Distinguishing Disruptive Trolls from Automated Bots[⋆]

Joshua Uyheng[1][0000−0002−1631−6566] and Kathleen M. Carley[1][0000−0002−6356−0238]

CASOS Center, Institute for Software Research
Carnegie Mellon University, Pittsburgh PA 15213, USA
{juyheng,kathleen.carley}@cs.cmu.edu

Over the past decade, the vast literature on digital disinformation has highlighted the important role played by specialized agents in the spread of falsehoods, hate speech, and other distortions of free and open conversation in cyberspace. Inorganic actors such as bots and trolls have been studied extensively in the recent literature, spanning predictive modelling of digital trace data to detect inauthentic accounts, observational studies that quantify their activity and impacts in the context of coordinated information operations or in digital conversations more generally, and even human studies of personality traits that predict pathological participation in online social networks. A diverse menagerie of other such agent types has likewise been documented, such as sybils, cyborgs, astroturfers, buzzers, and state-sponsored accounts.

In many of these cases, the use of distinct terminologies has often been predicated on prevailing cultural or political contexts. However, their translation into etic computational modelling efforts has in various scenarios resulted in a range of conceptual confusions. Ambiguities may arise regarding what aspects of inauthentic behavior are empirically measurable in the first place; how to measure them; and which measures are identical, correlated, or otherwise unique.

In this paper, we posit and verify an orthogonal framework that may be used to organize the diversity of disinformation agents. Our framework is structured by behavioral dimensions exemplified by two exemplary types of disinformation agents: bots and trolls. We specifically define bot behavior in terms of *automation*, and troll behavior in terms of *disruption*. We demonstrate the utility of this taxonomic scheme across a series of studies, highlighting convergent evidence from statistical, predictive, and practical perspectives. By advancing this two-construct framework of agent categorization, we advocate a data-driven lens

[⋆] This work was supported in part by the Knight Foundation and the Office of Naval Research grants N000141812106 and N000141812108. Additional support was provided by the Center for Computational Analysis of Social and Organizational Systems (CASOS) and the Center for Informed Democracy and Social Cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research or the U.S. government. We thank Xuanye Li and Max Dunaevschi for assistance in data collection and processing.

for operationalizing the diversity of disinformation behaviors and delineating key empirical signals for computational modelling.

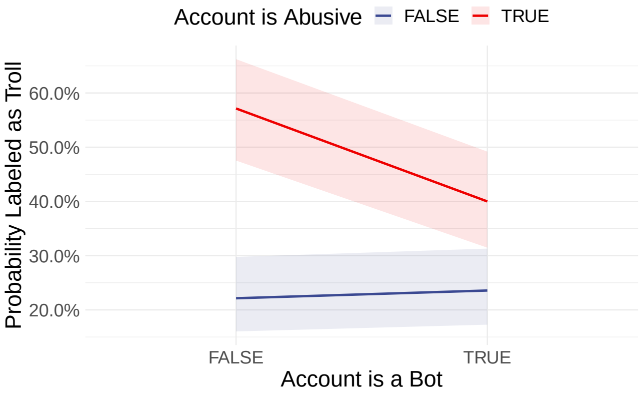

Fig. 1. Predicted likelihood that an account is manually classified as a troll, given that it is predicted to be a bot and whether it uses abusive language. Bands represent 95% confidence intervals.

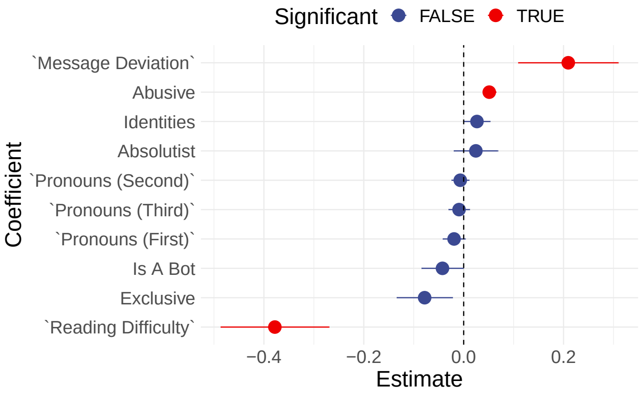

Fig. 2. Coefficients of a logistic regression model that predicts the likelihood that a message is troll-like. Estimates that are statistically significant at $\alpha = .05$ are colored red. Error bars represent 95% confidence intervals.

We first show with statistical significance that manually labeled troll messages are not more likely to come from bots; in fact, abusive human-like accounts are more likely to be trolls (Figure 1). A regression analysis with additional psycholinguistic covariates further revealed that troll-like messages are strongly as-

Table 1. Performance of individual-based and relational-based machine learning and deep learning models in troll prediction task. Bolded values are the highest achieved performance metrics for individual and relational models, while underlined values are the second highest.

Model	Individual		Relational	
	Accuracy	F1	Accuracy	F1
Logistic Regression	0.5883	0.6037	0.5893	0.5957
Random Forest	0.6468	0.6298	0.6817	0.6500
SVM	0.6499	0.5859	0.6797	0.5910
LSTM	**0.6809**	**0.6584**	**0.6839**	**0.6771**

sociated with abusive language, low reading difficulty, and off-topic interactions, yet again independent of the likelihood that they come from bots (Figure 2).

Finally, a modelling study shows the utility of relational modelling using both machine learning and deep learning approaches in troll prediction, as summarized in Table 1. The gains in relational modelling signal the importance of context in determining trolling behavior in particular, and disruption as a concept more generally. In addition, the small performance disparity between handcrafted psycholinguistic features in the machine learning setup and the high-dimensional embedding used in the deep learning setup also indicate the relative strength of these socio-scientifically grounded signals at detecting disruptive behaviors.

Collectively our findings suggest the importance of empirically clarifying measurable aspects of disinformation agents of concern, and shifting from dichotomous views to seeing disinformation activities along multiple dimensions. We conclude with directions for further research in this area alongside considerations for platform regulation and policy-making to curtail the coordinated spread of disinformation.