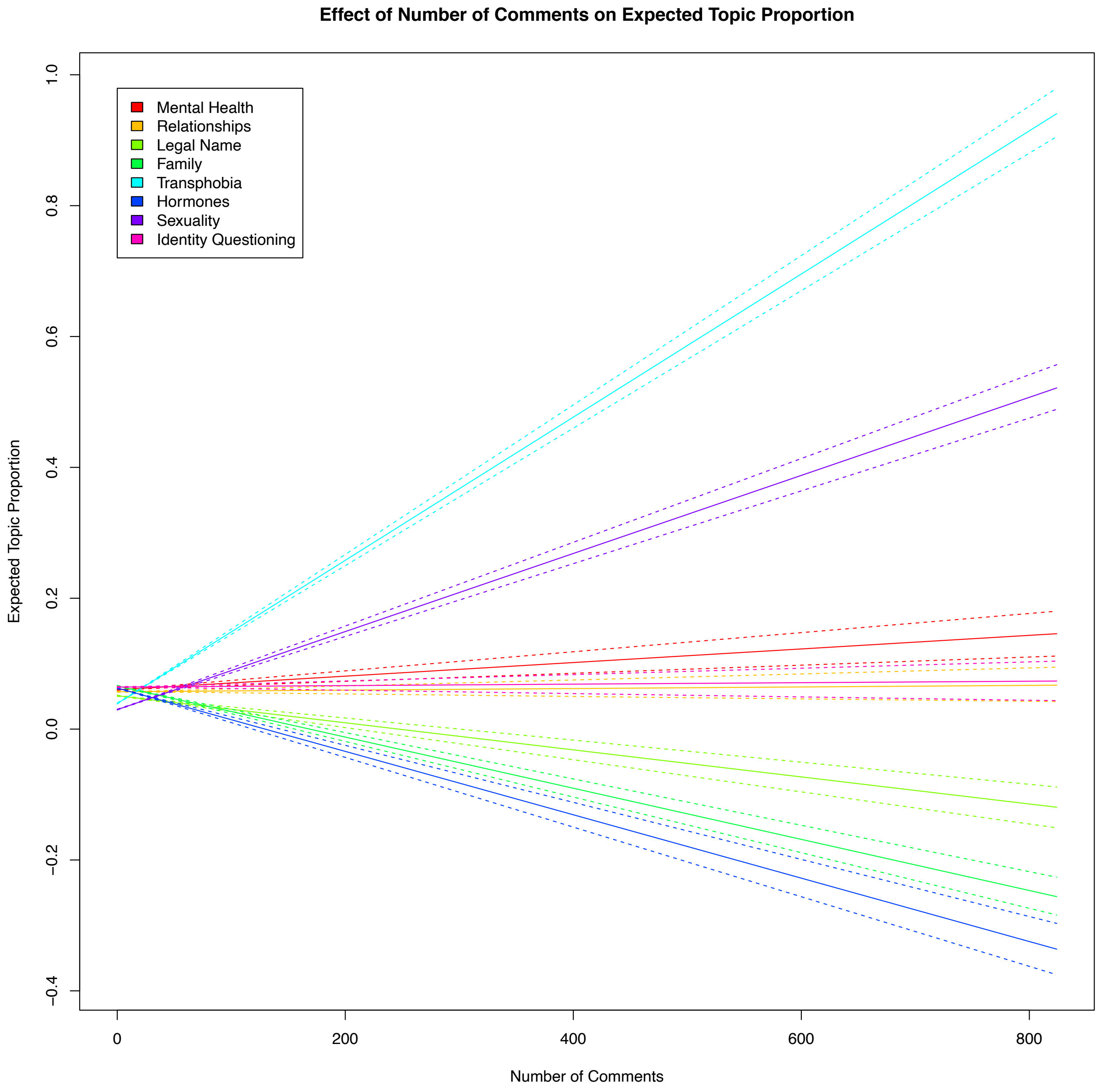
The effect of including the `Gay_Marriage` covariate on topic prevalence is depicted in the following graph:



As we can see, it seems like Obergefell v. Hodges did have an effect on the topics of concern for the transgender community. While mental health, changing one's legal name, transphobia, and hormones were more talked about before the Supreme Court ruling, relationships, family, sexuality, and identity questioning occur more often in posts after Obergefell. As hypothesized earlier, this indicates a shift in discourse towards the transgender community and an increase in individuals feeling like they are supported by their families, friends, and significant others. Additionally, in line with the fact that the ruling emphasized that the constitution protects peoples' ability to express their gender identity, we can see that individuals are now more open to exploring their identity and sexuality online with their peers.

Lastly, the following two graphs depict whether the proportion of topics within documents vary with increasing number of comments and higher scores.





In terms of score, we can see that for the Family, Mental Health, and Legal Name topics, as a post's score increases, the expected proportion of that topic being in a given document increases, while it decreases for the Relationships, Transphobia, Sexuality, Identity Questioning, and Hormones Topics. This generally makes sense given that posts on mental health issues such as depression, anxiety, and suicidal thoughts and legal issues regarding name change (which is a long process) are more likely to get support from the community when compared to posts on starting hormones, which are likely to be less downvoted or to just receive less votes overall. We see a different pattern when it comes to number of comments; for the Transphobia and Sexuality topics, as the number of comments increases, the expected proportion of that topic being in a given document also increases, while it decreases for the Legal Name, Family, and Hormones topics. This also makes sense since posts related to transphobia are likely to spark a lot of anger and sadness, thus having more comments, while posts on family tends to have less comments as they tend to be more personal.

All in all, this project has shown, via LDA and STM topic models, that the discourse surrounding the transgender community, along with their topics of concern, effectively changed after the Obergefell Supreme Court ruling. Moving forward, it would be interesting to investigate whether the content for these topics varies based on the type of transition an individual is experiencing (male to female vs. female to male). To do this, I could look at two subreddits dedicated to these types /r/MtF and /r/FtM/.

References

Devin Soni (22 March, 2018). Supervised vs. Unsupervised Learning. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d

David B. Cruz (2015). Transgender Rights After Obergefell. https://towardsdatascience.com/supervised-vs-unsupervised-learning-14f68e32ea8d