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

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

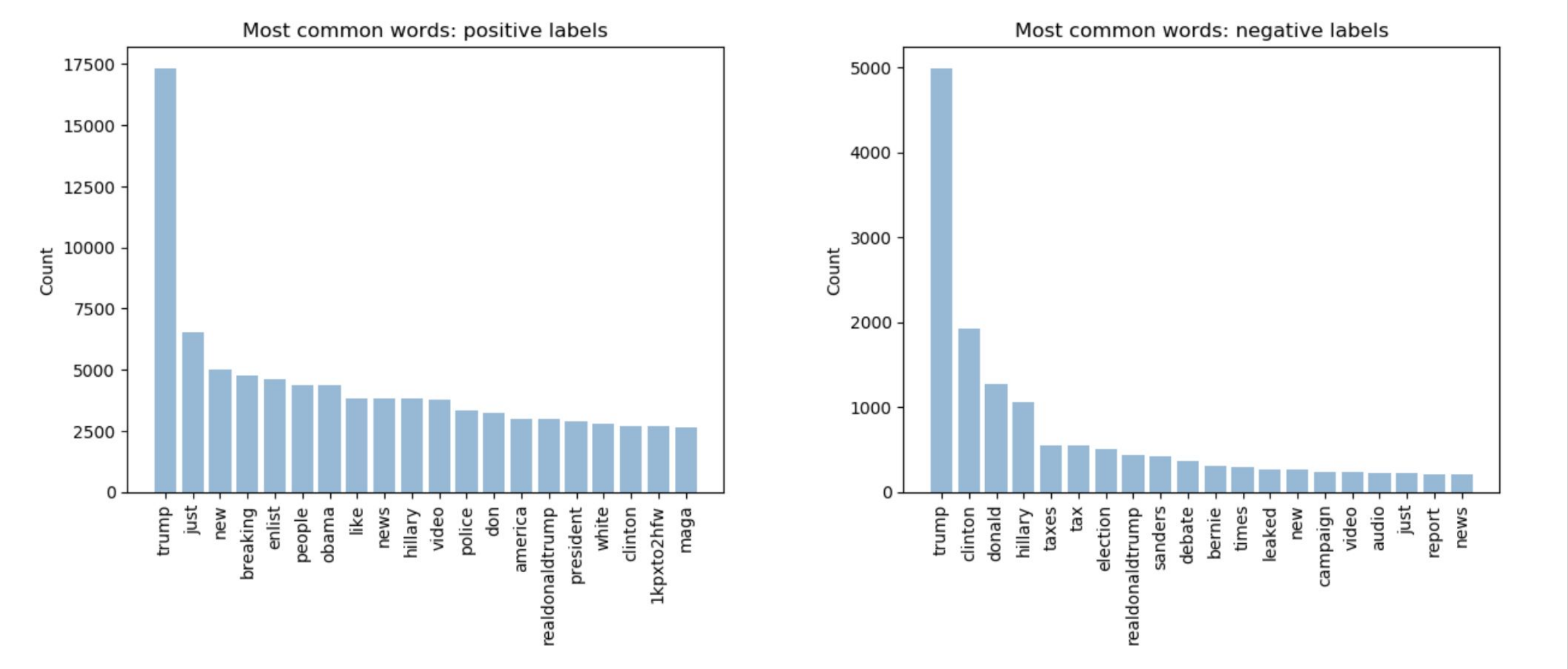
Stanford
Computer Science

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.



Most common words across positive and negative datasets

Featurization

- Word count: Mapped tweet to a word vector, capped at 5,000 features
- TF-IDF: Extracted word, character, and ngram TF-IDF scores in order to better identify uncommon words, capped at 5,000 features
- Word embeddings: Mapped tweets to higher-dimensional vectors for LSTM neural network, capped at 10,000 features

Models

Naive Bayes

$$p(y=1|x) = \frac{(\Pi_{j=1}^{d} p(x_j|y=1))p(y=1)}{(\Pi_{j=1}^{d} p(x_j|y=1))p(y=1) + (\Pi_{j=0}^{d} p(x_j|y=1))p(y=0)}$$

Logistic Regression

$$\ell(\theta) = \sum_{i=1}^{n} y^{(i)} \log g(\theta^T x^{(i)}) + (1-y^{(i)}) \log(1 - g(\theta^T x^{(i)}))$$

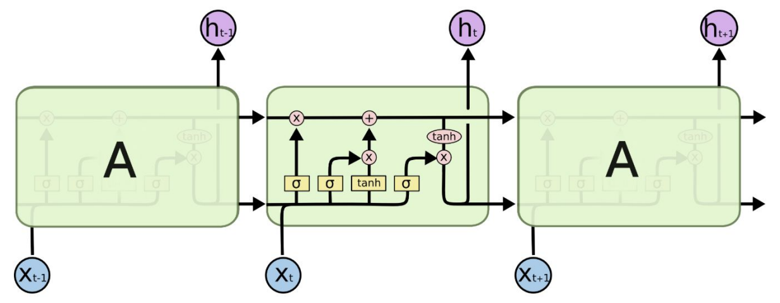
RBF Kernel SVM

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{s.t. } y^{(i)}(w^T x^{(i)} + b) \geq 1, i = 1, \ldots, n$$

Random Forest

$$GINI(t) = 1 - \sum_{i=1}^{j} P(i|t)^2$$

LSTM Neural Network



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

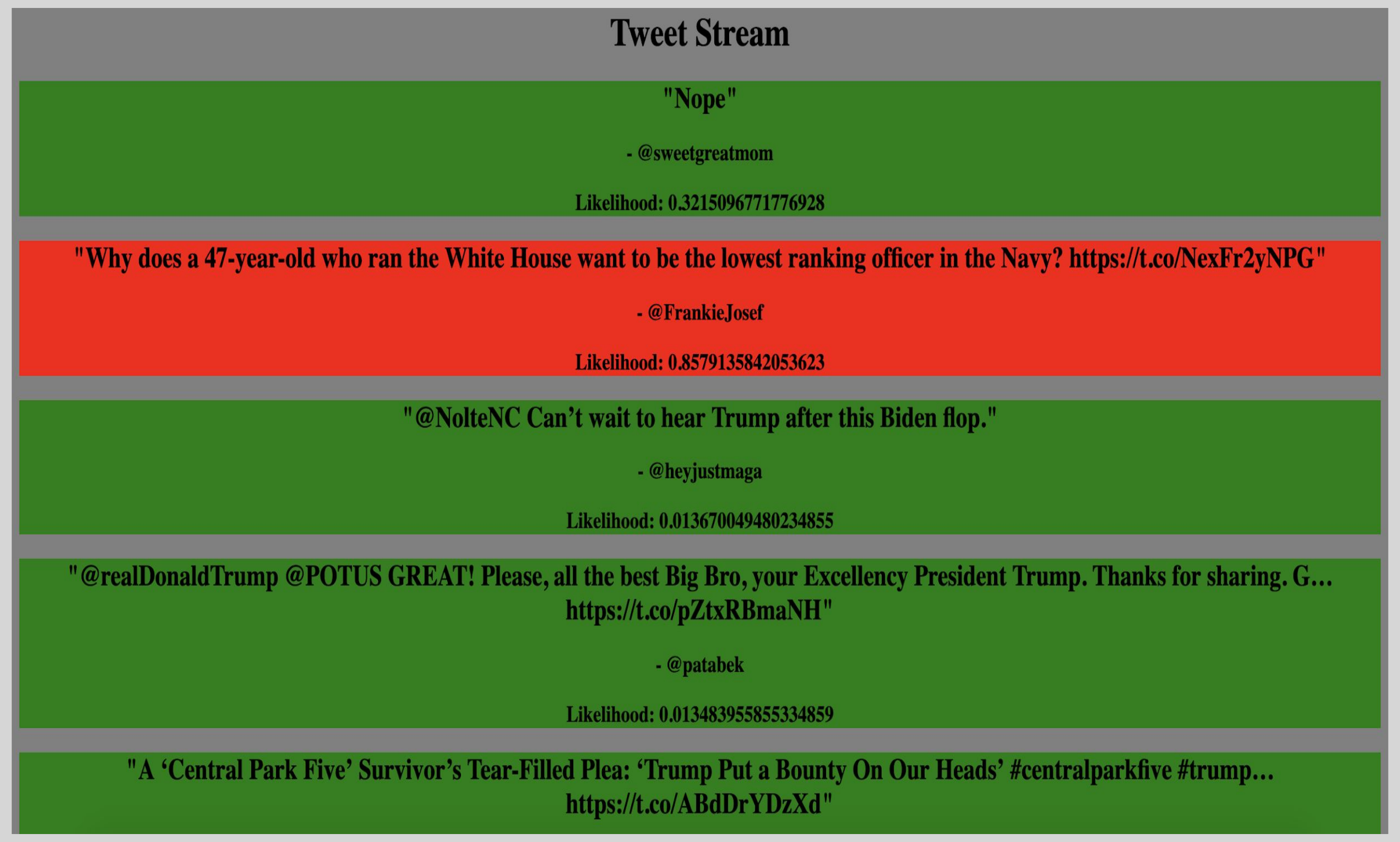
Method	Binary Count		Word Count		Word TF-IDF		Ngram TF-IDF		Char TF-IDF	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Naive Bayes	0.953	0.898	0.951	0.901	0.883	0.874	0.814	0.801	0.884	0.882
Logistic Regression	0.977	0.939	0.977	0.936	0.934	0.926	0.835	0.818	0.948	0.940
Kernel SVM	1.00	0.947	1.00	0.944	0.987	0.959	0.917	0.861	0.995	0.964
Random Forest	0.993	0.915	0.993	0.906	0.991	0.930	0.930	0.826	0.998	0.955

LSTM: Train 0.981, test 0.957

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.



Tweet Stream

"Nope"

- @sweetgreatmom

Likelihood: 0.3215096771776928

"Why does a 47-year-old who ran the White House want to be the lowest ranking officer in the Navy? https://t.co/NexFr2yNPG"

- @FrankieJosef

Likelihood: 0.8579135842053623

"@NolteNC Can't wait to hear Trump after this Biden flop."

- @heyjustmaga

Likelihood: 0.013670049480234855

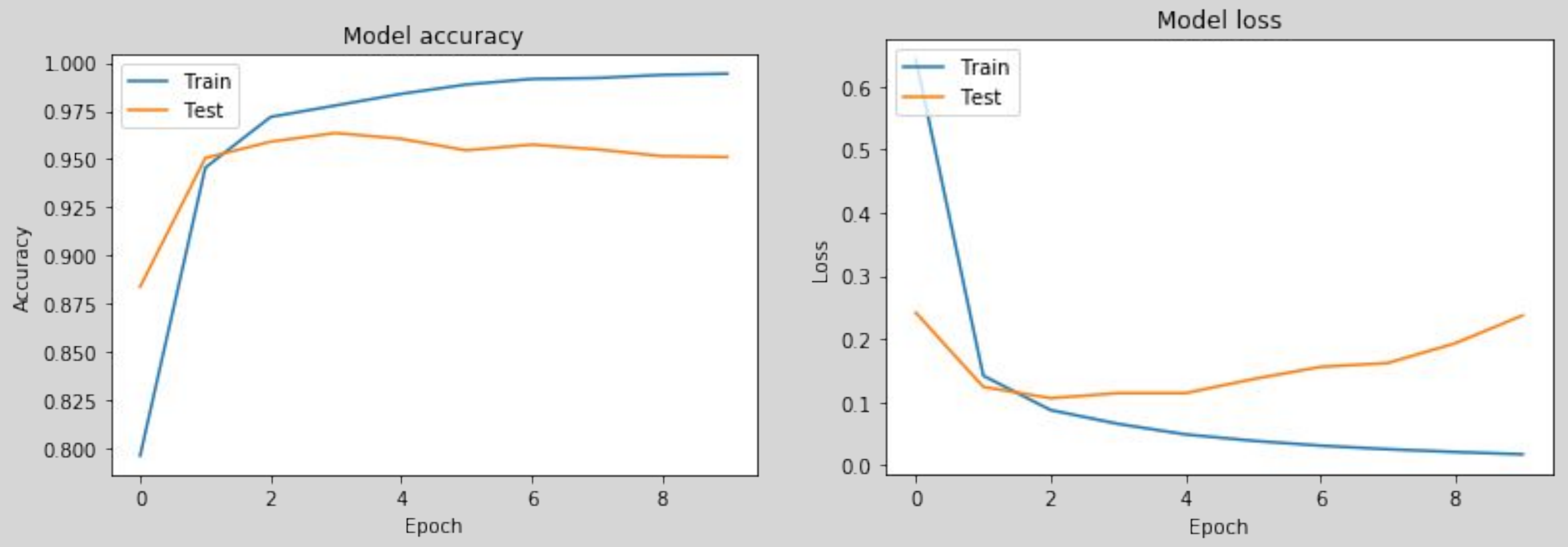
"@realDonaldTrump @POTUS GREAT! Please, all the best Big Bro, your Excellency President Trump. Thanks for sharing. G... https://t.co/pZtxRBmaNH"

- @patabek

Likelihood: 0.013483955855334859

"A 'Central Park Five' Survivor's Tear-Filled Plea: 'Trump Put a Bounty On Our Heads' #centralparkfive #trump... https://t.co/ABdDrYDzXd"

Classification tool built for live feed of tweets



LSTM neural network loss and accuracy across epochs

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.

References

FiveThirtyEight. (2018). *3 million Russian troll tweets*. [online] Available at: https://github.com/fivethirtyeight/russian-troll-tweets [Accessed 12 Jun. 2019].
Justin Littman. (2018). TweetSets. Zenodo. https://doi.org/10.5281/zenodo.1289426
Olah, C. (2019). *Understanding LSTM Networks*. [online] Colah.github.io. Available at: https://colah.github.io/posts/2015-08-Understanding-LSTMs/ [Accessed 12 Jun. 2019].