Using Sentiment Induction to Understand Variation in Gendered Online Communities

Li Lucy
Symbolic Systems Program
Department of Computer Science
Stanford University
lucy3@stanford.edu

Julia Mendelsohn
Department of Linguistics
Department of Computer Science
Stanford University
jmendels@stanford.edu

Abstract

We analyze gendered communities defined in three different ways: text, users, and sentiment. Differences across these representations reveal facets of communities' distinctive identities, such as social group, topic, and attitudes. Two communities may have high text similarity but not user similarity or vice versa, and word usage also does not vary according to a clearcut, binary perspective of gender. Community-specific sentiment lexicons demonstrate that sentiment can be a useful indicator of words' social meaning and community values, especially in the context of discussion content and user demographics. Our results show that social platforms such as Reddit are active settings for different constructions of gender.

1 Introduction

Social groups can be described by many factors, such as the demographics of its participants or its physical location. To detect sociolinguistically significant groups, linguists have built upon the concept of *communities of practice*, which are characterized by their participants' shared actions, beliefs, values, and language styles (Eckert and McConnell-Ginet, 1992; Eckert, 2006). This concept initially emerged to understand the complex interplay between language and gender, and has been applied to study social identities in numerous communities (e.g. Eckert, 1989; Mendoza-Denton, 1996; Hall, 2009). While previous variationist work has primarily studied traditional physical communities, we focus instead on online ones.

Online communities have been shown to form collective linguistic norms, which give rise to a rich amount of language variation across communities, even on the same website (Danescu-Niculescu-Mizil et al., 2013; Yang and Eisenstein, 2015). One of the largest content aggregator and discussion platforms, Reddit, contains thousands of unique communities, known as *subreddits*. These subreddits vary in topic, such as r/sport and r/history, content type, such as r/pics and r/videos, and format, such as the Q&A style of r/IamA and the narratives on r/confession. Online communities such as subreddits are often characterized by language use and user membership (Hamilton et al., 2016; Datta et al., 2017; Bamman et al., 2014; Martin, 2017).

Sociolinguists have primarily analyzed phonological and syntactic variables (Eckert, 2012), though some have studied lexical variables (e.g. Wong, 2005). Previous computational work also focuses on lexical variation (Bamman et al., 2014). We approach variation from a new direction, where we examine the salient semantic dimension of sentiment to understand *how* users use the same words to convey different meanings. We also create representations for subreddits that encode text and user membership to situate insights gained from sentiment-based representations and to understand the intersection of speaker identity (user), content (text), and affect (sentiment). This paper focuses on explicitly gendered subreddits, which cater towards masculine- or feminine-identifying groups.

We provide two main contributions:

1) Salient aspects of social group identities, such as gender, can produce low user overlap in communities sharing similar topics.

2) Sentiment-based representations of communities can reveal a type of variation across social groups that word-choice alone cannot. An in-depth study of words' sentiments in gendered subreddits reveals patterns of how linguistic resources construct a wide array of gendered identities in the online sphere.

Proceedings of the Society for Computation in Linguistics (SCiL) 2019, pages 156-166.
New York City, New York, January 3-6, 2019

2 Previous Work

The study of online communities is highly interdisciplinary, spanning machine learning, natural language processing, social network analysis, communications, and sociolinguistics. Twitter, online news, and other websites have been a rich source of data for computational social scientists, and Reddit in particular has been of interest to much previous work (Althoff et al., 2014; Kumar et al., 2018; Newell et al., 2016; Jaech et al., 2015; Hamilton et al., 2017). Research in this area expands beyond quantitative measures, by comparing results with social science theories and employing qualitative thinking to highlight trends of individual words and communities (Bamman et al., 2014; Danescu-Niculescu-Mizil et al., 2013; Zhang et al., 2017; Althoff et al., 2014).

In particular, the characterization and identification of gender is a common use case for online data. Language is often used to infer demographics of users, especially in classification tasks that tend to find clear distinctions between men and women (e.g. Burger et al., 2011; Argamon et al., 2007; Schler, 2006; Rao et al., 2010). Previous work have found strong patterns of gender differences in language (Newman et al., 2008; Mulac et al., 2001). For example, Volkova et al. (2013) showed that the sentiment of words, hashtags, and emoticons vary between men and women on Twitter and used gender-dependent features to improve sentiment classification of tweets. Our work aims to look beyond a straightforward divide between men and women, particularly because gender is not a fixed biological variable, but rather a dynamic, social one that is actively created and reinforced through repeated behaviors (Butler, 1988; Nguyen et al., 2014; Herring and Paolillo, 2006).

Language is used to simultaneously co-construct multiple identities, so the plethora of gendered identities that emerge from communities of practice may substantially differ from mainstream stereotypes of "femininity" and "masculinity" (Eckert, 1989; Mendoza-Denton, 1996; Hall, 2009). On Twitter, Bamman et al. (2014) investigated cross-community lexical variation and the varied ways of constructing gender identities. They clustered users based on bag-of-words representations of their posts, and the resulting clusters corresponded not only to topical interest but also gender. Some clusters had language patterns that were orthogonal to expected language differences between men and women, demonstrating diversity in gendered language styles.

We compare text, user, and sentiment representations of communities by their predictions of similarity and identify cases where these predictions agree or disagree. Pavalanathan et al. (2017) suggested that subreddits with similar topics can have dissimilar user groups due to differences in preferred interactional styles. Datta et al. (2017) introduced a method for finding misalignments of inferred user-based and text-based networks on Reddit. They found that pairs with high text but low user similarity tend to be communities that conflict (such as political subreddits) as well as communities with hierarchical relationships (such as a niche subreddit with a more generic one). High user but low text similarity suggested a single overarching community scattered across multiple subreddits. We assessed Datta et al. (2017)'s z^2-score method in section 5.1, but found that its results hid important misalignments.

We use domain-specific lexicon induction techniques for creating sentiment-based subreddit representations. Previous work has built or adapted word embeddings or scores to flexibly encode semantic dimensions or specific domains, and some have applied their techniques to social media communities (Rothe et al., 2016; Yang and Eisenstein, 2015; Hamilton et al., 2016). Rothe et al. (2016)'s DENSIFIER model involves dense word embeddings created by mapping generic word embeddings into meaningful subspaces. These learned embeddings may even contain a single dimension, which can act as labels for an induced lexicon, and are best applied to corpora with several billion tokens, which is far greater than any subreddit that we study.

While Rothe et al. (2016) learned lexicons for generic domains such as news and Twitter, Hamilton et al. (2016)'s SENTPROP method solves a similar task across fine-grained domains, including Reddit communities. They applied a label propagation method to 250 Reddit communities, and found a wide range of sentiment variation. For example, *insane* is negative in r/twoxchromosomes but positive in r/sports, while *soft* shows the opposite pattern. SENTPROP was the most suitable approach for our purposes due to its ability to operate on smaller datasets. We extend this work by examining how vector representations created from subreddit-specific sen-

Subreddit	Description
r/actuallesbians	"A place for cis and trans lesbians, bisexual girls, chicks who like chicks..."
r/askgaybros	"Where you can ask the manly men for their opinions on various topics."
r/mensrights	"For those who wish to discuss men's rights and the ways said rights are infringed upon."
r/askmen	"A semi-serious place to ask men casual questions about life, career, and more."
r/askwomen	"Dedicated to asking women questions about their thoughts, lives, and experiences."
r/xxfitness	"For women and gender non-binary redditors who are fit, want to be fit..."
r/femalefashionadvice	"A subreddit dedicated to learning about and discussing women's fashion."
r/malefashionadvice	"Making clothing less intimidating and helping you develop your own style."
r/trollxchromosomes	"A subreddit for rage comics and other memes with a girly slant."

Table 1: Gendered communities with descriptions from their sidebars or subreddit search listings.

timent lexicons compare to text-based and user-based representations, with gendered communities as a case study.

3 Data

In order to gain a broad perspective of how gendered subreddits relate to each other and other communities within the larger Reddit context, we consider data from subreddits with 50,000+ subscribers, as provided in a user-curated list[1]. These subreddits span topics ranging from plants to cryptocurrency, and provide a glimpse into Reddit's long tail of diverse niche communities, which is a primary draw to the platform (Newell et al., 2016). The most popular subreddits tend to be part of a set of "default" subreddits to which users have historically been auto-subscribed (Newell et al., 2016; Datta et al., 2017). To focus on more niche and non-artificially inflated communities, we filtered out about 50 default subreddits based on lists in r/defaults created during May 07 2014, May 26 2016, and March 26 2017.

We took the top 400 remaining subreddits and used their comments created between May 2016 and April 2017. The vast majority of these subreddits contain between 10^7 and 10^8 tokens, with r/politics (the largest) containing over 764 million tokens to r/accidentalwesanderson (the smallest) containing over 7 million. From these subreddits we manually selected nine subreddits with clearly gender-oriented names (Table 1).

4 Approach

4.1 Text & User Representations

To provide a basis of comparison for our sentiment-based representations, we created term frequency-inverse document frequency (tf-idf) vectors for each subreddit using user and unigram frequencies. Here, we define subreddit user frequencies as the number of times a user comments to a specific subreddit. For unigram or user t in subreddit d, its tf-idf weighted frequency is

$$w_{t,d} = (1 + \log tf_{t,d}) \log(N/df_t),$$

where $tf_{t,d}$ is the frequency of t in d, df_t is the number of subreddits in which t appears, and N is the total number of subreddits.

We filtered out rare users and bots, with $1 < df_t \leq 380$ for users, and filtered out rare words and stop words, with $5 < df_t \leq 380$ for unigrams. We used truncated singular value decomposition (SVD) to reduce these vectors to 100 dimensions and normalized them to each have a unit norm (Pedregosa et al., 2011; Halko et al., 2011).

Positive	love, loved, loves, awesome, nice, amazing, best, fantastic, correct, happy
Negative	hate, hated, hates, terrible, nasty, awful, worst, horrible, wrong, sad

Table 2: Positive and negative Twitter seed words from Hamilton et al. (2016)

4.2 Sentiment Representations

We induced community-specific sentiment lexicons using the SentProp method introduced by Hamilton et al. (2016). This framework was demonstrated to perform well on moderately sized domains of 10^7 tokens, which matches the majority of our subreddits.

SentProp begins by creating community-specific word embeddings. All comments from a given subreddit were first concatenated into a single document, separated by 5 dummy tokens so adjacent comments did not influence the linguistic contexts of the first and last words. Following Hamilton et al. (2016), word co-occurrence matrices for each subreddit were created with a symmetric context window of 4 words and

[1] Available here.

reweighted using positive pointwise mutual information (PPMI) with context distribution smoothing $c = 0.75$ (Levy et al., 2015). The dimensionality of each word embedding was then reduced to 100 using SVD.

After obtaining subreddit-specific word embeddings, we introduce a small set of seed words with positive and negative polarity. We used the same seed words as Hamilton et al. (2016) did for Twitter, another social media platform (Table 2). SENTPROP runs a series of random walks from both the positive and negative seed words, and the resulting sentiment value for each word is based on the probabilities that the word was hit by the positive random walk versus the negative one. We used SENTPROP's default parameters for the Reddit lexicon induction portion of their paper, setting $\beta = 0.9$ and $K = 25$, where K is the number of nearest neighbors in the semantic space to which edges are drawn in graph construction. A higher β favors similar labels for neighbors and a lower β favors correct labels on seed words. We induced sentiment for the top 5000 words by frequency in each subreddit.

Adjusting the parameters β and K did not change our main conclusions or observations. Lowering β from 0.9 to as far as 0.5 shrinks the overall range of sentiment values from -3 to 3 to about -2 to 2. Words with neutral sentiment tend to be slightly more positive or negative with lower values of β, but words with the highest polarities are consistent. Sentiment scores also remain steady when varying K. With $\beta = 0.9$, the Pearson correlation of sentiment scores between $K = 25$ and $K = 15$ is 0.9183 ($p < 0.001$) and 0.9668 between $K = 25$ and $K = 35$ ($p < 0.001$) for r/xxfitness.

We standardized values for each word to have zero mean and unit variance. The resulting sentiment vectors were then an array of negative and positive values corresponding to sentiment, averaged over 50 bootstrap-sampled runs. Each index in these vectors maps to a word in the vocabulary, which is the union of all subreddits' vocabularies. If a word's sentiment was not induced in a certain subreddit, its sentiment value is set to a neutral zero.

4.3 Metrics

We performed agglomerative clustering on all 400 subreddits' user- and text-based representations to see where gendered subreddits' users and content are situated within Reddit. We fixed the number of clusters to 20 and compared cluster sets provided by different representations by calculating their adjusted mutual information (AMI), where the possible range of values is 0 for random cluster and 1 for identical ones (Vinh et al., 2010). To further compare the different representations, we calculated the Spearman correlation between subreddits' pairwise similarities. We identify misalignments as subreddit pairs that have high similarity for one representation but low similarity for another.

We also implemented the misalignment identification method proposed by Datta et al. (2017). This method subtracts two pairwise similarity rank matrices created by two different representation types, such as text and user, and z-score normalizes the difference matrix's columns and rows. Values in the final misalignment matrix are called z^2-scores. Pairs of subreddits with a high positive z^2-score have a higher similarity than expected with the first representation compared to the second, while a large negative z^2-score signifies the opposite.

5 Analysis

5.1 Text & Users

The clusterings of user- and text-based representations are similar, with an AMI of 0.5610, and the text-based clusterings are more topically coherent. For example, r/femalefashionadvice and r/malefashionadvice are in the same text-based cluster but different user-based clusters, since both are about fashion but cater towards different genders. Subsets of the clusters in which gendered subreddits appear can be found in Table 3. The feminine subreddits are all in the same user-based cluster, while the masculine ones are more scattered. Female Reddit users may find themselves pushed into this cluster because of an overall predominant masculine culture throughout the platform which can be hostile to women (Massanari, 2017). A majority of the gendered subreddits occur in the text-based cluster containing those related to personal topics, such as families and relationships. These clusters situate our gendered communities based on *what* they talk about and *who* is talking, and changing the representation type alters the perceived geography of Reddit.

The Spearman correlation between text and

User-based Clusters	Text-based Clusters
femalefashionadvice askwomen xxfitness trollxchromosomes actuallesbians weddingplanning makeupaddiction justnomil skincareaddiction raisedbynarcissists dogs childfree vegan parenting running teachers unresolvedmysteries	**askwomen actuallesbians askmen askgaybros trollxchromosomes** suicidewatch justnomil deadbedrooms babybumps seduction raisedbynarcissists dogs childfree casualiama legaladvice parenting foreveralone dating_advice teachers polyamory
mensrights sandersforpresident changemyview neutralpolitics forwardsfromgrandma the_donald anarchism economics atheism subredditdrama	**mensrights** thathappened teenagers forwardsfromgrandma niceguys blackpeopletwitter roastme trashy facepalm photoshopbattles outoftheloop 4chan cringe
askgaybros askmen suicidewatch teenagers sex bodybuilding depression seduction offmychest ama advice foreveralone dating_advice tinder polyamory	**xxfitness** fatlogic loseit keto cooking vegan running bodybuilding
malefashionadvice houston cooking churning entrepreneur seattle financialindependence investing travel homeimprovement jobs photography homebrewing bicycling personalfinance	**femalefashionadvice malefashionadvice** weddingplanning makeupaddiction skincareaddiction sneakers streetwear asianbeauty fashionreps

Table 3: The gendered subreddits and a subset of the subreddits that occur in the same clusters as them.

user vectors' pairwise similarities is 0.549 ($p < 0.0001$), plotted in Figure 1. The vast majority of the pairs with high user and low text similarity are those pertaining to European countries, where comments are in different languages. The pairs with high text and low user similarity include subreddits pertaining to cities, as well as those divided by gender, sexual orientation, or personal topics. Thus, user demographics are important motivators for community formation on Reddit. User-based similarities for gendered subreddits tend to vary based on whether they cater towards the same gender, though some subreddits act as bridges between them: r/askmen and r/askwomen have a user similarity of 0.417, which is above the average among gendered subreddits (0.2880). The structure of these communities facilitates this by encouraging questions posted by users of any gender, and follow-up dialogue across groups accompanies the targeted gender's answers in the comments.

The misalignments identified by raw similarities are more intuitive than those identified by z^2-scores. Pairs in the top 20 with high text and low user similarity based on z^2-scores included understandable ones such as r/casualconversation-r/tinder, but also many pairs on very different topics such as r/bodybuilding-r/learningprogramming and r/makeupaddiction-r/legaladvice. Pairs in the top 20 with low text and high user similarity were easier to interpret and included r/truegaming-r/askmen and r/vegan-r/askscience. The z^2 score method normalizes any subreddit's skewed distribution of similarities. For example, country subreddits in general have lower text similarity to other subreddits since Reddit is mostly in English, and normalization causes inter-country similarities to match their expected similarity differences. Thus, normalization steps may reduce meaningful variation in how some subreddit similarities deviate from the norm of Reddit as a whole.

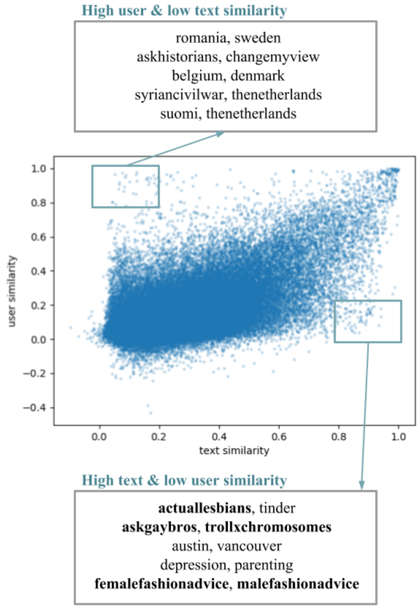

Figure 1: Subreddit similarities, with text similarities on the horizontal axis and user similarities on the vertical axis. Each point represents a pair of subreddits. The listed subreddit pairs are examples of outliers, where one type of similarity is very small (< 0.2) and another is large (> 0.8).

5.2 Sentiment

We calculated the pairwise cosine similarities between each of the nine explicitly gendered sub-

Highest	Similarity	Lowest	Similarity
askmen, askwomen	0.6702	femalefashionadvice, mensrights	0.1802
askgaybros, askmen	0.6144	askwomen, malefashionadvice	0.1876
askwomen, trollxchromosomes	0.6003	malefashionadvice, trollxchromosomes	0.2162
actuallesbians, trollxchromosomes	0.5462	malefashionadvice, mensrights	0.2170
askgaybros, askwomen	0.5310	mensrights, xxfitness	0.2181

Table 4: Highest and lowest sentiment similarities between gendered subreddits, with a mean of 0.3701.

	Spearman ρ	p
text, user	0.4268	< 0.01
text, sentiment	0.6371	< 0.0001
user, sentiment	0.4219	$= 0.01$

Table 5: Correlations of text, user, and sentiment representations using pairwise cosine similarity between nine gendered subreddits. Though sentiment and text are related, they provide different information.

reddits using text, user, and sentiment community representations. Table 5 shows the resulting Spearman correlation between these representations. Sentiment representations correlate more strongly with those of text, which could be explained by how they are both linguistically motivated. However, this correlation is far from 1, suggesting that sentiment representations capture some aspects of communities that weighted word counts do not.

The pairs of subreddits with highest and lowest sentiment similarity can be found in Table 4. Some of the highest similarities are between subreddits oriented towards the same gender, and while some of the lowest are between those of different genders, but there are several exceptions. Therefore, sentiment does not divide itself evenly based on gender. The high sentiment similarity between r/askmen and r/askwomen misaligns with their low text similarity (0.2874) and near average user similarity (0.4168). Another outlier across text, user, and sentiment similarities is r/actuallesbians and r/trollxchromosomes, which have high user similarity (0.8856), above average sentiment similarity (0.5462), and high text similarity (0.9415).

The most positive and negative non-seed words in each subreddit are consistent with the concepts we expect to be relevant to them (some examples[2] in Table 6.). Many negative words in r/trollxchromosomes and r/femalefashionadvice revolve around pain and health, while those in r/mensrights refer to gender bias (within the top fifteen most negative words are *misandrist* and *manhating*). The most positive words in r/xxfitness are similar to those in other subreddits since many are adjectives such as *great* and *fun*, but its most negative words almost entirely focus on physical ailments, such as *flu*, *infection* and *headache*. The words *brothers* and *brother* have highest sentiment in r/mensrights compared to other gendered subreddits, suggesting that this community values masculine solidarity. Likewise, even though the words *troll* and *trolls* are predominantly negative, they have high positive sentiment in r/trollxchromosomes, as users in this community have re-appropriated these terms to refer positively to themselves.[3]

Subreddits with high text similarity such as r/malefashionadvice and r/femalefashionadvice still contain distinct cultures. The highly positive words in r/femalefashionadvice reflect their custom of referencing daily outfits using days of the week, while users on r/malefashionadvice do not follow this format. The expressive elongation in r/femalefashionadvice's highly positive *loooove* has previously been shown to be a female marker (Bamman et al., 2014; Rao et al., 2010). Sentiment is a helpful but sometimes superficial metric for determining community values, and its interpretation is best understood in context with topic and users. For example, *men* is most negative in r/mensrights compared to other gendered subreddits, but that does not mean these users dislike men, since the opposite is actually the case. The strong negativity here is instead associated with how their discussions center on injustices towards men.

Table 7 shows the words with the highest variance in sentiment with the subreddits in which they have the most positive and negative polarity. Much of the cross-community variation in

[2] The token *lt3* is a punctuation-stripped HTML heart. Similarly, *d* is often a happy emoji *:D*.

[3] The full lexicons can be found in our Github repo here.

TrollXChromosomes		FemaleFashionAdvice		MaleFashionAdvice		MensRights	
Positive	Negative	Positive	Negative	Positive	Negative	Positive	Negative
lovely	infections	gorgeous	gross	sweet	gross	wonderful	vile
gorgeous	yeast	adore	marks	beautiful	annoying	favorite	evil
beautiful	pain	lovely	blood	cool	dirty	excellent	disgusting
wonderful	dealing	stunning	rough	those	stupid	watch	horribly
congratulations	mess	thursdays	painful	vouch	armpits	enjoyed	cruel
fabulous	horrific	mondays	worse	plus	shitty	honey	hating
congrats	painful	tuesdays	messed	dig	crap	clip	misogynistic
yay	minor	killer	causing	perfect	garbage	enjoying	hateful
d	uti	loooove	horribly	makes	crappy	fan	sexist
lt3	infection	fabulous	poor	interesting	sweaty	episode	bigots

Table 6: Most positive and negative non-seed words for a selected set of subreddits.

Word	Variance	Positive Subreddits	Negative Subreddits
sounds	2.775	femalefashionadvice, actuallesbians	askgaybros, askmen
smells	2.652	actuallesbians, malefashionadvice	askgaybros, askmen
hilarious	2.547	askwomen, actuallesbians	askgaybros, malefashionadvice
absolutely	2.231	femalefashionadvice, malefashionadvice	askwomen, askmen
obsessed	2.094	askwomen, femalefashionadvice	mensrights, askmen
sharp	2.087	actuallesbians, femalefashionadvice	xxfitness, trollxchromosomes

Table 7: Words with greatest variance in sentiment across all gendered subreddits, along with the subreddits in which they are the most positive and most negative.

sentiment is likely due to polysemy, where different senses of a given word are predominant in different communities. For example, when calculating the sentiment of *sick* (13th highest variance) in each subreddit's semantic space, SENTPROP encounters neighbors such as *nauseous* (r/xxfitness), *disgusting* (r/mensrights), or *dope* (r/malefashionadvice).

> *i never thought id see sitting on a tricycle look so badass looks* ***sick*** *love it*[4] (r/malefashionadvice)
>
> *...what kind of* ***sick*** *twisted person could do that to another human being truly disgusting what some people are willing to do to others* (r/mensrights)

Furthermore, words may adopt the sentiment polarities that reflect the overall discourse style. In r/femalefashionadvice, seemingly negative words such as *jealous* (10th highest variance) take on a positive meaning as they are used to compliment the original poster:

> *that is amazing congratulations i am so* ***jealous*** (r/femalefashionadvice)

This relationship with overall discourse style may be even more pronounced for judgment-related words such as *sounds*, whose polarity reflects whether communities tend to use it to evaluate other users or entities positively or negatively.

> *...that* ***sounds*** *like a really polite and productive way to deal with the gift issues* (r/femalefashionadvice)
>
> ***sounds*** *like you just have shitty friends* (r/askgaybros)

Finally, many affective differences emerge depending on whether a subreddit's users talk about their own feelings, beliefs, and passions (personal) or make claims about other people's mental states (impersonal). In particular, sentiment of words such as *jealous* or *obsessed* varies depending on whether one uses it to describe themselves or somebody else (Figure 2).

> *...recently became* ***obsessed*** *with this podcast this is super cool* (r/trollxchromosomes)
>
> *so sad that so many female teachers are feminist* ***obsessed***... (r/mensrights)

Volkova et al. (2013) studied sentiment variation using Twitter data, but treated gender as a binary variable of male versus female. They listed examples with large gender differences: *weakness* is used positively by women and negatively by men, while *overdressed* is used positively by men and negatively by women. Their most polarized words, with hashtags removed, do not split according to a binary in our subreddits. We observe diverse language styles: *weakness* is strongly negative in r/xxfitness (-2.0634 ± 0.5577) but positive in r/actuallesbians (0.9112 ± 0.2100), and the sentiment score of *overdressed* is similar in r/malefashionadvice (-0.6016 ± 0.5273) and r/femalefashionadvice (-0.6696 ± 0.8512). This implies that though gender can be a helpful variable for improving sentiment analysis, its expression is not fixed across multiple contexts.

The word *omg* seems to be used in far broader

[4] Examples transcribed as they appear after preprocessing.

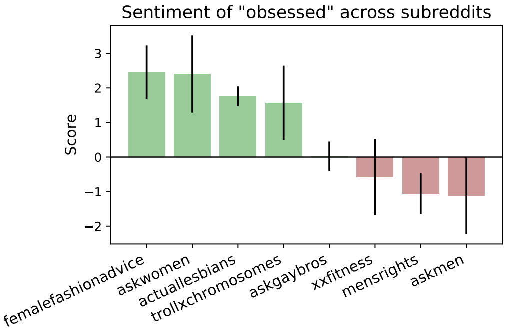

Figure 2: Words with high variance across communities such as *obsessed* can be strongly positive or negative.

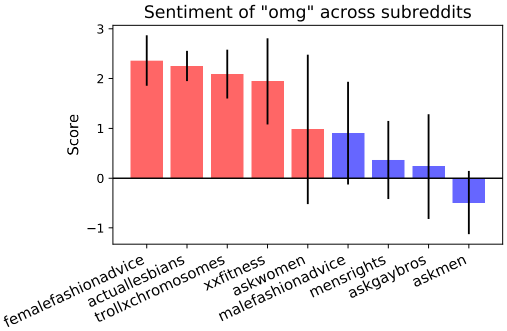

Figure 3: Average sentiment scores and standard deviations of "omg" in explicitly-gendered subreddits, sorted in decreasing order, based on 50 SENTPROP runs. Women-oriented subreddits are marked in red, and men-oriented are in blue.

contexts and have a substantially different meaning than its origin phrase, *oh my god*. In Figure 3, the sentiment of *omg* is highest among the five women-oriented subreddits, and lower in the men-oriented subreddits. This finding is consistent with prior results that show that forms predominant in computer-mediated communication are more commonly used by and highly associated with women (Bamman et al., 2014; Carpenter et al., 2017). However, these results have implications beyond just associating forms such as *omg* with women. In particular, *omg* is not a filler word devoid of meaning in women-oriented communities. Rather, it conveys highly positive affect and may also indicate cooperativeness and engagement in a conversation. It seems to not play the same role in men-oriented communities, where *omg* is used frequently in indirect quotes of others' speech.

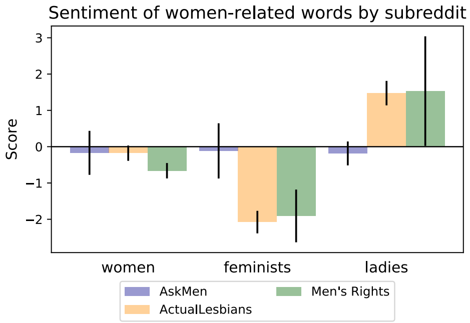

Figure 4: Sentiment scores of the words "women", "feminists", and "ladies" across three subreddits, with error bars showing the standard deviation of 50 bootstrap-sampled SENTPROP runs.

> ***omg*** *love her did you see that shoeprinted dress she had on this weeks episode* (r/femalefashionadvice)

> *...if anything it was usually* ***omg*** *you two would have the most beautiful babies...* (r/askmen)

Sentiment-based representations can also detect words that may be denotationally similar but have different social meanings due to repeated associations with certain beliefs and stereotypes. Figure 4 demonstrates community variation in the affective meaning of the denotationally similar terms *women* and *ladies* and the related but semantically distinct *feminists* across three subreddits. This variation demonstrates the potential for denotationally similar terms to acquire community-dependent connotational and affective meanings.

Unsurprisingly, *feminists* is highly negative in r/mensrights. However, it is similarly negative in r/actuallesbians, while neutral in r/askmen. Members of r/actuallesbians tend to not disapprove of feminists in general; instead, much of the discourse that includes this word focuses on the perceived exclusivity of many feminist movements towards LGBTQ individuals.

> *...didnt realize that transphobia was such an organized and politically influential problem especially from some* ***feminists*** *and not just old white guys...* (r/actuallesbians)

The words *women* and *ladies* are associated with many distinct social meanings. For example, *women* is seen as both a neutral and cold label, while *ladies* can be seen as traditional, patronizing, and sexual (Cralley and Ruscher, 2005; Friedman, 2013). More recently, *ladies* has also been reclaimed as an age-agnostic label popular with modern feminists (Friedman, 2013).

Both *women* and *ladies* are used in r/actuallesbians to name the target group of romantic and sexual attraction. However, the much more positive sentiment of *ladies* may be due to its additional meaning as an in-group label.

> *its a pretty slow reddit but* ***ladies*** *do say hello and interact* (r/actuallesbians)

Even though *ladies* also has positive sentiment in r/mensrights, the word is used very differently there. Instead of being used to sexualize women, *ladies* was far more commonly written in a patronizing manner.

> *...****ladies*** *this is what equality looks like time to give up some of your numerous privileges* (r/mensrights)

This reveals one limitation of a sentiment-only approach to analyzing sociolinguistic variation. The patronizing and sometimes sarcastic usage of *ladies* is a complex phenomenon that cannot be easily captured by methods based on vector space models. SENTPROP mistakenly arrives at a positive polarity for *ladies* in r/mensrights, although its high standard deviation hints at some underlying source of inconsistency.

6 Conclusion

Sentiment representations are useful for understanding variation both on a broad scale as well as among specific lexemes, particularly when combined with in-depth qualitative analyses. We focused on only explicitly gendered subreddits, but other subreddits can also be implicitly gendered. From the user-based clusters, we may be able to infer that subreddits like r/weddingplanning and r/makeupaddiction have mostly feminine users. In the future, we would like to situate our sentiment analysis of gendered subreddits in the larger context of Reddit. For example, comparing r/xxfitness with its fitness-related neighbors may allow a understanding of how explicitly targeting some demographic changes words' sentiment.

Sentiment is a salient semantic dimension, but it may also be illuminating to define subreddits along some other dimension, such as arousal, concreteness, or various emotions (Brysbaert et al., 2014; Warriner et al., 2013; Mohammad and Turney, 2010). Rarely are subreddit communities redundant. Though two subreddits may align with high similarities with some type of representation, they should differ in some other one. Still, a comparison of sentiment and emotion could result in strong alignment; negativity underlies anger and sadness, while positivity is fundamental to happiness and surprise.

7 Acknowledgements

We would like to thank Chris Potts, Will Hamilton, and Bill MacCartney for their helpful ideas and comments. Additional thanks to Will for Reddit data.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proceedings of International Conference on Web and Social Media*.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2007. Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, 12(9).

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1301–1309. Association for Computational Linguistics.

Judith Butler. 1988. Performative acts and gender constitution: An essay in phenomenology and feminist theory. *Theatre journal*, 40(4):519–531.

Jordan Carpenter, Daniel Preotiuc-Pietro, Lucie Flekova, Salvatore Giorgi, Courtney Hagan, Margaret L Kern, Anneke EK Buffone, Lyle Ungar, and Martin EP Seligman. 2017. Real men dont say cute using automatic language analysis to isolate inaccurate aspects of stereotypes. *Social Psychological and Personality Science*, 8(3):310–322.

Elizabeth L Cralley and Janet B Ruscher. 2005. Lady, girl, female, or woman: Sexism and cognitive busyness predict use of gender-biased nouns. *Journal of Language and Social Psychology*, 24(3):300–314.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd international conference on World Wide Web*, pages 307–318. ACM.

Srayan Datta, Chanda Phelan, and Eytan Adar. 2017. Identifying misaligned inter-group links and communities. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW):37:1–37:23.

Penelope Eckert. 1989. *Jocks and burnouts: Social categories and identity in the high school*. Teachers College Press.

Penelope Eckert. 2006. Communities of practice. *Encyclopedia of language and linguistics*, 2(2006):683–685.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Penelope Eckert and Sally McConnell-Ginet. 1992. Think practically and look locally: Language and gender as community-based practice. *Annual review of anthropology*, 21(1):461–488.

Ann Friedman. 2013. Hey "ladies": The unlikely revival of a fusty old label. *The New Republic*.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. 2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288.

Kira Hall. 2009. Boys talk: Hindi, moustaches and masculinity in new delhi. In *Gender and spoken interaction*, pages 139–162. Springer.

William L Hamilton, Kevin Clark, Jure Leskovec, and Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 595–605.

William L Hamilton, Justine Zhang, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Loyalty in online communities. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, volume 2017, page 540. NIH Public Access.

Susan C Herring and John C Paolillo. 2006. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031.

Srijan Kumar, William L Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 933–943. International World Wide Web Conferences Steering Committee.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Trevor Martin. 2017. community2vec: Vector representations of online communities encode semantic relationships. In *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 27–31.

Adrienne Massanari. 2017. # gamergate and the fappening: How reddits algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 19(3):329–346.

Norma Mendoza-Denton. 1996. muy macha: Gender and ideology in gang-girls discourse about makeup. *Ethnos*, 61(1-2):47–63.

Saif M Mohammad and Peter D Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics.

Anthony Mulac, James J Bradac, and Pamela Gibbons. 2001. Empirical support for the gender-as-culture hypothesis: An intercultural analysis of male/female language differences. *Human Communication Research*, 27(1):121–152.

Edward Newell, David Jurgens, Haji Mohammad Saleem, Hardik Vala, Jad Sassine, Caitrin Armstrong, and Derek Ruths. 2016. User migration in online social networks: A case study on reddit during a period of community unrest. In *Tenth International AAAI Conference on Web and Social Media*.

Matthew L Newman, Carla J Groom, Lori D Handelman, and James W Pennebaker. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3):211–236.

Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1950–1961.

Umashanthi Pavalanathan, Jim Fitzpatrick, Scott Kiesling, and Jacob Eisenstein. 2017. A multidimensional lexicon for interpersonal stancetaking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 884–895.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM.

Sascha Rothe, Sebastian Ebert, and Hinrich Schütze. 2016. Ultradense word embeddings by orthogonal transformation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 767–777.

Jonathan Schler. 2006. Effects of age and gender on blogging. In *Proceedings of AAAI Symposium on Computational Approaches for Analyzing Weblogs, 2006*, pages 199–205.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207.

Andrew D Wong. 2005. The reappropriation of tongzhi. *Language in society*, 34(5):763–793.

Yi Yang and Jacob Eisenstein. 2015. Putting things in context: Community-specific embedding projections for sentiment analysis. *Arxiv-Social Media Intelligence*.

Justine Zhang, William L Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky, and Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In *Proceedings of the... International AAAI Conference on Weblogs and Social Media. International AAAI Conference on Weblogs and Social Media*, volume 2017, page 377. NIH Public Access.