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 defenses 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:

![Dataset Distribution Chart]

[Graph showing distribution of non-toxic and toxic comments]

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 (MaLI-GAN) and a multi-layer RNN consisting of RNNLSTM and GRU (Textgenmn).

MODEL ARCHITECTURE

- **Generator**
  - MultiGAN
  - Textgenmn

- **Classifier**
  - BERT
  - LSTM

ANALYSIS

- Classification: training results on original dataset:

  ![Classification Training Results Chart]

- Visualization of BERT Attention Layer:

  ![Visualization of BERT Attention Layer]

RESULTS

- Original Data
  - BERT
  - GloVe
  - CNN
  - LSTM

- Original + MaLI-GAN
  - BERT
  - GloVe
  - CNN

- Original + Textgenmn
  - BERT
  - GloVe
  - CNN

Examples of generated text:

**MaLI-GAN:** 1. Sorry do you think it makes me really stupid and?

**Textgenmn:** 1. You are a dishonest idiot. Stop censoring this site.

**Discussion:**

- The prowess of the Attention is obvious - both the BERT (Bidirectional Encoder Representations from Transformers) and BLSTM 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 MaLI-GAN and Textgenmn, it will be difficult for malsuited malicious actors to trick abusive language classifiers.

REFERENCES


---

Jaeying Huang hjy1227@stanford.edu
Zhizhao Lin zhl@stanford.edu