Identifying Tweets Written by Russian Troll Accounts
Ethan Brown, Brendan Edelson, Elijah Taylor
esbrown@stanford.edu, bedelser@stanford.edu, elijahb@stanford.edu

In the wake of Department of Justice investigations and general public weariness towards cybersecurity, we decided to investigate how deep learning may help in identifying Russian “Twitter bots”. Given the text of a tweet, our neural network outputs a binary classification (Russian Bot / Not a Russian Bot).

Data

- 200,000 Russian Bot tweets (ground truth)
  - Released by NBC for public analysis
- Over 1 million politically-themed tweets from the 2016 election season (assumed not Russian bots)
  - Collected through a Harvard research project

Features

- 200 GloVe Vectors
  - trump
  - is
  - loud
  - ...
  - (0.067)
  - (0.063)
  - (0.988)
  - ...
  - (0.905)
  - (0.094)
- 200 dimensional feature vectors trained on a Twitter corpus
- Used word embeddings to convert word indexes to GloVe vectors

Models

- LSTM Hyperparameter
  - GloVe Embedding Dimensions: 200
  - LSTM Hidden Nodes: 130
  - Fully Connected Nodes: 6
  - Dropout Rate: 25
  - Epochs: 10
  - Minibatch Size: 128

Results

<table>
<thead>
<tr>
<th>Model</th>
<th># Train Examples</th>
<th># Test Examples</th>
<th>Train Accuracy</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>3-Layer NN w/ 25D Glove vectors</td>
<td>35,000</td>
<td>5,000</td>
<td>68.12%</td>
<td>62.15%</td>
</tr>
<tr>
<td>LSTM w/o Fully Connected Layer</td>
<td>1.1 Million</td>
<td>12,000</td>
<td>97.45%</td>
<td>94.93%</td>
</tr>
<tr>
<td>LSTM with Fully Connected Layer*</td>
<td>1.1 Million</td>
<td>12,000</td>
<td>96.69%</td>
<td>95.07%</td>
</tr>
</tbody>
</table>

*Final model

Discussion

Better than expected results!!
- NLTK tokenizer
- Twitter corpus trained GloVe vectors
- Fully connected layer reduces overfitting

Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>9843</th>
<th>153</th>
</tr>
</thead>
<tbody>
<tr>
<td>161</td>
<td>161</td>
<td>1843</td>
</tr>
</tbody>
</table>

Correct Classification Rates
- Non-Russian: 98.9% (True Positive)
- Russian: 91.97% (True Negative)

Future

- Further analyze misclassified examples to find trends
- Implement other features into the model (e.g. time posted, user account data, etc.)
- Look into creating a GAN to simulate Russian tweets
- Classifying tweets/users with probabilities of various political biases using a softmax output

References