Learning to Adapt Domain Shifts of Moral Values via Instance Weighting

Xiaolei Huang[1], Alexandra Wormley[2], Adam Cohen[2]

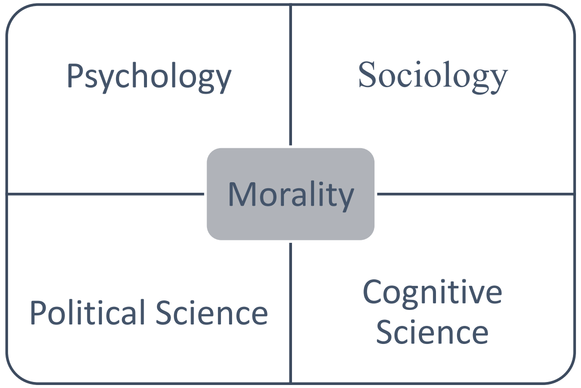
xiaolei.huang@memphis.edu

1. University of Memphis, 2. Arizona State University

Morality

- Increasing studies have extracted moral values in user-generated text from social media in understanding community cultures and interpreting user behaviors of social movements.

Moral Foundations[1] → Moral values	
Virtue 🤗	Vice 👿
authority	subversion
care	harm
fairness	cheating
loyalty	betrayal
purity	degradation



Psychology

Sociology

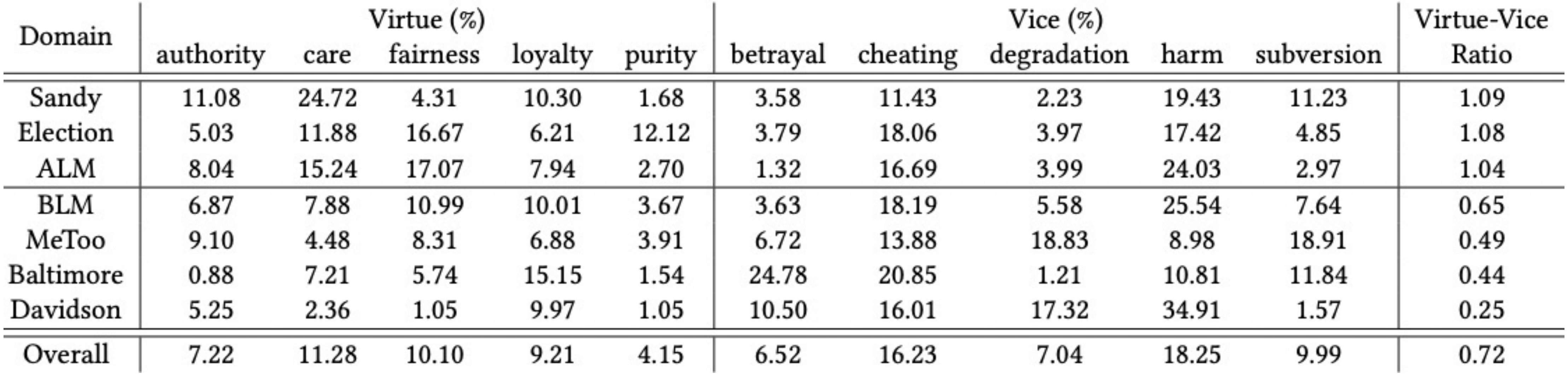
Morality

Political Science

Cognitive Science

[1] Jonathan Haidt. 2007. The New Synthesis in Moral Psychology. Science 316, 5827 (may 2007), 998–1002. https://doi.org/10.1126/science.1137651

Morality Data of Social Topics[2]



Domain	Virtue (%)					Vice (%)					Virtue-Vice
	authority	care	fairness	loyalty	purity	betrayal	cheating	degradation	harm	subversion	Ratio
Sandy	11.08	24.72	4.31	10.30	1.68	3.58	11.43	2.23	19.43	11.23	1.09
Election	5.03	11.88	16.67	6.21	12.12	3.79	18.06	3.97	17.42	4.85	1.08
ALM	8.04	15.24	17.07	7.94	2.70	1.32	16.69	3.99	24.03	2.97	1.04
BLM	6.87	7.88	10.99	10.01	3.67	3.63	18.19	5.58	25.54	7.64	0.65
MeToo	9.10	4.48	8.31	6.88	3.91	6.72	13.88	18.83	8.98	18.91	0.49
Baltimore	0.88	7.21	5.74	15.15	1.54	24.78	20.85	1.21	10.81	11.84	0.44
Davidson	5.25	2.36	1.05	9.97	1.05	10.50	16.01	17.32	34.91	1.57	0.25
Overall	7.22	11.28	10.10	9.21	4.15	6.52	16.23	7.04	18.25	9.99	0.72

Distributions of moral values across social topics (domains). The virtue-vice ratio is a divide operation between virtue-related and vice-related moral value counts. Overall, the data shows higher ratio of the vice-related moral values.

Significant variations of moral values exist across 7 domains.

- degradation and subversion → top 2 in the MeToo.
- cheating and harm → top 2 in the Election.

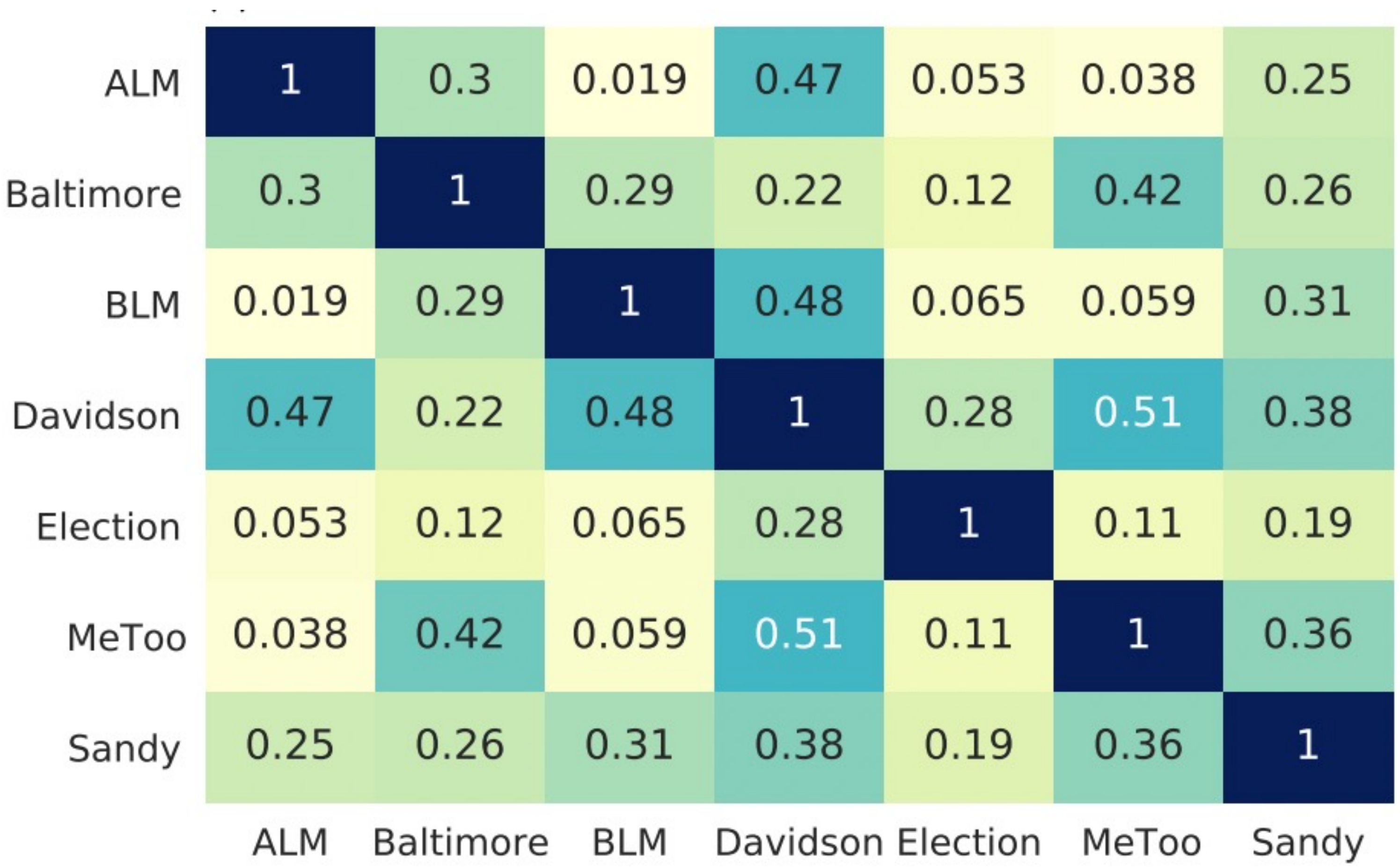
[2] Hoover, Joe, et al. "Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment." *Social Psychological and Personality Science* 11.8 (2020): 1057-1071.

Moral Variation Qualification

Top tokens in each social movement domain			
ALLLivesMatter	BlackLivesMatter	Election	MeToo
people, god, justice, police, love, black, respect, human, racist, violence	blm, injustice, solidarity, police, black, people, ferguson, iuic, respect, blacktwitter	realdonaldtrump, trump, potus, president, justice, gop, maga, america, obey, donaldtrump	sandy, hurricanesandy, liberty, hurricane, holy, bitch, frankenstorm, god, love, obama

The qualitative results show the most predictable words (ranked by mutual information) towards the 10 moral values. We notice that the top features may reflect the societal and cultural variations of the social issues.

Moral Variation Quantification

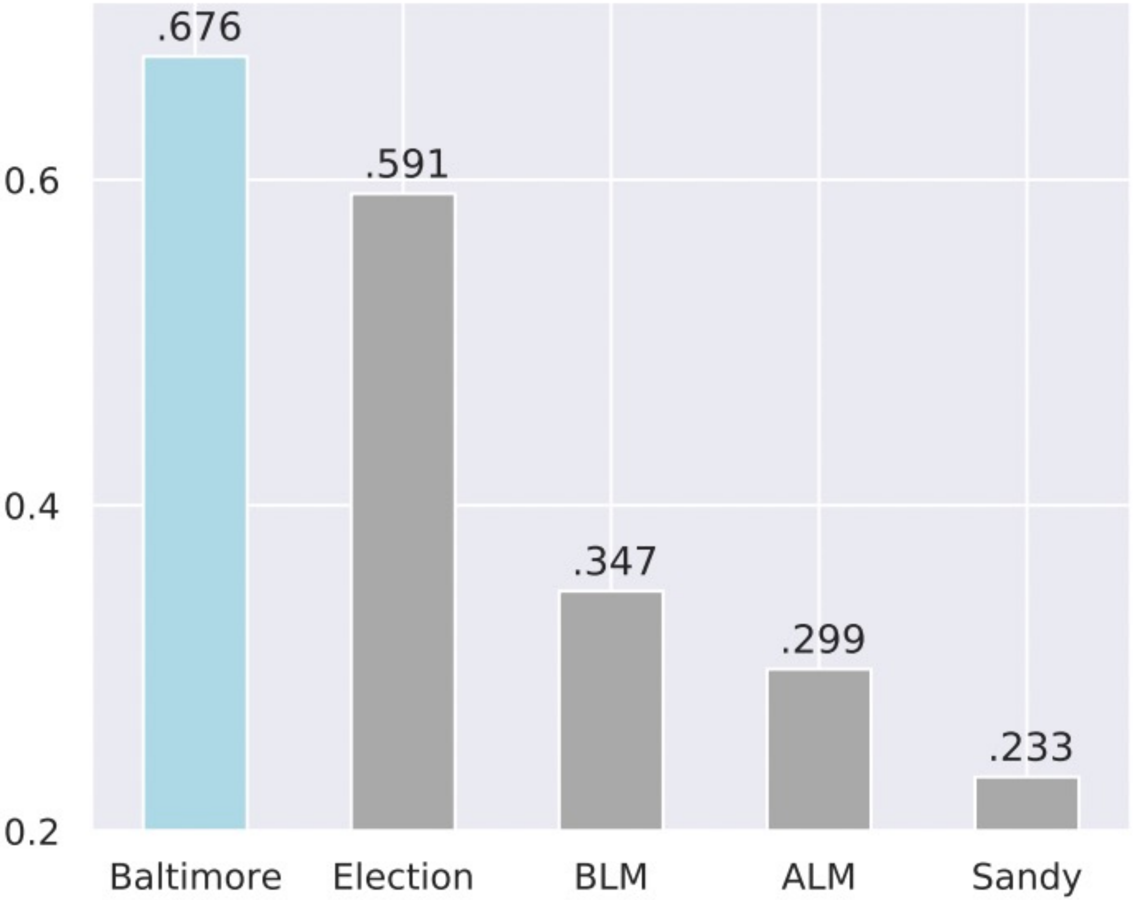


	ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
ALM	1	0.3	0.019	0.47	0.053	0.038	0.25
Baltimore	0.3	1	0.29	0.22	0.12	0.42	0.26
BLM	0.019	0.29	1	0.48	0.065	0.059	0.31
Davidson	0.47	0.22	0.48	1	0.28	0.51	0.38
Election	0.053	0.12	0.065	0.28	1	0.11	0.19
MeToo	0.038	0.42	0.059	0.51	0.11	1	0.36
Sandy	0.25	0.26	0.31	0.38	0.19	0.36	1

- Extract linguistic patterns of language usage by a topic model[3], which abstracts language usage into thematic vectors.
- Compute cosine similarities of the topic distributions between every two domains.
- The low similarities indicate language regarding moral values shift significantly across domains.

[3] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

Moral Variation Impact on Classification

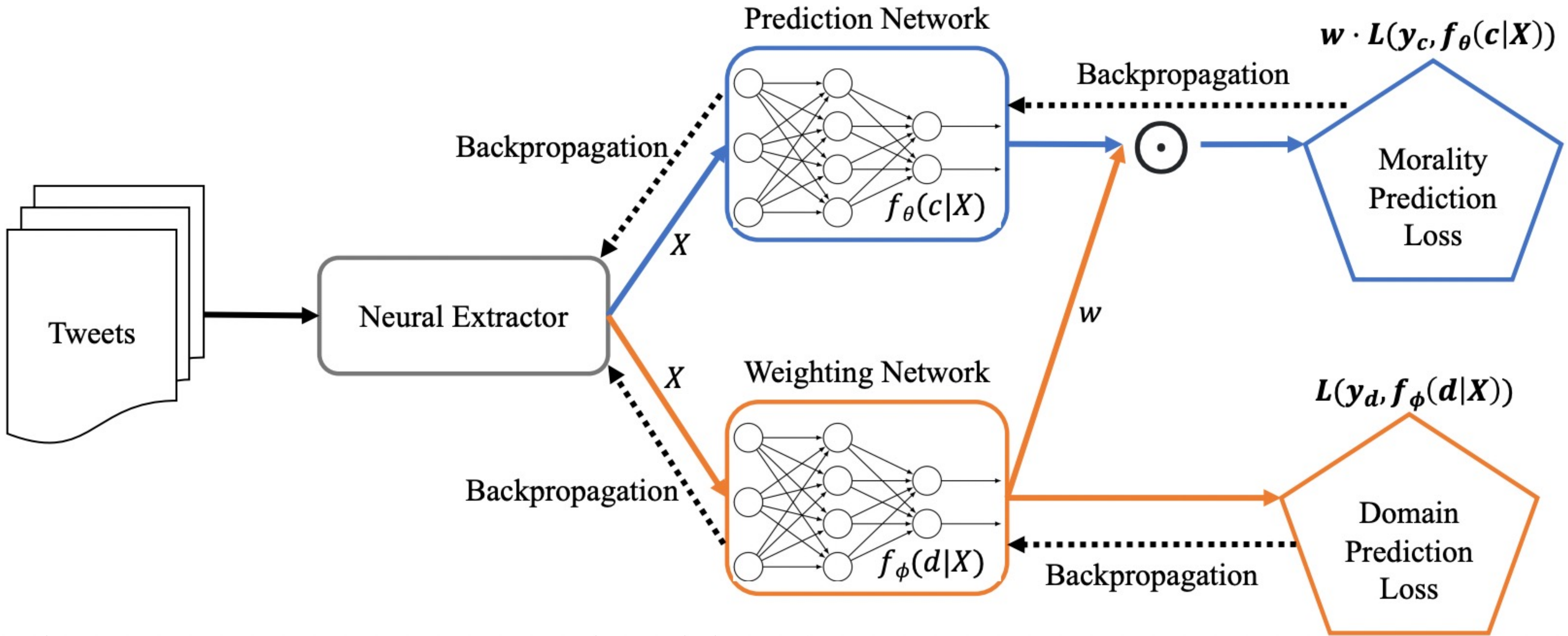


In-domain (Baltimore in light blue) vs. Cross-domain (others in grey) classification performance.

- Split 80% of documents as the training set and hold out 20% of documents as the testing set for each domain corpus.
- Build a logistic regression with n(1-,2-,3-)-gram features.
- Evaluate the classifier across each domain’s test set using the F1 score.

Moral variations across social movements can impact morality classifier performance when training and testing sets of classifiers are from different social movements.

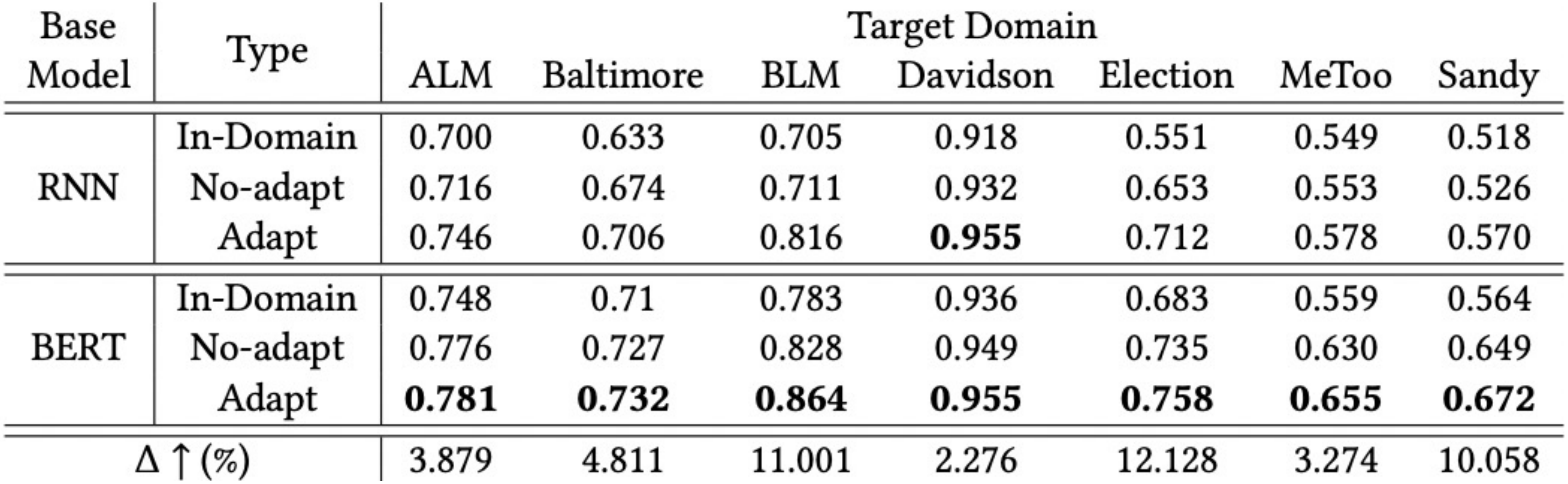
Adaptation Framework via Instance Weighting[4]



1. Neural Feature Extractor: treats neural models as feature extractors.
2. Prediction Network: takes the document representations to predict moral values.
3. Weighting Network: dynamically adapts the domain shifts of language and moral values.
4. Joint Optimization: jointly trains both prediction and weighting networks by the two losses

[4] Jiang, Jing, and ChengXiang Zhai. "Instance weighting for domain adaptation in NLP." ACL, 2007.

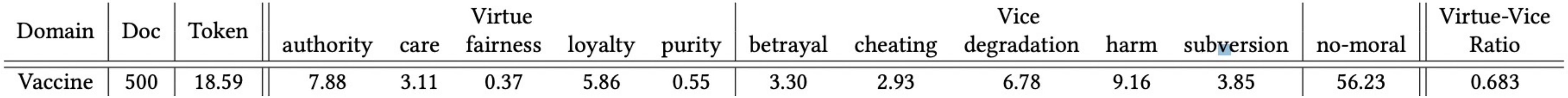
Results on Classification



Base Model	Type	Target Domain						
		ALM	Baltimore	BLM	Davidson	Election	MeToo	Sandy
RNN	In-Domain	0.700	0.633	0.705	0.918	0.551	0.549	0.518
	No-adapt	0.716	0.674	0.711	0.932	0.653	0.553	0.526
	Adapt	0.746	0.706	0.816	**0.955**	0.712	0.578	0.570
BERT	In-Domain	0.748	0.71	0.783	0.936	0.683	0.559	0.564
	No-adapt	0.776	0.727	0.828	0.949	0.735	0.630	0.649
	Adapt	**0.781**	**0.732**	**0.864**	**0.955**	**0.758**	**0.655**	**0.672**
Δ ↑ (%)		3.879	4.811	11.001	2.276	12.128	3.274	10.058

The improvements indicate that adapting domain shifts can improve morality classifiers and demonstrate that our approach can effectively adapt shifts in moral values and language usage across social events.

COVID-19 Case Study



Domain	Doc	Token	Virtue					Vice					no-moral	Virtue-Vice Ratio
			authority	care	fairness	loyalty	purity	betrayal	cheating	degradation	harm	subversion		
Vaccine	500	18.59	7.88	3.11	0.37	5.86	0.55	3.30	2.93	6.78	9.16	3.85	56.23	0.683

- We randomly sampled 500 vaccine-related tweets from a public repository that has retrieved COVID-19 data[5] using the Twitter streaming API since 2020.
- Different from the existing 7 domains of social movements: 1) authority and loyalty → top two for virtue-related morality, 2) and degradation and harm → top two for vice-related morality.
- Classification evaluations show the effectiveness of our proposed adaptation framework.

Traditional interventions of vaccine hesitancy focusing on harm and fairness moralities may not fit for the new domain of COVID-19 vaccine hesitancy.

[5] Huang, Xiaolei, et al. "Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations." Retrieved from10 5281 (2020).

Takeaways

- Language reflecting human ideologies (moral values) may shift across social movements. The variations can impact extracted feature representations and weaken morality classifiers for new target domains.

- Our approach dynamically adjusts weights on training instances from source domains regarding the new target domain and demonstrates its effectiveness on the public data as well as our new released COVID-19 data.

- The first pilot work investigating the moral values of the COVID-19 vaccine may generate alternative aspects to reduce vaccine hesitancy.

- https://github.com/xiaoleihuang/MoralCausality