Estimating Semantic Similarity between In-Domain and Out-of-Domain Samples

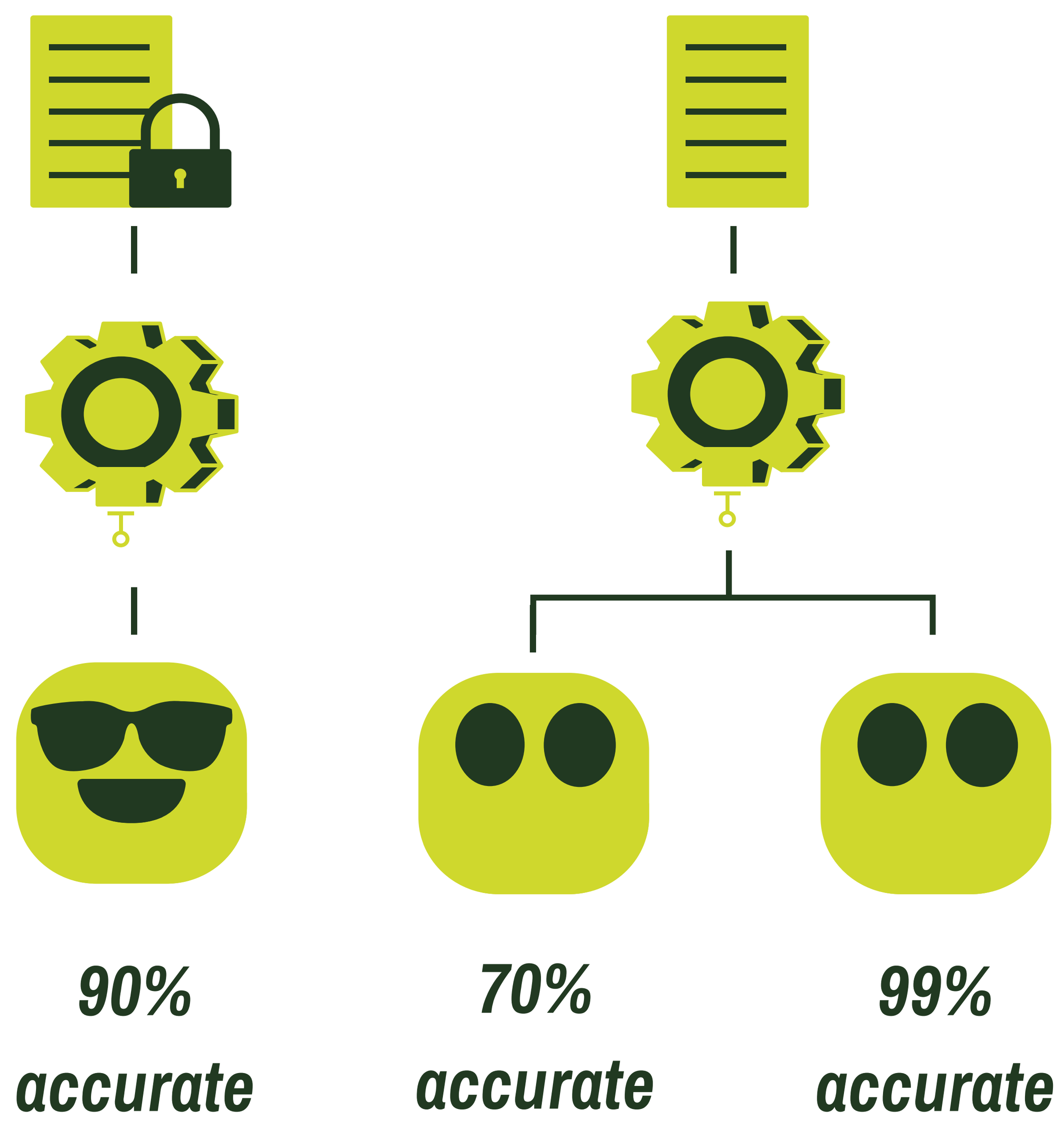
Rhitabrat Pokharel and Ameeta Agrawal

PortNLP Lab, Department of Computer Science, Portland State University

***SEM, ACL 2023, Toronto**

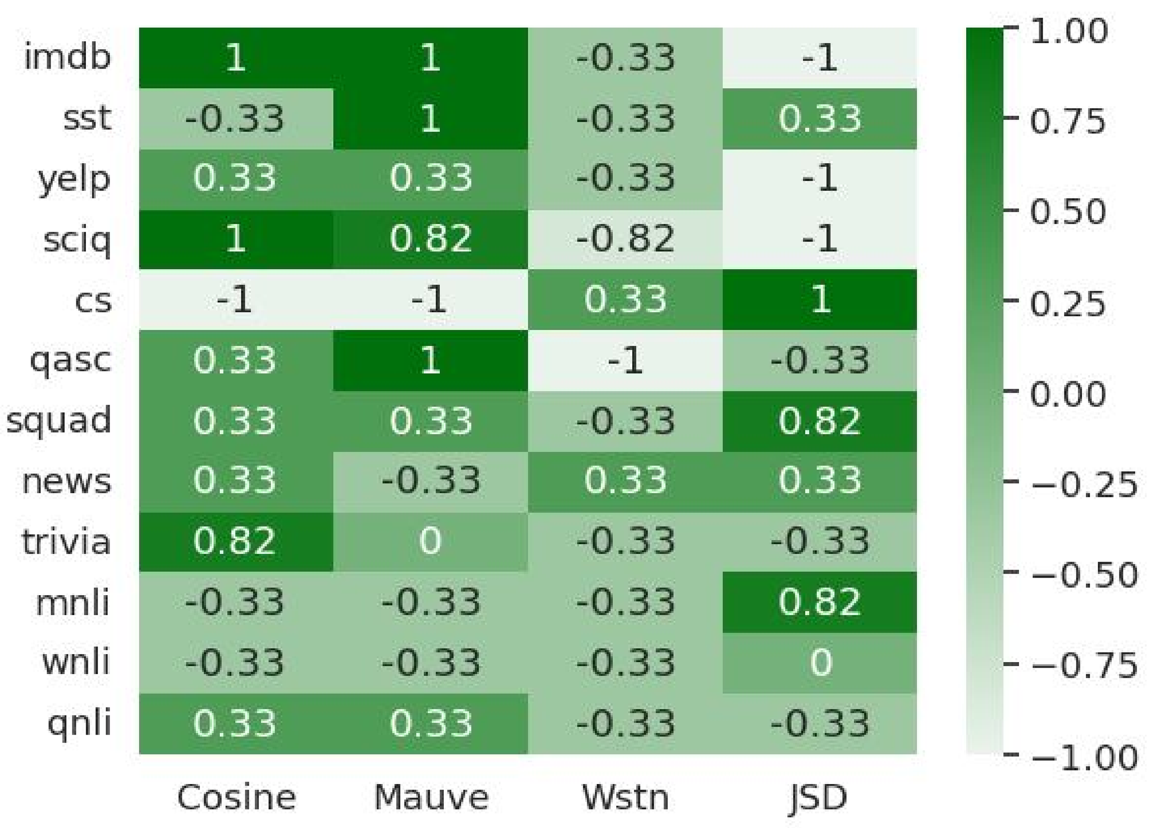
MOTIVATION

Models that demonstrate strong performance on carefully curated test/train sets may not necessarily showcase equivalent levels of effectiveness on real-world datasets.



Out-of-domain (OOD) vs Out-of-distribution (OODist)

RESULTS



Wstn and cosine show the most consistent correlation

OOD vs OODist

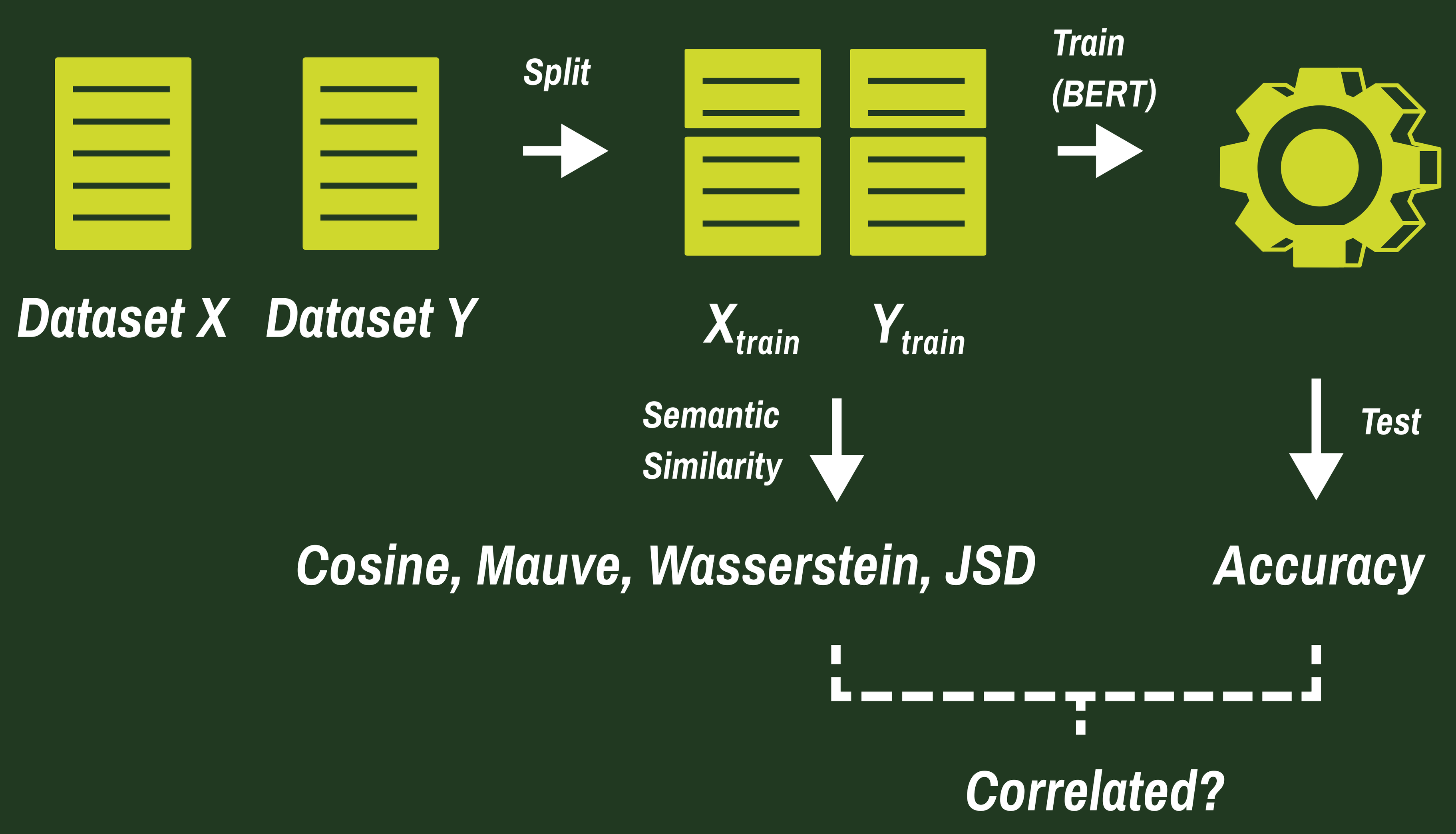
- *Data from a related but different domain[1] (Amazon vs Twitter sentiment)*
- *Different datasets for the same task[2] (SST, IMDb, and Yelp for sentiment classification)*
- *Data collected at a different time[3] maybe under different settings*
- *Datasets that are not in the training set[4]*

DATASETS

Datasets	Task
IMDb, SST2, Yelp	Sentiment Analysis
SCIQ, Commonsense, QASC	MCQ
SQUAD, News, Trivia	Extractive Question Answering
MNLI, WNLI, QNLI	Natural Language Inference

For each of train, validation (when available), and test sets, we downsampled to the size of the smallest dataset.

METHODOLOGY



CONCLUSION

- *Wasserstein could be a potential metric for determining OOD samples*
- *Model does not always perform worse on OOD samples*

CODE & PAPER

pokharel@pdx.edu

[1] Dai, Wenyuan, Gui-Rong Xue, Qiang Yang, and Yong Yu. "Co-clustering based classification for out-of-domain documents." [2] Chrysostomou, George, and Nikolaos Aletras. "An empirical study on explanations in out-of-domain settings." [3] Ovadia, Yaniv, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift." [4] Lin, Bill Yuchen, Sida Wang, Xi Victoria Lin, Robin Jia, Lin Xiao, Xiang Ren, and Wen-tau Yih. "On continual model refinement in out-of-distribution data streams."