Hate speech is expressed through language, particularly sensitive to the target communities, which makes it harder to understand and perceive.

**PROBLEM**

Hate speech falls flat. It’s been 3 months and these words have only proven to be true. Suck it "Butch" who can’t Deadlift. It’s History you know. Mary Beard told me so and she has a vagina from Oxford.

**EXISTING APPROACHES**

Our framework integrates knowledge about the target groups to ground predictions to the entities supporting a decision, while preserving optimal predictive quality.

**HYBRID MACHINE LEARNING**

- **System A**
  - Deep Learning Based
  - Neural Network
  - Performance: 0.95
  - Interpretable

- **System B**
  - Linguistic Based
  - Regular Expressions
  - Performance: 0.44
  - Interpretable

**EVALUATION**

<table>
<thead>
<tr>
<th>Category (FP)</th>
<th>Definition</th>
<th>A/E</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographic descriptor</td>
<td>Direct explicit reference to a member of the group.</td>
<td>X</td>
<td>117</td>
</tr>
<tr>
<td>Targeted language</td>
<td>Insult, sexually explicit, or topics related to the group.</td>
<td>X</td>
<td>20</td>
</tr>
<tr>
<td>Implicit reference</td>
<td>Refers to a group member using pronouns.</td>
<td>X</td>
<td>10</td>
</tr>
<tr>
<td>False match</td>
<td>Incorrectly flagged due to polysemy.</td>
<td></td>
<td>3</td>
</tr>
</tbody>
</table>

**Error Analysis**

- No reference: No language related to the group.
- Missed at content: Not identified at validation, due to misspellings or being out-of-training.
- Missed by method: Mention not correctly found or given importance by model.

**Knowledge-grounded Predictions Analysis**

- woman, man, LGBTQ, transgender person. Black, man, gay, home, heterosexual, homosexuality, feminist, lesbian, queer sexual orientation, same-gender marriage, pedestrian, sexuality, homophobia, queer identity, sex, feminist, mentor, partner, sex worker, lag, laggot, masculine, pussy, thot, rape, menstruation, bitch
- woman, man, LGBTQ, gay, heterosexual, homosexual, feminist, lesbian, queer sexual orientation, same-gender marriage, womanism, sexuality, homophobia, queer identity, abusive person, sex, partner, laggot, semen, pussy, bitch
- feminist, homophobia, sex, sex work client, mental partner, parent, sex worker, lag, laggot, raped, female gender role, society, abortion, vaginism, that, panic, rape, domestic violence, she, bimbo, sexual assault, cunt, bitch, whore, alt, fuck, Mrs., rainbow flag

**LESSONS LEARNED**

- Our hybrid models are effective as the state-of-the-art models (System A), while adjusting better to domain and data changes.
- Entities in a Knowledge Graph (KG) are a richer representation than simple terms (System B), resulting in more transparency and competitive performance with language models.
- Grounding predictions in explicit knowledge is key to understand model outcomes, the data used to train hate speech detection systems, and the ambiguous cases in human annotations.