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.

Figure 1
Tweet Content

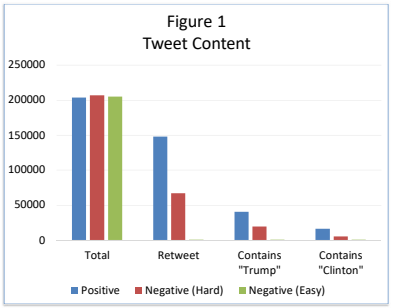
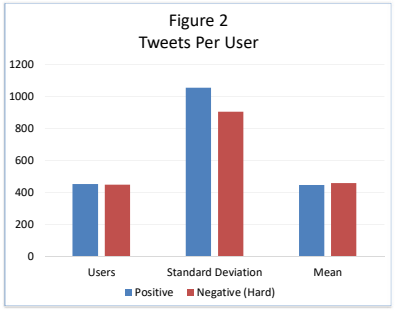


Figure 2
Tweets Per User



Data Processing

Tweets were stripped of all metadata other than time and converted into ASCII to ensure the model did not learn to identify Tweets based on unintended artifacts in the data. They were then organized by user and time and split into groups of 20. Section tags were added, the groups were padded for input into the CNN, and then entered as a character vector.

Example of a group of two:
<S> <ST>11:10<ET><SW>Tweet text<EW> <ST>11:15<ET><SW>RT @User Tweet text<EW> <E>>>>>

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 its own character vector and the previous node. See Figure 4.

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 4.

Figure 3

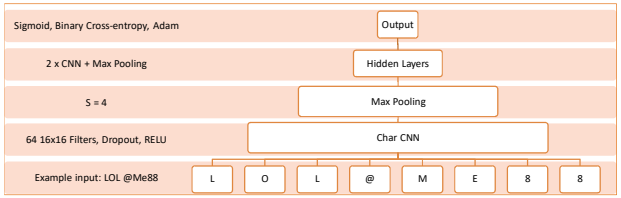
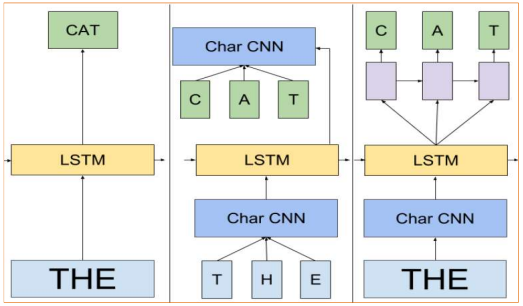


Figure 4 (Jozefowicz, Rafal, et al.)



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.

Results (Hard)

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.

Figure 5
Easy Dataset

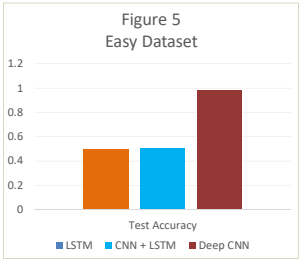
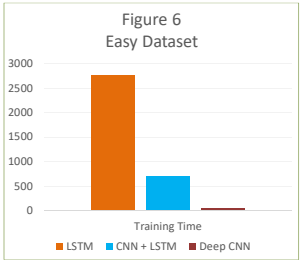


Figure 6
Easy Dataset



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

18-cr-00032-DLF. United States District Court. 22 Feb. 2018. Internet Research Agency Indictment. United States Department of Justice, n.d. Web. 15 Apr. 2018.
Popken, Ben. "Twitter Deleted Russian Troll Tweets. So We Published More than 200,000 of Them." NBCNews.com. NBCUniversal News Group, 14 Feb. 2018. Web. 14 Apr. 2018.
Jozefowicz, Rafal, et al. "Exploring the limits of language modeling." arXiv preprint arXiv:1602.02410 (2016)
"USA - Free Twitter Dataset. 300,000 Free USA Tweets." Followthehashtag. DNOiSE. Web. 2 May 2018.
King, Ed. "Election Day Tweets." Countries of the World. Kaggle. 26 Nov. 2016. Web. 10 May 2018.