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

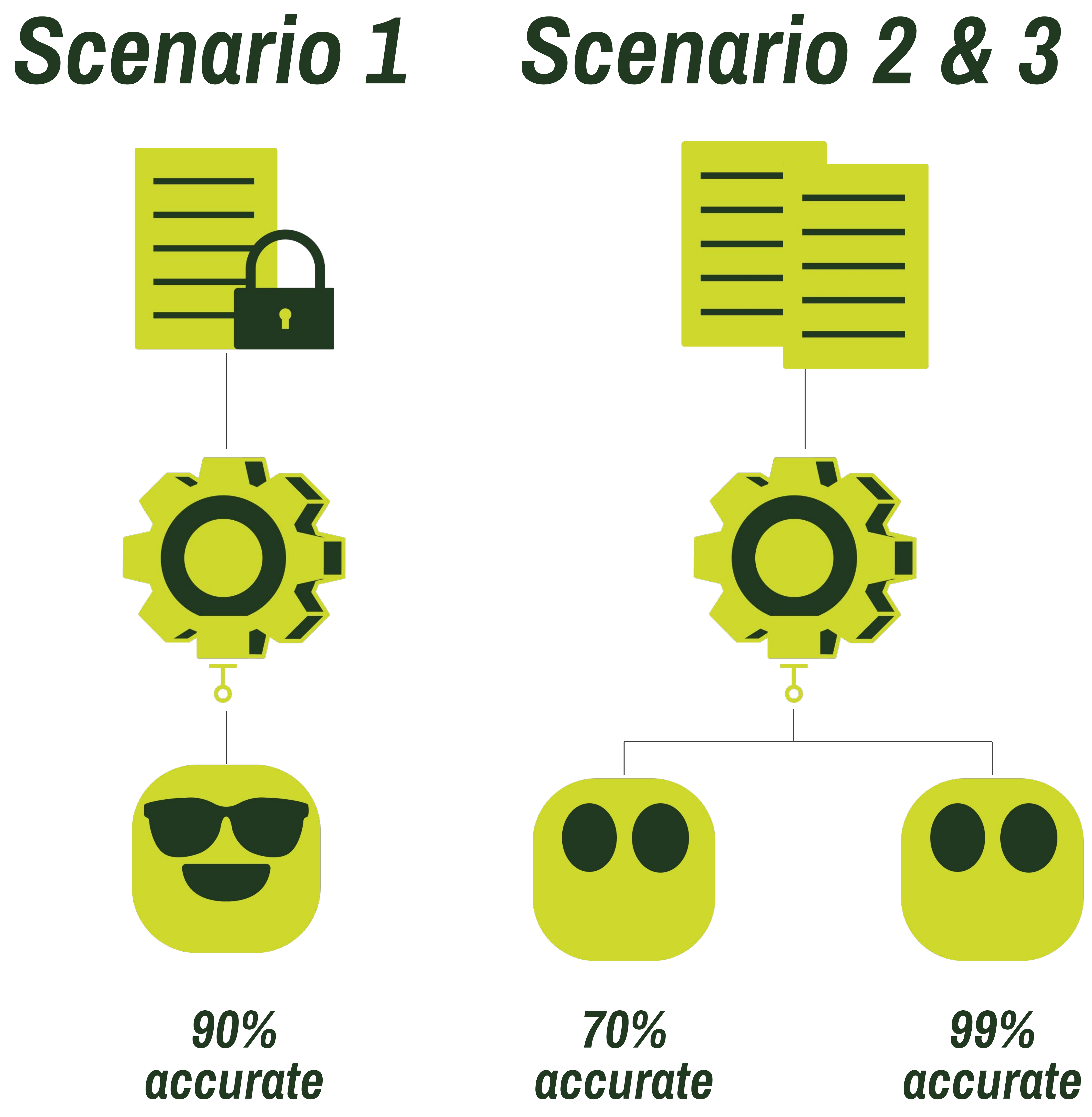
Rhitabrat Pokharel
Ameeta Agrawal

PortNLP Lab
Department of Computer Science
Portland State University

***SEM, ACL 2023**
Toronto

Model Performance on Unseen Data

- Models that demonstrate strong performance on carefully curated test/train sets may not necessarily showcase equivalent levels of effectiveness on real-world datasets.
- In real-world scenarios, false predictions or misclassified results by machine learning models can have severe consequences[1].



[1] Gokhale, Tejas, Swaroop Mishra, Man Luo, Bhavdeep Singh Sachdeva, and Chitta Baral. "Generalized but not robust? comparing the effects of data modification methods on out-of-domain generalization and adversarial robustness."

Outline of the Presentation

- Introduction & Related Work
- Problem Description
- Datasets
- Methodology
- Results and Discussion
- Conclusion

INTRODUCTION & RELATED WORK

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

- sometimes interchangeably, other times to mean different things

OOD

- Data from a related but different domain[2] (Amazon vs Twitter sentiment)
- Different datasets for the same task[3] (SST, IMDb, and Yelp for sentiment classification)

OODist

- Data collected at a different time[4] maybe under different settings
- Datasets that are not in the training set[5]

[2] Dai, Wenyuan, Gui-Rong Xue, Qiang Yang, and Yong Yu. "Co-clustering based classification for out-of-domain documents." [3] Chrysostomou, George, and Nikolaos Aletras. "An empirical study on explanations in out-of-domain settings."
[4] 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."
[5] 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."

Usage of the terms OOD and OODist under different Scenarios

A = train set is from one dataset, and the test set from another dataset

B = train and test sets are two subsets of the same dataset

C = a combination of both A and B

Paper	Setup	Term	Metrics	Task
Chrysostomou and Aletras (2022)	A	OOD	-	Sentiment classification
Le Berre et al. (2022)	A	OOD	Accuracy	MCQ
Lin et al. (2022)	A	OODist	-	Extractive QA
Nejadgholi et al. (2022)	A	OOD	AUC, F1	Sentiment classification
Chiang and Lee (2022)	A	OODist	Cosine similarity, Confidence score, Probability distribution	Sentiment classification
Mishra and Arunkumar (2022)	A	OODist	NLI diagnostics	NLI
Varshney et al. (2022)	A	OOD	Accuracy	NLI, Duplicate detection, Sentiment analysis, MCQ, Commonsense Reasoning
Omar et al. (2022)	A	OODist	Accuracy, Success rate, Error rate, Diversity, Fairness, IBP tightness, Robustness	Classification, Paraphrasing, NLI
Adila and Kang (2022)	A	OODist	Confidence, Variability	NLI
Singhal et al. (2022)	A	OOD	Accuracy	NLI, Phrase identification
Agrawal et al. (2022)	A	OOD	Accuracy	Visual QA
Aghazadeh et al. (2022)	A, B	OODist	Accuracy	Metaphorical knowledge
Chen et al. (2023)	A, B	OODist	Accuracy	Sentiment analysis, Toxicity detection, News Classification, Dialogue Intent Classification
Mai et al. (2022)	B	OODist	-	Anomaly detection
Garg et al. (2022)	B	OOD	Accuracy	Rating generation, Toxicity classification
Jin et al. (2021)	B	OOD	False Positive Ratio, AUROC, AUPR	Text Classification
Atwell et al. (2022)	C	OOD	h-discrepancy	Discourse parsing
Gokhale et al. (2022)	C	OOD	Accuracy, EM	NLI, QA, Image classification

Existing works on OOD/OODist

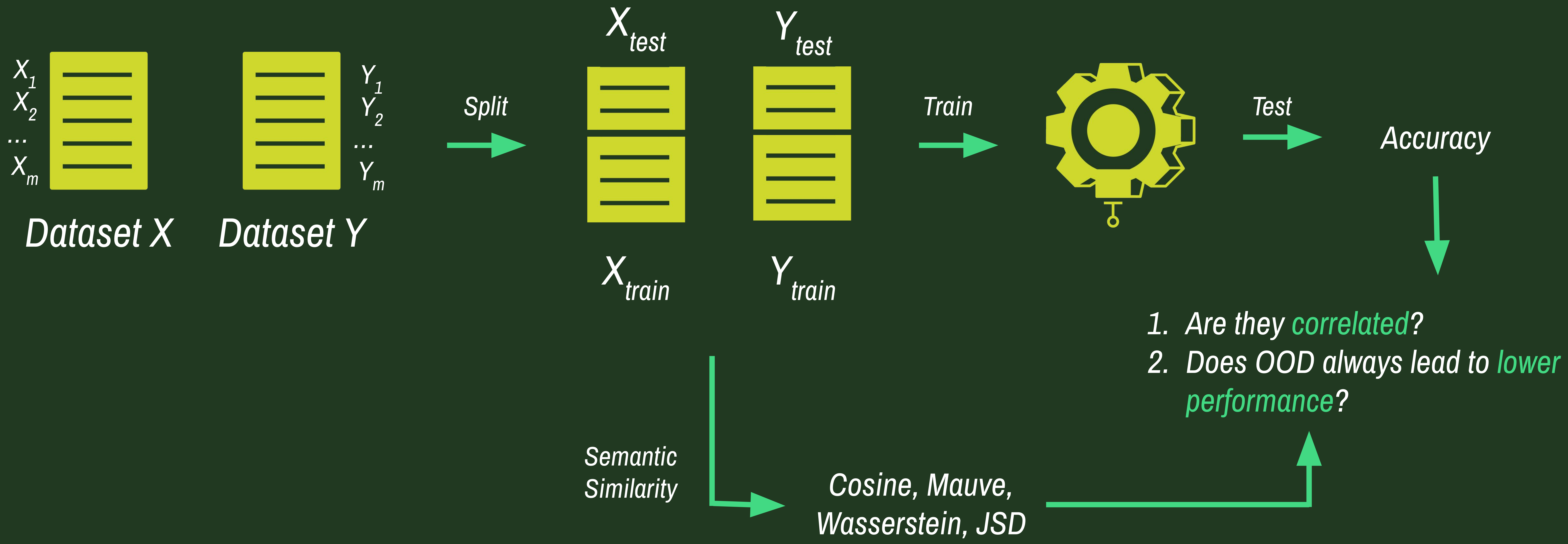
1. Detection of OOD/OODist samples (a vast majority of work)
2. Generalization: improving the performance of a model for OOD samples
3. Study of the types of OODist shifts
4. Various metrics that have been used for detection
    a. Model's accuracy
    b. Input features, hidden representation, & probability distributions of the network layers
    c. F1 & AUC scores

None of them discuss OOD/OODist detection when the model is not provided.

PROBLEM DESCRIPTION

Problem Description

- Investigate whether a trained model's performance on test set is correlated to the semantic similarity between the training data and testing data.



DATASETS

Datasets

Task	Datasets	Dataset Description
Sentiment Analysis	IMDb, SST2, Yelp	Classification of sentences into positive/negative category
Multiple Choice Question Answering	SCIQ, Commonsense QA, QASC	Given a context, choose a correct answer to a question
Extractive Question Answering	SQUAD, News, Trivia	Given a context, answer a question
Natural Language Inference	MNLI, WNLI, QNLI	Given 2 sentences, determine how they are related to eachother (neutral, entrailment, contradiction)

Data Preparation

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

For instance, all the splits of all three sentiment analysis datasets are downsampled to be of equal size. Additionally, we **balance the number of instances** for each class when possible.

Task	Datasets	train/ val/ test
Sentiment	IMDb, SST2, Yelp	3310/ 428/ 909
MCQ	SCIQ, CS, QASC	8134/ 926/ 920
Extractive QA	SQUAD, News, Trivia	61688/ -/ 4212
NLI	MNLI, WNLI, QNLI	635/ 71/ 146

METHODOLOGY

Measure of Performance and Similarity

Performance

- Finetuned the $BERT_{base}$ uncased model for 2 epochs on each X_{train}
- Tested on X_{test} and Y_{test}

Similarity

- Randomly sampled two sets of 20 instances from X_{train} and Y_{test}
- Estimated the pairwise similarity between all these samples - 400 similarity scores - averaged

For Instance

Train on	Test on	Type of data
IMDb-train	IMDb-test SST-test Yelp-test	ID OODist OODist
SST2-train	IMDb-test SST-test Yelp-test	OODist ID OODist
Yelp-train	IMDb-test SST-test Yelp-test	OODist ID OODist

Metrics

1. Performance Metrics
   - Accuracy/F1 score
2. Similarity Metrics
   - Cosine Similarity, Mauve Score, Wasserstein Distance, Jensen Shannon Distance
3. Correlation Metrics
   - Kendall Tau, Pearson

Emeddings used with the similarity metrics - word2vec

RESULTS AND DISCUSSION

Performance Evaluation

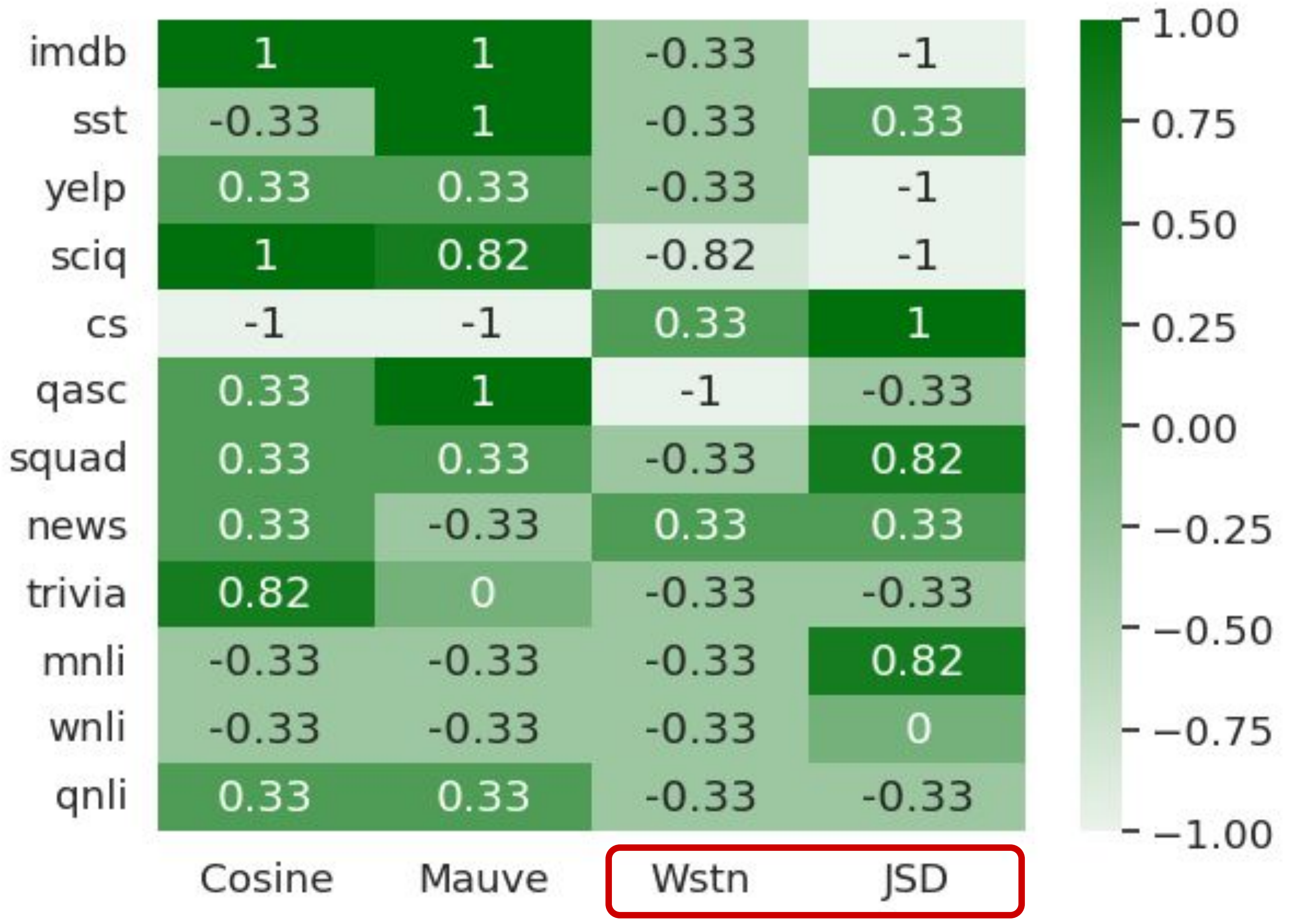
- As concluded by previous researches, the model performed better under ID settings.
- Exceptions - 3 cases
- "OOD accuracy is less than the ID accuracy" does not always hold true.

Trained on	Tested on	Performance
IMDb-train	IMDb-test	0.90
	Yelp-test	0.87
	SST2-test	0.17
SST2-train	SST2-test	0.89
	IMDb-test	0.21
	Yelp-test	0.16
Yelp-train	Yelp-test	0.93
	IMDb-test	0.86
	SST2-test	0.19
SCIQ-train	SCIQ-test	0.64
	QASC-test	0.18
	CS-test	0.34
CS-train	CS-test	0.49
	SCIQ-test	0.58
	QASC-test	0.84
QASC-train	QASC-test	0.92
	SCIQ-test	0.51
	CS-test	0.48

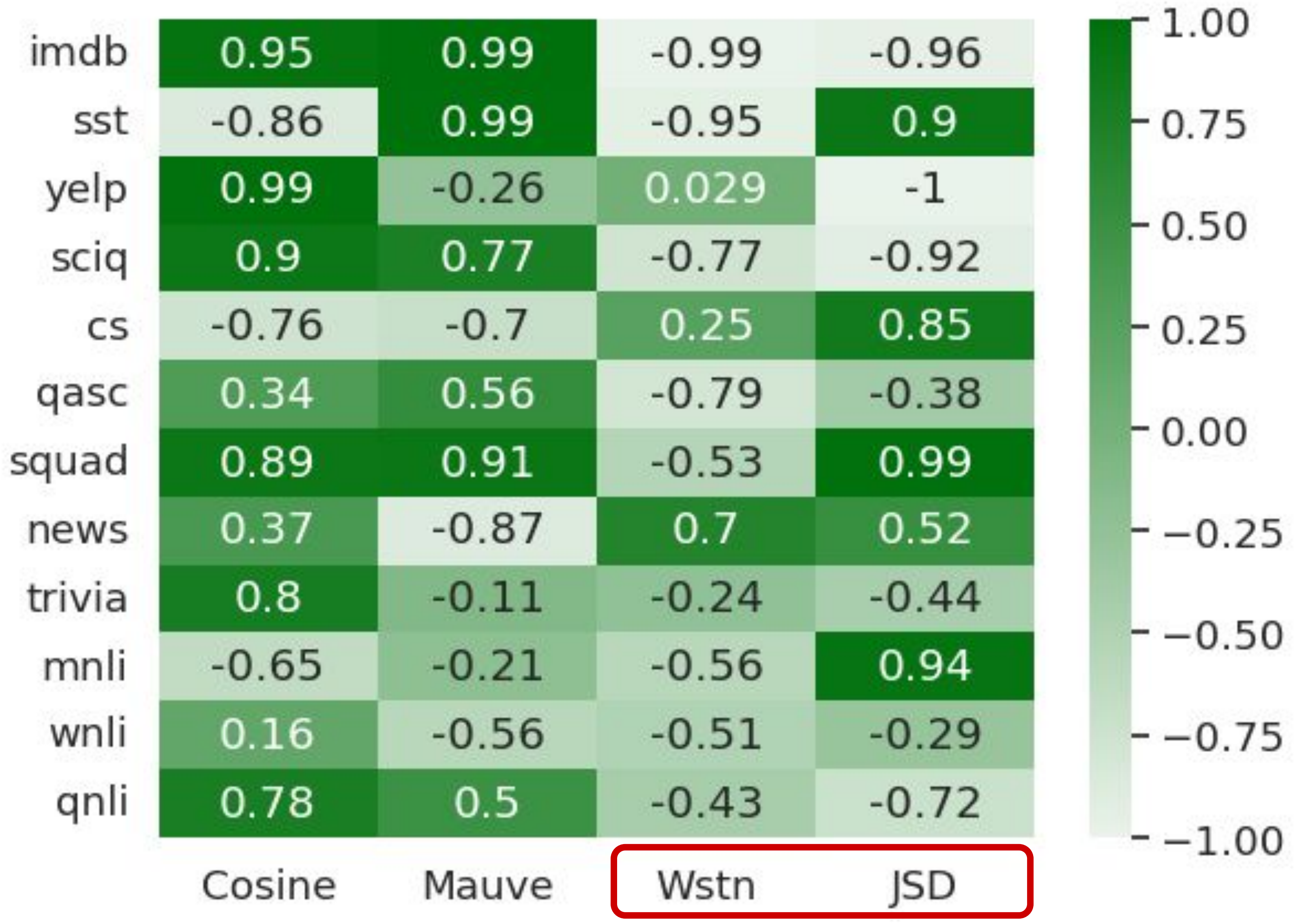
Trained on	Tested on	Performance
SQUAD-train	SQUAD-test	0.86
	News-test	0.51
	Trivia-test	0.55
News-train	News-test	0.66
	SQUAD-test	0.77
	Trivia-test	0.56
Trivia-train	Trivia-test	0.66
	SQUAD-test	0.52
	News-test	0.31
MNLI-train	MNLI-test	0.57
	WNLI-test	0.56
	QNLI-test	0.54
WNLI-train	WNLI-test	0.42
	MNLI-test	0.26
	QNLI-test	0.47
QNLI-train	QNLI-test	0.83
	MNLI-test	0.43
	WNLI-test	0.56

Performance vs Similarity

- Kendall - Wasserstein distance (Wstn) shows the most consistent correlation
- Pearson - Both Wstn and cosine show the consistent correlation
- JSD is least correlated



Kendall Correlation between performance and similarity metrics



Pearson Correlation between performance and similarity metrics

CONCLUSION

Conclusion

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

Future Work

- Determine the threshold for OOD
- Performance under different embeddings
- In some datasets, ID performance was worse than OOD. Why?
- Analysis on non-English languages

THANK YOU

QA

pokharel@pdx.edu
ameeta@pdx.edu

Link to GitHub