Evading hate-speech detection

Hate-speech
- Attacks or threatens an individual or group.
- Classified with word- and character-based features in prior work [1 - 4].
- Various challenges: e.g. how to distinguish from offensive speech [2]?

Our evasion attacks
- Two easily implementable methods tested on five models and three datasets.

1. Typos
   - I hate you
   - I htae you

2. Word appending (10 to 50 words)
   - Common English words [5]:
     - I hate you
     - I hate you [ make people thing (…) ]
   - Words from non-hate class of training set
     - I hate you
     - I hate you [ good nice sweet (…) ]

Adversarial training
- Augmenting training set with similar examples as used in evasion.
- Common word appending deteriorated performance of word-based logistic regression.
- No marked negative effect on other models.
- Typo-augmentation improved original test data performance of character models.
- Evasion susceptibility reduced in 12/15 tests.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Original</th>
<th>Appending common</th>
<th>Appending non-hate</th>
<th>Typos</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR characters [1]</td>
<td>D1</td>
<td>0.76</td>
<td>-0.28 + 0.20</td>
<td>-0.29 + 0.20</td>
<td>-0.15 + 0.11</td>
</tr>
<tr>
<td>MLP characters [1]</td>
<td>D1</td>
<td>0.76</td>
<td>-0.26 + 0.22</td>
<td>-0.27 + 0.19</td>
<td>-0.21 + 0.16</td>
</tr>
<tr>
<td>LR words [2]</td>
<td>D2</td>
<td>0.51</td>
<td>-0.03 - 0.04</td>
<td>-0.06 - 0.15</td>
<td>-0.21 + 0.12</td>
</tr>
<tr>
<td>CNN+GRU [4]</td>
<td>D2</td>
<td>0.35</td>
<td>-0.32 + 0.23</td>
<td>-0.35 + 0.30</td>
<td>+0.01 + 0.13</td>
</tr>
<tr>
<td>LSTM [3]</td>
<td>D3</td>
<td>0.74</td>
<td>-0.22 + 0.23</td>
<td>-0.49 + 0.46</td>
<td>-0.59 + 0.54</td>
</tr>
</tbody>
</table>

F1-scores for the hate class. Added number shows the effect of adversarial training.

LR = logistic regression
MLP = multilayer perceptron
CNN = convolutional neural network + gated recurrent unit
LSTM = long short-term memory network

Discussion
- Word-based approaches more vulnerable to typos: misspelled words often unrecognized.
- Both character- and word-based models are vulnerable to word appending attack.
- Adversarial training helps, but does not fully mitigate either attack.
- Hate-speech detection as anomaly detection rather than text classification.