Label Agnostic Pre-training for Zero-shot Text Classification

Christopher Clarke[1], Yuzhao Heng[1], Yiping Kang[1], Krisztian Flautner[1], Lingjia Tang[1], Jason Mars[1]
[1]University of Michigan, Ann Arbor, MI.

Introduction

Conventional approaches to text classification typically assume the existence of a fixed set of predefined labels to which a given text can be classified. However, in real-world applications, there exists an infinite label space for describing a given text. In addition, depending on the aspect (sentiment, topic, etc.) and domain of the text (finance, legal, etc.), the interpretation of the label can vary greatly.

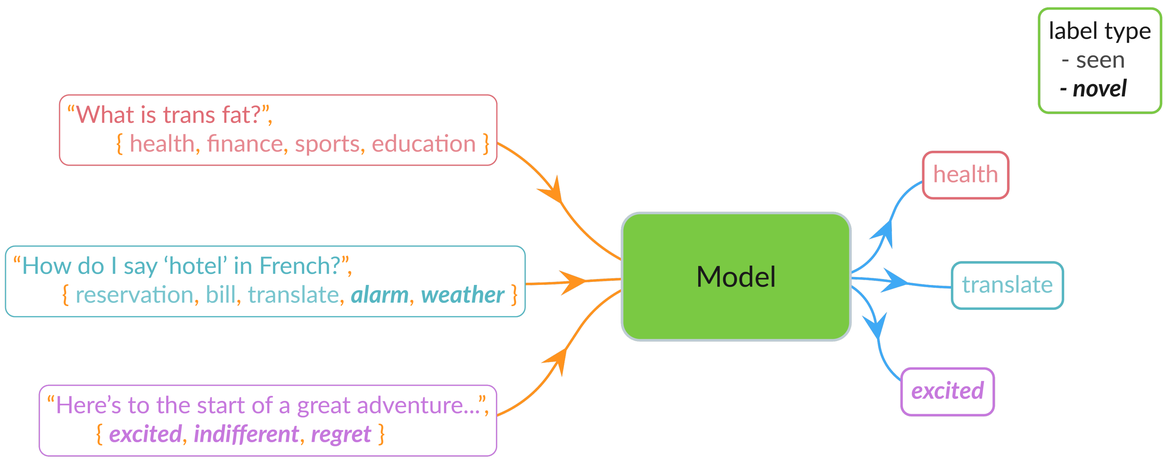

Figure 1: In real-world applications, the model needs to adapt to unseen labels. For a given aspect and domain, the interpretation of a given text-label pair can vary greatly.

Task Formulation

Zero-shot Learners are models capable of predicting unseen classes. When applied to text classification, these models aim to associate a piece of text with a given label without the need for having been trained on that text label pair. We canvas the range of zero-shot formalizations for enabling zero-shot text classification on PLMs and conduct our study.

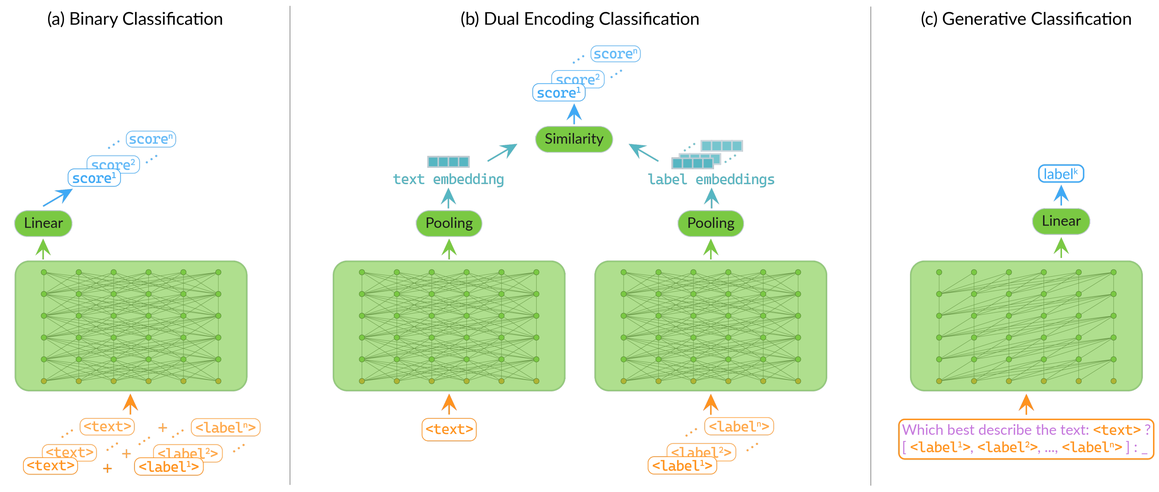

Figure 2: Zero-shot Text Classification Formalizations: Part (a) illustrates the binary classification formalization. Part (b) illustrates dual encoding. Part (c) illustrates generative text classification.

Methods and Approaches

We investigate the challenge of reducing the performance gap present in zero-shot models compared to their supervised counterparts on unseen data. We theorize that the poor generalization of these zero-shot models is due to their lack of aspect-level understanding during their training process. To alleviate this, we introduce two new simple yet effective pre-training strategies, **Implicit and Explicit pre-training** which specifically inject aspect-level understanding into the model.

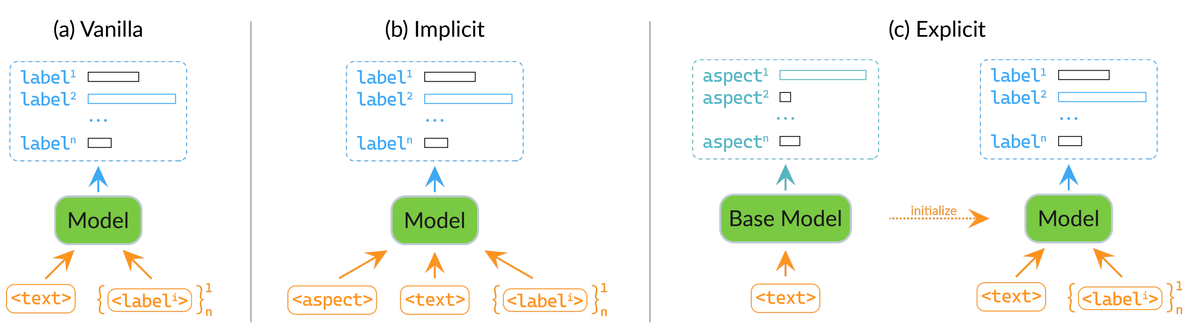

Figure 3: Zero-shot Text Classification Training Strategies. Part (a) shows standard model training where a text and the set of label options are passed to the model. Part (b) illustrates implicit training where the aspect is additionally passed as input. Part (c) shows injecting aspect knowledge to the model explicitly through gradient update, to initialize subsequent training.

Results & Contributions

Model	Training Strategy	Sentiment			Intent			Topic			Average
		Amazon Polarity	Fin. Phrase Bank	Yelp	Banking 77	SNIPS	NLU Eval	Multi EURLEX	Patent	Consumer Finance	
BERT Seq-CLS*	individual	96.0	97.2	84.8	88.6	99.0	88.9	94.8	64.1	82.6	88.4
	full	93.1	24.9	79.0	84.7	97.3	87.4	81.4	50.2	76.9	75.0
Binary BERT	vanilla	**80.7**	**68.9**	58.5	**51.4**	82.9	71.6	28.7	13.6	22.3	53.2
	implicit (ours)	80.1	66.0	**59.8**	51.3	82.5	**73.1**	30.3	15.2	23.4	53.5
	explicit (ours)	76.1	66.7	56.0	49.8	**83.8**	69.6	**44.5**	**19.5**	**30.2**	**55.1**
Bi-Encoder	vanilla	69.9	**71.7**	46.5	9.4	70.4	**71.1**	33.5	**11.7**	18.4	44.7
	implicit (ours)	**79.6**	64.0	**56.8**	**21.1**	**72.5**	61.9	**35.4**	9.6	11.3	**45.8**
	explicit (ours)	71.5	63.6	52.1	9.7	71.9	70.0	27.4	9.3	**27.0**	44.7
GPT-2†	vanilla	88.3	71.1	70.9	**22.8**	52.2	61.7	22.3	23.5	12.6	47.3
	implicit (ours)	89.3	61.4	**71.9**	16.5	33.7	**63.1**	18.6	**25.8**	12.2	43.6
	explicit (ours)	**89.7**	**75.9**	71.5	22.4	**54.1**	60.7	**23.5**	21.6	**13.9**	**48.2**
BART‡	Zero-shot	91.0	40.2	75.2	42.2	61.4	40.1	19.8	8.9	24.6	44.8
GPT-3‡	Zero-shot	54.4	52.8	77.0	23.7	13.9	37.9	-	-	-	43.3

Table 1: Aspect-Normalized out-of-domain accuracy. *Supervised upper bound, not a zero-shot framework.

By conducting Implicit and explicit pre-training, we are able to outperform vanilla on generalizing to unseen data on 6, 6, and 8 of the 9 datasets in out-of-domain UTCD across Binary BERT$_{[1]}$, Bi-encoder$_{[2]}$, and GPT-2$_{[3]}$ models respectively. For explicit training on Binary BERT, we achieve a massive improvement in zero-shot generalization (as much as +%16 for the topic aspect, +9% on average). Additionally, in comparison to the popular zero-shot baselines of BART and GPT-3 our models are able to outperform on 7 and 8 of the 9 datasets respectively.

Universal Text Classification Dataset (UTCD)

In order to test the zero-shot generalization of these NLP models we introduce v1 of the **Universal Text Classification Dataset (UTCD)**. UTCD is a large-scale compilation of 18 classification datasets spanning 3 main categories of Sentiment, Intent/Dialogue, and Topic classification. UTCD focuses on the task of zero-shot text classification where the candidate labels are descriptive of the text being classified. UTCD consists of **~6M/800K** train/test examples.

In order to make NLP models more broadly useful, zero-shot techniques need to be capable of label, domain & aspect transfer. As such, in the construction of UTCD we enforce the following principles:

- **Textual labels:** In UTCD, we mandate the use of textual labels. While numerical label values are often used in classification tasks, descriptive textual labels such as those present in the datasets across UTCD enable the development of techniques that can leverage the class name which is instrumental in providing zero-shot support.
- **Diverse domains and Sequence lengths:** In addition to broad coverage of aspects, UTCD compiles diverse data across several domains such as Banking, Finance, Legal, etc. each comprising varied length sequences (long and short).

HuggingFace Dataset

HuggingFace Models

GitHub Code

Acknowledgements

We thank our anonymous reviewers for their feedback and suggestions. This work is supported in part by award NSF1539011 by the National Science Foundation.

References

[1]: Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.
[2]: Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT