Bot Recognition of Tweets in Political and Nonpolitical Contexts

William Johnson and Andrew Hatfield

Overview

Broad Goals

Are LMs able to categorize tweets as human or bot-made when trained on similar types of tweets?

This is an interesting task as apps like Twitter / X need to be able to identify and remove bots who can potentially incite violence or spread propaganda

Specific Question

How successful are LMs trained on political or non-political tweets in identifying and classifying these same types of tweets?

Do models trained solely on political bot tweets and general human tweets perform better than a model trained more generally on both political and non-political tweets?

Our aim is to see whether bot detectors trained on more specific contexts (i.e. political) perform better than more generally trained models. While this experiment is evaluated on both a masked and causal model, future experiments would need to extrapolate these results to see if they generalize to other tweet databases and "categories"

Methods and Data

Both human and bot tweets (not labeled as political or non-political) are taken from Twitter Bot Detection Project (TBD)

Politically and non-politically labeled bot tweets are taken from Golbeck et al. (2021)

Three models are separately fine-tuned on Distilbert and Distilgpt2 to generalize results across masked and causal model types

One model is trained on political bot tweets, a second on non-political bot tweets, and a third on both political and non-political bot tweets from Golbeck et al. Each of these models are trained on 35,000 bot tweets (from Golbeck) and the same set of 35,000 human tweets from TBD

Methods and Data (2)

5,000 human tweets and 5,000 bot tweets are used for validation of parameters for each model

For evaluation, each model is evaluated on the same 5,000 political bot tweets, 5,000 nonpolitical bot tweets, 5,000 bot tweets from TBD (without categorical labels), and 5,000 human tweets from TBD. TBD bot tweets were used to assess generalization of bot detection across datasets for each model

Tweets from Golbeck et al. were originally organized by user (with bot or human labels)

These tweets were aggregated by type and training examples were randomized and scrambled in order to eliminate user-specific bias

All models were trained on the same set of human tweets from TBD and no models were trained on TBD bot tweets in order to assess generalization across tweet databases (which may take tweets from different time horizons, regions, etc.)

The combined model is trained on a subset of the exact same tweets utilized in the training of the other two models so performance differences are simply due to pattern dilution

Distilbert Results

Model	Data	Precision	Recall	F1	Accuracy
combined	nan	0.716	0.68	0.667	**0.681**
combined	nonpolitical	1.0	0.86	0.92	**0.86**
combined	political	1.0	0.846	0.916	**0.846**
nonpolitical	nan	0.719	0.67	0.65	**0.6761**
nonpolitical	nonpolitical	1.0	0.8694	0.93	**0.8694**
nonpolitical	political	1.0	0.6172	0.763	**0.6172**
political	nan	0.698	0.6596	0.641	**0.6596**
political	nonpolitical	1.0	0.7396	0.85	**0.7396**
political	political	1.0	0.8714	0.931	**0.8714**

Future Considerations

In order to assess result generalization across model types, the experiment will be repeated with distilgpt2

This will allow for a comparison between Distilbert (a masked model) with Distilgpt2 (a causal model)

Pre-processing of data was a complex process because of the different domains of training and evaluation tweets

We will also re-run the experiment, separating the “nan” label for TBD evaluation tweets into human and bot categories in order to establish more concrete and visual comparisons between the domains