Main Ideas

- Online communities could be better understood through vector-based representations that capture users, text, and affect.
- Sentiment can provide insight into variation that word-choice alone cannot. Members of online communities exploit linguistic resources to construct a wide array of gendered identities.

Background

- **Communities:** In sociolinguistics, communities of practice are communities that focus on sharing knowledge and skills. Practices have to do with how people do things in different domains. Communities of practice in online settings can be understood through the lens of sociolinguistics.
- **Gender:** NLP often treats gender as a fixed biological variable, rather than a dynamic and social one. Gender varies within and between groups, and can be expressed through language. 
- **Sentiment Variation:** Community-specific lexicon induction has revealed cross-community variation in small domains. 

Approach

We represent subreddits in three different ways.

**Text & Users**

- The geography of Reddit differs from that of other online communities. Reddit is a platform for discussion, while other platforms may be focused on content creation. 
- User demographics are important motivators for community formation on Reddit.

**Sentiment Similarities**

<table>
<thead>
<tr>
<th>highest pairs</th>
<th>cosine</th>
<th>highest pairs</th>
<th>cosine</th>
</tr>
</thead>
<tbody>
<tr>
<td>askmen, askwomen</td>
<td>0.6702</td>
<td>femalefashionadvice, mensrights</td>
<td>0.1802</td>
</tr>
<tr>
<td>askgaybros, askmen</td>
<td>0.6144</td>
<td>askwomen, malefashionadvice</td>
<td>0.1876</td>
</tr>
<tr>
<td>askwomen, trollchromosomes</td>
<td>0.6003</td>
<td>malefashionadvice, trollchromosomes</td>
<td>0.2162</td>
</tr>
<tr>
<td>actuallesbians, trollchromosomes</td>
<td>0.5462</td>
<td>malefashionadvice, mensrights</td>
<td>0.2170</td>
</tr>
<tr>
<td>askgaybros, askwomen</td>
<td>0.5310</td>
<td>mensrights, softness</td>
<td>0.2181</td>
</tr>
</tbody>
</table>

- r/askmen and r/askwomen have high sentiment similarity, low text similarity (0.942 and 0.886), above average sentiment similarity (0.637). 
- Some of the variation is due to polysemy: s

**Future Work**

- Expanding to other communities: implicitly gendered subreddits (e.g. high user-overlap with gendered ones), comparing gendered subreddits (r/xxfitness) to non-gendered ones (r/fitness).
- Examining other semantic dimensions: arousal, emotions.

Conclusion

Online language does not vary according to a clearcut, binary perspective of gender. Sentiment can be a useful indicator of words’ social meaning and community values, especially in the context of discussion content and user demographics.