Bots in the Net: Applying Machine Learning to Identify Social Media Trolls

Jack Cable (cablej@stanford.edu) and Grant Hugh (ghugh@stanford.edu)

Introduction

Social media is becoming an increasingly attractive target for foreign actors to spread disinformation. For our project, we implemented and compared different machine learning techniques to classify whether tweets were posted by trolls. On a dataset of confirmed Russian troll tweets and normal tweets, we were able to achieve 96.4% accuracy on test data. We then applied our algorithms to build a live classifier that can aid social media companies in moderating content.

Dataset

We compiled a dataset of 142,560 tweets from a combination of FiveThirtyEight’s database of Russian troll tweets and election-related tweets from George Washington University’s TweetSets database. Data was preprocessed to filter for tweets in English and to parse tweet contents. Finally, we split our tweets 80/20 for the training and test datasets.

Results

(114,048 training samples, 28,512 testing samples)

<table>
<thead>
<tr>
<th>Method</th>
<th>Binary Count</th>
<th>Word Count</th>
<th>Word TF-IDF</th>
<th>Ngram TF-IDF</th>
<th>Char TF-IDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.955</td>
<td>0.998</td>
<td>0.991</td>
<td>0.853</td>
<td>0.874</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>0.977</td>
<td>0.995</td>
<td>0.979</td>
<td>0.934</td>
<td>0.926</td>
</tr>
<tr>
<td>RBF Kernel SVM</td>
<td>1.00</td>
<td>0.947</td>
<td>1.00</td>
<td>0.987</td>
<td>0.959</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.993</td>
<td>0.915</td>
<td>0.993</td>
<td>0.991</td>
<td>0.930</td>
</tr>
<tr>
<td>LSTM:</td>
<td>0.981</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LSTM: Train 0.981, test 0.957

Future Work

- Improve dataset of positive labels by collecting known troll tweets from a wider range of sources.
- Apply other more advanced forms of deep learning, such as C-RNN-GAN to adapt to possible novel attacks by adversaries.

Discussion

Basic methods including Logistic Regression and Naive Bayes achieved high accuracy on the test data, indicating that trolls could be easily detected from common word usage. Applying the kernel SVM and the LSTM neural network further improved results. This is expected as both methods can capture deeper complexities, such as the relationships between words, not attainable via simpler methods.

One limitation is that our troll test data belongs to the same FiveThirtyEight dataset as the troll training data. As a result, we may be overfitting to this specific dataset, and could have trouble classifying new types of troll accounts.

Models

- Naive Bayes: $p(y = 1|x) = \frac{(BF_{y=1}(x/y=1) \cdot p(x/y=1)) \cdot p(y=1)}{\sum_{y=0}^{1} (BF_{y=1}(x/y=1) \cdot p(x/y=1) + BF_{y=0}(x/y=0) \cdot p(x/y=0))}$
- Logistic Regression: $\log \frac{y}{1-y} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}$
- RBF Kernel SVM: Minimize $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$
- Random Forest
- LSTM Neural Network

References