Label Agnostic Pre-training for Zero-shot Text Classification

Christopher Clarke Yuzhao Heng Yiping Kang Krisztian Flautner
Lingjia Tang Jason Mars

Computer Science & Engineering
University of Michigan
Ann Arbor, MI
{csclarke, stefanhg, ypkang, manowar, lingjia, profmars}@umich.edu

Abstract

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. This makes the task of text classification, particularly in the zero-shot scenario, extremely challenging. In this paper, we investigate the task of zero-shot text classification with the aim of improving the ability of pre-trained language models (PLMs) to generalize to both seen and unseen data across varying aspects and domains. To solve this we introduce two new simple yet effective pre-training strategies, *Implicit* and *Explicit pre-training*. These methods inject aspect-level understanding into the model at train time with the goal of conditioning the model to build task-level understanding. To evaluate this, we construct and release UTCD, a new benchmark dataset for evaluating text classification in zero-shot settings. Experimental results on UTCD show that our approach achieves improved zero-shot generalization on a suite of challenging datasets across an array of zero-shot formalizations.

1 Introduction

Text classification is the process of categorizing text into sets of organized groups where each set consists of similar content in a well-defined manner (Minaee et al., 2021; Joulin et al., 2016). Supervised approaches have achieved great success in recent years due to the availability of rich training data and the advent of large pre-trained language models such as BERT (Devlin et al., 2018). These conventional approaches typically assume the presence of a pre-defined set of labels to which a given text can be classified. However, in real-world applications, several challenges emerge:

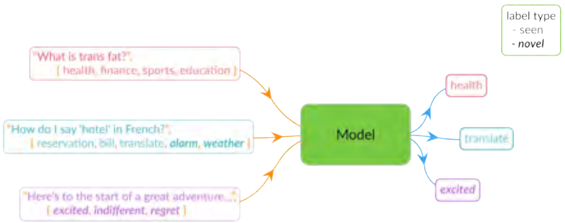
label type
- seen
- *novel*
"What is trans fat?",
[health, finance, sports, education]
"How do I say 'hotel' in French?",
[reservation, bill, translate, *alarm*, *weather*]
"Here's to the start of a great adventure...",
[*excited*, *indifferent*, *regret*]
Model
health
translate
excited

Figure 1: Zero-shot Text Classification Problem: 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.

1) The label space is constantly evolving. Over time, new labels are constantly emerging and the definition of the label space is constantly being refined. For example, intent classification systems such as those used in chatbots and dialogue systems are constantly introducing new intents as their range of supported features increases. Social networks such as Twitter encounter new and emerging topics on a daily basis from massive amounts of content that need to be classified. Figure 1 shows an example of this emerging label space.

2) The range of applications for text classification is vast. Text classification is pivotal to many different application areas from sentiment analysis to topic labeling, etc, and is used in a variety of domains such as finance, health, etc. When applied to this conglomeration of uses, it is typically assumed that there exists a comprehensive dataset of well-defined text-label pairs for each use case. However, in many real-world settings, annotated data is either scarce or unavailable entirely. Additionally, the use of dedicated models for each task is impractical due to the additional compute overhead and maintenance, thus making it difficult to scale over time.

Zero-shot learning (ZSL) is aimed at addressing these constraints. 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

Findings of the Association for Computational Linguistics: ACL 2023, pages 1009–1021
July 9-14, 2023 ©2023 Association for Computational Linguistics

for having been trained on that label. However, despite recent advancements in the capabilities of PLMs, zero-shot models still vastly underperform their supervised counterparts (Pushp and Srivastava, 2017; Puri and Catanzaro, 2019; Brown et al., 2020). As such, this remains an open research problem.

In this paper, we investigate the challenge of reducing the aforementioned performance gap present in these 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.

In order to evaluate these strategies, we canvas the range of zero-shot formalizations for enabling zero-shot text classification on PLMs and apply our techniques. Additionally, we introduce the Universal Text Classification Dataset (UTCD), a large-scale text classification dataset for evaluating zero-shot text classification. UTCD is a compilation of 18 classification datasets spanning 3 main aspects of Sentiment, Intent/Dialogue, and Topic classification. Our results on UTCD show that by employing both our implicit and explicit pre-training strategies we can achieve improved zero-shot performance on a suite of challenging datasets for which the model was not trained on.

Specifically, this paper makes the following contributions:

- We introduce *Implicit & Explicit* pre-training, two new simple yet effective pre-training strategies for improving zero-shot performance.
- We construct and release UTCD, a new benchmark dataset for evaluating text classification systems across a suite of diverse tasks and domains. We release our models and dataset[1].
- We conduct a thorough evaluation of various zero-shot text classification formalizations showing the effectiveness of our training strategies on each as well as insights gained.

2 Task Formulation

In this section, we introduce the task of zero-shot text classification and describe a set of formalizations for facilitating the classification of text in a zero-shot manner, i.e. being able to predict unseen labels.

Conventional Text Classification Text classification approaches using PLMs assume the existence of a pre-defined set of labels $\{y_i\}_n^1$ where for a given input sequence X, the model outputs a representation of that sequence as a sequence of hidden states $\{h_i\}_l^1$. Hidden states in the final layer are pooled to a single vector. In the case of BERT (Devlin et al., 2018), the $[\text{CLS}]$ token is taken, and a linear softmax layer is added to predict the probability distribution of the label set:

$$\vec{\mathrm{P}}\left(\{y_i\}_n^1 \mid h\right) = \mathrm{softmax}(Wh) \tag{1}$$

For the zero-shot scenario, this approach breaks since the output class set $\{y_i\}_n^1$ is fixed. This prevents the classification of text to new labels unless the model is re-trained with the new label set or a mapping of existing labels to unseen labels is built, both of which are impractical and cumbersome for real-world scenarios.

2.1 Binary Zero-shot Classification

To facilitate zero-shot classification of PLMs, Halder et al. (2020); Pushp and Srivastava (2017); Yin et al. (2019) formulate text classification as a series of binary classification tasks:

$$f(\mathrm{label}(y_i), x) = \mathrm{P}\left(\mathrm{True} \mid y_i, x\right) \tag{2}$$

The model is provided with a concatenation of the class label $\mathrm{label}(y_i)$ and input text and the output layer generates a binary $\mathrm{True}/\mathrm{False}$ prediction with a confidence score P. The True-prediction class with the highest confidence is selected as the final prediction, that is,

$$\hat{y} = \underset{i \in \{1...n\}}{\arg\max}\ f(\mathrm{label}(y_i), x) \tag{3}$$

where n is the number of classes/labels. Such cross-attention (CA) models apply attention layers on the text and labels jointly, which intuitively allows for rich interactions. This architecture is shown in part (a) of Figure 2.

[1] https://github.com/ChrisIsKing/zero-shot-text-classification

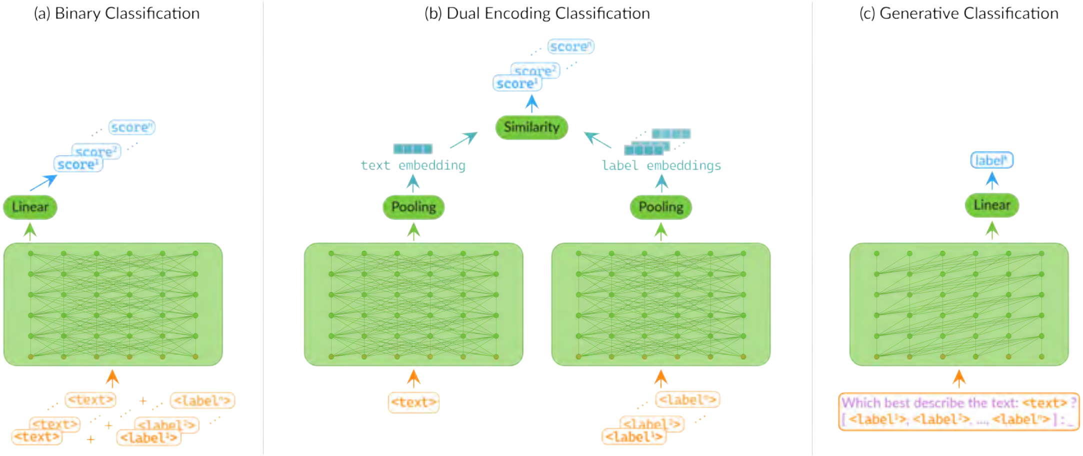
(a) Binary Classification
score^n
score^2
score^1
Linear
<text>
<label^n>
<label^1>
+
(b) Dual Encoding Classification
Similarity
text embedding
label embeddings
Pooling
Pooling
<text>
<label^n>
<label^2>
<label^1>
(c) Generative Classification
label^k
Linear
Which best describe the text: <text> ?
[<label^1>, <label^2>, ..., <label^n>] : _

Figure 2: Zero-shot Text Classification Formalizations: **Part (a)** illustrates the binary classification formalization described in section 2 where concatenated <text, label> pairs are passed as input to the model. **Part (b)** illustrates dual encoding where text label pairs are encoded separately and scored via a distance metric. **Part (c)** illustrates text classification where the model generates desired label based on a natural language instruction template.

2.2 Dual Encoding Zero-shot Classification

In contrast to cross-attention based architectures, Dual Encoder models (Reimers and Gurevych, 2019; Casanueva et al., 2020a; Clarke et al., 2022) instead focus on learning representations for a given text and label independently. They separately embed the text and label, via an encoder Φ and compute pair-wise scores S based on the encoded representations with a distance metric $Dist$, such as dot-product or cosine similarity:

$$S(x, y_i) = Dist\left(\Phi(x), \Phi(y_i)\right) \quad (4)$$

Sentence-Bert (Reimers and Gurevych, 2019) takes PLMs such as BERT and RoBERTa as the base encoder and use siamese networks to derive sentence embeddings by comparing similarities between sentence pairs as shown in part (b) of Figure 2. For text classification, this architecture can be used to derive latent representations for a given text and label and classify a sequence x according to:

$$\hat{y} = \underset{i \in \{1 \ldots n\}}{\arg\max}\; S(x, y_i) \quad (5)$$

2.3 Generative Classification

Lastly, the generative formulation of zero-shot text classification uses autoregressive language models by passing in text and label sets as natural language prompts and training the model to generate the target label token by token. As described in Puri and Catanzaro (2019), we reformulate the text classification problem as a multiple choice question answering problem. The model is provided with a multiple-choice question description containing each class label in natural language, and trained to generate the correct answer, as shown in part (c) of Figure 2. The intuition behind this approach is to train the model to use common sense reasoning to select the most probable description of the text data from a provided list of rich natural language classes. Given some input text t, the model is optimized with the next token prediction language modeling loss:

$$\sum_t \mathcal{L}(w_t, P(\hat{w}_t | w_{[1,t-1]})) \quad (6)$$

3 Method

In this section, we outline the methodology for our *Implicit* & *Explicit* pre-training strategies which allow us to inject aspect-specific knowledge into PLMs to improve generalization to unseen data. We first define the term aspect and outline the gap between the performance of the zero-shot models shown in section 2 on seen data compared to that of unseen data. Lastly, we describe our intuition behind why localization of aspect knowledge helps to bridge this gap.

Aspect Definition In the scope of this work, we define an aspect as the type of task to which a given set of datasets belong too. For example, sentiment is considered an aspect because it cleanly defines a task definition of understanding the emotion conveyed in a given text. This definition holds true even if the domain of the data changes. e.g senti-

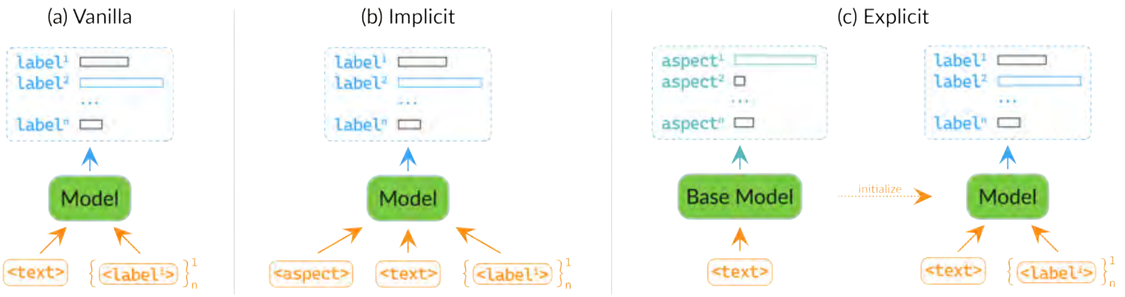

Figure 3: Zero-shot Text Classification Training Strategies. **Part (a)** shows standard model training where a text and the set of label options are passed to the model. **Part (b)** illustrates implicit training where the aspect is additionally passed as input. **Part (c)** shows injecting aspect knowledge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

ment detection of news data vs sentiment of social media tweets. In addition to having a clean task definition, we stipulate that the set of labels considered in a given aspect must convey that aspect. e.g For intent, the label *"turn off alarm"* conveys that the text describes the intention to do something.

3.1 Transfer Learning for Text Classification

The prevailing method for training models to perform classification tasks is to add a linear head on top of a pre-trained language model and fine-tune the entire network on labeled data (Devlin et al., 2018). However, when scaled to multi-task, multi-domain applications these models suffer from issues such as catastrophic forgetting and conflicting knowledge transfer across tasks (Aribandi et al., 2021; Geva et al., 2021; Clark et al., 2019; Alonso and Plank, 2016). We observe a similar trend in the Bert Seq-CLS row of Table 3 and 2, where despite the overarching task of text classification remaining the same when scaling the output space of the classification head to more labels across aspects, we see heavy performance degradation compared to having individual dataset models. For example, in table 3 training a multi-dataset BERT sequence classifier performs worse for every benchmark dataset compared to its single-dataset counterpart. Additionally, for the zero-shot formalizations, we observe the lowest positive transfer on datasets with the lowest level of token overlap between labels seen during training and out-of-domain labels, as shown in Figure 4. We theorize that the reason for this phenomenon is that the model is over-fitting to the specific labels seen during training instead of generalizing to the "aspect".

3.2 Implicit Training

In order to introduce aspect specification into our zero-shot models, we take inspiration from T5's (Raffel et al., 2019) text-to-text framework for multi-task generalization. In this framework, the model is fed some text for context and is then asked to produce some output text. As an example, to ask the model to translate the sentence "That is good." from English to German, the model would be fed the sequence "translate English to German: That is good." and would be trained to output "Das ist gut." Similarly, for each aspect (as defined in section 4), we introduce a conditional aspect token to the model input that acts as a context for that specific aspect. As such, in addition to learning the best contextual representation for the <text, label> input pair, the model implicitly learns a higher level understanding of the underlying aspect. By adding this conditional representation, even as the label space changes, the model is better able to understand the aspect at hand. This is shown in part(b) of figure 3. In the case of implicit binary zero-shot classification, the model is additionally provided with a concatenation of the aspect token and the output is selected as:

$$\hat{y} = \underset{i \in \{1...n\}}{\arg\max}\ f(\text{label}(y_i), \text{aspect}(a_{y_i}), x) \quad (7)$$

3.3 Explicit Training

Given our hypothesis that these language models will be able to generalize to unseen labels as a result of implicitly learning the task at hand, we explore the idea of explicitly training this generalization in a supervised manner. Instead of adding a conditional aspect token, we add an additional pre-training step in which the model is trained on aspect

Dataset	Aspect	Train/Test	#labels	Dataset	Aspect	Train/Test	#labels
in-domain				*out-of-domain*			
GoEmotions	sentiment	43K/5.4K	28	Amazon Polarity	sentiment	3.6M/400K	2
TweetEval	sentiment	45K/12K	3	Fin. Phrase Bank	sentiment	1.8k/453	3
Emotion	sentiment	16K/2K	6	Yelp	sentiment	650K/50K	3
SGD	intent	16K/4.2K	26	Banking77	intent	10K/3.1K	77
Clinc-150	intent	15K/4.5K	150	SNIPS	intent	14K/697	7
SLURP	intent	12K/2.6K	75	NLU Eval	intent	21K/5.2K	68
AG News	topic	120K/7.6K	4	MultiEURLEX	topic	55K/5K	21
DBpedia	topic	560K/70K	14	Patent	topic	25K/5K	9
Yahoo	topic	1.4M/60K	10	Consumer Finance	topic	630K/160K	18

Table 1: Universal Text Classification Dataset (UTCD) consists of the following datasets: Demszky et al. (2020); Barbieri et al. (2020); Saravia et al. (2018); Rastogi et al. (2020); Larson et al. (2019); Bastianelli et al. (2020); Zhang et al. (2015); Auer et al. (2007); Malo et al. (2014); Casanueva et al. (2020b); Coucke et al. (2018); Xingkun Liu and Rieser (2019); Chalkidis et al. (2021); Sharma et al. (2019); Bureau (2012)

detection. This step acts as an initialization process whereby the model representations are tuned at the aspect level first. Once this step is completed the model is then fine-tuned for its respective zero-shot classification objective. This process is shown in part (c) of figure 3. For a given text x this explicit training step is defined as:

$$\vec{\mathrm{P}}\left(\{a_j\}_m^1 \mid h\right) = \mathrm{softmax}(Wh) \qquad (8)$$

4 UTCD: Universal Text Classification Dataset

In order to test the zero-shot generalization of these NLP models we introduce UTCD. UTCD is a compilation of 18 classification datasets spanning 3 main aspects of Sentiment, Intent/Dialogue, and Topic classification. A breakdown of each dataset is provided in appendix A. UTCD focuses on the task of zero-shot text classification where the candidate labels are descriptive of the text being classified. 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. As such, for each of the compiled datasets, labels are standardized such that the labels are descriptive of the text in natural language.

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). The datasets are listed in Table 1.

As described in section 3, we define aspect as the sub-task type to which a given set of datasets can belong too. We simulate the Zero-shot learning case by splitting UTCD into *in-domain*, data a given model would be trained on, and *out-of-domain*, data with novel classes unseen during training. Additionally, to prevent data imbalance across aspects, we sub-sample the *in-domain* datasets such that the total number of unique text in each aspect is the same while maintaining class label distribution for each dataset. Class imbalance is known to degrade performance in deep learning models (Buda et al., 2018; Ochal et al., 2021). We observe a similar trend where aspect normalization results in performance improvement.

5 Experimental Setup

Model Architectures For binary classification, we use $\mathrm{BERT}_{\mathrm{BASE}}$ with sentence pair classification as in Devlin et al. (2018). For dual encoding classification, we use Sentence-BERT (Reimers and Gurevych, 2019) with $\mathrm{BERT}_{\mathrm{BASE}}$ as the base

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 2: Aspect-Normalized out-of-domain accuracy. *Supervised upper bound, not a zero-shot framework. †In case none of the given labels are generated at inference, the generated text is embedded and compared with label embeddings. ‡Out-of-the-box zero-shot classifier.

encoder, mean pooling, and cosine similarity as the distance metric. For generative classification, we use the 345M GPT-2 (Radford et al., 2019) as the language model and the input representation described in Puri and Catanzaro (2019). These models are denoted Binary BERT, Bi-Encoder, and GPT-2 respectively.

Training We train all models with AdamW (Loshchilov and Hutter, 2019) and weight decay of 0.01 on all *in-domain* data for 3 epochs, for both pre-training and fine-tuning stages. For explicit pre-training, we use a learning rate of 2e-5, batch size of 16, and linear learning rate warmup over the first 10% steps with a cosine schedule. For binary and dual encoding we use a learning rate of 2e-5, batch size of 16, with 10% warmup and a linear schedule. For generative classification fine-tuning, we use a learning rate of 4e-5, batch size of 128, with 1% warmup and a cosine schedule as reported in Puri and Catanzaro (2019). We pre-process data and train all models with different random seeds over multiple runs.

6 Results & Discussion

In this section we present and analyze the results of our experiments, detailing our insights and discussing the implications of each of our techniques.

Evaluation Task We report accuracy on the test set of all *in-domain* and *out-of-domain* datasets. In multi-label cases where there is more than one valid label, the prediction is considered correct if the model predicts any one of the correct labels. For generative classification, we observe instances in which GPT-2 may not generate one of the label options, a known problem for PLM generation (Radford and Narasimhan, 2018; Pascual et al., 2021). In such cases, we consider the label option most similar to the generated answer as prediction, by mapping the generated output and the valid classes to an embedding space. For this encoding, we use the pre-trained model MPNet (Song et al., 2020) with mean pooling encoder from Sentence-BERT (Reimers and Gurevych, 2019) for mapping the labels and cosine similarity as the distance metric. This ensures the consistency of GPT-2's output with the other zero-shot formalizations.

Upper-bound & Zero-shot Baselines To gauge the ability of our models to generalize to unseen data, we establish our upper-bound as the performance of a fully supervised model on the target data. Specifically, we fine-tune two variations of $\text{BERT}_{\text{BASE}}$ for sequence classification which we denote as *"individual"* and *"full"*. For *individual*, we fine-tune a dedicated classification model for each dataset in UTCD. For *full*, we fine-tune a single model for all datasets. Additionally, we compare the zero-shot performance of our models to the popular LLM GPT-3 (Brown et al., 2020), and BART MNLI (Yin et al., 2019) which is the most popular and widely downloaded zero-shot model on

Model	Training Strategy	Sentiment			Intent			Topic			Average
		Go Emotions	Tweet Eval	Emotion	SGD	Clinc -150	SLURP	AG News	DBpedia	Yahoo	
BERT Seq-CLS*	individual	63.0	69.5	92.9	78.7	95.2	85.5	94.1	99.2	73.4	83.4
	full	56.7	55.4	91.1	80.5	82.9	77.3	86.7	98.6	66.6	77.3
Binary BERT	vanilla	59.3	**67.6**	**92.4**	91.5	87.8	**81.8**	**90.0**	98.9	67.9	81.9
	implicit (ours)	59.9	67.2	91.8	**93.5**	87.1	81.8	89.2	**98.9**	**68.1**	**82.0**
	explicit (ours)	**60.2**	66.6	91.8	93.4	**88.0**	80.4	88.7	98.9	67.8	81.7
Bi-Encoder	vanilla	**59.2**	65.7	92.8	82.2	**84.4**	79.9	89.3	99.0	67.4	80.0
	implicit (ours)	56.9	66.0	90.9	**81.3**	82	78.9	88.8	**99.0**	**67.9**	79.1
	explicit (ours)	58.8	**66.8**	**91.8**	82.7	83.3	**79.9**	**89.5**	98.9	67.7	**80.0**
GPT-2[†]	vanilla	58.8	**70.6**	75.9	84.2	81.4	75.3	86.7	98.5	68.3	77.7
	implicit (ours)	59.0	70.3	71.4	**84.7**	81.7	73.1	87.7	98.4	68.3	77.2
	explicit (ours)	**60.1**	70.1	**76.4**	84.3	**81.9**	**76.7**	**87.9**	**98.6**	**68.6**	**78.3**
BART[‡]	Zero-shot	24.2	47.8	37.7	41.4	50.4	27.5	71.7	65	49.2	46.1

Table 3: Aspect-Normalized in-domain accuracy.

Huggingface Hub[2].

6.1 Out-of-domain Performance

In table 2, we report results on the out-of-domain test set for UTCD. To evaluate the ability of our zero-shot models to adapt to unseen data, we evaluate our fine-tuned models from table 3 on the out-of-domain test set without training on any out-of-domain data. Across the zero-shot formalizations, we observe that our explicit Binary BERT achieves the best performance with a 2% increase over its vanilla counterpart. Thus showing the power of the explicit pre-training strategy for binary classification formalization.

When compared to the "full" supervised out-of-domain model, despite having not been trained on any data from the target dataset, across the aspects of sentiment and intent, our models are able to generalize well. Specifically, across all formalizations, our models are able to outperform the supervised model on the financial phrase bank dataset. We observe that this drop is due to conflicting domain data. UTCD's out-of-domain set consists of similar financial datasets in the other aspects of intent and topic. Given that examples from the finance phrase banks dataset are general in nature, without seeing the label, it is difficult for the sequence classifier to understand the task at hand, thus causing it to classify to conflicting labels from similar datasets. This showcases the need to include aspect-specific knowledge.

Lastly, when inspecting the performance of vanilla fine-tuning compared implicit and explicit 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, Bi-encoder, and GPT-2 models respectively. In particular, 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 massive zero-shot baselines of BART and GPT-3 our models are able to outperform on 7 and 8 of the 9 datasets respectively.

6.2 In-domain Performance

In table 3, we report results on the in-domain test set for UTCD. For in-domain, we conduct implicit & explicit training across each zero-shot formalization. We observe that when compared with the "full" supervised model, our zero-shot models are more performant while maintaining the flexibility of facilitating zero-shot. When compared with the "individual" variation, as our zero-shot models are trained jointly across different datasets, we achieve better performance than the single supervised model on datasets such as SGD, showing the power of knowledge transfer from other intent datasets such as Clinc-150 & SLURP.

For vanilla fine-tuning without implicit or explicit training, we observe that across zero-shot formalizations, injecting task specification through implicit and explicit pre-training preserves performance for in-domain data. Showing that while achieving better zero-shot transfer ability our models do not suffer performance loss on data already

[2] `https://huggingface.co/facebook/bart-large-mnli`

seen during training.

6.3 Importance of Label token overlap

In addition to the need for aspect-specific knowledge, we also observe a high correlation in zero-shot generalization results between the overlap of tokens seen during training and those evaluated on the out-of-domain test. Figure 4 shows the pairwise overlap of label tokens across the in-domain and out-of-domain datasets. When inspected across aspects, we see that our models are able to achieve the best out-of-domain performance on datasets with the most overlapping label tokens to those seen during training.

7 Related Work

Zero-shot text classification is the task of classifying text into novel categories unseen during training. Early zero-shot classification studies frame the problem as binary classification on whether a given label describes the text (Pushp and Srivastava, 2017; Yin et al., 2019). With the advancement of PLMs, subsequent works (Yin et al., 2019; Puri and Catanzaro, 2019) rely on transformer architectures to learn representations from descriptive labels passed in. In particular, Puri and Catanzaro (2019) fine-tune an autoregressive language model to generate titles based on a prompt template containing Tweet articles and a list of title options. Though the model is trained on a great variety of title options, the approach limits the learning to topic classification only, as the authors only analyze performance on topic datasets, unlike our approach which considers a wide array of aspects, each requiring focus on different sections of a given text.

Yin et al. (2019) similarly categorize zero-shot text classification by aspects and implicitly introduce aspects during training with a dedicated template for each aspect. They further propose the classification of a text, label pair as a logic entailment problem. However, the authors analyze a less challenging zero-shot case where a model is trained on a subset of text, label pairs, and evaluated on the remaining text with unseen labels in the same domain. Additionally, the authors introduce WordNet definition of the labels as the labels are all single words. This process requires manual intervention and is not applicable for multiple-word label sequences common in intent classification, such as "Check Balance". Our work evaluates a more diverse set of datasets for each aspect and a more comprehensive set of zero-shot architectures.

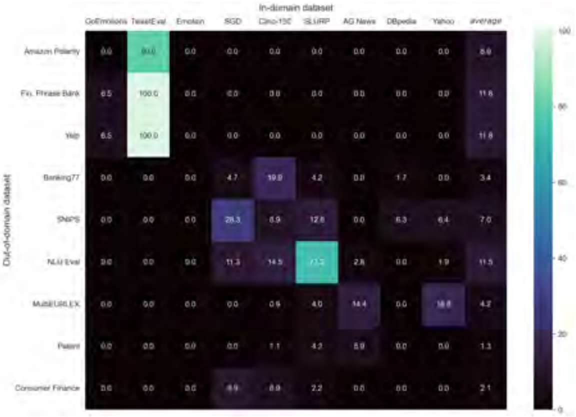

Figure 4: UTCD Out-of-domain Dataset Label Pair-wise Overlap with In-domain Dataset. 0 is no overlap, 100 is exactly the same label set. From sentiment to intent to topic, label overlap decreases in general.

8 Conclusion

In this paper, we investigate the task of zero-shot text classification with the aim of improving the ability of PLMs to generalize both seen and unseen data across domains without the need for additional training. We introduce two new simple yet effective pre-training strategies, *Implicit training* & *Explicit pre-training* which specifically inject aspect-level understanding into the model at train time. To evaluate this, we release UTCD, a new benchmark dataset for evaluating text classification in zero-shot settings. Experimental results on UTCD show that our approach achieves improved zero-shot generalization on a suite of challenging datasets in UTCD and across many zero-shot formalizations.

9 Limitations

While our approach is shown to be effective in improving the zero-shot adaption ability of these PLMs, the scope of this work has only been extended to English languages and has not been tested on other languages. In addition, another limitation of this work is the scope of the aspect. Aspect is defined across 3 main categories of intent, sentiment, and topic in the work. However, given the massive space of text label interpretations, our aspect range can be refined and expanded even further, lending to more analysis of the stability of implicit & explicit training as the number of aspects grows. We do not investigate this scenario in this work.

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

Héctor Martínez Alonso and Barbara Plank. 2016. When is multitask learning effective? semantic sequence prediction under varying data conditions.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. Ext5: Towards extreme multi-task scaling for transfer learning.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The Semantic Web*, pages 722–735, Berlin, Heidelberg. Springer Berlin Heidelberg.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Consumer Financial Protection Bureau. 2012. Consumer complaint database. https://www.consumerfinance.gov/data-research/consumer-complaints/. Accessed: Jun. 24th, 2022.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020a. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temcinas, Daniela Gerz, Matthew Henderson, and Ivan Vulic. 2020b. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. Multieurlex – a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D. Manning, and Quoc V. Le. 2019. BAM! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, Florence, Italy. Association for Computational Linguistics.

Christopher Clarke, Joseph Peper, Karthik Krishnamurthy, Walter Talamonti, Kevin Leach, Walter Lasecki, Yiping Kang, Lingjia Tang, and Jason Mars. 2022. One agent to rule them all: Towards multi-agent conversational AI. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3258–3267, Dublin, Ireland. Association for Computational Linguistics.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *ArXiv*, abs/2005.00547.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mor Geva, Uri Katz, Aviv Ben-Arie, and Jonathan Berant. 2021. What's in your head? emergent behaviour in multi-task transformer models.

Kishaloy Halder, Alan Akbik, Josip Krapac, and Roland Vollgraf. 2020. Task aware representation of sentences for generic text classification. In *COLING 2020, 28th International Conference on Computational Linguistics*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3).

Mateusz Ochal, Massimiliano Patacchiola, Amos Storkey, Jose Vazquez, and Sen Wang. 2021. Few-shot learning with class imbalance.

Damian Pascual, Béni Egressy, Clara Meister, Ryan Cotterell, and Roger Wattenhofer. 2021. A plug-and-play method for controlled text generation. *ArXiv*, abs/2109.09707.

Raul Puri and Bryan Catanzaro. 2019. Zero-shot text classification with generative language models. *ArXiv*, abs/1912.10165.

Pushpankar Kumar Pushp and Muktabh Mayank Srivastava. 2017. Train once, test anywhere: Zero-shot learning for text classification. *CoRR*, abs/1712.05972.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc.

Pawel Swietojanski Xingkun Liu, Arash Eshghi and Verena Rieser. 2019. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages xxx–xxx, Ortigia, Siracusa (SR), Italy. Springer.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A UTCD Datasets

UTCD is a compilation of 18 classification datasets spanning 3 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.

For sentiment we have the datasets Go Emotion (Demszky et al., 2020), TweetEval (Barbieri et al., 2020), Emotion (Saravia et al., 2018), Amazon Polarity (Zhang et al., 2015), Finance Phrasebank (Malo et al., 2014) and Yelp (Zhang et al., 2015). The GoEmotions dataset contains 58k carefully curated Reddit comments labeled for 27 emotion categories or Neutral. The TweetEval dataset consists of seven heterogenous tasks in Twitter, all framed as multi-class tweet classification. The tasks include - irony, hate, offensive, stance, emoji, emotion, and sentiment. We used the sentiment portion of this dataset for UTCD. Emotion is a dataset of English Twitter messages with six basic emotions: anger, fear, joy, love, sadness, and surprise. The Amazon Polarity dataset consists of reviews from Amazon. The data spans a period of 18 years, including 35 million reviews up to March 2013. Reviews include product and user information, ratings, and a plaintext review. The Finance Phrasebank dataset consists of 4840 sentences from English language financial news categorised by sentiment. The Yelp dataset consists of over 600k reviews for the task of sentiment classification.

For the intent/dialogue aspect we have the datasets: Schema Guided Dialgoue (Rastogi et al., 2020) is an annotated multi-domain, task-oriented conversations between a human and a virtual assistant. Clinc-150 (Larson et al., 2019) is an intent classification (text classification) dataset consisting of 150 in-domain intent classes. SLURP (Bastianelli et al., 2020) is dialuge dataset derived from SLU systems English spanning 18 domains. Banking77 (Casanueva et al., 2020b) is an intent classification dataset for the banking domain. It comprises 13,083 customer service queries labeled with 77 intents. Snips is an NLU dataset of over 16,000 crowdsourced queries distributed among 7 user intents. NLU Evaluation (Xingkun Liu and Rieser, 2019) is an NLU dataset from the conversational domain annotated with corresponding intents and dialogue scenarios.

Lastly, for the topic aspect we have the datasets: AG News (Zhang et al., 2015) is a topic classification dataset extract from the AG News article corpus. It consist of 4 classes from the original corpus. Each class contains 30,000 training samples and 1,900 testing samples. Yahoo Answers dataset (Zhang et al., 2015) contains 4,483,032 questions and their answers across 10 categories. Each class contains 140,000 training samples and 5,000 testing samples. DBpedia (Auer et al., 2007) dataset is a topic classification dataset constructed from picking 14 non-overlapping classes from DBpedia 2014. Multi Eurlex (Chalkidis et al., 2021) is a multilingual dataset for topic classification of legal documents. The dataset comprises 65k European Union (EU) laws, officially translated in 23 languages, annotated with multiple labels from the EUROVOC taxonomy. Big Patent (Sharma et al., 2019) is a topic classification dataset for the legal domain consisting of 1.3 million records of U.S. patent documents along with human written abstractive summaries. Consumer Finance (Bureau, 2012) dataset is a collection of complaints about consumer financial products and services sent to companies for response.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
9

□ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3

☑ B1. Did you cite the creators of artifacts you used?
Left blank.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

Left blank.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.