The subtle language of exclusion: Identifying the Toxic Speech of Trans-exclusionary Radical Feminists

Christina Lu
Dartmouth College*
christinalu@deepmind.com

David Jurgens
University of Michigan
jurgens@umich.edu

Abstract

Toxic language can take many forms, from explicit hate speech to more subtle microaggressions. Within this space, models identifying transphobic language have largely focused on overt forms. However, a more pernicious and subtle source of transphobic comments comes in the form of statements made by Trans-exclusionary Radical Feminists (TERFs); these statements often appear seemingly-positive and promote women's causes and issues, while simultaneously denying the inclusion of transgender women as women. Here, we introduce two models to mitigate this antisocial behavior. The first model identifies TERF users in social media, recognizing that these users are a main source of transphobic material that enters mainstream discussion and whom other users may not desire to engage with in good faith. The second model tackles the harder task of recognizing the masked rhetoric of TERF messages and introduces a new dataset to support this task. Finally, we discuss the ethics of deploying these models to mitigate the harm of this language, arguing for a balanced approach that allows for restorative interactions.

1 Introduction

Transgender individuals are frequent targets of toxic language in online spaces (Craig et al., 2020; Haimson et al., 2020). Multiple approaches to recognizing such abusive language have focused on identifying explicit forms of abuse, such as using trans-specific slurs (Waseem et al., 2017; Schmidt and Wiegand, 2017; Fortuna and Nunes, 2018). However, not all verbal abuse directed towards the transgender community is so explicit. Within those transphobic groups, trans-exclusionary radical feminists (TERFs) are a community who is critical of the notion of gender, and position the existence of trans women as antithetical to "womanhood."[1]

I find it increasingly harder to believe that the people saying this nonsense actually believe it. A man is a woman because he wears some lipstick and says he's a woman, but a woman isn't a woman because of biology??
Some would say that LGB have already been "thrown under the bus" to accommodate an ideology that relies heavily upon gender stereotypes and "being in the wrong body." I hear there're a lot of lesbians who feel like this.
Guarantee they'll expect more rigorous research to debate the ethics of fancy shoes than they did for men in women's sports

Figure 1: Examples of harmful rhetoric by TERFs which reference notions of biological essentialism in defining gender and exclusion of transgender women from sports. While offensive, we include the examples here to highlight the subtlety in their exclusionary messages. Throughout the paper, all messages are lightly paraphrased for privacy.

As such, the language of their attacks is frequently couched in arguments promoting women's safety and rights—nominally positive language. TERF groups maintain an active presence across public social media and are often a source of transphobia online (Pearce et al., 2020). However, their masked rhetoric is unrecognized by current models for hate speech detection, and indeed, identifying TERFs in general can be difficult if one is not familiar with their lines of argumentation, as seen in the examples in Figure 1. Interacting with individuals propagating these beliefs can be materially harmful and as a result, multiple transgender communities and allies have established lists of known TERF accounts to help individuals block or avoid abuse. However, the recruitment of new individuals with TERF beliefs as well as sockpuppet accounts make

* Work performed in part at the University of Michigan

[1] We acknowledge that the use of the term TERF is potentially contentious, as some individuals who identify these views consider it derogatory. Nonetheless, our use follows current academic practice in naming (e.g., Williams, 2020).

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 79 - 91

manually keeping these lists up-to-date a challenge for mitigating their impact. In this paper, we widen the scope of abusive detection online by demonstrating a model for detecting both TERFs and nuanced TERF rhetoric on Twitter by analyzing their tweets and community features.

Work in abusive language detection for social media has become more widespread (Fortuna et al., 2020; Zampieri et al., 2020), but more subtle forms of hate speech such as dog whistles are notoriously difficult to capture (Caselli et al., 2020). TERF rhetoric directly falls into this category, as it consists of a particular brand of transphobia that employs dog whistles and bad faith argumentation. Prior work has only begun to address these subtle form of offensive such as microaggressions (Breitfeller et al., 2019; Han and Tsvetkov, 2020), condescension (Wang and Potts, 2019; Perez Almendros et al., 2020), and other social biases (Sap et al., 2020). Our work identifying TERFs and their rhetoric extends this recent line of research by filling the gap into an under-researched but important area of transphobic hate speech.

We introduce the first computational method for detecting TERF accounts on Twitter, which combines information from user messages and network representations. Using community-sourced data of over 22K users, we show that social and content information can accurately identify TERF accounts, attaining a F1 of 0.93. To support identifying TERF messages directly, we introduce a new dataset of gender and trans-identity related messages annotated for TERF-specific rhetoric, showing that despite the challenging nature of the task, we can obtain 0.68 F1. Together, these methods allow individuals to recognize and screen out the uniquely transphobic rhetoric of TERFs.

This paper provides the following contributions. First, little computational attention has been paid to TERFs and transphobic speech in previous work within the realm of abusive content detection. Our model is the first to tackle the challenge of capturing nuanced, transphobic rhetoric from TERFs, and leveraging it to identify TERFs on Twitter. Second, we introduce a new dataset for recognizing TERF-specific rhetoric, allowing the community to expand current efforts at combating abusive language. Finally, acknowledging the dual use of NLP (Hovy and Spruit, 2016), we consider the ethics of deploying these technologies in the risks and benefits of censuring versus allowing engagement with TERFs, arguing for a balanced approach that facilitates restorative justice.

2 TERFs in Online Spaces

Feminist ideals aim to promote women's rights and mainstream feminism is considered inclusive of transgender women (Williams, 2016). However, a small number of individuals claiming to be feminists have taken an opposite stance, arguing for transphobic views that push for biological essentialism and criticizing the notion of gender (Williams, 2020). This group was given the name "trans-exclusionary radical feminists" or TERFs as a way of separating their views. Drawing in part upon feminist arguments in Raymond (1979), TERFs argue that gender derives fully from the biological sex, which is dependent on a person's chromosomes and thus is binary and immutable (Riddell, 2006; Serano, 2016); it follows in their biological reductivist reasoning that a transgender woman is a man. As a result, TERFs frequently make claims seeded with anxiety about the encroachment of transgender women into women's spaces and rights (e.g., participation in sports or use of restrooms), as well as the need for biological tests of gender (Earles, 2019).[2]

For many TERFs, their rationale is embedded with real but misdirected fear of violence against and subjugation of women. Regardless, such harmful rhetoric directly marginalizes and excludes transgender women (Hines, 2019; Vajjala, 2020), often invalidating their very existence. These arguments frequently follow the subtle language of microaggressions (Sue, 2010, Ch.2). TERFs themselves are also not a monolithic bloc; individuals may vary in their stances towards transgender people, from claiming to openly support them as a separate group to radically opposing them and arguing such identities themselves are flawed. While all such attitudes are harmful, this range suggests that some viewpoints could be changed.

Less prevalent in the United States and Canada, TERFs within the United Kingdom hold an unfortunately mainstream position within feminism (Lewis, 2019), with a notable proponent being J.K. Rowling (Kelleher, 2020), author of the Harry Potter series. TERFs are present on multiple platforms; TERFs maintained an active community of over

[2]We note that recent proponents of this ideology have adopted the name "gender critical" but espouse the same offensive beliefs of biological essentialism (Tadvick, 2018).

64K users on the r/gendercritical subreddit, until June of 2020, after which it was banned by Reddit for the promotion of hate speech.

The presence of TERFs in online communities represents a significant risk to transgender individuals, as they perpetuate targeted harassment and doxxing. Online spaces are particularly critical for transgender individuals due to their role in facilitating the transition experience (Fink and Miller, 2014) and seeking social support during the coming out process (Haimson and Veinot, 2020; Pinter et al., 2021). As some individuals may not have publicly come out to family and coworkers (but do so online, potentially anonymously), targeted harassment poses risks for some individuals (Kade, 2021). Potential interactions between TERFs and transgender individuals can further marginalize individuals and reduce the perceived support.

3 A Dataset for Recognizing TERFs

As frequent targets of abusive language, transgender individuals and their allies have curated lists of known TERF users on Twitter in attempts to mitigate the harm they cause. These user lists form the basis for our dataset, described next.

3.1 User Lists

Our ultimate goal is to identify TERF users and their rhetoric. Prior work has shown that user-created lists on Twitter are reliable signals of identity that can be used for classification tasks (Kim et al., 2010; Faralli et al., 2015). Accordingly, we collect curated lists from two communities, along with a random sample of users as a control set.

First, TERFblocklist is a manually-curated list of TERF accounts by trans women and activists. The block list uses a third-party Twitter API web app, Block Together,[3] which enables users to screen out content and interaction from users on shareable, custom block lists. Potential additions to this list are sent to the maintainer who verifies the accusations of transphobia before they are added. Through manual submissions, users identified 13,399 TERF accounts, which forms the basis for our list of Twitter users who are TERFs.[4]

[3]As of June 2020, Block Together shut down but other alternatives such as Block Party and Moderate have the same functionality.

[4]We recognize that block lists are themselves products of exclusion that can potentially include users who do not have a particular view or identity. However, we still use such lists here, as they have been curated by members of the trans community we trust their judgments in who poses risks.

Category	No. users	No. tweets	Description
TERF	8,631	13,508,673	TERFs
Trans-friendly	14,827	1,291,908†	Explicitly trans-friendly
Control	11,510	33,573,308	Random English speakers

Table 1: Summary of the sizes of the datasets used in these studies, reflecting only English-language tweets per category. †Only up to 100 recent tweets were collected for each user in the Trans-friendly category.

Second, as a direct response to TERFblocklist, TERF users created a separate block list of their own on Block Together, which contained 17,091 "transactivists and transcultists," as a way of identifying users whom they could actively target or selectively ignore. While initially designed for unethical reasons (targeting users), this data forms the basis for our list of trans-friendly users. Because both TERF and trans-friendly users share high-level themes in their discussion around transgender issues, having representation of both groups is essential for ensuring that trans-friendly accounts are not being mistakenly labeled as TERFs.

Third, as not all users discuss transgender issues, we randomly sample 13,152 "control" English-speaking users from the Twitter decahose in May 2020 and retain all users who are not on either of the two blocklists. As some users had private Twitter accounts, the final number of users in our corpus is a subset of these original lists.

3.2 Linguistic and Social Data

For each user, we collect two types of data that we hypothesize will capture whether they are a TERF or not: tweet text and the user's friends (i.e., the Twitter users they follow). While the text of a tweet carries the most information about the stance of the user, the people they follow are also strong signals for both the community they are a member of and what content they willingly engage with. This task is particularly context-sensitive due to the dog whistles employed by TERFs, and necessitates both types of data.

Through Tweepy and the Twitter API, we collect all recent (2019 onward) tweets from each user in the TERF (13,508,673 tweets), trans-friendly (1,291,908 tweets), and control (33,573,308 tweets) groups and discard non-English tweets using the language classifier of Blodgett et al. (2016) for labeling social media English. Due to API limitations

when retrieving tweets, we keep only up-to-100 recent tweets for each user in the Trans-friendly category to maximize the diversity in that sample, without overrepresenting any one user. We also collect the list of user IDs belonging to each user's friends using the Twitter API. At the time of collection, some users had taken their accounts private, which prevented collecting all data. Table 1 shows the statistics for our final dataset.

4 Building a TERF classifier

To recognize TERF users, we use a multi-stage approach that combines information from individual messages on topics discussed by TERFs with social features representing who they follow. Following, we describe the three stages: how we (1) recognize topics closely related to TERF rhetoric, (2) identify individual messages likely to come from TERFs, and (3) combine textual and social features to detect TERF users themselves.

4.1 Identifying TERF Topics

Despite espousing harmful rhetoric, individuals with TERF beliefs routinely engage in conversations about commonplace topics. As a result, training any TERF-specific classifier is likely to mistakenly pick up on idiosyncratic content not related to TERF rhetoric. Therefore, in the first stage, we build a topic model to identify content themes that are related to TERF rhetoric and focus our later analysis primarily on this content.

To identify potentially TERF content, we fit a STTM topic model (Qiang et al., 2019), which suits the brevity of character-limited tweets. Prior to fitting the model, tweets are preprocessed to remove links and tokens under three characters and to filter out tokens appearing in fewer than 10 tweets or more than half of all, as these words are either unlikely to be content words related to our target construct or too rare to aid in topic inference. All remaining tweets with four or more tokens are used to fit the topic model. The number of topics is determined using topical coherence and we vary the number from 5 to 80 in 5-topic increments. Coherence was maximized at 15 topics; following best practice from Hoyle et al. (2021), a separate human evaluation was also done by the authors who also found 15 topics resulted in the most-coherent, least-redundant themes. As a robustness test, this procedure was replicated three times in each configuration to manually ensure that topical themes were roughly consistent across runs.

Topic	Top words
0	people like police country know trump illegal think right state want border time iran years world government going need america
2	labour brexit vote party people like think corbyn deal leave want know voted election time tory right boris tories remain
5	jesus like love people church life christ know lord good world time think catholic bible christian great right said family
8	like movie good think time people love know watch great character best star film thing going better movies shit story
9	**women trans people male female gender woman rights like think males know right want girls spaces need biological lesbians females**
14	twitter people like tweet know read think account news time media video tweets good said youtube right women article going

Figure 2: The most probable words for a sample of topics learned from TERF tweets. Topic 9 (bolded) reflects the content most likely to pertain to transgender issues and contain transphobic messages.

All runs demonstrated a manually-identified topic that contained content about trans women, gender, and other common transphobic TERF talking points. The most-probable words for a sample of topics are shown in Figure 2, where Topic 9 was identified by experts as most related to TERF-related rhetoric. Across all content, approximately 7.4% of tweets from TERFs are from this topic, compared to 4.3% for transgender individuals and 0.2% for individuals from the randomly-sampled control group. The use of this topic by non-TERF users underscores that the topic itself is broad and not necessarily solely TERF rhetoric, but rather a more general topic that includes material related to gender and trans issues (both appropriate and abusive). We refer to this topic as the *trans topic* in later sections. Finally, we note that the topic models consistently identified topics relating to British-specific content (e.g., Brexit), shown in Topic 2 in Figure 2, underscoring the association of TERFs with the UK (Hines, 2019; Lewis, 2019).

4.2 Classifying TERF-signaling Tweets

Using the topic model, the subsequently-identified trans topic act as an initial feature for helping distinguish TERF users. To identify whether messages with this topic are offensive, we fine-tune a language model to identify trans topic tweets from TERF users, using the topic as a weak label on

whether the content is offensive—i.e., that content from TERF users in this topic is likely to be offensive, while content from others would not be. We train a BERT model (Devlin et al., 2019) to recognize whether a tweet with this topic came from a known-TERF user versus a user in our control set, which includes transgender individuals, their allies, and a sample of English-speaking users. Because of the heuristic labeling of data, this classifier's decisions are intended to act as features for the downstream task of recognizing users, rather than being designed for recognizing TERF rhetoric (which is addressed later in §5).

Tweets were selected for the training set as follows. To avoid potential confounds from multiple tweets from a single user, we partition users 90:10 into training and test sets.[5] We added all TERF-topic tweets across the three groups of training users into the training set, so the model could learn to distinguish when TERF-topic tweets came specifically from TERFs. We also supplemented the corpus with a sample of other tweets from non-TERFs, in order to make the model more robust against unrelated tweets. In total, this yielded 491,998 TERF-topic tweets from TERFs and 275,189 and 315,202 mixed topic tweets from the transgender and control user sets, respectively, which reflect inoffensive content in this topic. The BERT model is fine-tuned for four epochs using AdamW (η=2e-5, ϵ=1e-8) on a batch size of 32.

Results The classifier ultimately had high performance on the test set, attaining an F1 of 0.98 on identifying control tweets from non-TERFs and an F1 of 0.96 on recognizing that a TERF-topic tweet came from a TERF.[6] Such tweets were labeled as TERF 92% of the time, while signal tweets from non-TERFs (which are supposed to be the most difficult to distinguish) were labeled as TERF approximately 45% of the time. This result points to strong linguistic differences in the language of the two groups and that the BERT classifier can potentially be useful for distinguishing the two user types. However, the high false-positive rate for signal tweets from non-TERFs (i.e., those not espousing such rhetoric) underscores the risks in using single-tweet classifications alone to label a user as a TERF; great care is needed to reduce the rate of false positives at the user label. We refer to this classifier as the TERF-*signal* classifier in later analyses.

4.3 Identifying TERF users

In the final phase, we aim to identify TERF users themselves through their linguistic and social features. While linguistic features such as those of our BERT and STTM models identify TERF-related content, extra-linguistic features of accounts can also be powerful signals of the account type (Al Zamal et al., 2012; Lynn et al., 2019) and can even help identify accounts known to engage in abusive behavior (Abozinadah and Jones Jr, 2017). In particular, the social network aspect of Twitter allows us to use particular frequently-followed accounts as features—e.g., accounts by high-profile users that promote TERF ideology. Following, we build a classifier to identify these users using linguistic and network features. Our ultimate goal is to help supplement existing TERF user lists to mitigate the users' effect on the transgender community.

Experimental Setup Information on who a person follows on Twitter is potentially informative of their world view and what information they are regularly exposed to. We encode a user's social network as a set of binary features corresponding to whether the user follows specific accounts on Twitter. We include features for (i) each of the thousand most-followed users overall in our training data and (ii) each of the thousand most-followed accounts by users in our TERF list.

Our linguistic features combine different aspects of the STTM and BERT models, computed over the 100 most-recent tweets from each user. Six features are used: (1, 2) the mean posterior probability of a tweet being from the trans topic and the max across all tweets, (3) the percentage of tweets that are from the transgender topic, (4) the mean probability of a transgender-topic tweet being a signal tweet, (5) the mean probability of a tweet in any other topic tweet being a signal tweet, and (6) the maximum probability of any tweet being a signal tweet.

A logistic regression model is trained on these network and linguistic features using the same train and test partitions in previous experiments to avoid data leakage. To test the contribution of each feature type, we evaluate ablation models that reflect using (i) only features from the STTM topic model, (ii) only features from the signal classifier, (iii) all the text-based features from the STTM and signal

[5] No hyperparameter optimization was performed, so no development set was used.

[6] Throughout the paper, we use Binary F1 with the TERF-related category as the positive class.

Model	AUC	Prec.	Rec.	F1
Random	0.50	0.18	0.53	0.27
LR Baseline	0.92	0.64	0.68	0.66
Topic Feats.	0.70	0.55	0.29	0.38
BERT Feats.	0.89	0.89	0.68	0.77
Topic & BERT Feats.	0.91	0.94	0.78	0.85
Network Feats.	0.95	0.92	0.80	0.86
All Features	**0.98**	**0.96**	**0.90**	**0.93**

Table 2: Performance at recognizing TERF accounts from different feature types. The Logistic Regression (LR) baseline was trained solely on unigram and bigram features of the text; The All Features model does not include the baseline's lexical features, only those of the non-baseline models.

models, and (iv) only the network features (no text-related features). Finally, as a test for whether this high-level aggregation is needed to improve performance, we include a Logistic Regression baseline trained on unigrams and bigrams from the concatenated messages of a user.[7] Models are compared with a random baseline.

Results The combined model was highly accurate at identifying TERF accounts, attaining an F1 of 0.93 as shown in Table 2. Models trained on individual feature categories outperformed the random baseline, indicating they each contained meaningful signals. Only the signal features and network features were able to outperform the Logistic Regression text-based baseline ($p<0.01$ using McNemar's test). However, the transgender topic features still capture complementary information as the signal features, where combining them still improves performance ($p<0.01$) over models trained on each feature individually.

The social network features and combined-linguistic features provided similar performance, with network features outperforming slightly ($p=0.04$). This network result suggests that many TERF users actively engage in strategic social networking to the point that the users they follow are reliable indicators of their underlying attitudes on transgender issues. This high performance of network features mirrors similar types of inferences for social attitudes like political affiliation (Barberá et al., 2015) and topical stance (Lynn et al., 2019).

Ultimately, the combination of all features was essential for high performance and significantly improved ($p<0.01$) over any individual feature type. Performance gains over both feature types came from increased Recall, which indicates that not all TERF users engage in following prominent TERF accounts or frequently share TERF rhetoric.

The act of classifying users as TERFs potentially carries a risk of harm. While the model's performance is notably high, misclassifications can potentially disenfranchise users who are mistakenly labeled as TERFs—e.g., labeling an individual from the transgender community as a TERF themself—or lead to ostracizing. The best model's performance indicates that most errors are of omission, not labeling a TERF as such, which we view as the appropriate type of error to avoid the risk of harm.[8] While the model is highly accurate, we explicitly call for avoiding its use in fully automated settings, e.g., automatically banning or censuring users; instead, this classification tool is only meant to help humans identify accounts among the huge search space and then manually review such accounts.

Compared to users in the random sample portion of our dataset, both TERFs and transgender individuals likely have overlap in their topical content. As a result, errors that are introduced through the topic model and signal tweets could potentially bias the model so that most false positive errors are made for transgender users. However, examining the false positive error rates shows that between these groups, individuals from the random sample are more likely to be labeled as TERFs (1.9%) versus those in the trans-friendly group (1.3%), suggesting the features are not biased due to shared topicality.

5 Recognizing TERF Rhetoric

When making transphobic statements, TERFs employ regular arguments that delegitimize the status and inclusion of transgender women in the definition of woman. While recent work has aimed to identify explicit slurs used against transgender individuals (Kurrek et al., 2020), the TERF rhetoric is more subtle. However, the high performance of our signal classifier (§4.2) indicates TERF users can be accurately identified when discussing transgender topics. Now, we test whether we can explicitly recognize which statements contain harmful TERF rhetoric. We first create a topically-focused dataset of transgender-related content and label messages

[7] Minimum ngram frequency was set to 50, with limited hyperparameter tuning on the development set showing lower performance for including higher-order ngrams or when using a lower (25) or higher (100) minimum frequency threshold.

[8] We also note that because these labels are derived through public lists, we speculate that some noise may exist due to misunderstanding or even users changing beliefs over time.

by whether they contain a TERF rhetoric, and then use this corpus to train classifiers.

Data and Annotation Data was sampled from the transgender topic (§4.1) from a balanced number of TERF-identified, transgender, and control users. Content labeled with the topic represents an ideal dataset for recognizing TERF language, as it focuses primarily on trans and gender-related discussion (not necessarily TERF-related) and likely contains both TERF arguments and rebuttals to TERF arguments.

The two authors first reviewed hundreds of messages as an open coding exercise to identify salient themes used in TERF arguments. Salient categories included (a) bad-faith arguments, (b) concerns about transgender women competing in women's sports, (c) and biological essentialist exclusion of transgender women; these three themes were sufficient to cover all TERF arguments seen in the reviewed data. Following the construction of the categories, the authors completed two rounds of training annotation where each independently labeled 50 tweets and then discussed all labels. Comments were labeled as either (i) not TERF-related or (ii) having any of the three different categories of TERF rhetoric.

Annotators completed 580 items and attained a Krippendorff's α of 0.53, reflecting moderate agreement. Disagreements often stemmed from the difficulty of interpreting the intention of the message. For example, the tweet "Gender is a form of oppression, which only serves the patriarchy" could be viewed through the lens of TERF rhetoric that defines gender fully as a biological construct; alternatively, such a message could be promoting gender fluidity and the rejection of hegemonic norms of gender, which is not a TERF argument. Other disagreements were due to ambiguity around sarcasm or whether the perceived attack on women was related to transgender issues. Disagreements were adjudicated and ultimately 34.4% of the instances were labeled as transphobic arguments in the final dataset.

Experimental Setup Our task mirrors analogous work on stance detection, which aims to identify a user's latent beliefs towards some entity, which may or may not be present in the message. Recent work has shown that pretrained language models are state of the art for stance detection (Samih and Darwish, 2021), so we test one such model here.

Model	AUC	Prec.	Rec.	F1
Random	0.50	0.23	0.54	0.32
Perspective API	0.52	0.45	0.43	0.44
Logistic Regression	0.63	0.17	0.08	0.11
RoBERTa	**0.76**	**0.67**	**0.70**	**0.68**

Table 3: Performance on recognizing TERF rhetoric.

Data was split into train, development, and test sets using an 80:10:10 percent random partitioning. We test two models: a RoBERTa model (Liu et al., 2019) initialized with the `roberta-base` parameters and a Logistic Regression model. The RoBERTa model was fine-tuned using AdamW with ϵ=1e-8 and η=4e-5 and a batch size of 32; the model was fine-tuned over 10 epochs, selecting the epoch that performed highest on the development data (#6). The logistic regression model used unigram and bigrams with no minimum token frequency due to the dataset size. We compare these against a uniform random baseline and a competitive baseline of a commercial model for recognizing toxic language, Perspective API using 0.5 as a cut-off for determining toxicity.

Results The RoBERTa model was effective at recognizing the rhetoric of tweets, attaining an F1 of 0.68 (Table 3), which is slightly above inter-annotator agreement. This performance suggests that the model is near the upper bound for performance in the current data (due to IAA) and that TERF rhetoric can be easily recognized by deep neural models. In contrast, the simple lexical baseline performed poorly and, surprisingly, below chance. When viewed in contrast to a similar baseline for recognizing TERF users in §4.3, this low performance suggests that simple lexical features alone are insufficient for recognizing TERF rhetoric specifically due to their nuance, even if they may be useful for identifying TERF users themselves or identifying other kinds of more explicit hate speech (e.g., Waseem and Hovy, 2016). The competitive baseline of Perspective API was not able to recognize the subtle offensive language of TERF rhetoric, though it does surpass chance; as Perspective API is widely deployed, this result suggests TERF rhetoric is unlikely to be flagged for review.

The RoBERTa model was robust to hard cases such as paraphrased TERF arguments by non-TERF as a rebuttal to strong rhetoric, which included the language of the rhetoric itself. Examining the error shows that the model struggled with cases where

Label	Pred.	Tweet
TERF	NOT	Definitive signs of an unbearable human: using queer as an umbrella category. That's it.
TERF	NOT	The ease with which women's rights can be sidelined by the government underscores the vulnerability of those rights: we can't take anything for granted
NOT	TERF	Talking about gender "incongruence" as well as dysphoria is never limited to the body of the trans-identified person. They describe misery within their gender roles. Men are tired of demands for invulnerability while women want to be looked in the eye and spoken to like adults.
NOT	TERF	How do you know for sure Yaniv isn't trans? How does anyone tell whether someone is a "genuine" trans identifying male and a predator?

Table 4: Examples of misclassifications by the model for recognizing TERF rhetoric show false negatives from subtle arguments (top two) and false positives likely-innocuous questions (bottom two).

the interpretation of the message could be ambiguous. Table 4 shows a sample of four misclassifications; the first two false negatives highlight subtle arguments that the model misses, while the last two suggest the model is overweighting arguments that could appear to be made in bad faith. Overall, the moderately-high performance suggests that TERF rhetoric can be recognized but represents a challenging NLP task if deployed solely in a manner designed to censure such content.

6 Values and Design Considerations

The computational tools developed in this paper in §4 and §5 facilitate the detection of TERFs and their rhetoric. To what end should these tools be used? The majority of antisocial or toxic language detectors are used punitively for censure or removal—uses of toxic speech are removed from public visibility and the transgressing individuals are potentially subject to temporary suspensions or even account removals. Given that at their core, many TERFs are feminists who are primarily concerned with women's rights and safety (albeit mistakenly latching onto a biological essentialist definition of "women"), we view the application and deployment of our tools as an ideal ethical case study for alternatives to the traditional punitive uses of abusive language detection. As NLP moves from focusing on the language of bad actors to examining nuanced discourse in a gray area, we must rethink how our methods are deployed and what the ultimate goals of such tools are: reconciliation and rehabilitation, or potential radicalization through alienation.

Due to the political nature of a TERF detector, it is worth critically examining such work through contemporary lenses of "cancel culture" (Bouvier, 2020) and restorative justice (Braithwaite, 2002). This work intends to provide a useful tool allowing marginalized people in the trans community to curate their online experiences and avoid doxxing and harassment at the hands of TERFs. However, examining its impact could raise concerns of censorship or evoke the echo chambers of algorithmically-constructed Facebook feeds—which we explicitly acknowledge and seek to avoid.

"Cancel culture" is a contemporary form of ostracism that straddles online and real-world spheres and often leads to material loss for the "cancelled" (Bouvier, 2020). The phenomenon is largely punitive and, combined with other forms of online censorship such as deplatforming, generates further polarization; it pushes people away to be radicalized in remote spaces. Online moderation tools have typically relied on these types of actions to remove content (Srinivasan et al., 2019). While community-level bans have been effective at reducing harm without creating spill-over into other communities (Chandrasekharan et al., 2017), such actions still run the risk of removing the possibility of further engagement that leads to a change in underlying views. Thus, we do not label people as TERFs in order to silence or "cancel" them. Rather, we consider it a tool to better engage, understand, and ultimately find a path to reconciliation.

We reiterate that the methods outlined in this paper should *not* supersede human judgment, but rather be used in tandem to best inform the user. It is worth being cautious of the fact that people take AI models to be objective arbiters when in reality, they can and do embed bias in many facets (e.g., Sap et al., 2019; Ghosh et al., 2021). Such a system should not be viewed as the end-all-be-all in decision-making.

The ideal use-case of TERF detection should be grounded within a framework of restorative justice (Schoenebeck and Blackwell, 2021); instead of punitive retribution, we seek rehabilitation through mutual engagement, dialogue, and consensus. Users should be able to decide how to engage upon encountering a TERF guided by an assessment of TERFs stance (e.g., transphobic severity)

and whether they are equipped and able to put in the labor of understanding and addressing their fears.

As potential next steps for deploying our models in a manner to minimize risk, Kwon et al. (2018) and Im et al. (2020) have proposed visual mechanisms for displaying "social signals" of other individuals on social media to create an informed decision about potential interactions; our tool could easily lend itself to such mechanisms by identifying users by their likelihood of being a TERF and also, if the user is willing, to show content our model has identified as being TERF rhetoric to assess their stance. While promoting interactions between the transgender community and TERFs poses risks, we retain some optimism for establishing shared common ground to facilitate dialogue. Indeed, as our topic model showed, the bulk of TERF users' message is *not* about transgender issues and much of this content overlaps with that written by transgender women; for those willing to engage, new NLP methods could be used to (i) identify particular non-confrontational topics to foster an initial dialogue, (ii) suggest potential counterspeech, building upon recent work on counterspeech for hate speech (Garland et al., 2020; Mathew et al., 2019; Chung et al., 2019; He et al., 2021), and (iii) analyze their statements to identify those TERFs whose stances signal they could be open to change (Mensah et al., 2019).

7 Conclusion

Online communities serve essential roles as places of support and information. For transgender individuals, these spaces are especially critical as they provide access to accepting and supportive communities, which may not be available locally. However, the public forums of social media can also harbor less than welcoming users. Trans-exclusionary radical feminists (TERFs) promote a harmful rhetoric that rejects transgender women as women, pushes an agenda that reduces gender to biology, and seeks to invalidate transgender women in policy and practice. As a result, transgender individuals and their allies have adopted technological solutions to limit interactions with TERFs by manually curating block lists, which require frequent updating and currently rely only on self-reporting to recognize those users who pose harm.

This paper introduces new datasets and models for supporting the trans community through automatically identifying TERF users and their rhetoric. We present a new multi-stage model that identifies salient themes in TERF users' content and show that these signals, when combined with social network features, result in a highly accurate classifier (0.93 F1) that reliably identifies TERF users with minimal risk of mistakenly labeling trans-friendly users as TERFs, despite sharing similar content themes. Further, we introduce a new dataset for directly identifying the often-subtle rhetoric of TERFs and show that despite the challenging task, our model can attain moderately high performance (0.68 F1). Together, these two tools can aid the trans community in mitigating harm through preemptive identification of TERFs. All data, code, models, and annotation guidelines will be available at `https://github.com/lu-christina/terfspot`.

Acknowledgments

We thank the members of the Blablablab for their helpful thoughts and comments as well as the WOAH reviewers for their thoughtful critiques—with a special shout out to R3 for an exceptionally helpful and detailed review. Finally, we also thank the work of the trans women and activists who have curated the initial TERFblocklist and their work in helping keep the community safe.

8 Ethics

Data Privacy Our data includes lists of Twitter users who belong to marginalized categories, notably transgender individuals. This data is obtained from entirely public sources of Twitter lists and is not directly maintained by the research team. While we are not able to minimize the privacy implications of this public data, the research team took additional steps to maintain the privacy of the data on our servers. Further, this data will only be shared further to researchers who agree to ensure future privacy and use the data in ethical ways.

Using TERF as a term The TERF acronym has been considered by some to be a derogatory term directed at a group of people and some have called for the term not to be used (e.g., Flaherty, 2018). While recognizing these views, we opt to follow common scholarly practice and use the term. However, we took additional precautions when writing to ensure that the framing of such users was from a neutral point of view.

Do we *need* to predict TERF users? Labeling a user as a TERF is a potentially risky act. Misclassifications could lead to being socially ostracised

by peers and increased mistrust. However, this risk is offset, in part, by the risk of *not* developing such technology. Transgender individuals actively and manually identify TERF users to minimize their interactions with such toxic content. However this identification is labor intensive and (i) exposes users to TERF content, increasing harm and (ii) is likely to miss some users due to the scale of finding TERF users on social media. As a result, inaction increases the harm to transgender users. Recognizing this trade-off, we have performed additional analyses to minimize the risk of false positive classifications of users as a TERF, showing that our model has a low false positive rate (§4.3).

Who should be on a block list? Our models are trained on community-curated block lists, with a goal of helping individuals identify others who might be engaged in harmful TERF rhetoric. Yet, it is worth considering whether such actions potentially perpetuate harm by minimizing discourse, increasing polarization, or even serving as a "marker of success" for antagonistic users to aim for. We explicitly do not advocate automatically including any user on a block list and, instead, as outlined in §6, argue for more nuance and consideration in how users apply this technology. We view an ideal application of our model as one that allows each person to define their own comfort level in exposure and engagement in an informed manner. Our tool can serve as a social signal to help others guide their decision but should not be taken as ground truth for blocking anyone.

Dual-use Risks Many NLP methods, including those presented here, have dual-use for good and bad purposes. Our models could be used to deployed to identify and "cancel" TERF users, cutting them off from the larger social media community. Further, TERF users could use our models adversarially to test how their own accounts are classified and systematically change their behavior to avoid future detection. Yet, in our setting, the technology offers substantial benefits for a marginalized group, transgender individuals, who have been overlooked by NLP methods for identifying transgender-targeted content. Our models augment their ability to identify TERF users and use this knowledge as they see fit. Given the harm faced by transgender individuals, we view the benefits as substantially outweighing risks.

References

Ehab A Abozinadah and James H Jones Jr. 2017. A statistical learning approach to detect abusive twitter accounts. In *Proceedings of the International Conference on Compute and Data Analysis*, pages 6–13.

Faiyaz Al Zamal, Wendy Liu, and Derek Ruths. 2012. Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In *Proceedings of the International Conference on Web and Social Media (ICWSM)*.

Pablo Barberá, John T. Jost, Jonathan Nagler, Joshua A. Tucker, and Richard Bonneau. 2015. Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber? *Psychological Science*, 26(10):1531–1542.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Gwen Bouvier. 2020. Racist call-outs and cancel culture on twitter: The limitations of the platform's ability to define issues of social justice. *Discourse, Context & Media*, 38:100431.

John Braithwaite. 2002. *Restorative justice & responsive regulation*. Oxford University press on demand.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, Inga Kartoziya, and Michael Granitzer. 2020. I feel offended, don't be abusive! implicit/explicit messages in offensive and abusive language. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6193–6202, Marseille, France. European Language Resources Association.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):1–22.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate

speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Shelley L Craig, Andrew D Eaton, Lauren B McInroy, Sandra A D'Souza, Sreedevi Krishnan, Gordon A Wells, Lloyd Twum-Siaw, and Vivian WY Leung. 2020. Navigating negativity: a grounded theory and integrative mixed methods investigation of how sexual and gender minority youth cope with negative comments online. *Psychology & Sexuality*, 11(3):161–179.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Earles. 2019. The "penis police": Lesbian and feminist spaces, trans women, and the maintenance of the sex/gender/sexuality system. *Journal of lesbian studies*, 23(2):243–256.

Stefano Faralli, Giovanni Stilo, and Paola Velardi. 2015. Large scale homophily analysis in twitter using a twixonomy. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 2334–2340. AAAI Press.

Marty Fink and Quinn Miller. 2014. Trans media moments: Tumblr, 2011–2013. *Television & New Media*, 15(7):611–626.

Colleen Flaherty. 2018. "TERF" War. *Inside Higher Ed*.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):85.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794, Marseille, France. European Language Resources Association.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Sayan Ghosh, Dylan Baker, David Jurgens, and Vinodkumar Prabhakaran. 2021. Detecting cross-geographic biases in toxicity modeling on social media. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 313–328, Online. Association for Computational Linguistics.

Oliver L Haimson, Justin Buss, Zu Weinger, Denny L Starks, Dykee Gorrell, and Briar Sweetbriar Baron. 2020. Trans time: Safety, privacy, and content warnings on a transgender-specific social media site. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–27.

Oliver L Haimson and Tiffany C Veinot. 2020. Coming out to doctors, coming out to "everyone": Understanding the average sequence of transgender identity disclosures using social media data. *Transgender health*, 5(3):158–165.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against disguised toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2021. Racism is a virus: anti-asian hate and counterspeech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 90–94.

Sally Hines. 2019. The feminist frontier: On trans and feminism. *Journal of Gender Studies*, 28(2):145–157.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in Neural Information Processing Systems*, 34.

Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu, HI, USA, April 25-30, 2020*, pages 1–12. ACM.

Tristen Kade. 2021. "hey, by the way, i'm transgender": Transgender disclosures as coming out stories in social contexts among trans men. *Socius*, 7:23780231211039389.

Terri M Kelleher. 2020. Jk rowling: Guilty, of crime of stating that sex is determined by biology. *News Weekly*, (3072):10.

Dongwoo Kim, Yohan Jo, Il-Chul Moon, and Alice Oh. 2010. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *ACM CHI workshop on microblogging*, volume 6. Citeseer.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Saebom Kwon, Puhe Liang, Sonali Tandon, Jacob Berman, Pai-ju Chang, and Eric Gilbert. 2018. Tweety holmes: A browser extension for abusive twitter profile detection. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 17–20.

Sophie Lewis. 2019. How british feminism became anti-trans. *The New York Times*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Veronica Lynn, Salvatore Giorgi, Niranjan Balasubramanian, and H. Andrew Schwartz. 2019. Tweet classification without the tweet: An empirical examination of user versus document attributes. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 18–28, Minneapolis, Minnesota. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 369–380.

Humphrey Mensah, Lu Xiao, and Sucheta Soundarajan. 2019. Characterizing susceptible users on reddit's changemyview. In *Proceedings of the 10th International Conference on Social Media and Society*, pages 102–107.

Ruth Pearce, Sonja Erikainen, and Ben Vincent. 2020. Terf wars: An introduction. *The Sociological Review*, 68(4):677–698.

Carla Perez Almendros, Luis Espinosa Anke, and Steven Schockaert. 2020. Don't patronize me! an annotated dataset with patronizing and condescending language towards vulnerable communities. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5891–5902, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anthony T Pinter, Morgan Klaus Scheuerman, and Jed R Brubaker. 2021. Entering doors, evading traps: Benefits and risks of visibility during transgender coming outs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–27.

Jipeng Qiang, Qian Zhenyu, Yun Li, Yunhao Yuan, and Xindong Wu. 2019. Short text topic modeling techniques, applications, and performance: A survey. *ArXiv preprint*, abs/1904.07695.

Janice G Raymond. 1979. *The Transsexual Empire the Making of the She-Male*. Beacon Press (Ma).

Carol Riddell. 2006. *Divided sisterhood: a critical review of Janice Raymond's*. Routledge London and New York.

Younes Samih and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646, Online. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sarita Schoenebeck and Lindsay Blackwell. 2021. Reimagining social media governance: Harm, accountability, and repair. *Yale Journal of Law and Technology*, 23(1). Justice Collaboratory Special Issue.

Julia Serano. 2016. *Whipping girl: A transsexual woman on sexism and the scapegoating of femininity*. Hachette UK.

Kumar Bhargav Srinivasan, Cristian Danescu-Niculescu-Mizil, Lillian Lee, and Chenhao Tan. 2019. Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–21.

Derald Wing Sue. 2010. *Microaggressions in everyday life: Race, gender, and sexual orientation*. John Wiley & Sons.

Teresa Tadvick. 2018. Practicing gender in online spaces. Bachelor's thesis, University of Colorado, Boulder.

Emily Vajjala. 2020. *Gender-critical/Genderless? A Critical Discourse Analysis of Trans-Exclusionary Radical Feminism (TERF) in Feminist Current*. Ph.D. thesis, Southern Illinois University, Carbondale.

Zijian Wang and Christopher Potts. 2019. TalkDown: A corpus for condescension detection in context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3711–3719, Hong Kong, China. Association for Computational Linguistics.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Cristan Williams. 2016. Radical inclusion: Recounting the trans inclusive history of radical feminism. *Transgender Studies Quarterly*, 3(1-2):254–258.

Cristan Williams. 2020. The ontological woman: A history of deauthentication, dehumanization, and violence. *The Sociological Review*, 68(4):718–734.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.