PROJECT PROPOSAL

Unmasking Misinformation Bots on Twitter using Deep Learning

Jamison MacFarland

Today, social bots make up between 9-15% of all Twitter accounts. This is a direct threat to democracy, as an explicit attempt to influence public opinion and sow dissent. Techniques do exist for discovering these accounts through data analysis and AI . More trivial bots are easily found, but there exists a class of bots using sophisticated algorithms to evade detection and continue to spread their propaganda (*Varol et al.)*

I plan to train an Artificial Neural Network on existing public datasets containing confirmed and suspected bot accounts.

==========

STATE SPACE

- All input nodes
  - Age of account
  - Ratio followers:followed
  - Hashtags/topics tweeted
  - Accounts tweeted at (journalists, politicians, bot networks)
  - Timing of tweets
- A number of hidden intermediary nodes
- Two output nodes, Bot or Not (with percentage score based on confidence)

==========

PROBLEM DEFINITION

The problem is partially observable, and uses a single agent. It is stochastic, as there is no factor that solely determines whether an account is counterfeit or not. This determination is based on a large number of factors, only a few of which will be measured for the sake of simplicity and time.

This ANN will be dealing with discrete observations of a somewhat-continuous world. Data points such as word choice, timing, etc. are continuous, but hashtags, names, and account network connections are not.

I do not yet know the inner workings of an Artificial Neural Net well enough to properly lay out my process going forward, but I will be in close contact with the Professor as I develop my approach.

EXISTING DATA

There exist several **high-quality datasets containing known and suspected bot accounts**, many including additional annotated data. These datasets are organized by type (e.g. political bot, retweet spambots, etc). These sets are presented in standard TSV format and are compatible with many common analysis tools in Python.

The end product of this project will be an **easy-to-use tool for constructing social network representations, gathering tweet data, and scoring accounts on a percentage scale of Bot or Not.** This will take an account ID and attempt to discover whether or not it is a political propaganda bot.

This result will be manually evaluated for suspicious activity, as well as submitted to the https://botometer.iuni.iu.edu bot checker en masse. **This will give a total accuracy score, and allow evaluation of the ANN's algorithm for correctness and accuracy of prediction.**

==========

COMPONENTS AND SCHEDULE

DATA INTAKE/NORMALIZATION (weeks 1-2):

My input datasets only contain Twitter account IDs. It will be necessary to produce a data-collection scraper to gather user metadata and tweet history for each of the input IDs, and present this data in a format that can be analyzed and used as input for the ANN. In addition, the project will require negative data (real, normal human accounts) to reduce false positives.

NEURAL NET (weeks 2-3):

This section will consist of producing a neural net catered to the input data, and training the net on both positive and negative bot datasets.

==========

REFERENCES

Alessandro Balestrucci, Rocco De Nicola, Marinella Petrocchi and Catia Trubiani, Do You Really Follow Them? Automatic Detection of Credulous Twitter Users, Intelligent Data Engineering and Automated Learning – IDEAL 2019, 10.1007/978-3-030-33607-3_44, (402-410), (2019).

Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in online social networks. *Computer Communications*, 36(10–11), 1120– 1129.

Broniatowski, D. A., Jamison, A. M., Qi, S., AlKulaib, L., Chen, T., Benton, A., … Dredze, M. (2018). Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American Journal of Public Health*, 108(10), 1378– 1384.

Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, Alessandro Flammini. Online Human-Bot Interactions: Detection, Estimation, and Characterization. ICWSM 2017