Cyberbullying Classifiers are Sensitive to Model-Agnostic Perturbations

Chris Emmery, Ákos Kádár, Grzegorz Chrupała, Walter Daelemans

Tilburg University
Department of Cognitive Science & Artificial Intelligence

What did we do?

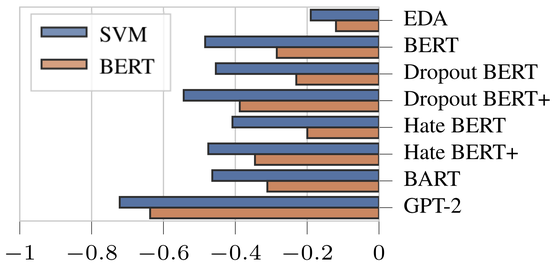
- Show transformer-based model-agnostic lexical substitutions [1, 3] severly impact performance of cyberbullying classifiers.
- Tested various models to generate candidates for augmentation, and fine-tune after augmentation.
- We show there is a trade-off between robustness against lexical variation and task performance.

(Augmented) Data

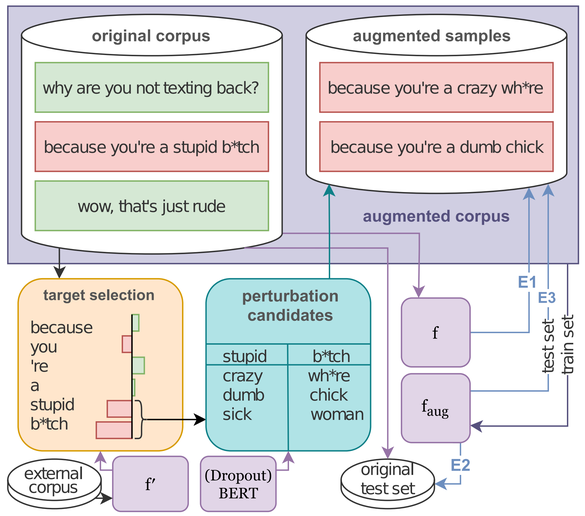
			TTR	AVG TOK/MSG
Ask.fm	5,001	89,404	.154	12 ($\sigma = 23$)
	426	1,627	.016	391 ($\sigma = 285$)
Twitter I	237	5,258	.154	14 ($\sigma = 8$)
Twitter II	281	4,654	.221	18 ($\sigma = 8$)
YouTube	417	3,045	.063	239 ($\sigma = 252$)
Formspring	1,025	11,742	.060	27 ($\sigma = 29$)

	TRAIN		TEST	
	😈	😎	😈	😎
Merged	4,789	72,243	561	8,001
Augment Train	28,148	72,243	561	8,001
Augment Test	4,789	72,243	3,283	8,001

Adversarial Power



Experimental Setup



Method

Substitute Model f'

- Provides the omission scores to determine which words to augment.
- Uses disjoint data and architecture, collected and trained by user (can thus facilitate human-in-the-loop attack).

Candidate Generation C

- Dropout(BERT embedding(w_i)) [4], predict top-k words (dropout substitution).

Ranking

- Concatenate last four BERT layers [4], use as contextualized representation h. SIM score is given by:

$$\text{SIM}(D, D'; t) = \sum_{i}^{n} \alpha_{i,t} \times \Lambda\left(\boldsymbol{h}(D_i), \boldsymbol{h}(D'_i)\right)$$

Augmentation Examples

							TARGETS						
PROMPT	You	are	a	retarded	dweeb	and	stupid	af	.	Go	fuck	yourself	.
TOKENS	You	are	a	re ##tar ##ded	d ##we ##eb	and	stupid	a ##f	.	Go	fuck	yourself	.
#1	You	are	a	silly	baby	and	silly	af	.	Go	screw	yourself	.
#2	You	are	a	useless	teenager	and	dumb	af	.	Go	duck	yourself	.
#3	You	are	a	sick	bitch	and	foolish	af	.	Go	shit	yourself	.
#4	You	are	a	crazy	dog	and	useless	af	.	Go	damn	yourself	.
#5	You	are	a	dumb	idiot	and	ignorant	af	.	Go	piss	yourself	.

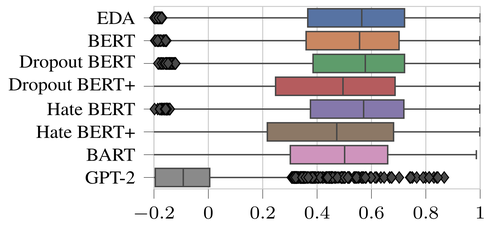
Experiments 1 and 2

X_{train}	Plain	EDA	BERT	Dropout BERT	Dropout BERT+	Hate BERT	Hate BERT+	BART	GPT-2
$f(X'_{\text{pos}})$	.614 .009	.598 .012	.458 .002	.491 .005	.439 .002	.520 .005	.478 .004	.436 .007	.334 .007
$f_{\text{aug}}(X_{\text{test}})$	.563 .014	.553 .007	.538 .014	.546 .012	.523 .004	**.562** .007	.535 .009	.536 .013	.550 .017

Experiment 3

	INITIAL TPR	f_{aug}	
	Plain	EDA	BERT
$X'_{\text{test}} \downarrow$		Δ TPR	
EDA	.498	.270	.106
BERT	.390	-.033	.195
Dropout BERT	.421	-.017	**.183**
Dropout BERT+	.362	.015	.228
Hate BERT	.444	-.020	**.170**
Hate BERT+	.394	.034	.188
BART	.378	-.010	**.200**
GPT-2	.303	-.098	.031
MEAN	.399	-.114	**.158**

Semantic Scores



References

[1] Emmery, C., Kádár, Á., & Chrupała, G. (2021, April). Adversarial Stylometry in the Wild: Transferable Lexical Substitution Attacks on Author Profiling. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume (pp. 2388-2402).

[2] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT*.

[3] Zhou, W., Ge, T., Xu, K., Wei, F., & Zhou, M. (2019). BERT-based Lexical Substitution. *ACL*.