The Benefits of Label-Description Training

for Zero-Shot Text Classification

Lingyu Gao[1], Debanjan Ghosh[2†], Kevin Gimpel[1†]

1 TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO, 2 EDUCATIONAL TESTING SERVICE

† Co-senior authors

The 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)

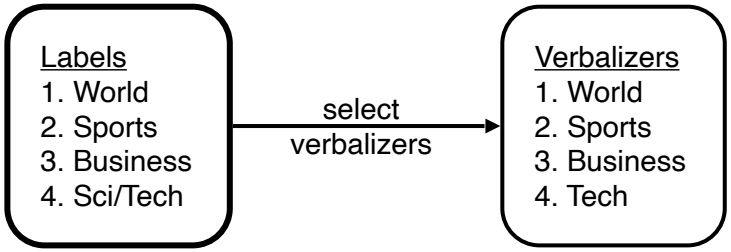
Motivation

- Zero-Shot Text Classification: No data available for fine-tuning
- Standard classifier: Needs finetuning for the classification head

- The pattern-verbalizer approach
  - Input: Overpriced, salty and overrated! **The restaurant is [MASK].**
  - Output with MLM head: great/awful
  - Sensitive to **the choice of specific pattern/verbalizer pairs**

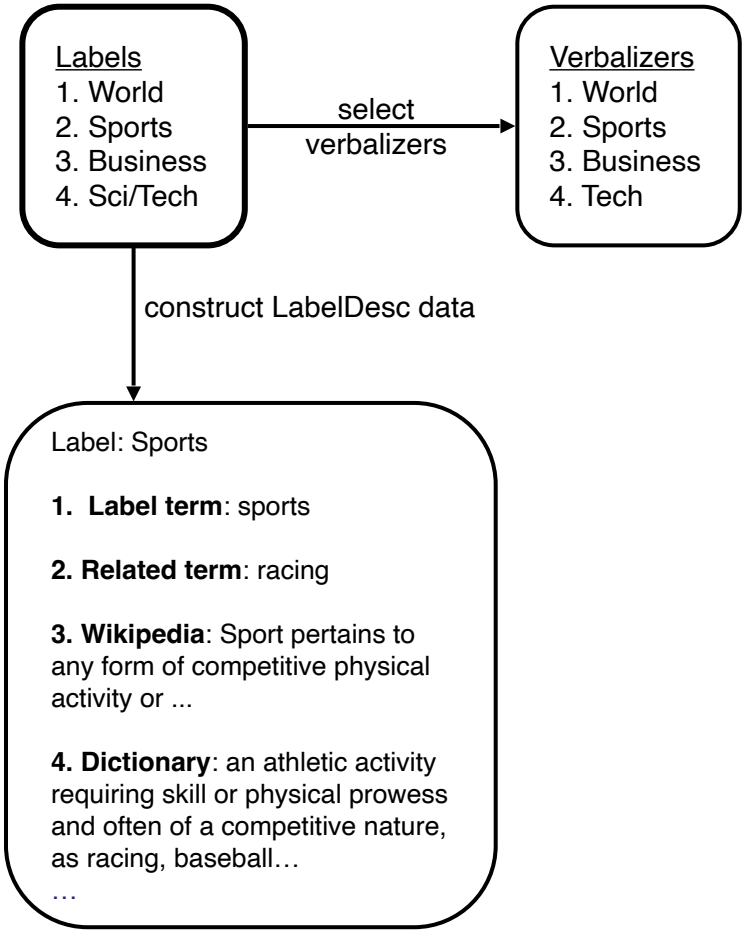
- Question: Could we curate datasets with **label descriptions** to improve zero-shot text classification performance for this approach?

Overview (Data Construction + Finetuning)



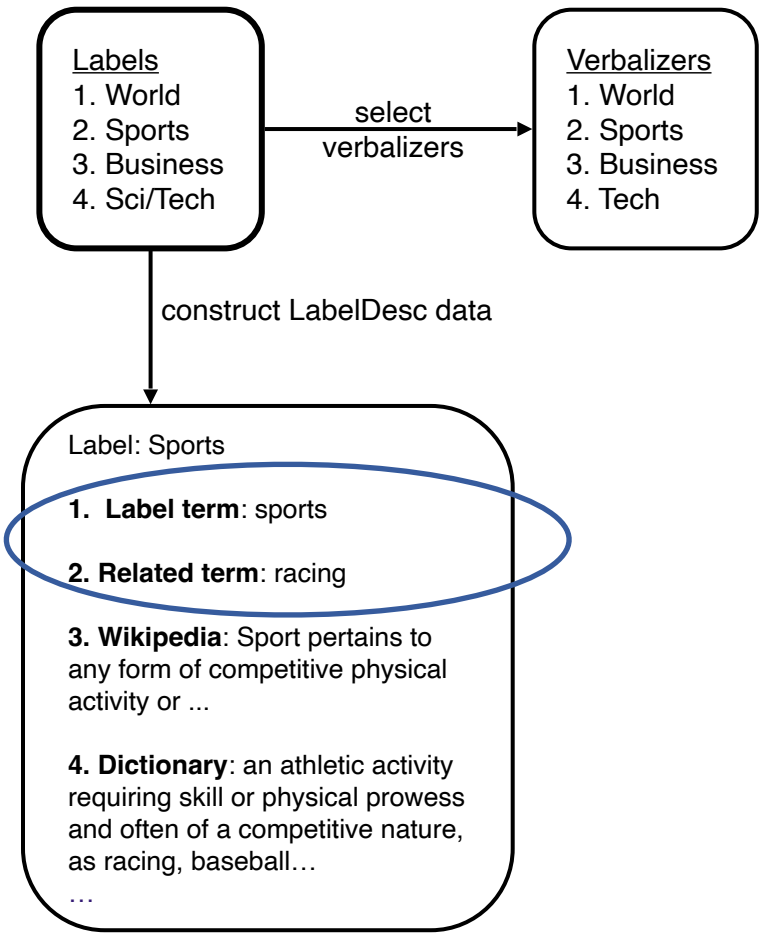
Select Verbalizers

Overview (Data Construction + Finetuning)



Construct LabelDesc Data

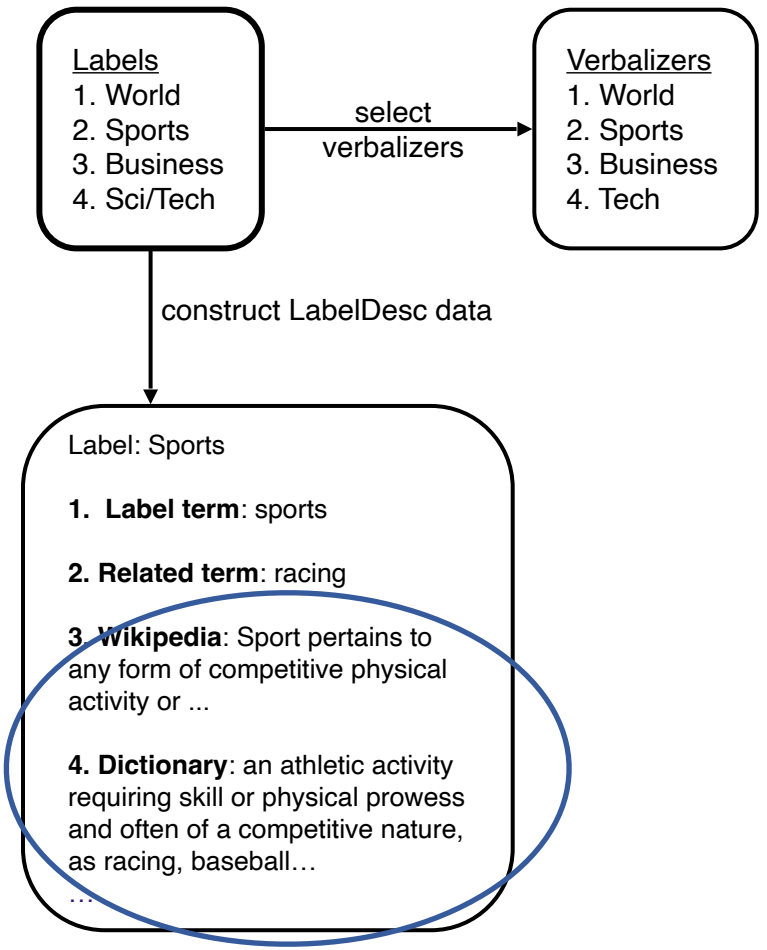
Overview (Data Construction + Finetuning)



Construct LabelDesc Data

- Subjective descriptors
  - Label term
  - Related term

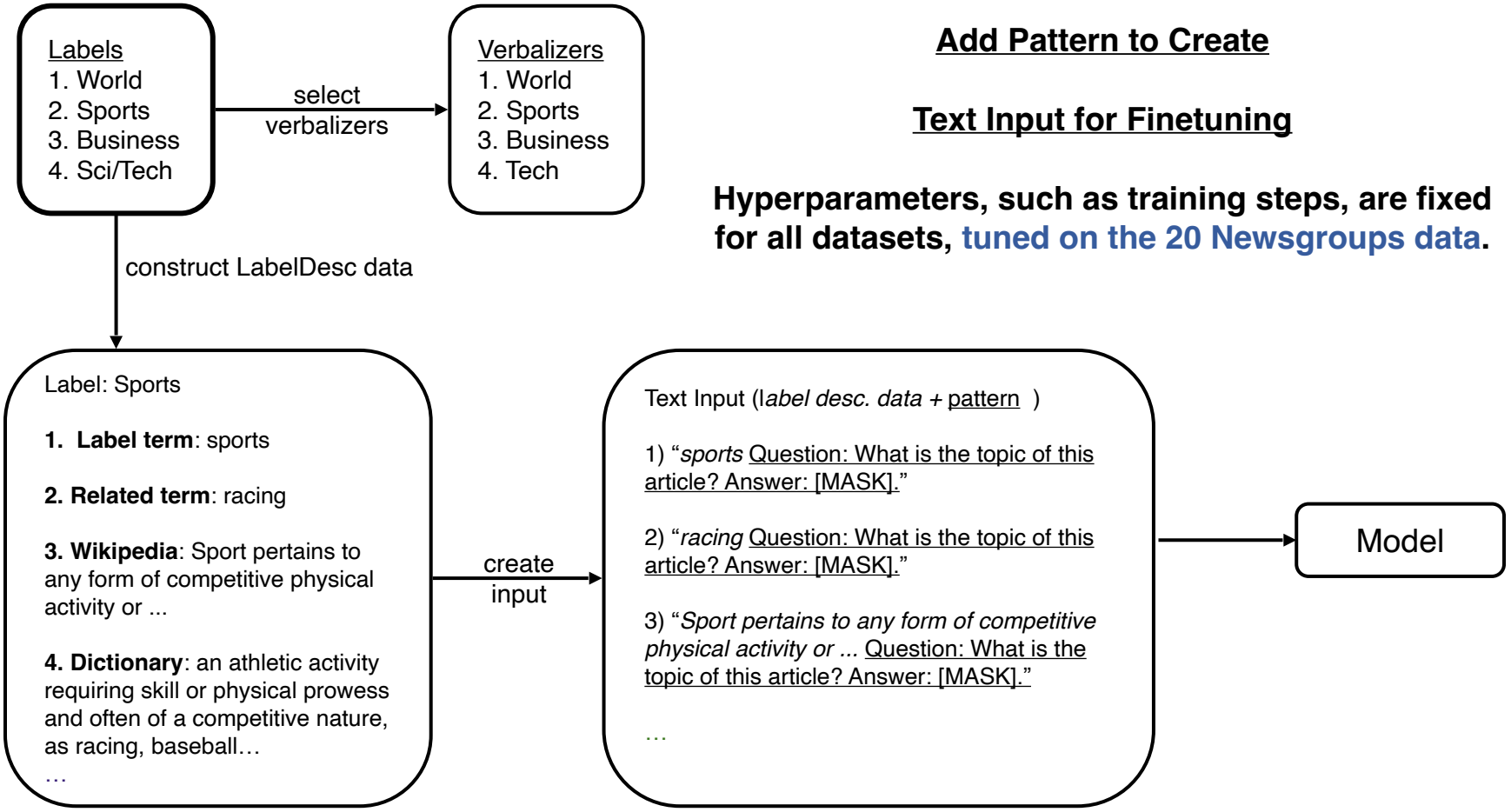
Overview (Data Construction + Finetuning)



Construct LabelDesc Data

- Subjective descriptors
- Objective sources of information
  - Wikipedia sentences
  - Dictionary definitions

Overview (Data Construction + Finetuning)



Add Pattern to Create

Text Input for Finetuning

Hyperparameters, such as training steps, are fixed for all datasets, tuned on the 20 Newsgroups data.

Overview (Inferencing)

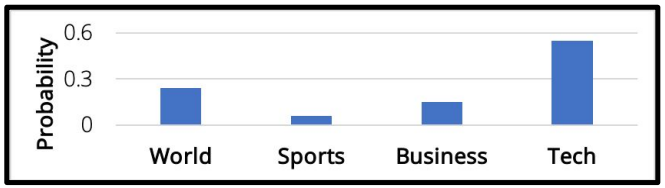
Test data from AGNews

"Need for carbon sink technologies Climate scientists tell a conference that greater efforts should be made to pull CO2 from the atmosphere."

Test data + pattern

"Need for carbon sink technologies Climate scientists tell a conference that greater efforts should be made to pull CO2 from the atmosphere. Question: What is the topic of this article? Answer: [MASK]."

Model



Prediction: Sci/Tech

Examples of LabelDesc data

- Examples of LabelDesc data for sentiment classification

Label	Input
Very Negative	awful It was *terrible*. A *horrendous* experience. Just *horrible*. Overall, it was *dreadful*.
Very Positive	great It was *amazing*. An *excellent* experience. Just *fantastic*. Overall, it was *outstanding*.

- Related terms to the label:
  - awful
  - terrible
  - …
- Simple hand-crafted templates:
  - It was t.
  - t could be replaced by the terms above.

Results and Evaluations

- Comparison against SOTA results (RoBERTa-base) using a single pattern with LabelDescTraining

	AGNews	Yahoo	DBPedia	Yelp-2	SST-2	Amz-2	IMDB
LabelDescTraining	84.6±0.3	59.9±0.3	82.4±1.2	84.8±0.6	88.2±0.2	89.6±0.4	83.4±0.4
Chu et al. (2021a)	68.8	57.8	81.9	67.3	65.0	66.8	-
Chu et al. (2021b)	75.1	60.0	88.6	-	-	-	-
van de Kar et al. (2022)	79.2	56.1	80.4	92.0	85.6	92.0	86.7

- **Sentiment classification**: Our method is better than dataless classification (Chu et al. 2021a) and competitive with mining-based approach, van de Kar et al. (2022)

- **Topic classification**: Our method is better than that of van de Kar et al. (2022)

*All our results are averaged over 3 random seeds

Results and Evaluations

- **LDT: LabelDescTraining**

		AGNews	Yahoo	DBPedia	Yelp-5	SST-5	Yelp-2	SST-2	Amz-2	IMDB	Avg.
zero-shot	b	62.7±7.4	41.5±7.0	54.6±18.9	38.0±4.3	35.6±4.3	63.6±10.7	62.6±11.0	64.0±10.3	69.9±13.2	54.7±9.7
	l	68.0±7.8	47.7±8.2	63.9±9.7	38.7±7.8	35.0±7.7	70.6±15.7	63.7±14.3	67.5±13.7	74.1±17.0	58.8±11.3
LDT_{20NG}	b	61.8±7.0	49.4±5.2	72.9±7.8	34.6±4.6	36.5±3.7	67.7±10.3	63.4±9.7	67.2±9.6	72.5±10.5	58.4±7.6
	l	72.4±6.8	54.4±4.3	71.9±10.8	36.3±5.7	36.6±7.1	63.4±13.0	56.9±8.7	60.9±10.2	67.5±15.2	57.8±9.1
LDT	b	77.4±4.9	58.8±1.6	79.5±4.4	43.6±2.1	42.0±1.6	88.3±2.5	84.5±2.2	88.6±1.4	86.9±1.8	72.2±2.5
	l	79.4±5.0	60.8±2.1	86.6±3.0	51.3±2.4	49.2±1.6	94.6±1.8	91.3±2.0	94.1±1.3	92.1±1.2	77.7±2.3
MLM_r	b	77.3±4.0	54.3±3.9	81.3±7.3	38.1±3.8	37.0±3.2	78.4±10.0	73.3±7.9	80.0±9.9	73.8±9.6	65.9±6.6
	l	75.2±5.0	58.0±3.0	85.4±13.0	46.4±3.3	43.4±2.9	90.8±7.6	84.1±6.8	90.2±7.1	87.4±6.2	73.4±6.1
MLM_m	b	73.1±5.6	50.1±5.4	72.6±8.1	36.8±2.8	35.8±2.5	80.1±7.2	75.8±5.0	81.8±6.8	76.7±6.0	64.8±5.5
	l	66.4±8.6	44.5±4.9	73.1±7.3	41.9±4.0	38.7±4.2	83.6±6.5	78.1±6.0	85.0±6.0	77.7±6.9	65.4±6.0
classifier	b	72.5±5.5	57.1±0.7	87.7±2.6	40.3±1.3	39.4±2.5	86.9±2.9	79.7±1.1	89.1±0.9	80.6±3.6	70.4±2.3
	l	77.8±1.5	50.9±7.3	78.2±1.0	42.4±1.6	35.3±9.2	93.3±0.9	86.6±1.4	93.7±0.5	85.7±2.0	71.5±2.8

- Zero-shot v.s. **LDT** (averaged across 3 random seeds, 14 patterns)

Results and Evaluations

- **LDT: LabelDescTraining**
- LDT_{20NG}: LDT finetuned on 20Newsgroup data
- MLM_r: verbalizer embedding randomly initialized
- MLM_m: mismatched label and verbalizers
- classifier: classifier without patterns

		AGNews	Yahoo	DBPedia	Yelp-5	SST-5	Yelp-2	SST-2	Amz-2	IMDB	Avg.
zero-shot	b	62.7±7.4	41.5±7.0	54.6±18.9	38.0±4.3	35.6±4.3	63.6±10.7	62.6±11.0	64.0±10.3	69.9±13.2	54.7±9.7
	l	68.0±7.8	47.7±8.2	63.9±9.7	38.7±7.8	35.0±7.7	70.6±15.7	63.7±14.3	67.5±13.7	74.1±17.0	58.8±11.3
LDT_{20NG}	b	61.8±7.0	49.4±5.2	72.9±7.8	34.6±4.6	36.5±3.7	67.7±10.3	63.4±9.7	67.2±9.6	72.5±10.5	58.4±7.6
	l	72.4±6.8	54.4±4.3	71.9±10.8	36.3±5.7	36.6±7.1	63.4±13.0	56.9±8.7	60.9±10.2	67.5±15.2	57.8±9.1
LDT	b	77.4±4.9	58.8±1.6	79.5±4.4	43.6±2.1	42.0±1.6	88.3±2.5	84.5±2.2	88.6±1.4	86.9±1.8	72.2±2.5
	l	79.4±5.0	60.8±2.1	86.6±3.0	51.3±2.4	49.2±1.6	94.6±1.8	91.3±2.0	94.1±1.3	92.1±1.2	77.7±2.3
MLM_r	b	77.3±4.0	54.3±3.9	81.3±7.3	38.1±3.8	37.0±3.2	78.4±10.0	73.3±7.9	80.0±9.9	73.8±9.6	65.9±6.6
	l	75.2±5.0	58.0±3.0	85.4±13.0	46.4±3.3	43.4±2.9	90.8±7.6	84.1±6.8	90.2±7.1	87.4±6.2	73.4±6.1
MLM_m	b	73.1±5.6	50.1±5.4	72.6±8.1	36.8±2.8	35.8±2.5	80.1±7.2	75.8±5.0	81.8±6.8	76.7±6.0	64.8±5.5
	l	66.4±8.6	44.5±4.9	73.1±7.3	41.9±4.0	38.7±4.2	83.6±6.5	78.1±6.0	85.0±6.0	77.7±6.9	65.4±6.0
classifier	b	72.5±5.5	57.1±0.7	87.7±2.6	40.3±1.3	39.4±2.5	86.9±2.9	79.7±1.1	89.1±0.9	80.6±3.6	70.4±2.3
	l	77.8±1.5	50.9±7.3	78.2±1.0	42.4±1.6	35.3±9.2	93.3±0.9	86.6±1.4	93.7±0.5	85.7±2.0	71.5±2.8

- Zero-shot v.s. **LDT** (averaged across 3 random seeds, 14 patterns):
  - Across a range of topic and sentiment datasets, our method is more accurate than zero-shot by **17-19% absolute**.

Results and Evaluations

- **LDT: LabelDescTraining**
- LDT_{20NG}: LDT finetuned on 20Newsgroup data
- MLM_{r}: verbalizer embedding randomly initialized
- MLM_{m}: mismatched label and verbalizers
- classifier: classifier without patterns

		AGNews	Yahoo	DBPedia	Yelp-5	SST-5	Yelp-2	SST-2	Amz-2	IMDB	Avg.
zero-shot	b	62.7±7.4	41.5±7.0	54.6±18.9	38.0±4.3	35.6±4.3	63.6±10.7	62.6±11.0	64.0±10.3	69.9±13.2	54.7±9.7
	l	68.0±7.8	47.7±8.2	63.9±9.7	38.7±7.8	35.0±7.7	70.6±15.7	63.7±14.3	67.5±13.7	74.1±17.0	58.8±11.3
LDT_{20NG}	b	61.8±7.0	49.4±5.2	72.9±7.8	34.6±4.6	36.5±3.7	67.7±10.3	63.4±9.7	67.2±9.6	72.5±10.5	58.4±7.6
	l	72.4±6.8	54.4±4.3	71.9±10.8	36.3±5.7	36.6±7.1	63.4±13.0	56.9±8.7	60.9±10.2	67.5±15.2	57.8±9.1
LDT	b	77.4±4.9	58.8±1.6	79.5±4.4	43.6±2.1	42.0±1.6	88.3±2.5	84.5±2.2	88.6±1.4	86.9±1.8	72.2±2.5
	l	79.4±5.0	60.8±2.1	86.6±3.0	51.3±2.4	49.2±1.6	94.6±1.8	91.3±2.0	94.1±1.3	92.1±1.2	77.7±2.3
MLM_r	b	77.3±4.0	54.3±3.9	81.3±7.3	38.1±3.8	37.0±3.2	78.4±10.0	73.3±7.9	80.0±9.9	73.8±9.6	65.9±6.6
	l	75.2±5.0	58.0±3.0	85.4±13.0	46.4±3.3	43.4±2.9	90.8±7.6	84.1±6.8	90.2±7.1	87.4±6.2	73.4±6.1
MLM_m	b	73.1±5.6	50.1±5.4	72.6±8.1	36.8±2.8	35.8±2.5	80.1±7.2	75.8±5.0	81.8±6.8	76.7±6.0	64.8±5.5
	l	66.4±8.6	44.5±4.9	73.1±7.3	41.9±4.0	38.7±4.2	83.6±6.5	78.1±6.0	85.0±6.0	77.7±6.9	65.4±6.0
classifier	b	72.5±5.5	57.1±0.7	87.7±2.6	40.3±1.3	39.4±2.5	86.9±2.9	79.7±1.1	89.1±0.9	80.6±3.6	70.4±2.3
	l	77.8±1.5	50.9±7.3	78.2±1.0	42.4±1.6	35.3±9.2	93.3±0.9	86.6±1.4	93.7±0.5	85.7±2.0	71.5±2.8

- Zero-shot v.s. **LDT** (averaged across 3 random seeds, 14 patterns):
  - Across a range of topic and sentiment datasets, our method is more accurate than zero-shot by **17-19% absolute**.
  - LDT is also **more robust** to choices regarding patterns and verbalizers.

Multi-Domain Evaluation

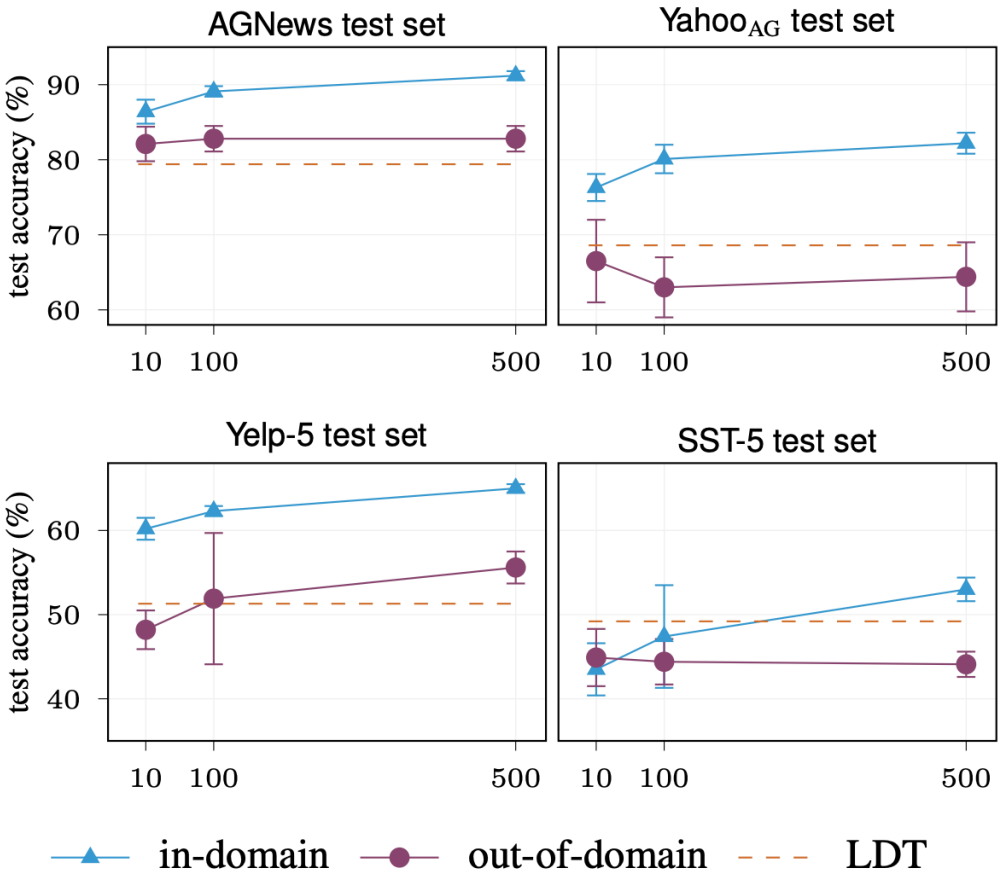


Figure 1: Domain transfer results, where the X-axis shows the number of training examples per label.

Our method even improves over **few-shot out-of-domain** classification in multiple settings.

*Yahoo$_{AG}$ is a sampled version of Yahoo dataset to match classes of AGNews

Conclusion

- Our method:
  - Achieves **17 - 19% accuracy gains** across 9 topic/ sentiment datasets over zero-shot setting
  - More **robust to pattern/verbalizer choices**
  - Domain agnostic, **robust across domains**

Thank you!