6-4-2020

A computational approach to analyzing and detecting trans-exclusionary radical feminists (TERFs) on Twitter

Christina T. Lu
Dartmouth College

Recommended Citation

Lu, Christina T., "A computational approach to analyzing and detecting trans-exclusionary radical feminists (TERFs) on Twitter" (2020). *Dartmouth College Undergraduate Theses*. 165.
https://digitalcommons.dartmouth.edu/senior_theses/165

A computational approach to analyzing and detecting trans-exclusionary radical feminists (TERFs) on Twitter

Christina T. Lu
Advisor: Saeed Hassanpour
Dartmouth College, Department of Computer Science

TR2020-900

June 4, 2020

Abstract

Within the realm of abusive content detection for social media, little research has been conducted on the transphobic hate group known as trans-exclusionary radical feminists (TERFs). The community engages in harmful behaviors such as targeted harassment of transgender people on Twitter, and perpetuates transphobic rhetoric such as denial of trans existence under the guise of feminism. This thesis analyzes the network of the TERF community on Twitter, by discovering several sub-communities as well as modeling the topics of their tweets. We also introduce TERFSPOT, a classifier for predicting whether a Twitter user is a TERF or not, based on a combination of network and textual features. The contributions of this work are twofold: we conduct the first large-scale computational analysis of the TERF hate group on Twitter, and demonstrate a classifier with a 90% accuracy for identifying TERFs.

Introduction

Abusive content on social media has become a more salient issue to computer scientists working in natural language processing (NLP). While work has been done to study hate speech (particularly in the misogynist and racist vein) as well as the rise of radicalized internet groups (such as the alt-right), little computational attention has been paid to online instances of transphobia and communities of trans-exclusionary radical feminists (TERFs).

Outline and contributions

Within the introduction, we provide a sociological background on gender and trans people, along with a brief summarization of transphobia and the trans-exclusionary radical feminist (TERF) community.

Chapter 1 provides a computational analysis of the TERF community on Twitter. We utilize two network clustering algorithms to detect intra-community submodules. Then, we build a topic model of all tweets within our corpus to better understand the distribution of topics that the community discusses.

Chapter 2 describes a classifier for detecting whether a Twitter user is a member of the TERF community or not. Our features can be divided into two categories: network features and tweet text attributes. Using these features, we test several models for this classification task.

We conclude by summarizing our results and discussing future steps for this work. The contributions of this thesis are as follows:

- Conducting the first computational study of the TERF community at large on the microblogging site, Twitter.
- Creating a classifier to distinguish whether a Twitter user is a TERF or not, based on a combination of network features and tweet text attributes.

Gender, the transgender community, and transphobia

Modern trans-inclusive conceptions of gender indicate an epistemology that generally encompasses three aspects that do not always align: gender as a set of imposed sociocultural norms, gender as performed by an individual, and gender as internally felt. This differs from traditional "folk" models of gender which position it as a male-female binary, physiologically determined, and immutable.

"Transgender" is used as an umbrella term to capture the variation of queered gender experiences that do not fit under the cisgender binary (cisgender referring to people whose gender corresponds with their sex). Members of the transgender community either have a gender identity that is different from their birth sex, or identify as outside the binary (including non-binary, gender non-conforming, and genderfluid).

The transgender community sees mental health risks and rate of suicide which far exceed the norm, even in comparison to the rest of the queer community [6]. Studies point to causes including gender dysphoria but also the lack of societal acceptance; instances of transphobia are often manifested as physical violence [15]. Violence against trans people occurs in disproportionately high numbers even as many cases go unreported; though there is a dearth of statistics, it is claimed that the life expectancy of trans women in the Americas is between 30 and 35.

This illuminates the critical state of the safety of the transgender community as well as the necessity of identifying and combating instances of transphobia—the prejudice and hatred against transgender people. Transphobic instances are varied and can include misgendering (referring to one by the incorrect gender, often in speech such as through incorrect pronouns), insisting that transgender people are merely homosexual or mentally ill, and outright denying their existence.

In a survey of recent work on abusive content online, Vidgen et al. describe the challenges and frontiers of the field [31]. Research focus has been unevenly distributed; the majority of work focuses on detecting racist or sexist content, with little attention paid to other flavors of prejudice including transphobia. A critical step forward is to diversify the study of the targets of abusive content, towards those that are less prevalent, and this work aims to address that.

TERFs: Trans-exclusionary radical feminists

A marginalized community such as transgender people is already subject to hegemonic sociocultural norms, but they also face targeted harassment and hatred from trans-exclusionary radical feminists (TERFs). Members of this group claim to be feminists who are "gender-critical," but are known for their transphobia masked in the language of feminism. Their hate is frequently targeted towards transgender women, and spans from denying their right to exist to perpetuating biological essentialism.

Coined in 2008 by a feminist blog responding to the rising trend, the phrase TERF generally refers to self-proclaimed feminists who view the existence of transgender people as encroaching upon some conception of "womanhood." While they like to engage in seemingly logical arguments about women's rights or deconstructing gender entirely, their dependency on biological essentialism and conflicting stances belie the simple transphobia at their core.

It is worth distinguishing TERFs and their behavior from general instances of transphobia in order to clarify our target group. While "TERF" and "transphobe" are often used interchangeably online, such usage dilutes the meaning of the term and does not account for their unique behavior of infiltrating feminist and queer spaces. TERFs are a subset of transphobes, which encompasses those who hold all transphobic viewpoints. Uniquely TERF specific rhetoric is often overly occupied with a delineation between ciswomen and transgender women, though they do parrot classic transphobic talking points as well.

This thesis deals with TERF communities within the Anglosphere, although we note that their specific flavor of transphobia and masked feminist rhetoric likely occurs in other languages as well. Less prevalent in the United States and Canada, TERFs within the United Kingdom hold an unfortunately mainstream position within feminism [20], endangering the trans community there and eroding their rights. Online communities of TERFs are visibly concentrated on several social media sites including Reddit (r/gendercritical) and Twitter, as well as independent forums, under the moniker of "gender-critical."

We define a TERF online by their (a) particular strain of transphobic beliefs as well as their (b) membership within such an online community that perpetuates and reinforces such viewpoints and targeted harassment of trans people.

Abebe et al. [1] describe several roles for computing in social change, including serving as a diagnostic—describing problems with clarity, which is what we seek to do here. It is critical to conduct an empirical investigation of TERF communities online in order to better understand their behavior. Only by doing so can we develop informed ways of computationally identifying Twitter users that are TERFs and combating their abusive content online.

Chapter 1: Analysis of the TERF community on Twitter

A more comprehensive understanding of the trans-exclusionary radical feminist (TERF) community on the microblogging site, Twitter, is necessary to underpin the following work. We utilize a couple of network clustering algorithms in order to model intra-community clusters, and then employ Latent Dirichlet allocation (LDA) to build a topic model of their tweet contents.

1.1 Data collection and corpus building

We built our dataset of TERF tweets using TERFblocklist [35], a publicly available community resource on the application “Block Together.” This application allows Twitter users to cultivate lists of users and block the entire list through the Twitter API. Block lists have been a tool for the trans community and other interested parties to minimize interactions with known TERFs on Twitter.

TERFblocklist is not the only available block list of TERFs on the application Block Together (another popular one is TerfBlocker), but was chosen due to its transparency in method and size of approximately 13,000 users. The list is built by hand and maintained by a trans woman, and likely TERF accounts are crowdsourced by the community. Links to accounts or tweets that contain transphobia are sent to the maintainer, who verifies the accusations and adds them to the block list if the account “uses transphobic slurs, denies or polices trans identities, and various other dog whistles.” The process is not automated, as the maintainer notes that attempting to do so yielded many false positives. For this study, we did not run a second verification of the given list of users, but future studies could look into different methods of culling this existing list.

Through Block Together, we obtained the list of Twitter user IDs on TERFblocklist, which contained at time of download 13,472 users. Using the Twitter API, we collected all publically available tweets from users on the list from January to December 2019, for a total of 15,469,346 tweets. No pre-processing was done at the time of collection, besides discarding emojis. Data per tweet includes date and time; the user it is in response to (if applicable); number of replies, retweets, and favorites; the text of the tweet; the geotag (optionally available); other mentions within the tweet; hashtags; and tweet ID. Though we are primarily concerned with tweet text content within this study, the remaining grey data may prove useful in future work.

For each public Twitter user on our list, we also collected the list of Twitter users they follow through the Twitter API, for network analysis purposes. We chose to only collect following data one layer deep, as we are most interested in seeing if any patterns emerge from the users our list chose to follow and further layers expand the size of the list exponentially. This resulted in a list of 5,087,581 total Twitter users, comprising our target list of users along with their immediate followees. We did not collect follower data, as a Twitter user has little to no choice over who follows them, and such data is therefore a weaker signal for the kind of content they choose to engage with.

1.2 Clustering

1.2.1 Methods

In order to detect community patterns and verify the existence of sub-communities within our larger group of users, we employed two different network clustering algorithms, Louvain and Infomap, with demonstrated performance on large graphs.

There are several community detection methods that perform efficiently on comparative analysis of synthetic networks. Previous work by Lancichinetti and Fortunato [19] test a series of algorithms on several benchmark and random graphs. They introduce a new class of benchmark graphs called the Lancichinetti-Fortunato-Radicchi (LFR), which generalize upon the Girvan and Newman (GN) benchmark [14] by introducing power law distributions of community size and degree. They tested a wide spectrum of community detection methods, using the metric of normalized mutual information [9] to measure cluster quality. After performing a comparative analysis on several benchmark graphs of a few thousand nodes, they conclude that the Infomap algorithm performs best on benchmark graphs, with the Louvain method performing well also. Both methods exhibit low computational complexity and therefore can be used on graphs of millions of nodes and edges such as ours.

A more recent study by Emmons et al. [11] examines the relationship between several cluster quality metrics and information recovery metrics by analyzing the performances of four network clustering algorithms, including Infomap and Louvain. They use synthetic and natural graphs ranging from 1,000 to 1,000,000 nodes, and consider the cluster quality metrics of modularity, conductance, and coverage, along with the information recovery metrics of adjusted Rand score, and normalized mutual information. While they declare another algorithm—smart local moving—to perform the best overall, they do not classify it as absolutely superior due to discrepancies in cluster evaluation metrics. They do note that Louvain performed better than Infomap in nearly all networks, a contradiction to other work including the previously discussed Lancichinetti study.

Previous applications of network clustering on Twitter user following graphs also frequently use both the Infomap and Louvain algorithms [29]. We thus select both algorithms to perform network clustering on our graph due to their demonstrated performance efficiency, partition quality, and information recovery.

Infomap. The Infomap algorithm, first described by Rosvall and Bergstrom [27], detects community structure in directed graphs by utilizing the probability flow of random walks as a proxy for information flow in a system. It seeks to decompose the network into modules containing nodes which information flows quickly between, while also correctly modeling inter-module information flow. Such modeling is thus an optimal compression problem of the random walk, intuitively so that the original structure is retained as much as possible when decompressed. The Infomap algorithm utilizes Huffman encoding in order to do so while maximizing the objective function called the maximum description length.

We use the *infomap* package on python, part of the MapEquation software package, to run the Infomap algorithm. The partitioning is done with default parameters, on a directed version of the graph.

Louvain. On the other hand, the Louvain algorithm, devised by Blondel et al. [5], utilizes Newman-Girvan modularity maximization. Its greedy optimization method for modularity, a measure of relative inter- and intra-connectedness within modules in a graph, runs by assigning nodes to modules and re-calculating the change in overall modularity by moving it to a neighboring module. It then creates super nodes from the modules of the previous step, and iteratively repeats these two steps until modularity can no longer be improved. The Louvain algorithm uses an efficient heuristic to solve the underlying NP-hard problem in $O(n\ log\ ^2 n)$ time.

We use the python package *networkx* and *python-louvain* to perform the Louvain partition on an undirected representation of our graph. We also ran it using default parameters, and used the lowest hierarchical clustering returned by the algorithm.

In order to construct our graph, we compile our target list of TERFs as well as the users they immediately follow, which amounted to 5,087,581 total users. We removed users who were followed by less than one target user and were not in our target list, for a reduced total of 1,360,281 user nodes in our graph. The directed edges in our graph indicate the following direction. After clustering, we record the cluster labels for the 13,070 users on our target list (slightly reduced from the total length of the original block list due to some private accounts).

1.2.2 Results

After running both network clustering algorithms on our 1,360,281 user nodes, we examine the assigned partitions for the 13,070 users on our list and assess them using several cluster quality metrics. Since there is no ground-truth for our clustering, we only analyze metrics such as modularity which do not require it. We also analyze mutual agreement between both clustering results to measure their similarity.

For members of our target list, the Infomap algorithm identified 155 network clusters in total, ranging from the largest cluster of size 7,258 to the smallest of size 1. On the other hand, the Louvain algorithm identified 11 network clusters in total, ranging from the largest cluster of size 7,398 to the smallest of size 4.

Table 1.1: Top 10 cluster sizes.

Infomap	Louvain
7258	7398
3382	2647
1547	1129
280	953
186	734
59	93
50	48
29	34
26	17
19	13

We calculate several standalone cluster quality metrics for the two partitions: modularity, performance, and coverage. For mutual agreement, we use two information recovery metrics: adjusted Rand index and normalized mutual information (NMI) [12].

Cluster quality metrics aim to indicate whether a partition of a graph is *good* or not, with several metrics arising due to the lack of agreement on what constitutes good. High *modularity* [24] indicates that a graph has dense intra-module connections and sparse inter-module connections, with a maximum value of 1. It is calculated by comparing the existence of each intra-module edge to the probability that the edge would exist in a random graph. *Performance* represents the ratio of the number vertex pairings that are correctly assigned (placed in the same partition if they share an edge, and placed in different communities if they do not) to the total number of possible pairs. Another similar metric, *coverage*, is the ratio of the number of intra-module edges in the graph to the total number of edges; a graph where all modules are completely separated would yield a coverage of 1.

To calculate the standalone quality metrics, we analyze the subgraph of only target users and the edges between them. The differences in all three metrics between the two graphs were not statistically significant, indicating that performance between the two clusters were roughly equal. A comparison can be seen in Table 1.2.

Table 1.2: Cluster metric comparison

Metric	**Infomap**	**Louvain**
Modularity	0.4306	0.4269
Coverage	0.8963	0.8946
Performance	0.6144	0.6279

Information recovery metrics measure the agreement or similarity between two partitions. Intuitively, the *adjusted Rand index* measures agreement by calculating the ratio of agreements between both partitions (pairs of elements in the same subset), to the total number of agreements and disagreements. *Normalized mutual information* is based on the notion of Shannon entropy from information theory. It seeks to capture the information overlap between the two partitions; or, how much you can know about partition X from partition Y (and vice versa). Labeling in perfect agreement for both metrics result in a score of 1.0.

We use the *scikit-learn* python package to calculate both information recovery metrics. The two partitions have an adjusted Rand index of 0.739 and a normalized mutual information score of 0.564.

Table 1.3: Mutual information scores

Adjusted Rand index	0.739
Normalized mutual information	0.564

1.2.3 Discussion

In terms of the cluster quality metrics, the scores for both the Louvain and Infomap algorithms did not differ in a statistically significant way, with neither better than the other. Both partitions exhibit modularity around 0.4, indicating that while the number of edges within modules exceeds that of chance, there are still a significant number of connections between modules. This suggests that while several large subdivisions appear in the Twitter TERF community, high overlap and information flow occurs between them, pointing to the interconnectedness of the entire community. The performance score of 0.6 reinforces this conclusion. Coverage was roughly equal at 0.9, with the majority of edges appearing within modules regardless.

The Infomap algorithm categorized 111 clusters of single users, indicating that areas of our following graph were extremely sparse. As our network only includes users in our target list of users and those they immediately follow, it makes sense that an algorithm using random walks would reflect the sparsity. The partition produced by the Louvain algorithm likely did not reflect such granular sparsity due to its method of reassigning nodes to modules only when a modularity improvement is gained.

The top three largest clusters are roughly the same size, and our two measures of mutual information indicate that both partitions are roughly in agreement, especially the adjusted Rand index of 0.739. The normalized mutual information score of 0.564 indicates that the two partitions are significantly similar enough to provide shared information and reduce mutual entropy.

From this analysis, we see that the majority of the network of TERFs on Twitter form three large clusters of thousands of users (85.4% and 93.2% of all users on our list, according to the Louvain and Infomap partitions, respectively), with the largest cluster containing over half of the users in our target list, according to both partitions. The rest of the users are spread among two medium-sized clusters with membership in the hundreds, and then scattered in much smaller clusters. Some are separated completely from the rest of the group.

Future work involving geographic tagging of users could help further explain the subdivisions. Comparisons with networks of Twitter users across different political affiliations and other interests could also provide more information about the clustering. In section 1.3, we provide a qualitative analysis of the top five Louvain clusters using the trained topic model. Since the performance of both partitions were statistically similar, we chose Louvain over Infomap arbitrarily.

In order to create more informative networks, future steps could also include building a network from more than just follow connections. Enriching the network connections using reply, retweet, and mentions could lead to more robust networks of information flow. It is worth watching out for reply interactions as they may indicate mutual agreement, but also disagreement.

1.3 Topic Modeling

1.3.1 Methods

We are also interested in understanding the contents of what TERFs within our target list talk about on Twitter. In order to do this, we apply Latent Dirichlet allocation (LDA) as described by Blei et al. [4] to obtain topic models from the text.

LDA is a generative probabilistic model which can build a topic model from a corpora of text documents. Each document is represented as a random mixture of various topics, and each topic is characterized by a distribution over words. The topic distribution is assumed to have a sparse Dirichlet prior, which models the topic-word distribution. The Dirichlet prior captures the intuition that topics utilize only a small set of words frequently and documents only cover a small set of topics.

As tweets on Twitter are limited to 280 characters, documents are extremely short and sparse. Several modifications of LDA have been proposed to address this issue and better model tweet topics.

The *Dirichlet Multinomial Mixture* (DMM) method is a variant that was designed to work better on shorter documents, such as tweets. In DMM, documents are only assigned to one topic, whereas LDA assumes a mixture over multiple topics. A comparison of LDA and DMM [22] found that DMM with Gibbs sampling outperformed LDA on topic stability while LDA out-performed DMM on topic coherence half the time on documents of length less than ten words.

Other methods involve pooling the tweets in various ways in order to create longer documents, including by user. Steinskog et al. [30] describe several pooling techniques including aggregating similar tweets sharing author or hashtag, while Alavarez-Mellis et al. [3] describe grouping together tweets in the same conversation. Hong and Davison [16] go in-depth on several schemes to aggregate text for topic modeling, and also demonstrate that the popular Author-Topic model fails at modeling social media hierarchical relationships.

Previous work applying topic modelling to tweets have ranged from work in clarifying public health discussions to evaluating its content with respect to traditional news media [13, 33]. Surian et al. characterized Twitter discussion of the HPV vaccine through community detection and both LDA and DMM topic models. They gathered tweets through selected keywords and concluded through manual intrusion as well as topic coherence metrics that DMM provided a more realistic clustering of tweets by topic.

Though we considered modifications of the LDA algorithm such as DMM, we ultimately chose to use the original algorithm due to its demonstrated topic coherence in the literature. We are interested in a topic model to first and foremost provide a human-comprehensible decomposition of the tweets from the TERF community. A possible future step would be re-generating topic models using DMM and comparing the performance.

We considered each tweet a document, and ran LDA over all tweets within our corpus of approximately 15 million tweets for 25, 50, 75, and 100 topics using the Python package *gensim*. The literature suggests that metrics such as topic coherence stabilize around 100 topics [28]. For preprocessing, we converted all text to lowercase, and removed all stopwords and tokens less than three letters long. We stemmed all tokens using the *nltk* package's implementation of the Snowball Stemmer [36]. We also filtered the frequency extremes of all tokens in our corpora; tokens that appeared in less than 1000 documents or more than 50% of all documents were removed and finally, we kept the top 100,000 most frequent remaining tokens to build our topic model.

In order to provide a clearer picture of the clusters obtained in section 1.2, we also analyze distribution of topics for the top five clusters as obtained by the Louvain clustering. We point out a few topics that are particularly indicative of TERF discussion points.

1.3.2 Results

We obtain four different topic models, varying on number of topics (*n*=25, 50, 75, and 100). In order to evaluate the performance of our various LDA topic models, we use two measures of topic coherence, UCI and UMass coherence.

Statements or facts are coherent if they support each other and can be interpreted in such a way that covers most of the facts. Topic coherence measures aim to capture how legible topics are to human judgement, and their algorithms can be categorized into intrinsic and extrinsic methods. Extrinsic methods use generated topics on external tasks and data, while intrinsic methods do not. A unifying model of coherence frameworks proposed by Roder et al. [28] breaks down each metric into four parts: segmentation, probability calculation, confirmation measure, and aggregation. Different methods of calculating coherence vary along these four dimensions.

Two state-of-the-art methods for calculating topic coherence are UCI coherence (Newman et al. [23]) and UMass coherence (Mimno et al., 2011). They calculate the topic coherence as the sum of pairwise similarity scores over the set of topic words. Both measures were designed for LDA topic models, and are found to often agree [28].

UCI coherence is based on the notion of pointwise mutual information, and estimates probabilities using word co-occurrence accounts obtained from a sliding window over an external reference corpus, such as Wikipedia. *UMass coherence* is based on document co-occurrence of top word pairs within a topic, calculated using the original corpus the topic model was trained on. Evaluations of these and several other coherence metrics against the gold standard of human judgement tasks indicate that they reflect human judgement better than perplexity measures.

We calculate both UCI and UMass coherence measures for all four models using the *gensim* python package.

Table 1.4: Coherence measures across topic models

Number of Topics	UCI	UMass
25	-0.0813	-5.4077
50	-0.1356	-5.8939
75	-0.1674	-6.1260
100	-0.3204	-6.5149

Since the topic model trained on 25 topics had the highest coherence measures, we proceed to do an in-depth examination of its generated topics. For top words and probability distributions of each topic in this model, see Appendix A.

Using this topic model, overall distributions across the entire tweet corpora as well as per Louvain cluster were calculated. We assign each tweet to the topic with the highest probability. If all topics had a less than 0.05 likelihood or multiple topics had the maximum probability, we did not assign the tweet.

Table 1.5: Statistics for the 5 largest clusters

Cluster	No. of users	No. of Tweets	Avg no. tweets/user
0	1,129	1,431,660	1,268
1	2,647	4,238,193	1,601
2	7,398	4,851,994	656
3	953	2,638,723	2,768
4	734	1,800,194	2,453
All	13,070	15,469,346	1,183

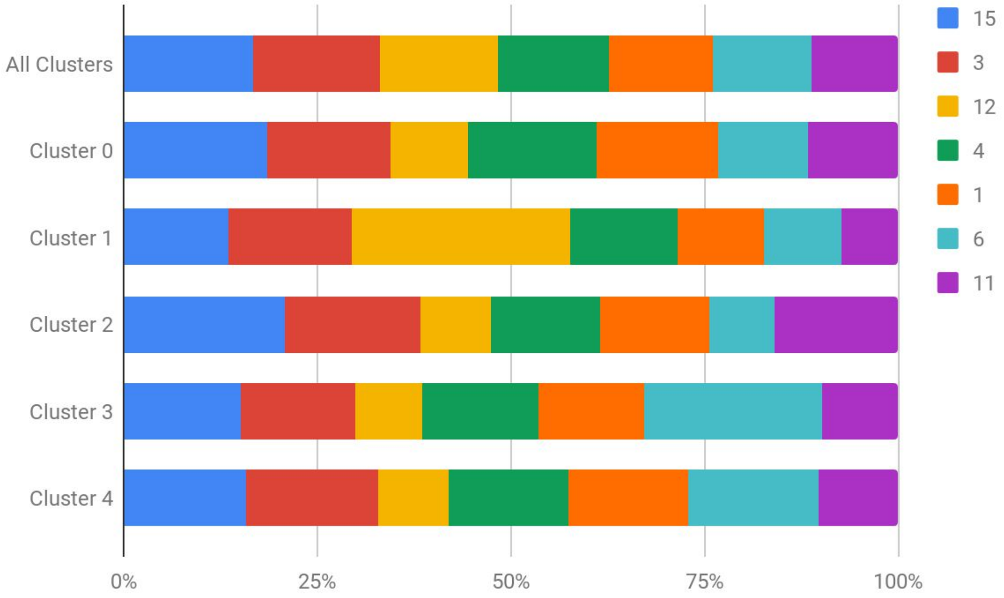
The top seven topics across all tweets were topics 15, 3, 12, 4, 1, 6, and 11 (in descending order).The top words for these prevalent topics are shown below.

Table 1.6: Top words for selected topics

Topic	**% of total**	**Words**
15	0.0785	like good look feel make wait can final understand peopl
3	0.0763	think thing you peopl mayb littl see mind they stuff
12	0.0707	right women tran woman gender true male femal charact sexual
4	0.0678	know want don talk happen peopl kid school call gonna
1	0.0627	time go sure hope real wrong that money long stupid
6	0.0592	trump leav vote white support state black democrat elect power
11	0.0524	game fuck tell stop yeah speak lie miss truth complet

The percentage distribution of these seven topics across the total corpus of tweets along with the per cluster breakdown for the top five Louvain clusters were also calculated and displayed below.

Figure 1.1: Topic distribution of top 7 topics across clusters



1.3.3 Discussion

The topic model built with 25 models had the best coherence scores. As the number of topics increased, coherence decreased steadily. The UCI coherence scores were particularly low across the board. However, it is worth noting that using extrinsic coherence measures such as UCI may present a noisy signal for true topic coherency; certain transphobic viewpoints discussed among the TERF community are not expected to be well represented on external reference sites such as Wikipedia.

Certain topics jump out as "TERF-like," particularly topic 12 from the 25-topic model, which accounts for approx. 7% of overall tweets. Topic 4 also likely reflects a common topic that TERFs discuss: trans-inclusive education in schools, a specter of “trans indoctrination” in children. Other topics scan as explicitly political, such as topic 6. A number of topics are less coherent, which we suspect is due to either the chosen method (LDA assigns multiple topics per document while the short length of tweets suggests that one topic per document may perform better) or reflects natural noise in social media tweets.

The difference in distribution of the top topics among clusters is particularly interesting for clusters 1 and 3. Cluster 1 sees 14.3% of tweets categorized as topic 12, the most explicitly TERF related topic. It is the second largest cluster, with users more active than average. We hypothesize that users within cluster 1 are likely a “nexus” of TERFs, and tweet the most about their viewpoints. On the other hand, 10.3% of tweets from cluster 3 belong to the explicitly political topic 6. While cluster 3 is the fourth largest cluster, it has the highest average number of tweets among the top five largest clusters, indicating particularly active users. Thus, cluster 3 is likely a group of active Twitter users who often tweet about politics.

Future work in topic modeling could involve using different methods for generating the model, with either DMM or some form of tweet aggregation. Word and topic intrusion methods as proposed by Chang et al. [8] could be used to measure the performance of the models, and a set of manual intrusion tests done with a human subject could also provide further insight on generated topics.

1.4 Analysis discussion

Overall, the computational analysis of the Twitter community of TERFs reveal several interesting qualities about the network as well as the topics they discuss. These findings present the first ever study on TERFs on Twitter, and lay a foundation for more granular analysis of their contents in the future.

The clustering obtained by both algorithms resulted in partitions that generally agreed, indicating a network of medium modularity; while there were several distinct clusters revealed in the graph, there was still significant information flow between clusters. There were sparse areas of the network, with some user nodes disconnected from the rest, but the majority of the TERF community was distributed across three large clusters. Within the clusters, we see slight variations in discussed topics according to the topic model obtained in the second part of our analysis.

We train several topic models with varying number of topics, and find the model with 25 topics performs the best. Several top topics within this model such as 12 and 4 immediately jump out as TERF-related topics, about transgender women and trans-inclusive education in schools. Other topics are identifiable as more about current politics, be it American or British. Using this model, we also take a closer look at our previously obtained clusters and see that while distributions of topics are generally similar, there are distinct clusters with larger percentages of TERF topics and political topics. This helps characterize the clusters further and explain their differences, with cluster 3 in particular appearing to contain a "nexus" of TERFs. We hypothesize that clusters may reflect different geographical or professional spheres (media, politics, etc).

Chapter 2: TERFSPOT, a Twitter user classifier

In the second half of this thesis, we describe TERFSPOT (Tracing Extremist Reprobates From Scanning Patterns On Twitter), a classifier to detect whether Twitter users are a TERF or not, and demonstrate its performance. Using three types of features (falling under network and text attributes) we construct and test across data collected from approximately 15,000 Twitter users.

2.1 Background

2.1.2 Motivations

The ultimate intentions of this study are to better understand the transphobic hate group of TERFs on Twitter, but also alleviate their impact on the trans community on the site. In order to do so, we build a classifier to detect whether a Twitter user is a TERF or not.

Tools such as Block Together from which our target list of TERF Twitter users was constructed have been essential for the trans community online to pre-empt harassment and targeted abuse. However, such lists have to be manually compiled and verified due to a lack of automation, and those most adept at identifying transphobia are unfortunately the ones who are the targets of it. Being exposed to such rhetoric in order to augment these lists places an undue amount of labor on an already marginalized community, with significant mental health repercussions.

Automating the process using a classifier such as TERFSPOT thus reduces the burden on community members to individually identify TERFs on Twitter by hand. We envision a future web application built on top of this model, where users can input a Twitter user and receive a qualified prediction of their label. Through evaluation of model features, we also hope to gain insight on what the most useful signals for detecting a TERF on Twitter might be.

2.1.3 Related work

Abusive content online. A range of techniques have been used to detect abusive Twitter users, primarily focusing on the content of their tweets. Abozinadah and Jones [2] estimated abusiveness of words in tweets using the PageRank algorithm and semantic orientation, and combined them with statistical user measures to predict whether a user was tweeting obscenities in Arabic. Qian et al. [25] leveraged inter- and intra-user representations to help better classify whether a tweet constituted hate speech. While research interest in extremist groups such as the alt-right have steadily increased over the years, most of the attention has been focused on the *content* these users produce, without taking into account the community aspect, which we hope to leverage.

Twitter user classification. Lynn et al. [21] compare the usefulness of user versus document attributes on several NLP tasks including stance detection, sarcasm detection, sentiment analysis, and prepositional phrase. Using Twitter user information like username and profile photo, along with inferred user attributes such as demographics and personality, they were able to do state-of-the-art stance detection without using the text of the Tweet itself. They found that user-level attributes were most helpful in tasks that predict "trait-like" attributes that are stable over time—like stance—over mercurial "state-like" attributes—such as sentiment. This is somewhat applicable for detecting TERFs on Twitter, as our characterization implies a certain "stance" on the trans community. Other stance detection methods involve weakly supervised models of Twitter activity [17]. However, our task is more complicated as it does not encompass solely transphobia, but also several semantically complex viewpoints.

While our task sits at an intersecting area of abusive content online and Twitter user classification, it presents challenges due to the nuanced flavor of abuse these users use. Jurgens et al. [18] describes the spectrum of abusive behavior online from microaggressions to doxxing (with the two axes being frequency and risk of physical danger), and while behavior from TERFs can fall across a wide range, they also include other types that are not easily identified. TERFs are perhaps best characterized by rhetoric that seemingly follows from a feminist viewpoint, but employs various straw man arguments and other logical fallacies. This is not easily represented through a Bag of Words (BOW) approach, so we turn to semantic sentence embeddings.

BERT sentence embeddings. In order to derive semantically meaningful embeddings from tweets, we use bi-directional encoder representations from transformers (BERT) [10]. Originally developed for tasks such as question answering and following sentence prediction, BERT produces sequence embeddings that capture semantic meaning and can be used for a variety of tasks through fine-tuning. The model is trained using a "masked language model," where the pre-training objective is to predict the randomly masked token correctly. This circumvents the limitations of unidirectional models such as Open AI's GPT, and allows the model to learn from the entire neighboring context of a word, both right and left. BERT comes in two model sizes, base and large, and we select base for our purposes due to it being computationally inexpensive with comparable performance.

BERT models can be fine-tuned for a variety of tasks, usually by adding a single untrained linear layer on top of a pre-trained model. Pre-training is done by classifying the final hidden layer output by the model, the output being either the classification token at the head (which is meant to capture the meaning of the entire sequence, in practice) or some other pooling method such as mean or max-pooling. Fine-tuning is inexpensive, compared to the process of pre-training BERT from scratch, and allows the encoded output of sentences to be specialized for a variety of semantic NLP tasks. One relevant application has been using BERT to identify offensive tweets, such as by Zhu et al. [34] for the SemEval 2019 Task6: OffensEval. Similarly, Bodapati et al. [7] use fine-tuned models to enhance the state of the art on several abusive language tasks such as detecting hate speech and toxicity.

2.2 Methods

2.2.1 Control group corpus building

In order to create the corpus of all Twitter users, TERF and non-, we augment our list by collecting control users from Twitter. From the Twitter API, we access the 1% gardenhose stream to access a random sample of live tweets, which previous studies indicate is a good sampling of Twitter as a whole [32]. Using Blodgett et al.'s extension [37] of langid [38] for recognizing social media English, we filter out all non-English tweets. For the remaining English tweets, we collect the usernames of their authors, for a total of 13,152 users.

For all users on the control list, we collect their public tweets from January to December 2019, the same time frame as that of the tweets collected from users in our target list built from TERFblocklist. We process these tweets as we processed the ones from our target users, by removing emojis and collecting thegrey data. For each user, we also collect the list of Twitter users they follow in order to generate our network features.

Combining the control users and previously collected TERFs from the TERFblocklist, we obtain our total pool of Twitter users and their tweets to train and test our classifier. For the final step, we remove all users with less than 100 tweets to obtain our final group of Twitter users. Our data for building a representation of a user comprises only their tweets and their followed users.

Table 2.1: Number of users

Group	**No. of users**
Control group	8,656
TERFblocklist	5,608
Total	14,264

2.2.2 Features

We propose a set of features to detect whether a Twitter user is a TERF, based on previous work in abusive content detection and Twitter user classification. The features we use for our classifier can be divided into two types. Network features reflect aspects of the network of users, particularly who they follow. Tweet text features are those directly constructed from the tweets themselves and aim to grasp the semantic meaning of the sequence.

Network features. Since membership of a user within the TERF community is integral to our definition of what a TERF is, not simply a user discussing similar topics or espousing transphobic rhetoric, this category of features indicate information about the network the Twitter user is a part of. We represent the network using a one-hot representation vector of top followed users, where 1 indicates that the user follows this person and 0 indicates they do not. The list of 2000 top followed users is constructed from the top 1000 accounts followed by the TERF users and the overall top 1000 accounts followed by all users in our corpus. Thus, the network is represented by a vector of length 2000, comprising 1s and 0s.

Text attributes. Text attributes are features that are derived from solely the tweets of users, that aim to capture the semantic meaning within them. The first category of text attribute features are *topic frequency* vectors. These represent the number of tweets made by the user that fall under a topic, as categorized by the 25-topic model trained in 1.3. The probability distribution across topics for the most recent 100 tweets by a user are calculated, and tweets are assigned to the topic with the highest probability, if the probability is greater than 10%. The final feature is a vector of integers holding the topic frequency of the users 100 most recent tweets across these 25 topics.

The second category of text attribute features are based on BERT *sequence embeddings* of user tweets, which seeks to capture semantic meaning. For each user, we select one tweet that is likely to be most indicative of TERF ideologies. This tweet is obtained by passing the most recent 100 tweets through the 25-topic model, and selecting the one with the highest probability of falling under topic 12 (which is most strongly TERF related, as discussed in 1.3.3). After parsing out links within the tweet, we use these tweets to fine-tune BERT using the *bert-base-uncased* model and train using the user labels in our corpus. This is done by classifying the final layer hidden-state of the first token in the sequence, the classification token, through processing by a linear layer and a tanh activation function. Table 2.2 includes a sample of selected tweets for users with high probabilities for topic 12, with more included in Appendix B.

Table 2.2: Sample of signal tweets to be embedded

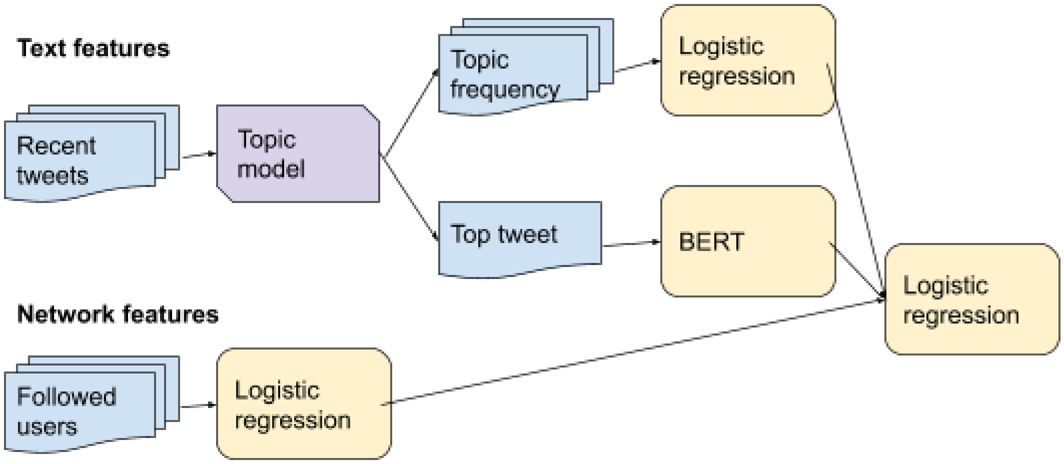
Tweet text	**% topic**	**Label**
Yup. Puberty blockers are horrendous.	0.68	1
Feminism is about fightinh for the rughts of FEMALES. By definition males are not females.	0.67	1
That old "you oppose trans rights" chestnut causes so much harm. No one oppose trans rights - and they share the same rights as everyone else in the UK. OJ is a nasty wee misogynist who will never consider what women are being expected to give up in the name of "trans rights".	0.56	1
@jk_rowling Men cannot become women	0.51	1
Biology deniers only have like 5 insults and just keep reusing them. Grow tf up dude. People are allowed to criticize an ideology that is helping male r*pists, and is hurting children, homeless women, gay women, and women in prisons.	0.34	1
CAN WE SAY BLACK EXCELLENCE!!!!!!!!!!	0.34	0
#Contagion is a must watch film. If @PrimeVideoIN had multiple language options for this film. This will become the most watched movie all over the world by now. Much relevant film . #coronavirusindia #COVID19outbreak	0.33	0
Ima drop this single in the next 2 weeks	0.25	0
snoo for skincare model please his skin is spotless and perfect for skincare products	0.20	0
Shit was all crazy I dipped they were burning the spot	0.17	0

These features have been chosen in order to capture the two most useful signals for detecting a TERF: their membership within a community and their unique brand of transphobic rhetoric. Network features simply aim to represent the community aspect of the TERF group on Twitter. Following is used as a strong signal for the types of information a user is interested in engaging with, and helps prevent false positives in which a user may tweet about similar topics but engage with them in a different way (such as trans activists or radical feminists who are not transphobic). We choose to use only the embedding of a single, most likely to be "TERF-like" tweet, with the logic that it only takes one tweet with such transphobic rhetoric to make it clear that a user is a TERF.

3.2.3 Classifiers

For the network and topic frequency features, we classify users using both logistic regression and a linear SVM, which output the probability of the user belonging in either class. In each case, the logistic regression slightly outperforms the linear SVM, so we choose it for our overall end-to-end model. We select these classifiers because they are computationally inexpensive without trading away performance. Meanwhile, the tweet embedding features, encoded by the pre-trained BERT model, are classified using a fully connected linear layer in order to fine-tune for this task.

Figure 2.1: Feature generation and model architecture



The overall model classifies users according to a logistic regression model over the probability the user is a TERF, according to each of the three categories of features. Each individual model produces a probability distribution across categories (0 - 1 for the logistic regression, -3 to 3 from the BERT logits) that is converted to the final predicted label. The pooling layer takes the probability for label 1 (the user is a TERF) produced by the three individual models for each feature category, and uses it as input for the overall logistic regression. Since we find that the network features and BERT tweet embeddings far outperform the topic frequency features, we also build a logistic regression model using only the label probability from those two feature categories to compare.

To build our fine-tuned BERT model, we use the *huggingface* library, which offers various NLP frameworks, along with *pytorch* for deep learning in python. For logistic regression, we use *scikit-learn*'s implementation. In each case, we split the data into 90% train and 10% test and perform k-fold cross-validation.

2.3 Results

The results from models trained on individual categories of features as well as models combining categories are compared here. A logistic regression trained solely on the topic frequency features performs the worst, with an F1 score of 0.4981, or worse than random. The network features indicating who a user follows provides the strongest signal for classifying TERFs out of all classifiers trained on a single category of features, including the tweet embeddings.

We see a slight improvement over using solely following features in our classifying task when we incorporate all three categories of features in our overall model, but it is not statistically significant. Since the topic frequency features had the weakest performance (as bad as random), we test one final model which combines both the tweet embedding and following features. This model performs the best overall, with an F1 score of 0.8793.

We report the confusion matrices in Appendix C (indicating number of false positives, true positives, false negatives, and true negatives) along with the averaged F1-score below, which is the harmonic mean of precision and recall. For each category, precision indicates how many users classified as that category are actually labeled as that category. Recall measures how many users are classified by the model correctly as belonging to that category, out of all users labeled as such. In Table 2.3, we detail the accuracy, precision, recall, and F1 for each model, averaged between both classes.

Table 2.3: Measures of model performance

Model features	**Accuracy**	**Precision**	**Recall**	**F1**[1]
Topic frequency	0.6622	0.6243	0.6583	0.4981
Following	0.9025	0.8915	0.9023	0.8700
Tweet embedding	0.7563	0.7474	0.7509	0.7055
Tweet embedding + following	**0.9078**	**0.9064**	0.9046	**0.8793**
All features	0.9014	0.8829	**0.9155**	0.8767

2.4 Discussion

Among the individual features, we see that the network category on its own provides a strong signal for classifying a TERF. This suggests that the

[1] F1 is averaged across runs while the other scores are calculated using true labels and classifications from the median k-fold run.

types of users one follows is a good indicator for members of a community-based hate group. The limitations of this method is that it remains static while communities are ever-changing: new users may become figureheads or communities may shift. This method is only useful for existing "clusters" of TERFs that are captured by the list we initially trained our model on. In the future, techniques for expanding the following list and validating a continuously growing list of TERFs could dynamically expand model coverage.

The topic model features performed the worst, with an accuracy of approximately 65% and an F1 score no better than random. This indicates that the topics are either not coherent enough to provide a good signal for the task, or our methods for measuring topic frequency should be adjusted. For each tweet, we calculated its topic distribution using our 25-topic model, and assigned it to the topic with the highest probability if it exceeded 10%. Either integrating this percentage into the network feature vector or setting a higher probability for assigning a tweet could lead to improvements in model performance. Another future area of exploration would be manually classifying topics as "TERF-indicative" or not. Generating the topic model with other methods besides LDA such as DMM or aggregating tweets into longer documents could also see an improvement in using topic distribution of tweets to classify users.

The tweet embedding only model, a fine-tuned BERT, performed better than the topic frequency feature but not as good as the network following model. This indicates that the semantic meaning of a selected tweet is generally a good signal for whether a user is a TERF or not. However, the single tweet selected for embedding may not be most indicative of the user's stance on trans people. We select the tweet most likely to be TERF-like by using the 25-topic model to find the one with the highest probability of falling under topic 12, which was hand-selected. This relies heavily on the topic model providing the most useful tweet, and changing the pipeline for obtaining the input could drastically affect the performance of the BERT fine-tuning.

Improving the tweet embedding feature could involve changing the way the "top tweet" is selected, or using different fine-tuning methods for BERT. As demonstrated by the topic frequency features, the topic distribution alone does not provide the best classifier for these users. Improving the topic model for these tweets could inform the selection process for the top tweet. Another option, instead of simply selecting a single tweet to represent the account, could involve selecting a handful of likely stance signifier tweets and aggregating their BERT encodings to be classified. Such an aggregation method could produce a more robust picture of the account's stance. Selecting a different pooling method for fine-tuning BERT could also improve the performance on the classification task. We currently use the [CLS] (classification) token to fine-tune the BERT model on an untrained linear layer, which is the part of the final hidden layer that is meant to represent the entire sequence. However, other work [26] in testing semantic similarity search using BERT has found that using other pooling methods on the final hidden layer sees a significant improvement in the task. Performing max-pooling or mean-pooling on the sequence of hidden-states for the entire input sequence may provide a better semantic representation instead of the [CLS] token. Changing the pooling method on what is being fine-tuned could also boost the performance of the BERT model.

Combining the features into the full model sees a slight improvement over the logistic regression built using solely the network features, while the model incorporating only the network and tweet embeddings performs the best. The performance of the logistic regression over solely the topic frequency features was the worst across the board, and adding its input likely only increases noise instead of providing helpful information. With a F1 score of 0.8793, we see extremely promising results from a Twitter user classification task that only requires two pieces of information from a user: their most recent 100 tweets and their following list.

Ultimately, we were able to build a model with an F1 of 0.8793 to classify whether a Twitter user was a TERF or not using only a few pieces of information from a user. From the 100 most recent tweets of a user, we obtain the tweet most likely to contain a hand-picked TERF-like topic by using a topic model to generate the topic probabilities. We encode this single tweet using BERT to obtain a vector embedding representing the semantic meaning, and use a fine-tuned linear layer to predict the probability of the tweet falling under the TERF label. On the other hand, we represent the Twitter user's following network using a one-hot vector of whether the user follows the top 2000 Twitter users followed by the TERF target list and overall Twitter users within our corpus (1000 from each). With these two categories of features, we are able to build a logistic regression model to predict whether a Twitter user is a member of the TERF hate group with 90% accuracy.

Conclusion

Within this study, we accomplished two tasks. We undertook the first computational study of the hate group known as trans-exclusionary radical feminists (TERFs) on Twitter, and built a model for classifying whether a Twitter user is a member of this group with an accuracy of 90%.

Though TERFs are likely the largest online group whose main tenets are underpinned by transphobia, little academic attention has been paid to their communities on social networks such as Twitter. Building upon a community-sourced list of approximately 13,000 TERFs on Twitter, we analyze the network structure of the online community as well as the textual content of their tweets. We find that while subgroups exist, the community as a whole is strongly connected and often tweet about topics in the same distribution; regardless, there are subgroups that are either more political in nature or more focused on traditional TERF-related topics.

As the group is known for targeted harassment and circulating hateful rhetoric regarding trans people on Twitter, we also build a model to predict whether a Twitter user would be a TERF or not in the interest of automating their detection. Our model ultimately uses two categories of features, relying on the following network of the user as well as the semantic information of one tweet, and labels users with an F1 of 0.8793.

Future work in this area could lie in finer-grained studies of the TERF community on Twitter and Reddit (where there are significant subreddits where they convene), enhancing the corpus of Twitter users that are TERFs from other publicly available lists, and improving the Twitter user classifier to release as a web application. We caution against a solely text-based approach to studying and detecting members of this community, as the risk of false positives are much higher without taking into account the network aspect. In particular, validating a classifier across certain groups on Twitter such as transgender social activists or regular radical feminists who are not transphobic is a necessary first step to ensure the model does not misclassify marginalized groups.

This thesis ultimately embarks on the first computational study on trans-exclusionary radical feminists (TERFs) on Twitter and demonstrates an automated process for detecting members of said hate group. We intend for this to be only the beginning of academic interest in this particular area of abusive content on social media, and for more research focus to attend to TERFs and other instances of transphobia online in the future.

Acknowledgements

I would like to thank my thesis advisor, Professor Saeed Hassanpour for his guidance on this project, even while the topic was a passion project and outside his area of expertise. I would like to thank other professors and mentors who have encouraged my research and lent their willing ear, particularly Professor Thomas Cormen. I would like to thank my peers who have shaped me and sharpened my pursuit for radical justice in the technological sphere, especially Os Keyes for bringing my attention to the dearth of research on TERFs online. I would like to thank my family and friends for their unconditional support and tactful rebukes. Most of all, I would like to acknowledge the trans folks who suffer from transphobic attacks and abusive harassment online and off; while I am able to approach the contents of this thesis from an academic standpoint, this is their lived experience that I only hope this work can reaffirm and alleviate.

Appendices

Appendix A: Top words per topic model, n=25

Topic	Words
0	0.052*"best" + 0.050*"help" + 0.039*"parti" + 0.038*"social" + 0.034*"media" + 0.033*"happi" + 0.024*"countri" + 0.019*"mental" + 0.018*"conserv" + 0.018*"nero"
1	0.128*"time" + 0.115*"go" + 0.049*"sure" + 0.040*"hope" + 0.035*"real" + 0.034*"wrong" + 0.033*"that" + 0.033*"money" + 0.027*"long" + 0.022*"stupid"
2	0.076*"better" + 0.045*"problem" + 0.042*"kind" + 0.036*"turn" + 0.029*"send" + 0.028*"plan" + 0.022*"beauti" + 0.022*"rule" + 0.021*"term" + 0.020*"defend"
3	0.153*"think" + 0.080*"thing" + 0.035*"you" + 0.032*"peopl" + 0.029*"mayb" + 0.027*"littl" + 0.021*"see" + 0.020*"mind" + 0.019*"they" + 0.017*"stuff"
4	0.142*"know" + 0.063*"want" + 0.055*"don" + 0.040*"talk" + 0.032*"happen" + 0.028*"peopl" + 0.027*"kid" + 0.023*"school" + 0.023*"call" + 0.023*"gonna"
5	0.038*"home" + 0.038*"face" + 0.034*"hell" + 0.031*"http" + 0.026*"christian" + 0.024*"order" + 0.024*"wear" + 0.023*"number" + 0.022*"citi" + 0.020*"rest"
6	0.060*"trump" + 0.055*"leav" + 0.048*"vote" + 0.035*"white" + 0.032*"support" + 0.027*"state" + 0.027*"black" + 0.026*"democrat" + 0.026*"elect" + 0.026*"power"
7	0.127*"say" + 0.062*"mean" + 0.039*"word" + 0.036*"pretti" + 0.026*"hand" + 0.021*"respons" + 0.018*"total" + 0.016*"obvious" + 0.015*"bear" + 0.015*"nada"
8	0.041*"case" + 0.033*"high" + 0.028*"damn" + 0.027*"control" + 0.027*"month" + 0.025*"build" + 0.024*"level" + 0.023*"muslim" + 0.023*"morn" + 0.017*"explain"
9	0.064*"read" + 0.053*"believ" + 0.034*"write" + 0.032*"book" + 0.030*"bring" + 0.028*"day" + 0.024*"imagin" + 0.021*"sign" + 0.019*"choos" + 0.019*"sell"
10	0.110*"thank" + 0.085*"watch" + 0.060*"https" + 0.036*"youtub" + 0.032*"liber" + 0.027*"youtu" + 0.022*"million" + 0.021*"class" + 0.021*"dead" + 0.018*"educ"

11	0.076*"game" + 0.073*"fuck" + 0.069*"tell" + 0.054*"stop" + 0.046*"yeah" + 0.029*"speak" + 0.023*"lie" + 0.021*"miss" + 0.020*"truth" + 0.019*"complet"
12	0.100*"right" + 0.097*"women" + 0.036*"tran" + 0.034*"woman" + 0.030*"gender" + 0.026*"true" + 0.025*"male" + 0.024*"femal" + 0.016*"charact" + 0.016*"sexual"
13	0.056*"differ" + 0.053*"follow" + 0.047*"away" + 0.044*"sorri" + 0.034*"wish" + 0.023*"drive" + 0.023*"respect" + 0.020*"deserv" + 0.017*"fund" + 0.017*"gotta"
14	0.105*"work" + 0.054*"like" + 0.042*"play" + 0.036*"girl" + 0.035*"hard" + 0.030*"chang" + 0.028*"sound" + 0.025*"open" + 0.024*"interest" + 0.021*"account"
15	0.212*"like" + 0.114*"good" + 0.097*"look" + 0.049*"feel" + 0.040*"make" + 0.025*"wait" + 0.021*"can" + 0.019*"final" + 0.019*"understand" + 0.018*"peopl"
16	0.038*"kill" + 0.036*"exact" + 0.028*"claim" + 0.028*"self" + 0.026*"pay" + 0.023*"anti" + 0.023*"stay" + 0.019*"pass" + 0.019*"especi" + 0.018*"creat"
17	0.071*"live" + 0.057*"life" + 0.050*"world" + 0.032*"hous" + 0.031*"head" + 0.027*"hold" + 0.025*"liter" + 0.024*"learn" + 0.023*"honest" + 0.020*"save"
18	0.047*"hate" + 0.041*"friend" + 0.034*"free" + 0.030*"report" + 0.027*"famili" + 0.026*"attack" + 0.023*"american" + 0.020*"illeg" + 0.018*"crime" + 0.017*"legal"
19	0.040*"place" + 0.037*"human" + 0.035*"nice" + 0.028*"clear" + 0.028*"public" + 0.026*"anim" + 0.024*"child" + 0.024*"abus" + 0.023*"protect" + 0.018*"fail"
20	0.102*"love" + 0.059*"great" + 0.047*"care" + 0.043*"post" + 0.038*"agre" + 0.030*"wonder" + 0.030*"rememb" + 0.022*"polic" + 0.020*"listen" + 0.019*"funni"
21	0.043*"fact" + 0.041*"question" + 0.035*"caus" + 0.032*"ask" + 0.027*"dude" + 0.027*"guy" + 0.025*"check" + 0.025*"forget" + 0.024*"instead" + 0.023*"ignor"
22	0.136*"https" + 0.100*"twitter" + 0.060*"tweet" + 0.050*"news" + 0.044*"video" + 0.042*"today" + 0.036*"status" + 0.036*"break" + 0.033*"week" + 0.024*"comment"
23	0.055*"hear" + 0.049*"stori" + 0.031*"base" + 0.030*"night" + 0.025*"share" + 0.021*"close" + 0.021*"second" + 0.021*"canada" + 0.020*"heart" + 0.019*"surpris"
24	0.073*"shit" + 0.059*"year" + 0.056*"take" + 0.050*"lose" + 0.045*"probabl" + 0.027*"absolut" + 0.026*"dont" + 0.025*"cours" + 0.023*"cool" + 0.022*"keep"

Appendix B: Sample of signal tweets selected for BERT embedding, topic 12

Label 1 (user is a TERF)

Tweet	**% topic**
Neither option. There will have to be provision made for 3rd spaces for trans	0.75
So men dressing as women is transphobic ie being trans is transphobic. Sounds about right.	0.63
Is carnage, regulations have already been changed to allow males to play in female teams.	0.41
This womans better have done this. OR ELSE.	0.34
I actually had a gender critical dream last night	0.47
The making of trans children	0.50
Butler - children are born without a biological sex. Nandy - Corbyn was not for the people. RLB - men are women. All - we'll remove safe spaces for women even with sky high violence and murder. Nandy - I don't know how we can pay for stuff. Unbelievable ignorance & stupidity	0.29
It'd just be "people who desire to radically alter their bodies" for some unspecified reason. A rather extreme desire, and explained by nothing. Blaire is right.	0.60
1. Pretty woman 2. Pretty B U F F woman	0.40
You can repeat that all you want. But it does not male it true.	0.40
There is no hate. There is no demonising of trans people. Trans people have rights! Same as all of us. Women have rights too. You’re a worryingly lazy thinker for an MP.	0.60
Stunning, brave and unbelievably powerful for a woman.	0.40
Nope. DNA says she was a woman. She. was. a .woman. A strong, fierce, woman.	0.43
Never forget why we fight	0.34
Just say "transgender", "transsexual", or "trans". We don't need the word "woman".	0.66

Of the gender ideology that Stonewall and TRAs are flogging, as opposed to a person having dysphoria, loathing their genitalia & body, and feeling they are in need of medical interventions. IMO, gender ideology is an insult to women, and Transexual people with dysphoria.	0.43
just wanna be clear, but your position is a 13-year-old NEVER decides they are trans because a lot of kids in their social circles are experimenting and exploring in this area? that just can't happen? it's horribly pseudoscientific to posit a situation where it happens?	0.25
Lesbian? I thought you were a dude	0.25
Definitely not. It's sexist and unfair on women to allow it.	0.50

Label 0 (user is not a TERF)

Tweet	**% topic**
One thing sis doesn't do is press the issue. It's me. I'm sis.	0.25
My mom forcing me to go in zoom for bible study	0.38
me?wanting to rave right about now	0.34
Someone take my phone from me I'm flirting with every single girl on my snap	0.51
Yes for both men and women	0.51
If you're active but have less than 15k followers Follow, retweet and drop your handle, let's follow you. Turn on my notification for more Gain...	0.09
me too omg every single time i saw it i'd just reply "sorry" like omg please just stop posting it	0.12
he betrayed all of us first when he dyed it blond again it's literally his fault	0.25
The Vashon community developed a model to test, trace and isolate — in essence, a coronavirus response plan that they call the Rural Test & Trace Toolkit. Their model could help in other isolated parts of the country.	0.37
I keep changing My mind about how to destroy Australia.	0.31
Next up... it becomes an Olympic Sport	0.34
#BREAKING: Maine is receiving $52,673,451 to purchase, administer, and expand capacity for COVID-19 testing. [link]	0.28

Gender dysphoria involves a conflict between a person's physical or assigned gender and the gender with which they identify. People with gender dysphoria may be very uncomfortable with the gender they were assigned, sometimes described as being uncomfortable with their +	0.69
They also mentor young adults from these communities to become coaches themselves, to drive the change further. A great tool in the hands of the right person, sport can be incredibly effective in driving social change & that is exactly what @[mention] is doing through @[mention] .	0.34
this is a true fact, scientist are stunned as well	0.40
Tbf I definitely have asked people in marginalized groups that I'm not a part of about their personal experience bc google will NOT answer my question but I do *try* google first	0.41
I think all you liberals Are dumb as a box of rocks. Republicans aren't trying to take abortion away from woman's rights. They're trying to do a stop late term abortion so you don't have the right to rip a baby limb for limb from the womb. Smh	0.21
Ima handsome ass mf, the women rather see me than the fit	0.36
Right on as a fellow Texan.	0.26

Appendix C: Sample of individual confusion matrices for user classification model predictions

Topic frequency

	Assigned Negative	**Assigned Positive**
Actual Negative	707	354
Actual Positive	128	238

Following

	Assigned Negative	**Assigned Positive**
Actual Negative	819	88
Actual Positive	51	468

Tweet Embedding

	Assigned Negative	**Assigned Positive**
Actual Negative	656	186
Actual Positive	158	412

References

[1] R. Abebe, S. Barocas, J. Kleinberg, K. Levy, M. Raghavan, and D. G. Robinson, “Roles for Computing in Social Change,” *In Conference on Fairness, Accountability, and Transparency (FAT* ’20)*, 2020.

[2] E. A. Abozinadah and J. H. Jones, “A Statistical Learning Approach to Detect Abusive Twitter Accounts,” *Proceedings of the International Conference on Compute and Data Analysis - ICCDA '17*, 2017.

[3] D. Alvarez-Melis and M. Saveski, “Topic Modeling in Twitter: Aggregating Tweets by Conversations,” *Proceedings of the Tenth International AAAI Conference on Web and Social Media - ICWSM*, 2016.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008.

[6] W. O. Bockting, M. H. Miner, R. E. S. Romine, A. Hamilton, and E. Coleman, “Stigma, Mental Health, and Resilience in an Online Sample of the US Transgender Population,” *American Journal of Public Health*, vol. 103, no. 5, pp. 943–951, 2013.

[7] S. Bodapati, S. Gella, K. Bhattacharjee, and Y. Al-Onaizan, “Neural Word Decomposition Models for Abusive Language Detection,” *Proceedings of the Third Workshop on Abusive Language Online*, 2019.

[8] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. M. Blei, “Reading Tea Leaves: How Humans Interpret Topic Models,” *Neural Information Processing Systems*, 2009.

[9] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, “Comparing community structure identification,” *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 09, 2005.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proceedings of NAACL-HLT*, 2019.

[11] S. Emmons, S. Kobourov, M. Gallant, and K. Börner, “Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale,” *Plos One*, vol. 11, no. 7, 2016.

[12] S. Fortunato, “Community detection in graphs,” *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.

[13] D. (D. Ghosh and R. Guha, “What are we ‘tweeting’ about obesity? Mapping tweets with topic modeling and Geographic Information System,” *Cartography and Geographic Information Science*, vol. 40, no. 2, pp. 90–102, 2013.

[14] M. Girvan and M. E. J. Newman, “Community structure in social and biological networks,” *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[15] K. Clements-Nolle, R. Marx, R. Guzman, M. Katz, “HIV prevalence, risk behaviors, health care use, and mental health status of transgender persons: implications for public health intervention,” *American Journal of Public Health*, vol. 91, no. 6, pp. 915–921, 2001.

[16] L. Hong and B. D. Davison, “Empirical study of topic modeling in Twitter,” *Proceedings of the First Workshop on Social Media Analytics - SOMA '10*, 2010.

[17] K. Johnson and D. Goldwasser, ““All I know about politics is what I read in Twitter”: Weakly Supervised Models for Extracting Politicians’ Stances From Twitter,” *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, 2016.

[18] D. Jurgens, L. Hemphill, and E. Chandrasekharan, “A Just and Comprehensive Strategy for Using NLP to Address Online Abuse,” *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[19] A. Lancichinetti and S. Fortunato, “Community detection algorithms: A comparative analysis,” *Physical Review E*, vol. 80, no. 5, 2009.

[20] S. Lewis, “How British Feminism Became Anti-Trans,” *The New York Times*, 07-Feb-2019. [Online]. Available: https://www.nytimes.com/2019/02/07/opinion/terf-trans-women-britain.html.

[21] V. Lynn, S. Giorgi, N. Balasubramanian, and H. A. Schwartz, "Tweet Classification without the Tweet: An Empirical Examination of User versus Document Attributes," *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, 2019.

[22] J. Mazarura and A. D. Waal, "A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text," *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2016.

[23] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic Evaluation of Topic Coherence," *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010.

[24] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[25] J. Qian, M. Elsherief, E. Belding, and W. Y. Wang, "Leveraging Intra-User and Inter-User Representation Learning for Automated Hate Speech Detection," *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018.

[26] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[27] M. Rosvall and C. T. Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences*, vol. 105, no. 4, pp. 1118–1123, 2008.

[28] M. Röder, A. Both, and A. Hinneburg, "Exploring the Space of Topic Coherence Measures," *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining - WSDM '15*, 2015.

[29] D. L. Sánchez, J. Revuelta, F. D. L. Prieta, A. B. Gil-González, and C. Dang, "Twitter User Clustering Based on Their Preferences and the Louvain Algorithm," *Advances in Intelligent Systems and Computing Trends in Practical Applications of Scalable Multi-Agent Systems, the PAAMS Collection,*

pp. 349–356, 2016.

[30] A. O. Steinskog, J. F. Therkelsen, and B. Gambäck, "Twitter Topic Modeling by Tweet Aggregation," *Proceedings of the 21st Nordic Conference of Computational Linguistics*, 2017.

[31] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," *Proceedings of the Third Workshop on Abusive Language Online*, 2019.

[32] Y. Wang, J. Callan, and B. Zheng, "Should We Use the Sample? Analyzing Datasets Sampled from Twitter's Stream API," *ACM Transactions on the Web*, vol. 9, no. 3, pp. 1–23, 2015.

[33] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing Twitter and Traditional Media Using Topic Models," *Lecture Notes in Computer Science Advances in Information Retrieval*, pp. 338–349, 2011.

[34] J. Zhu, Z. Tian, and S. Kübler, "UM-IU@LING at SemEval-2019 Task 6: Identifying Offensive Tweets Using BERT and SVMs," *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019.

[35] TERF Block List [https://twitter.com/terfblocklist]

[36] SnowballStem [https://snowballstem.org/]

[37] Demographic ensemble language identifier, extension of langid [http://slanglab.cs.umass.edu/TwitterLangID/]

[38] Langid [https://github.com/saffsd/langid.py]