A white hat approach to fighting online trolls

Experiments with BERT and GAN

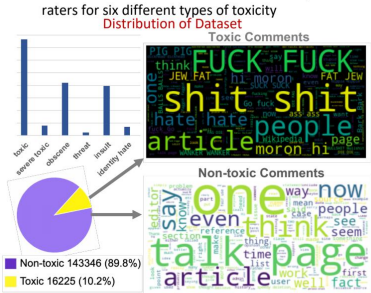
INTRODUCTION

- In cybersecurity, "white hat" refers to an ethical expert who tests an organization's security defences with the aim of strengthening them
- The very same deep learning tools that are used to improve human conditions are also available to malicious actors for automation of attacks
- Taking a leaf from cybersecurity, we first build classifiers to detect abusive online language and then use those toxic comments to generate new ones to test the robustness of our classifiers against machine generated negative language

DATASET

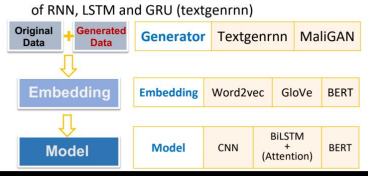
- The original dataset consists of 159,751 Wikipedia comments which have been labeled by human raters for six different types of toxicity

Distribution of Dataset



METHODS

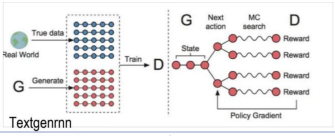
- **Classification**: permutations of 4 models (BERT, CNN, LSTM & Attention) and 3 word embeddings (BERT, GloVe and word2vec)
- **Generation of comments**: we used Maximum-Likelihood Augmented Discrete Generative Adversarial Networks (MaliGAN) and a multi-layer RNN consisting of RNN, LSTM and GRU (textgenrnn)



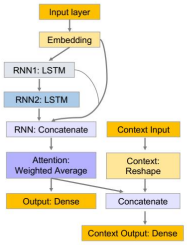
MODEL ARCHITECTURE

Generator

- MaliGAN



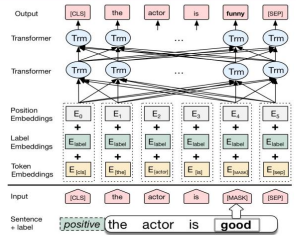
- Textgenrnn



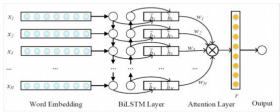
- Like most text GANs, Maligan trains a discriminator (D) to minimize binary loss between real and generated text
- What is unique about Maligan is a novel objective for the generator (G) to optimize, using importance sampling, which makes the training procedure closer to maximum likelihood (MLE) training of auto-regressive models, and thus more stable and with less variance in the gradients

Classifier

- BERT

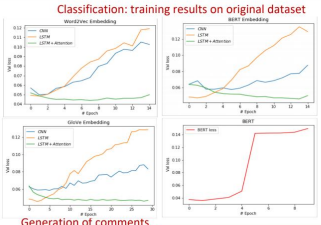


- BiLSTM Attention Model

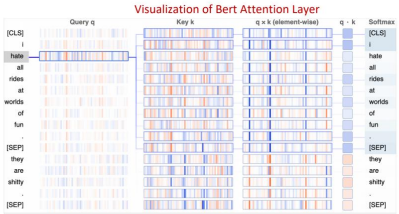


ANALYSIS

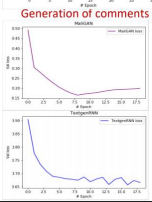
Classification: training results on original dataset



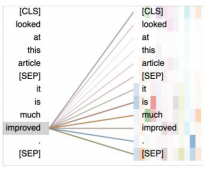
Visualization of Bert Attention Layer



Generation of comments



- **Classification:** all combination of model and word embeddings show signs of overfitting when the number of epochs is large
- In particular, LSTM + Attention is the least likely to overfit amongst all models
- BERT model has the lowest validation loss on real data
- **Generation of text:** As expected, the negative log likelihood (NLL) loss of Maligan decreases during pre-training and increases slightly during adversarial training



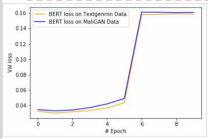
- Two sentences are chosen from the toxic and non-toxic class to visualize BERT's three attention patterns
- **Delimiter-focused attention**: to propagate sentence-level state to the word level, designating a set of [SEP]neurons
- **Bag of Words attention**: to distribute sentence-level state to the word level, taking nearly unweighted average of the word embeddings
- **Next-word attention**: all the attention is focused on the next word in the input sequence, make it able to capture sequential relationships

RESULTS

	Original Data			Original + MaliGAN				Original + Textgenrnn			
	Train loss	Val loss	Total AUC	Train loss	Val loss	Total AUC	Generate AUC	Train loss	Test loss	Total AUC	Generate AUC
W2V+CNN	0.012	0.050	0.953	0.041	0.057	0.976	0.990	0.031	0.045	0.978	0.987
W2V+LSTM	0.013	0.048	0.940	0.043	0.052	0.978	0.994	0.038	0.044	0.982	0.990
W2V+Attention	0.033	0.044	0.983	0.036	0.039	**0.986**	0.998	0.033	0.040	**0.985**	0.994
GloVe+CNN	0.029	0.059	0.954	0.050	0.049	0.975	0.987	0.052	0.051	0.976	0.984
GloVe+LSTM	0.019	0.046	0.953	0.039	0.042	0.981	0.991	0.037	0.045	0.982	0.998
GloVe+Attention	0.044	0.043	0.983	0.032	0.036	**0.989**	0.995	0.044	0.040	**0.985**	0.998
BERT+CNN	0.018	0.058	0.936	0.060	0.062	0.967	0.988	0.045	0.048	0.973	0.988
BERT+LSTM	0.004	0.047	0.933	0.379	0.505	0.977	0.986	0.033	0.043	0.979	0.998
BERT+Attention	0.029	0.041	0.985	0.035	0.040	**0.986**	0.994	0.047	0.045	**0.980**	0.999
BERT+BERT	0.027	0.035	**0.994**	0.027	0.033	**0.995**	0.999	0.023	0.032	**0.996**	0.999

Examples of generated text

MaliGAN 1. Sorry do you think it makes me really stupid shit?
2. You utter bitch get my real world abusing valid messages.
Textgenrnn: 1. You are a dishonest idiot. Stop censoring this cite.
2. Oh, I am going to vandalize Wikipedia, what the fuck.



- The word embeddings have marginal effect on this task with the BERT word embeddings performing the worst
- However, the BERT model reigns supreme on both real and generated comments with almost perfect AUC
- The toxic comments generated by textgenrnn are in general more believable than those by Maligan

DISCUSSION

- The prowess of the Attention is obvious - both BERT (Bidirectional Encoder Representations from Transformers) and BiLSTM Attention perform superior to LSTM and CNN across different word embeddings
- It is not surprising that the BERT model performed the best – the model is now considered the state-of-the-art in NLP as evident from its results on SQuAD v1.1
- Our results imply that even with access to modern text generation models such as MaliGan and Textgenrnn, it will be difficult for motivated malicious actors to trick abusive language classifiers

REFERENCES

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (2018)
[2] Maximum-Likelihood Augmented Discrete Generative Adversarial Networks (2017)
[3] Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification (2016)

Jiaying Huang hjy1227@stanford.edu Zhihao Lin zhl@stanford.edu