Latent Hatred: A Benchmark for Understanding Implicit Hate Speech

Mai ElSherief, Caleb Ziems, David Muchlanski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, Diyi Yang

1. Motivation
- Many hate speech datasets are highly explicit and overt
- Little is known about implicit hate

2. Taxonomy
- White Grievance
- Incitement to Violence
- Stereotypes + Misinfo
- Threatening + Intimidation
- Inferiority
- Irony

3. Data Annotation + Expansion
933 explicit | 6,346 implicit | 13,291 not hate

4. Detecting Implicit Hate

<table>
<thead>
<tr>
<th>Model</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>HateSonar</td>
<td>39.9</td>
<td>48.6</td>
<td>43.8</td>
</tr>
<tr>
<td>Perspective API</td>
<td>50.1</td>
<td>61.3</td>
<td>55.2</td>
</tr>
<tr>
<td>Linear SVM</td>
<td>61.4</td>
<td>67.7</td>
<td>64.4</td>
</tr>
<tr>
<td>BERT + Augmentation</td>
<td>67.8</td>
<td>72.3</td>
<td>70.4</td>
</tr>
</tbody>
</table>

5. Explaining Implicit Hate

6. Conclusion + Future Directions

GT-SALT/implicit-hate