Secure Systems Group, Aalto University

Tommi Gröndahl, Luca Pajola, Mika Juuti, N. Asokan

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]:

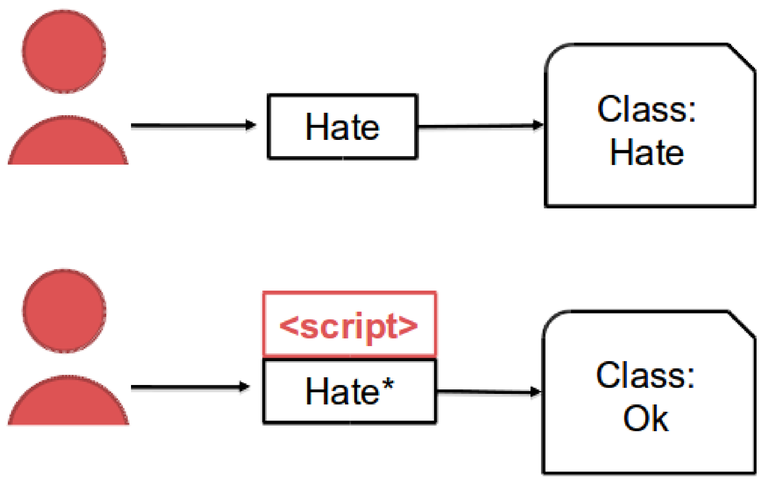
   I hate you ➡

   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.

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

F1-scores for the hate class.
Added number shows the effect of adversarial training.
LR = logistic regression
MLP = multilayer perceptron
CNN + GRU = 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.

Aalto University

[1] E. Wulczyn, N. Thain, L. Dixon. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*, 2017.

[2] T. Davidson, D. Warmslay, M. Macy, I. Weber. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th Conference on Web and Social Media*, 2017.

[3] P. Badjatiya, S. Gupta, M. Gupta, V. Varma. Deep Learning for Hate Speech Detection in Tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017.

[4] Z. Zhang, D. Robinson, J. Tepper. Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network. In *Proceedings of ESWC* , 2018.

[5] https://github.com/first20hours/google-10000-english