A Deep Learning Approach to Identifying Covert Disinformation Networks

Josh DuFault
JDuFault@stanford.edu

Introduction
Covert and organized networks can use social media platforms to launch large-scale disinformation campaigns and spread their influence undetected. Here I propose a data organization scheme and convolutional neural network (CNN) approach to identify these networks quickly and with high accuracy.

Data
Positive examples:
200,000 Tweets sent by the Internet Research Agency (IRA) which was charged with fraud for exerting undue influence on US elections on behalf of a foreign government.

Negative Examples (Easy):
200,000 random tweets geolocated in the US while the IRA was active. These Tweets represented typical Twitter conversations and were largely on different topics than the IRA Tweets. See Figure 1.

Negative Examples (Hard):
200,000 tweets selected to have similar content to the IRA Tweets. A distribution of Twitter users that sent the same median number of political Tweets during the 2016 Presidential Election as the IRA Tweets was selected. From these users, a collection of Tweets was randomly chosen in amounts to give a similar user distribution to the IRA Tweets.

These Tweets were chosen to ensure the model can differentiate between members of a specific influence network and non-members who Tweet about similar topics. They were largely similar to the IRA Tweets. See Figures 1 & 2.

Models

Deep CNN:
This network took in a character vector and performed three layers of convolution and max pooling. See figure 3.

LSTM:
LSTM’s are traditionally used for sequence data. This network consisted of a single layer LSTM with an equal number of nodes to embedded character vectors. Each node took input from both the context vector and the previous node. See Figure 1.

CNN + LSTM:
This network was identified as state of the art in previous literature at analyzing words from character vectors. It took in a character vector, performed a layer of convolution and max pooling, then fed the output into a traditional LSTM. See Figure 1.


Results (Easy)
LSTM and CNN + LSTM:
Despite being identified in previous literature as optimal for sequence data, these performed no better than chance. This was caused by the length of the input vectors which averaged over 1000 characters. See Figure 5.

Deep CNN:
Despite not being designed for sequence data, the CNN consistently delivered very high accuracy. These results indicate that identifying members of a network via Tweets requires different techniques than understanding language even though both tasks focus on text. See Figure 5.

The CNN was also significantly faster than the LSTM models. Based on these results, all testing on the hard dataset was done using a CNN. See Figure 6.

Training Accuracy:
0.967
Test Accuracy:
0.954

Initial results on the hard dataset were significantly lower. However, adjustment of hyperparameters resulted in accuracy levels within two percent of the results on the easy dataset.

Summary
By grouping Tweets by user into collections of 20 and using a deep CNN, it is possible to identify members of covert disinformation spreading networks. This result shows promise in identifying fake news networks where identifying content by individual Tweets requires background knowledge. This suggests that the optimal strategy to identifying misleading data is to identify the network spreading the data which can be done with high accuracy using a CNN.

References