Motivation & Goals

- **Goals:**
  - Use deep learning to identify Russian Trolls based on their comments
  - Use BERT\(^1\) to bypass significant data limitations
- **Motivation:**
  - Russian trolls are a major threat to social media platforms and remain at large on Reddit
  - Troll detection on Reddit is basically nonexistent!

Dataset

- ~7000 comments from over 900 banned Russian Trolls on Reddit from the 2017 Transparency Report
- ~7000 randomly sampled human comments from Reddit\(^2\)
- 64% Train, 16% Dev, 20% Test

Models

- BERT layer to find comment word embeddings
- Middle Layer Examples: RNN, CNN, RCNN\(^3\)
  - RCNN: bidirectional LSTM + Max pooling
- Dropout and early stopping for regularizations
- Binary linear classifier layer

Optimization

- Out of the 60 models we trained during Randomized Hyperparameter Search, we found a dropout rate of 0.445, comment size of 70 words, and hidden size of 415 to be optimal hyperparameters.

Output Analysis

- The model **performs better** at classifying troll accounts with **more comments** in the dataset
- The model also **performs marginally better** at classifying comments with **more words**, and is equally bad at misclassifying trolls and non-trolls on this metric alone
- Trigger words caused **classification to change** (e.g. Hillary).
  - Some were dependent on certain context words being present

Conclusions

- **Summary:**
  - Both CNN and RCNN models classify Trolls with > 84% AUC score, significantly outperforming both baselines
  - Troll detection on Reddit seems viable!
- **Limitations:**
  - Our dataset is very limited
  - Dataset distribution is skewed
  - Model responds too aggressively to “triggers”
- **Future Work**
  - Model adjustments based on error analysis
  - Significant and rigorous data collection

References: