

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

OOD vs OODist

 Data from a related but different domain^[1] (Amazon vs • Data collected at a different time^[3] maybe

showcase equivalent levels of effectiveness on realworld datasets.



accurate accurate accurate

Twitter sentiment)

Different datasets for the same task^[2] (SST, IMDb, and Yelp for sentiment classification)

under different settings

• Datasets that are not in the training set^[4]

DATASETS

DatasetsTaskIMDb, SST2, YelpSentiment AnalysisSCIQ, Commonsense, QASCMCQSQUAD, News, TriviaExtractive Question AnsweringMNLI, WNLI, QNLINatural Language Inference

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

RESULTS

| - 1 00 | | | | | |
|--------|-------|-------|-------|-------|--------|
| L 1.00 | -1 | -0.33 | 1 | 1 | imdb |
| - 0.75 | 0.33 | -0.33 | 1 | -0.33 | sst |
| 0.50 | -1 | -0.33 | 0.33 | 0.33 | yelp |
| -0.50 | -1 | -0.82 | 0.82 | 1 | sciq |
| - 0.25 | 1 | 0.33 | -1 | -1 | cs |
| 0.00 | -0.33 | -1 | 1 | 0.33 | qasc |
| - 0.00 | 0.82 | -0.33 | 0.33 | 0.33 | squad |
| 0.25 | 0.33 | 0.33 | -0.33 | 0.33 | news |
| 0.50 | -0.33 | -0.33 | 0 | 0.82 | trivia |
| 0.50 | 0.82 | -0.33 | -0.33 | -0.33 | mnli |
| 0.75 | Ο | -0.33 | -0.33 | -0.33 | wnli |
| | -0.33 | -0.33 | 0.33 | 0.33 | qnli |
| -1.00 | | | | | |

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

METHODOLOGY



Cosine Mauve Wstn JSD

| | | | | | 1 0 0 |
|--------|--------|-------|-------|-------|--------|
| imdb | 0.95 | 0.99 | -0.99 | -0.96 | - 1.00 |
| sst | -0.86 | 0.99 | -0.95 | 0.9 | - 0.75 |
| yelp | 0.99 | -0.26 | 0.029 | -1 | 0 5 0 |
| sciq | 0.9 | 0.77 | -0.77 | -0.92 | - 0.50 |
| cs | -0.76 | -0.7 | 0.25 | 0.85 | - 0.25 |
| qasc | 0.34 | 0.56 | -0.79 | -0.38 | - 0.00 |
| squad | 0.89 | 0.91 | -0.53 | 0.99 | - 0.00 |
| news | 0.37 | -0.87 | 0.7 | 0.52 | 0.25 |
| trivia | 0.8 | -0.11 | -0.24 | -0.44 | - 0.50 |
| mnli | -0.65 | -0.21 | -0.56 | 0.94 | 0.50 |
| wnli | 0.16 | -0.56 | -0.51 | -0.29 | 0.75 |
| qnli | 0.78 | 0.5 | -0.43 | -0.72 | - 1.00 |
| | Cosine | Mauve | Wstn | JSD | 1.00 |

Wstn and cosine show the most consistent correlation

Correlated?

CONCLUSION CODE & PAPER

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





[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."