A Compute and Environment

We estimate roughly 250 model training commands were executed for this study, running for a range of 10-120 minutes each, nearly all on a single local NVIDIA Tesla K40 GPU. The state of Georgia generates electricity from mostly Gas (43%) and nuclear (30%) resources; 9% from renewables.[9]

B All Results on Adversarial Setup

Table 5 presents results for all λ settings tested for the trained adversary experiment performed in §4. No post-selection was made, as these are all test set results.

C TVD/JSD Tradeoff by Class

Figure 6 breaks down the scatterplots from Figure 5 into those pertaining to each data class. We note the unique, near-concave shape of the positive class in the Diabetes dataset, which is a detection-type set biased towards the negative class. This is the setting where we would expect adversarial distributions to be most difficult to find, which is confirmed by this curve.

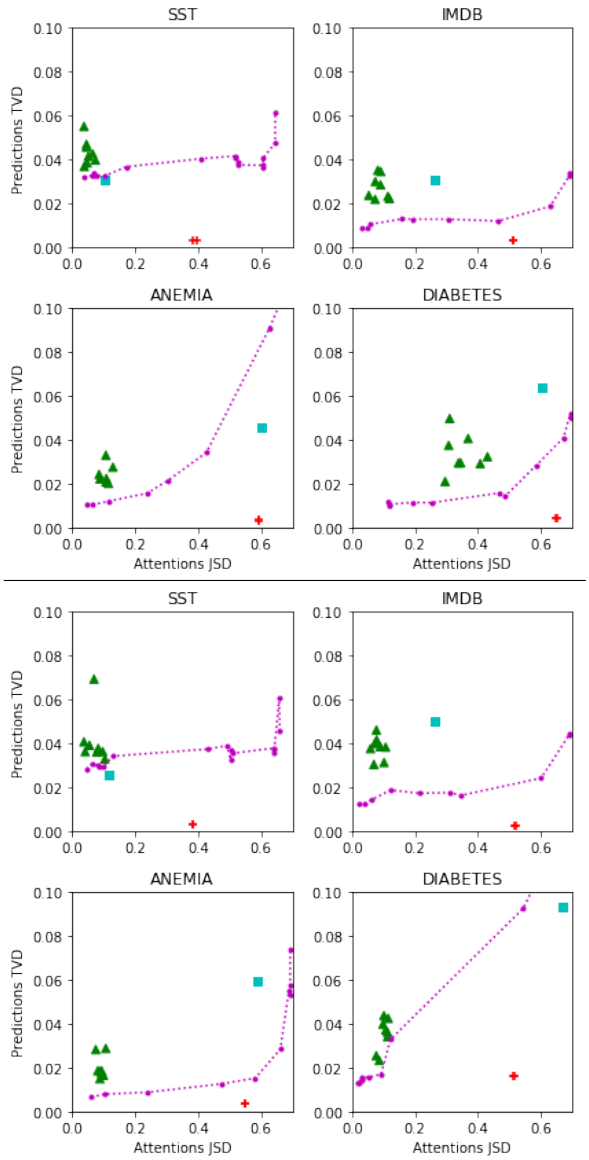

Figure 6: Per-instance test set JSD and TVD from base model on negative instances (top) and positive instances (bottom). ▲: random seed; ■: uniform weights; dotted line: our adversarial setup; +: adversarial setup from Jain and Wallace (2019).

[9] https://www.eia.gov/state/data.php?sid=GA#EnergyIndicators

Dataset	λ	F1	TVD	JSD	Dataset	λ	F1	TVD	JSD
Anemia	0	0.936	0.008	0.056	Diabetes	0	0.779	0.012	0.098
	1e-4	0.937	0.009	0.090		1e-5	0.769	0.011	0.098
	2e-4	0.938	0.010	0.194		2e-5	0.770	0.011	0.098
	3.5e-4	0.936	0.014	0.387		4e-5	0.780	0.012	0.162
	5e-4	0.942	0.017	0.481		5e-5	0.781	0.012	0.209
	0.001	0.938	0.030	0.576		1e-4	0.781	0.016	0.385
	0.002	0.895	0.068	0.666		2e-4	0.775	0.015	0.409
	0.004	0.888	0.074	0.690		5e-4	0.759	0.029	0.494
	0.005	0.875	0.079	0.692		0.001	0.690	0.051	0.646
	0.01	0.872	0.086	0.693		0.005	0.645	0.067	0.693
						0.01	0.643	0.069	0.693
SST	0	0.816	0.032	0.075	IMDb	0	0.906	0.011	0.043
	1e-5	0.822	0.030	0.042		1e-4	0.907	0.011	0.027
	5e-5	0.824	0.031	0.080		2e-4	0.906	0.012	0.059
	1e-4	0.823	0.032	0.064		4e-4	0.906	0.015	0.202
	5e-4	0.828	0.031	0.100		5e-4	0.905	0.016	0.141
	5.2e-4	0.827	0.036	0.150		7e-4	0.902	0.015	0.309
	5.25e-4	0.823	0.036	0.514		8e-4	0.906	0.014	0.405
	5.35e-4	0.814	0.039	0.420		0.001	0.905	0.022	0.615
	5.5e-4	0.809	0.040	0.505		0.005	0.888	0.038	0.691
	6e-4	0.813	0.039	0.513		0.01	0.885	0.039	0.691
	7.5e-4	0.811	0.037	0.518					
	0.001	0.815	0.038	0.623					
	0.01	0.821	0.036	0.624					
	0.1	0.811	0.047	0.653					
	0.5	0.799	0.061	0.652					
	1	0.819	0.039	0.624					

Table 5: All results for the Adversarial Setup. $\lambda = 0$ denotes a model where only minimum TVD is sought. SST models were trained for 80 epochs with best epoch selected based on loss objective over the test set; all other models for 40 epochs.