applied sciences

MDPI

Article

Improving Domain-Generalized Few-Shot Text Classification with Multi-Level Distributional Signatures

Xuyang Wang, Yajun Du *, Danroujing Chen, Xianyong Li , Xiaoliang Chen , Yongquan Fan, Chunzhi Xie, Yanli Li and Jia Liu

School of Computer and Software Engineering, Xihua University, Chengdu 610039, China
* Correspondence: duyajun@mail.xhu.edu.cn

Abstract: Domain-generalized few-shot text classification (DG-FSTC) is a new setting for few-shot text classification (FSTC). In DG-FSTC, the model is meta-trained on a multi-domain dataset, and meta-tested on unseen datasets with different domains. However, previous methods mostly construct semantic representations by learning from words directly, which is limited in domain adaptability. In this study, we enhance the domain adaptability of the model by utilizing the distributional signatures of texts that indicate domain-related features in specific domains. We propose a **Multi**-level **D**istributional **S**ignatures based model, namely MultiDS. Firstly, inspired by pretrained language models, we compute distributional signatures from an extra large news corpus, and we denote these as domain-agnostic features. Then we calculate the distributional signatures from texts in the same domain and texts from the same class, respectively. These two kinds of information are regarded as domain-specific and class-specific features, respectively. After that, we fuse and translate these three distributional signatures into word-level attention values, which enables the model to capture informative features as domain changes. In addition, we utilize domain-specific distributional signatures for the calibration of feature representations in specific domains. The calibration vectors produced by the domain-specific distributional signatures and word embeddings help the model adapt to various domains. Extensive experiments are performed on four benchmarks. The results demonstrate that our proposed method beats the state-of-the-art method with an average improvement of 1.41% on four datasets. Compared with five competitive baselines, our method achieves the best average performance. The ablation studies prove the effectiveness of each proposed module.

Keywords: domain-generalized few-shot learning; text classification; distributional signature; meta-learning

check for updates

Citation: Wang, X.; Du, Y.; Chen, D.; Li, X.; Chen, X.; Fan, Y.; Xie, C.; Li, Y.; Liu, J. Improving Domain-Generalized Few-Shot Text Classification with Multi-Level Distributional Signatures. *Appl. Sci.* **2023**, *13*, 1202. https://doi.org/10.3390/app13021202

Academic Editor: Rocco Furferi

Received: 20 December 2022
Revised: 10 January 2023
Accepted: 12 January 2023
Published: 16 January 2023

1. Introduction

Text classification [1,2] is a fundamental and crucial part of the NLP community. With the growth of deep neural networks, researchers have begun to focus on how to extend good classification performance to scenarios with only a small amount of labeled data.

Inspired by the advances of few-shot learning in CV [3,4], many studies [5,6] have leveraged the meta-learning based framework to tackle few-shot text classification (FSTC). Ohashi et al. [7] proposed a self-attention based encoder and a mutual information based loss function to obtain high-quality prototype representation. Sun et al. [8] proposed a data augmentation algorithm, which randomly generates instances within the smallest enclosing ball, to promote meta-learning methods. Unlike methods that learn directly from words, Bao et al. [9] proposed DS-FSL, which learns word importance via distributional signatures of texts. Distributional signatures mean characteristics of text distribution. A well-known instance of using these signatures is TF-IDF, which models the weights of words by their frequencies—a kind of explicit distributional signatures. These characteristics of text distribution imply underlying semantic knowledge, and their behaviors are consistent across classes. Bao et al. [9] suggest that learning from distributional signature is much more effective to generalize across different classes than directly learning from words.

However, the above approaches follow the in-domain setting. This setting assumes that the meta-training and meta-testing data are in the same data distribution, which is not feasible in real-world applications. Consequently, some researchers have begun to extend FSTC to a cross-domain setting. Zhang et al. [10] presented cross-domain few-shot text classification (CD-FSTC). In CD-FSTC, meta-training is conducted on one dataset in a single domain, and meta-testing on another in a different domain. Wang et al. [11] supposed that a meta-training dataset with only a single domain in CD-FSTC limits the generalization ability of the model, and proposed domain-generalized few-shot text classification (DG-FSTC). DG-FSTC replaces the single-domain dataset with a multi-domain dataset as the base dataset for meta-training. The meta-learning framework enables models to learn how to learn across different domains. Compared with conventional FSTC and CD-FSTC, DG-FSTC effectively improves the domain adaptability of the model, and is much closer to real application scenarios.

Currently, the DG-FSTC problem is still challenging for previous FSTC methods. Directly applying the FSTC methods to DG-FSTC will inevitably suffer from performance degradation. This is because the feature distribution changes significantly as the domain varies. The distribution variation makes those methods that learn directly from words struggle to capture discriminative features in specific domains under a domain shift. Therefore, we attempt to enhance the domain adaptability of the model by utilizing features that reflect specific domain information, such as distributional signatures. We find that in multi-domain data, the distributional signatures of different domains are highly domain-related. As shown in Figure 1, we utilize word frequency, an explicit distributional signature, to estimate word importance. The results show that, in a specific domain, the most informative words are highly domain-related. For instance, in travel domain, 'flight' and 'translate' are two of the most informative words. In a nutshell, the distributional signatures of texts are solid and annotation-free to reflect the word importance in specific domains.

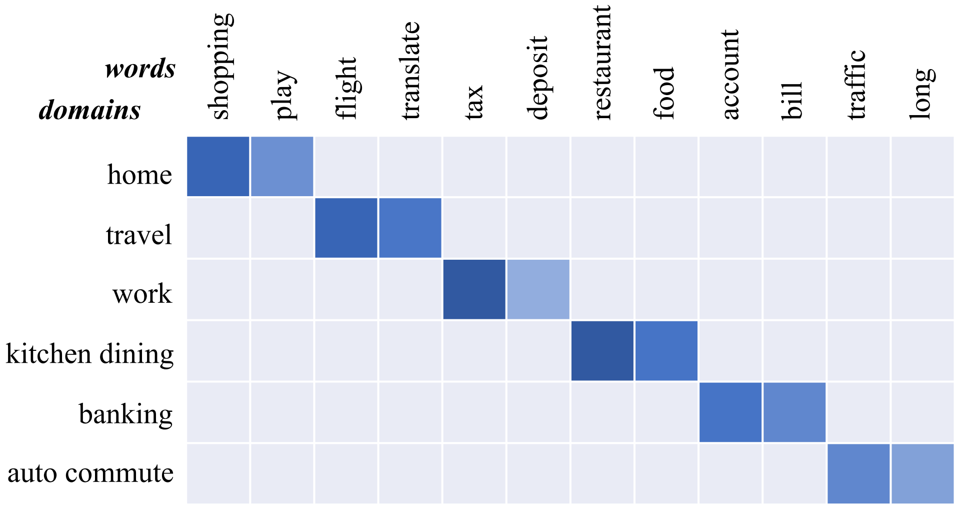

Figure 1. Word importance estimation on multi-domain data (Clinc150 [12]) using word frequency [13]. We list the top two words for each domain. The darker the color, the more important the word.

Consequently, this study introduces the distributional signature to solve DG-FSTC problem and propose a model based on Multi-level Distributional Signatures, namely MultiDS. MultiDS utilizes hierarchical distributional signatures to generate knowledge from three aspects, which are domain-agnostic, domain-specific, and class-specific, to improve the classification performance of the model under domain shift. Firstly, we argue that even textual representations of different domains have a kind of general and domain-agnostic features. Inspired by pretrained language models, we compute distributional signatures on a large news dataset containing information from multiple domains. The computed distributional signatures are used as the domain-agnostic features. Secondly, since the distributional signatures of different domains indicate specific domain information, we compute distributional signatures from texts of the same domain in each episode. These distributional signatures are treated as the domain-specific features. Thirdly, distributional signatures of different categories also show class-level differences, which can be used to model class-level word importance. In each episode, we calculate the distributional signatures of each class as class-specific features. After obtaining multi-level distributional signatures, we apply neural networks to translate them into word-level attention weights, which is able to help the model focus on informative features in different domains. In addition to utilizing distributional signatures to generate word importance, information indicated by distributional signatures are also beneficial in correcting feature distributions of different domains. Concretely, we think that domain-specific distributional signatures are also beneficial for neural networks to fit the feature distribution of specific domains. Therefore, we fuse domain-specific distributional signatures and word embeddings to generate instance-level calibration vectors. These calibration vectors effectively enable the model to adapt to different feature distributions. MultiDS thus obtains strong domain adaptability based on multi-level distributional signatures.

In summary, the main contributions of this study are as follows:

- We propose a simple yet powerful method based on multi-level distributional signatures to produce high-quality word-level attention values under domain shift. A large news corpus is firstly used to calculate domain-agnostic distributional signatures. Secondly, we compute domain-specific and class-specific distributional signatures from texts of the same domain and category, respectively. As a result, domain-adaptive word-level attentions are derived by translating multi-level distributional signatures using deep neural networks;
- We propose a domain calibration method based on domain-specific distributional signatures. By modeling the domain information indicated by domain-specific distributional signatures, the calibration method generates instance-level calibration vectors that are used to help the model fit the feature distributions of specific domains;
- We conduct extensive experiments on four datasets. The experimental results illustrate that our method outperforms the state-of-the-art method in DG-FSTC by 1.41% on average. Our method achieves the best average performance compared to five competitive baseline methods. Compared with DS-FSL [9], our method achieves an average improvement of 4.79%.

2. Preliminaries

2.1. Meta-Learning for Few-Shot Learning

In order to learn how to learn in the absence of a large number of annotated samples, some studies have introduced meta-learning framework to deal with the few-shot problem. The meta-learning framework simulates few-shot scenarios using a small amount of data, also referred to as an N-way-K-shot task. An N-way-K-shot task consists of N randomly selected classes, with each class containing $K + Q$ samples randomly sampled. It is called a labeled support set for the $N \times K$ samples, and an unlabeled query set for the $N \times Q$ samples. For each N-way-K-shot task, which is also called an episode, the model is trained with the annotated support set and tested on the query set. Using a number of episodes for meta-training, the model can learn how to learn in the low-resource scenario. When the meta-training is complete, a large number of N-way-K-shot tasks are sampled during meta-testing to evaluate the model performance.

2.2. Related Work

In this subsection, we introduce few-shot text classification, cross-domain few-shot text classification, domain-generalized few-shot text classification as well their recent trends.

2.2.1. Traditional Few-Shot Text Classification

Meta-learning based FSTC adopts the in-domain setting, which means that, in FSTC, the models are meta-trained and meta-tested on the data from the same distribution. FSTC ignores the distribution differences between meta-training and meta-validation data. This limits the application of FSTC to scenarios where new domain data emerges.

Two types of methods are always adopted to solve FSTC. The first one is transfer learning based methods [14–16]. These methods utilize specific algorithms to finetune pretrained language models to adapt to target data with few labeled samples. Wang et al. [17] leveraged two prompt encoders to learn task-agnostic and task-related features. They then design a task-level debiasing algorithm to alleviate task-level overfitting. Zhang et al. [18] proposed a contrastive pretraining algorithm, which reduces the distance between similar samples and enlarges the distance between samples from different classes simultaneously. Zhang et al. [19] suppose that a small set of annotated intent data makes a strong intent classification model. They pretrain their model on two intent datasets in a supervised manner, and further enhance the model via masked language modeling loss on target data.

The second type of methods is based on meta-learning framework. In this framework, models gradually learn to tackle few-shot problems through various N-way-K-shot classification tasks. Chen et al. [20] introduced the self-supervised objective function to learn discriminative semantic representation. Besides, they designed an unsupervised contrastive regularization to prevent overfitting at both task-level and instance-level. Luo et al. [21] suggested that the rich semantic information of labels helps meta-learners extract discriminative features. They simply augment feature representation by concatenating sentences with their corresponding labels.

2.2.2. Cross-Domain and Domain-Generalized Few-Shot Text Classification

In order to apply FSTC to real scenarios, Zhang et al. [10] proposed cross-domain few-shot text classification. CD-FSTC emphasizes the domain differences between meta-training and meta-testing data. It means that models are meta-trained and meta-tested on datasets in different domains. To solve CD-FSTC, they present a baseline method, which firstly conducts supervised pretraining for the model on the base dataset and then induces the classifier with few labeled samples.

Although CD-FSTC recognizes the importance of distribution differences between meta-training and meta-testing data, the way CD-FSTC models are trained on a single-domain dataset limits the generalization ability of the model. On the basis of CD-FSTC, Wang et al. [11] proposed a more promising setting, domain-generalized few-shot text classification. In addition to requiring different distributions between training and testing data, DG-FSTC uses a multi-domain dataset for meta-training. The combination of a meta-learning framework and a multi-domain dataset enables DG-FSTC models to learn better domain-agnostic meta-knowledge. Wang et al. [11] also designed a simple model that leverages two N-way-K-shot tasks in each episode to learn an enhanced domain knowledge generator.

3. Problem Definition

In this work, we focus on the DG-FSTC setting to solve few-shot problems. In DG-FSTC, models are meta-trained on a multi-domain dataset $D_{train} = \{d_i\}_{i=1}^{A}$. A is the numbers of domains. In each episode during meta-training, an N-way-K-shot task is sampled from a random domain. Each sampled task $t = \{t^s, t^q\} = \{(x_n^s, y_n^s)_{n=1}^{N\times K}, (x_m^q, y_m^q)_{m=1}^{N\times Q}\}$ contains data from N classes and each class contains $K + Q$ sentences, Then t is divided into support set $t^s = \{(x_n^s, y_n^s)_{n=1}^{N\times K}\}$ and query set $t^q = \{(x_m^q, y_m^q)_{m=1}^{N\times Q}\}$. The goal in each episode is to classify the unlabeled query set using the labeled support set. The model is updated by minimizing the following objective, as shown in Equation (1),

$$\mathcal{L}_{ce} = -\log(p(y^q|x^q; t^s, \theta)), \tag{1}$$

where θ is the parameter of the model; $p(y^q|x^q; t^s, \theta)$ denotes the probability of query set sample x^q belonging to label y^q and $\mathcal{L}_{ce}$ is the cross entropy loss. After meta-training, models are meta-tested on the dataset D_{test} with single or multiple domains. DG-FSTC expects that the models can learn transferable meta-knowledge on multi-domain datasets, then generalize to emerging classes from unseen domains. The notations in this study are listed in Table 1.

Table 1. Notations and explanations.

Notation	Explanation
N	Number of class in an N-way-K-shot task
K	Number of instance for each class in support set
Q	Number of instance for each class in query set
t	An N-way-K-shot task
t^s	Support set
t^q	Query set
x	Instance
y	Label
θ	Model parameter
D_{train}	Dataset for meta-training
D_{test}	Dataset for meta-testing
w	Word sequence
e	BERT embedding
$s(\cdot)$	General word importance matrix
$d(\cdot)$	Preliminary word weight matrix
$g(\cdot)$	Hidden state
$att(\cdot)$	Attention matrix
W	Weight
z	Bias
I	Identity matrix
$\widehat{y}$	One-hot label
η	Hyperparameter
$\alpha, \beta, \mu, \phi, \varepsilon$	Learnable parameter
r	Sentence representation
df	Domain feature
cal	Calibration vector

4. MultiDS Model

The main idea of this study is to empower DG-FSTC by introducing multi-level distributional signatures, which indicate rich and hierarchical information. To take full advantage of the power of distributional signatures, in our method, we mainly propose the following two modules.

- Hierarchical attention generator via multi-level distributional signatures (Section 4.2): We utilize three levels of distributional signatures, which are domain-agnostic, domain-specific, and class-specific, to approximate hierarchical word importance. To denoise distributional signatures and get more accurate attentions, this generator is devised to translate different levels of distributional signatures into word-level attention weights;
- Domain calibrator via domain-specific information (Section 4.3): In addition to accessing word importance, a domain calibrator is applied to calibrate sentence-level feature representations. By generating calibration vectors using domain-related information, this calibrator guides the model to adapt to specific domain distributions.

Overview: As illustrated in Figure 2, our method is divided into following steps: (1) we construct word embeddings by pretrained language model, BERT, in Section 4.1; (2) we calculate multi-level distributional signatures, and convert them to word-level attentions by hierarchical attention generator in Section 4.2; (3) we calibrate the feature representations by generating calibration vectors from domain-specific distributional signatures in Section 4.3; (4) we apply a ridge regressor as the classifier to categorize query samples in Section 4.4.

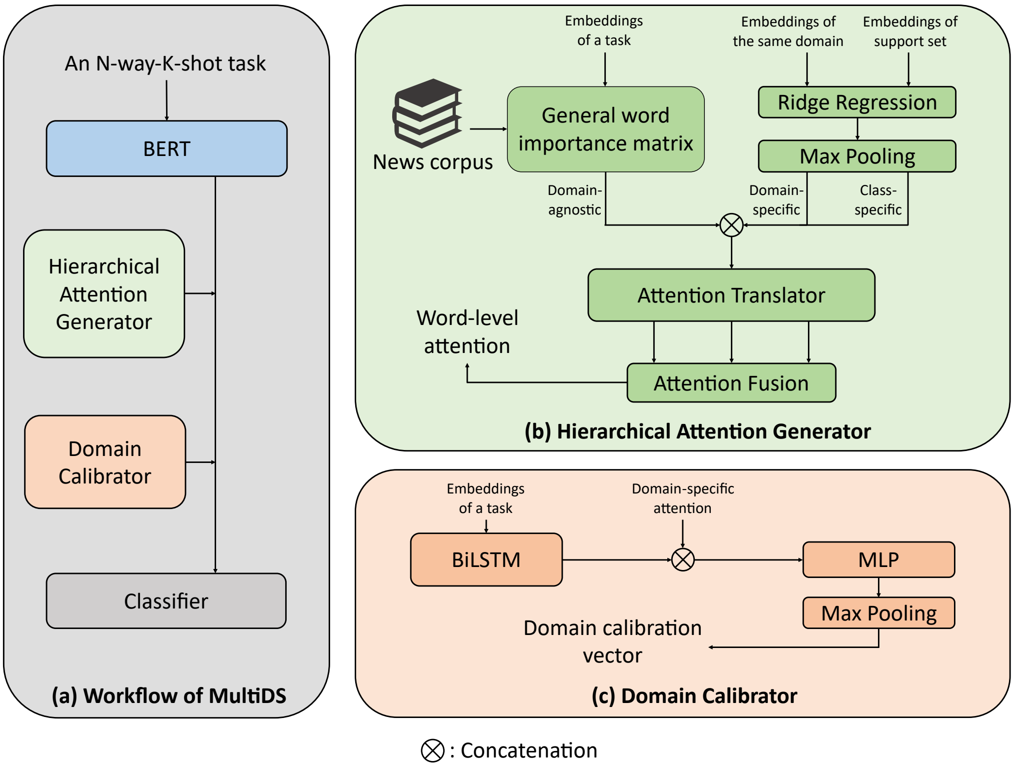

Figure 2. (**a**) The workflow of MultiDS model. (**b**) Hierarchical attention generator. (**c**) Domain calibrator.

4.1. Text Encoder

In our work, we choose BERT [22] as the text encoder for its formidable semantic representation. Due to pre-training on massive data, BERT excels at expressing and understanding textual information. Given an input sequence $w = \{w_1, w_2, \cdots, w_l\}$, the word embedding is derived according to Equation (2),

$$e = \text{BERT}(w), e \in \mathbb{R}^{l \times h_1}, \tag{2}$$

where l denotes the length of the sequence and h_1 is the dimension of BERT embedding.

4.2. Hierarchical Attention Generator via Multi-Level Distributional Signatures

In order to facilitate model's adaptation to various domains, we propose to utilize multi-level distributional signatures to generate hierarchical and domain-adaptive attentions. We find that different levels of distributional signatures exhibit hierarchical characteristics. A well-known point in recent years is that there are some general semantic features that can also be used to process other data, such as applying a pretrained model to downstream tasks. Consequently, we propose domain-agnostic attention, which can be transferred to various domains. Besides, distributional signatures in the same domain or category also contain rich semantic information. A good example is the phenomenon that the word frequency of data in the same domain shows unique characteristics that are relevant to that domain information. Similarly, data from the same category shows characteristics specific to that category. Based on the above phenomenons, we propose domain-specific attention and class-specific attention to help the model understand discriminative information at the domain- and class-level. In addition, to fuse three different levels of attentions and reduce informative redundancy, an attention fusion module is devised.

4.2.1. Domain-Agnostic Attention

Inspired by the paradigm of applying pre-trained models to downstream tasks, we believe that distributional signatures computed from large amounts of data can also be transferred to specific domains as domain-independent features. Specifically, we use a document-level news dataset, 20 Newsgroups [23], to calculate domain-agnostic distributional signatures. As shown in Table 2, 20 Newsgroups contains news articles from 6 domains, roughly including computers, recreation, science, politics, religion, and for-sale. This multi-domain document-level dataset enables the model to obtain general distributional signatures without domain bias.

Table 2. Statistics of 20 Newsgroups.

Example	Token/Example	Vocab	Domain	Example/Domain	Class
18,828	340	32,137	6	3138	20

Here, we leverage explicit distributional signatures, which mean word frequency, to infer word weights. The frequency of words appearing in documents is a kind of natural and annotation-free feature that effectively implies words importance in low-resource scenarios. For instance, frequently used words such as 'the', 'a', etc., are often considered less important, while those that rarely appear often contain more discriminative information.

Firstly, given a word sequence w^{news} in 20 Newsgroups, our approach is to utilize an existing method [13] to make a rough estimate of word weights via word frequency in Equation (3),

$$s^a(w_i^{news}) = \frac{\eta}{p(w_i^{news}) + \eta}, \tag{3}$$

where $p(w_i^{news})$ is the unigram likelihood of the i-th word over the dataset; η is a hyperparameter; $s^a(w_i^{news})$ represents the noisy weight of the i-th word. $s^a(\cdot)$ represents a mapping from words (within vocabulary) to their corresponding weights. According to the above hypothesis, the $s^a(\cdot)$ computed from a large document-level dataset contains generalized semantic knowledge that can be applied to the processing of other texts.

As the second step, we devise an attention translator that translates coarse word weights into fine-grained word-level attentions. Our attention translator consists of two components, a multi-layer perceptron (MLP) and a bidirectional LSTM (BiLSTM). Given an input sequence w^a, the MLP projects coarse word weights into a higher-dimensional space in Equation (4). Weights with higher dimensions contain richer semantic information.

$$d^a(w^a) = W_{att}s^a(w^a) + z_{att}, \tag{4}$$

where W_{att}, z_{att} are weight and bias of the MLP; $d^a(w^a) = \left\{d^a(w_1^a), \cdots, d^a(w_i^a), \cdots, d^a(w_l^a)\right\}$ are higher-dimensional word weights, and $d^a(w_i^a)$ means weight of the i-th word w_i^a. Since each word weight is contextually affected, we employ a BiLSTM, which is adept at processing sequence information, to further encode these weights in Equations (5)–(7), so that each word weight implies the correlation between contextual information.

$$g^a\overrightarrow{(w^a)} = \overrightarrow{\mathrm{LSTM}}\left(\left\{d^a(w_1^a), d^a(w_2^a), \cdots, d^a(w_l^a)\right\}\right) \tag{5}$$

$$g^a\overleftarrow{(w^a)} = \overleftarrow{\mathrm{LSTM}}\left(\left\{d^a(w_1^a), d^a(w_2^a), \cdots, d^a(w_l^a)\right\}\right) \tag{6}$$

$$att^a(w^a) = g^a(w^a) = [g^a\overrightarrow{(w^a)}; g^a\overleftarrow{(w^a)}] \tag{7}$$

The bidirectional hidden states $g^a\overrightarrow{(w^a)}$ and $g^a\overleftarrow{(w^a)}$ are calculated by Equations (5) and (6) using coarse word weights, respectively. Equation (7) concatenates two hidden states to get domain-agnostic attentions.

4.2.2. Domain-Specific Attention

In addition to domain-agnostic features, we are convinced that domain-related information is essential to guide the model in adapting to different domains. Since each N-way-K-shot task in DG-FSTC comes from a random domain, we propose to utilize the distributional signatures of each task to calculate domain-related information. However, each N-way-K-shot task contains only a small amount of data, and it is highly inaccurate to derive domain-related information using the word frequency of each small task. Consequently, we utilize a new policy to learn from distributional signatures. On the one hand, to capture more discriminative features with limited data, ridge regression is introduced, as shown in Equations (8) and (9), which admits a closed-form solution. On the other hand, we choose implicit distributional signatures, word embeddings, to infer word weights, which is more robust than explicit distributional signatures in low-resource scenarios. Given word sequences for an N-way-K-shot task w^b as well their BERT embeddings $e^b \in \mathbb{R}^{N(K+Q)\times l\times h_1}$, we derive the noisy attentions, which indicate domain information, below.

$$W_{rr}^b(w^b) = {e^b}^T\left(e^b{e^b}^T + \varepsilon I\right)^{-1}\widehat{y}^b, \tag{8}$$

$$s^b(w^b) = \max\left(\left|e^b W_{rr}^b(w^b)\right|\right), \tag{9}$$

where ε is a hyperparameter; I denotes the identity matrix; $\widehat{y}^b$ means the one-hot labels of w^b. In Equation (8), domain-specific features are extracted from texts of the same domain using ridge regression. Equation (9) derives noisy attentions by performing the multiplication of word embeddings with the weight matrix and calculating the most significant features in the product.

After that, we utilize our attention translator to produce accurate domain-specific attentions in Equations (10) and (11). According to Equation (10), the MLP is employed to project noisy attentions $s^b(w^b)$ into higher-dimensional space.

$$d^b(w^b) = W_{att}s^b(w^b) + z_{att}, \tag{10}$$

where $d^b(w^b) = \left\{d^b(w_1^b), d^b(w_2^b), \cdots, d^b(w_l^b)\right\}$ denote word weights with higher-dimension. According to Equation (11), we derive the domain-specific attentions after encoded by BiLSTM.

$$att^b(w^b) = \left[\overrightarrow{\mathrm{LSTM}}\left(d^b(w^b)\right); \overleftarrow{\mathrm{LSTM}}\left(d^b(w^b)\right)\right] \tag{11}$$

4.2.3. Class-Specific Attention

Apart from general semantic information and domain-specific information, class-specific information can also enhance the classification performance of models on multi-domain data. Unlike the way domain-specific information is processed, we calculate class-specific information using the support set in an N-way-K-shot task. It is considered that class-specific features extracted from the support set are applicable to the query set of the same task.

Given the support set in an N-way-K-shot task w^c and its BERT embeddings $e^c \in \mathbb{R}^{NK\times l\times h_1}$, the class-specific features are derived utilizing the ridge regression in Equation (12). The coarse attentions are obtained in Equation (13) by computing the most significant information in the product of class-specific features and word embeddings.

$$W_{rr}^c(w^c) = {e^c}^T\left(e^c{e^c}^T + \mu I\right)^{-1}\widehat{y}^c, \tag{12}$$

$$s^c(w^c) = \max\left(\left|e^c W_{rr}^c(w^c)\right|\right), \tag{13}$$

where μ is a trainable parameter and $\widehat{y}^c$ denotes the one-hot labels of w^c.

Similarly, we use an MLP to project coarse attentions into higher-dimensional space and get word weights $d^c(w^c) = \left\{ d^c(w_1^c), d^c(w_2^c), \cdots, d^c(w_l^c) \right\}$ in Equation (14). In Equation (15), a BiLSTM is utilized to encode word weights into class-specific attentions.

$$d^c(w^c) = W_{att} s^c(w^c) + z_{att} \tag{14}$$

$$att^c(w^c) = \left[\overrightarrow{\text{LSTM}}(d^c(w^c)); \overleftarrow{\text{LSTM}}(d^c(w^c)) \right] \tag{15}$$

4.2.4. Attention Fusion

After obtaining hierarchical attention weights, we design a neural module to fuse different levels of attentions. Due to the informative redundancy between different levels of attentions, simply concatenating them will harm the effect of hierarchical attentions. To alleviate the informative redundancy between different levels of attentions, consequently, an MLP is firstly employed to extract discriminative features from the concatenated multi-level attentions in Equation (16).

$$\tilde{att} = W_{fu}[att^b; att^b; att^c] + z_{fu}, \tag{16}$$

where W_{fu} and z_{fu} mean the weight and bias of the MLP.

We then use softmax function to convert attention features into word-level attention scores in Equation (17).

$$att = \text{softmax}(\tilde{att}) \tag{17}$$

Given a word sequence in an N-way-K-shot task w as well its word embeddings $e \in \mathbb{R}^{l \times h_1}$, we construct the sentence representation via word-level attentions in Equation (18).

$$r = \sum_{i=1}^{l} att(w_i) \cdot e_i \tag{18}$$

4.3. Domain Calibrator via Domain-Specific Information

In addition to deriving word importance from distributional signatures, we also focus on applying domain-specific distributional signatures to help the models adapt to specific feature distributions. In DG-FSTC, as each task comes from a different feature distribution, it is imperative for models to adapt effectively to various distributions. We propose to leverage features that indicate domain information, which are domain-specific distributional signatures and word embeddings, to calibrate feature distribution. Domain-specific attentions are regarded as features calculated from domain-specific distributional signatures. Besides, given a sequence in an N-way-K-shot task w and the embeddings e, we extract deep domain features from word embeddings of the same task by a BiLSTM in Equation (19).

$$df = \left[\overrightarrow{\text{LSTM}}(e)); \overleftarrow{\text{LSTM}}(e)) \right] \tag{19}$$

An MLP is then employed to fuse these two types of domain information, the domain features and domain-specific attentions, in Equation (20). We choose the most significant features from the fused results as the sentence-level calibration vectors.

$$cal = \max(W_{cal}[df; att^b] + z_{cal}), \tag{20}$$

where W_{cal} and z_{cal} are the weight and bias of the MLP.

Finally, we enhance the sentence representation via the calibration vectors and derive the final representation in Equation (21).

$$\tilde{r} = r \cdot cal \tag{21}$$

4.4. Ridge Regression Classifier

Here, we also use ridge regression [24] as a classification function due to its superiority in preventing overfitting. We let the final representation of the support set and query set be $s \in \mathbb{R}^{NK \times h_1}$ and $q \in \mathbb{R}^{NQ \times h_1}$, which are processed by hierarchical attentions and calibration vectors.

Firstly, we train the ridge regression with the annotated support set in Equation (22).

$$W_{rr} = s^T (ss^T + \phi I)^{-1} \widehat{y}^s, \tag{22}$$

where ϕ is a trainable parameter, and $\widehat{y}^s$ represents the one-hot label of the support set.

Secondly, as shown in Equation (23), the classifier is optimized by the following regularized squared loss.

$$\mathcal{L}_{rr} = \|W_{rr}s - \widehat{y}^s\|_F^2 + \phi\|W_{rr}\|_F^2, \tag{23}$$

where $\| \cdot \|_F$ denotes Frobenius norm.

Thirdly, the well-trained classifier is applied to categorize the query set samples. The classification loss based on cross entropy is derived in Equation (24).

$$\mathcal{L}_{cls} = CE(\alpha \cdot W_{rr}q + \beta, y^q), \tag{24}$$

where $CE(\cdot)$ represents the cross entropy function; α and β are trainable parameters; y^q means the true label of the query set.

The training procedure of MultiDS is concluded in Algorithm 1.

Algorithm 1 The training procedure of MultiDS

```
Require: Model parameter $\theta$; Meta-training episode $epi$; Dataset for meta-training $D_{train}$;
    Number of class $N$; Number of instance for each class in support set $K$; Number
    of instance for each class in query set $Q$; An N-way-K-shot task $t = \{t^s, t^q\} =
    \{(x_n^s, y_n^s)_{n=1}^{N\times K}, (x_m^q, y_m^q)_{m=1}^{N\times Q}\}$; Precomputed general word importance matrix $s^a(\cdot)$;
Ensure: Trained model parameter $\theta$;
Randomly initialize model parameter $\theta$;
for each $i \in [1, epi]$ do
    Randomly sample an N-way-K-shot task $t = \{t^s, t^q\}$ from $D_{train}$;
    Compute BERT embeddings of $t$ by Equation (2);
    Compute domain-agnostic attentions via general word importance matrix $s^a(\cdot)$ in
    Equations (4)–(7);
    Compute domain-specific attentions by Equations (8)–(11);
    Compute class-specific attentions by Equations (12)–(15);
    Fuse three kinds of attentions and derive hierarchical attentions by Equations (16) and (17);
    Construct domain calibration vector via word embeddings and domain-specific
    attentions by Equations (19) and (20);
    Train the classifier with support set $t^s = \{(x_n^s, y_n^s)_{n=1}^{N\times K}\}$, and classify query set
    instance $t^q = \{(x_m^q, y_m^q)_{m=1}^{N\times Q}\}$ by Equations (22)–(24);
    Update $\theta$ by minimizing classification loss in Equation (24);
end for
```

5. Experiments

In this section, we evaluate the effectiveness of our proposed method through comprehensive experiments. In Section 5.1, we firstly detail the setup of the experiments, including the datasets we choose, the baselines for comparison, and the details of the implementation. Secondly, we present the results of our method compared to multiple baselines on several datasets, and draw conclusions based on the experimental results in Section 5.2. Thirdly, ablation studies are conducted to explore the effectiveness and importance of key modules in our model in Section 5.3. Fourthly, we verify the stability of our method in more N-way-K-shot settings in Section 5.4. Finally, we present the computational overhead analysis for all methods in Section 5.5.

5.1. Experimental Setup

5.1.1. Datasets

We select the following five public datasets for the experiments.

- Clinc150 [12] is a multi-domain dataset for intent detection. It contains a total of 22,500 sentences from 10 domains. Each domain contains 150 classes;
- Banking77 [25] is a single-domain dataset with 77 fine-grained categories. These categories all belong to the banking domain;
- Huffpost [26] contains news headlines published on HuffPost from the year 2012 to 2018. These headlines cover a wide range of news varieties;
- Hwu64 [27] is also a multi-domain dataset, which contains fine-grained intents from 21 domains;
- Liu57 [27] contains 54 imbalanced categories. It brings challenges for models to reduce overfitting on major categories.

Details of the above datasets are shown in Table 3.

In DG-FSTC, we use a multi-domain dataset, Clinc150, for meta-training. The other different types of datasets are used to evaluate model performance. For instance, Clinc→Huffpost denotes meta-training on Clinc150 and meta-testing on Huffpost.

Table 3. Dataset statistics. unk denotes unknown.

	Dataset	Domain	Class	Example	Token/Example
Dataset for meta-training	Clinc150	10	150	22,500	150 ($\pm$0)
Dataset for meta-testing	Banking77	1	77	13,083	170 ($\pm$31)
	Huffpost	unk	41	36,900	900 ($\pm$0)
	Hwu64	21	64	11,036	172 ($\pm$40)
	Liu57	unk	54	25,478	472 ($\pm$823)

5.1.2. Baselines

In order to demonstrate the effectiveness of MultiDS, we select five competitive approaches as baselines to compare with the proposed method.

- ProtoNet [4]: This algorithm is a strong baseline for meta-learning based methods. ProtoNet proposes to average samples of the same class to obtain class center. It classifies samples based on distances between each class center and samples in feature space;
- HATT [28]: This model improves ProtoNet by a two-step attention mechanism, which is composed by feature-level and instance-level attentions. Feature-level attention focuses on more informative features. Instance-level attention assigns different weights to instances according to their significance;
- DS-FSL [9]: This approach firstly introduces distributional signatures into few-shot learning. It proposes to utilize precomputed distributional signatures as word weights and constantly update them to fit different data;
- MLADA [29]: This method combines adversarial learning with few-shot learning to solve the cross-domain problem in few-shot learning. The core idea is to produce knowledge by a generator to enhance the domain adaptability;

- DualAN [11]: This method is an improved version of MLADA. It enhances adversarial learning by introducing high-quality and stable data. These data come from two N-way-K-shot tasks from different domains.

5.1.3. Implementation Details

Parameter settings for model training and network architecture are given in Table 4. We choose the BERT-base-uncased version released by Hugging Face (https://huggingface.co/ accessed on 27 December 2017) as the BERT encoder. We replace word embeddings of all baselines with BERT embedding for fair comparison. For the implementation of ProtoNet, we employ a CNN with a global max-pooling after the BERT encoder.

Table 4. Parameter settings for model training and network architecture.

Parameter	Value
Meta-training episode	5000
Meta-validation episode	100
Meta-testing episode	1000
Learning rate	5×10^{-3}
Hyperparameter η	1×10^{-5}
Hidden size of MLP in attention translator	50
Hidden size of BiLSTM in attention translator	50
Hidden size of MLP in attention fusion	1
Hidden size of BiLSTM in domain calibrator	50
Hidden size of MLP in domain calibrator	768

For the meta-learning setup, we sample 5000, 1000 random tasks for meta-training and meta-testing. During meta-training, we meta-validate the model every 100 episodes with 100 tasks randomly sampled. We implement early stopping when the accuracy of 20 consecutive meta-validations do not improve. We select Adam [30] as the optimizer. The experiments are conducted on a single Geforce RTX 3090 GPU.

5.2. Experimental Results and Analysis

The experimental results of all methods on four dataset are shown in Table 5. Based on the above results, we get the following four observations.

Table 5. Experimental results of all methods on Clinc→Banking, Clinc→Huffpost, Clinc→Hwu and Clinc→Liu, respectively. The best and second best results are emphasized using bold fonts and underlines, respectively.

Model	**Clinc→Banking**		**Clinc→Huffpost**		**Clinc→Hwu**		**Clinc→Liu**		**Average**	
	10-Way-1-Shot	**10-Way-5-Shot**	**10-Way-1-Shot**	**10-Way-5-Shot**	**10-Way-1-Shot**	**10-Way-5-Shot**	**10-Way-1-Shot**	**10-Way-5-Shot**	**10-Way-1-Shot**	**10-Way-5-Shot**
ProtoNet	55.92	78.47	23.41	41.04	64.55	85.15	58.72	76.35	50.65	70.24
HATT	53.13	76.69	24.01	40.06	63.57	83.51	57.77	74.61	49.62	68.72
DS-FSL	53.85	82.13	29.10	46.61	62.70	86.45	57.45	82.17	50.78	74.34
MLADA	60.23	81.02	27.37	39.13	64.38	86.29	61.63	83.43	53.40	72.47
DualAN	**63.98**	85.61	28.88	45.15	66.45	86.43	63.65	82.90	55.74	75.02
MultiDS	63.84	**86.29**	**29.71**	**48.20**	**67.72**	**87.69**	**65.09**	**84.99**	**56.59**	**76.79**
Model	Clinc→Banking		Clinc→Huffpost		Clinc→Hwu		Clinc→Liu		Average	
	15-way-1-shot	15-way-5-shot	15-way-1-shot	15-way-5-shot	15-way-1-shot	15-way-5-shot	15-way-1-shot	15-way-5-shot	15-way-1-shot	15-way-5-shot
ProtoNet	49.61	73.91	17.60	33.78	58.16	81.10	52.65	69.64	44.51	64.61
HATT	47.60	72.95	18.50	33.20	57.00	79.59	51.28	68.01	43.60	63.44
DS-FSL	44.68	77.04	21.99	36.82	54.61	82.44	51.39	78.94	43.17	68.81
MLADA	54.62	76.79	20.82	32.97	57.47	81.36	54.85	78.44	46.94	67.39
DualAN	**58.58**	80.87	22.83	35.15	60.98	83.03	58.15	79.77	50.14	69.71
MultiDS	58.37	**81.75**	**23.53**	**40.97**	**62.07**	**84.28**	**59.38**	**81.11**	**50.84**	**72.03**

Our propose method exceeds all baseline methods in most datasets and settings. Compared to the second best methods, MultiDS achieves an average improvement of

2.69%. Compared to the five baseline methods, MultiDS outperforms them by 11.03% in Clinc→Banking, 4.38% in Clinc→Huffpost, 4.38% in Clinc→Hwu and 6.19% in Clinc→Liu on average. These above results illustrate the effectiveness and superiority of MultiDS.

Our proposed approach is able to surpass DS-FSL in all datasets and settings, improving by an average of 4.79%. This indicates that, compared to DS-FSL, the hierarchical attention generator we propose has better domain generalization and adaptability. The main reason is that MultiDS uses higher quality domain-agnostic attention and domain-specific attention that can constantly adapt to new domains.

Our proposed method outperforms adversarial learning based methods (MLADA, DualAN) in most cases, with an average lead of 2.71%. The main reason is that, in the few-shot scenario, it may be more accurate to directly use distributional signatures (such as word frequency) to derive word weights than adversarial training. A small amount of labeled data limits the effectiveness of adversarial training.

The average performance of adversarial learning based approaches (MLADA, DualAN) is superior to that of prototypical network based approaches (ProtoNet, HATT). Adversarial learning based methods are able to learn domain adaptability through adversarial training. However, when processing multi-domain data, the prototypical network based method lacks the ability to adapt to different domains, resulting in the generation of poor-quality class centers, which ultimately leads to the degradation of model performance.

5.3. Ablation Study

Here, to verify the effectiveness of each key module, we present the following variants of MultiDS.

- $-Att$: MultiDS without hierarchical attention;
- $-Att_{da}$: MultiDS without domain-agnostic attention in hierarchical attention generator;
- $-Att_{ds}$: MultiDS without domain-specific attention in hierarchical attention generator;
- $-Att_{cs}$: MultiDS without class-specific attention in hierarchical attention generator.
- $-Cal$: MultiDS without domain calibration vector;
- $-Cal_{df}$: MultiDS without domain features in domain calibrator;
- $-Cal_{ds}$: MultiDS without domain-specific attention in domain calibrator.

The experimental results are shown in Table 6. Based on the above results, we come to the following conclusions. (1) The effect of hierarchical attention is significant. When it is removed, model performance decreases by an average of 12.27%. This proves the effectiveness of hierarchical attention. (2) By removing domain-agnostic attention, model performance decreases by an average of 3.01%. We believe the reason is that the distributional signatures calculated from 20 Newsgroups contain high-quality knowledge of general semantics. This knowledge covers multiple domains with small domain biases and is therefore beneficial for processing information in different domains. (3) Domain-specific attention and class-specific attention also improve model performance. Besides, due to the relatively small number of domain information and class information in few-shot scenarios, the power of distributional signatures is limited. (4) Domain calibration vector is beneficial to model performance, and the average improvements is 0.35%. It proves that domain calibrator can extract informative features from domain information and domain-specific attention to calibrate domain distributions.

Table 6. Ablation studies on Clinc→Huffpost, Clinc→Hwu and Clinc→Liu. The best results are emphasized using bold fonts.

Model	**Clinc→Huffpost**	**Clinc→Hwu**	**Clinc→Liu**
	10-Way-1-Shot	**10-Way-1-Shot**	**10-Way-1-Shot**
MultiDS	**29.71**	**67.72**	**65.09**
$-Att$	22.86	54.40	48.45
$-Att_{da}$	27.47	64.82	61.19
$-Att_{ds}$	29.29	67.61	64.82
$-Att_{cs}$	29.40	67.33	64.83
$-Cal$	29.59	67.27	64.61
$-Cal_{df}$	29.42	67.49	64.76
$-Cal_{ds}$	29.59	67.58	64.92

5.4. Model Stability Verification in More Scenarios

Here we conduct extensive experiments to explore the performance and stability of models in more scenarios. The experimental results shown in Tables 7 and 8 illustrate the stability of our proposed model, which outperforms DS-FSL and DualAN in most cases. In addition, we have two more discoveries. (1) When K is fixed and N increases, more categories will increase the difficulty of classification. This requires the model to extract class-specific features to distinguish them from other categories. MultiDS outperforms other methods in most cases, indicating that MultiDS has better feature extraction capabilities. (2) When N is fixed and K increases, the challenge for the models is to focus effectively on crucial category information. Our model, which learns attention from distributional signatures, is better able to generalize in the classification task than those learning from words. MultiDS outperforms other methods in all scenarios, which proves the effectiveness of our hierarchical attention.

5.5. Computational Overhead Analysis

In addition to model performance, we also focus on the computational overhead of models. As shown in Table 9, we compare the computational time, including the time for meta-training and meta-testing, of all the methods. The results show that the total time of our method is less than that of four baseline methods, and is just a little longer than the total time of DS-FSL. Compared with DS-FSL, our method is able to significantly improve model performance with only a small increase in computational cost. This proves that our model design is reasonable and effective. Besides, our method shows a short inference time, which makes MultiDS valuable in real-world application scenarios.

Table 7. Model stability verification on Clinc→Liu. The best results are emphasized using bold fonts.

Model	***N*-Way-5-Shot on Clinc→Liu**					
	$N = 5$	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
DS-FSL	88.34	86.78	85.98	85.08	84.12	82.17
DualAN	90.08	88.64	85.95	86.07	85.16	82.90
MultiDS	**90.34**	**89.08**	**87.16**	**86.51**	**85.69**	**84.99**
Model	10-way-*K*-shot on Clinc→Liu					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
DS-FSL	57.45	70.90	77.24	80.90	83.43	84.55
DualAN	63.65	74.65	79.67	82.20	82.90	81.48
MultiDS	**65.09**	**75.55**	**80.37**	**83.28**	**84.99**	**86.03**

Table 8. Model stability verification on Clinc→Huffpost. The best results are emphasized using bold fonts.

Model	***N*-Way-5-Shot on Clinc→Huffpost**					
	$N = 5$	$N = 6$	$N = 7$	$N = 8$	$N = 9$	$N = 10$
DS-FSL	61.22	57.03	**54.70**	50.80	48.83	46.61
DualAN	58.45	55.21	52.36	45.97	43.90	45.15
MultiDS	**62.12**	**58.16**	53.15	**52.64**	**50.51**	**48.20**
Model	10-way-*K*-shot on Clinc→Huffpost					
	$K = 1$	$K = 2$	$K = 3$	$K = 4$	$K = 5$	$K = 6$
DS-FSL	29.10	36.43	41.06	44.85	46.61	49.62
DualAN	28.88	35.86	40.11	43.19	45.15	46.45
MultiDS	**29.71**	**37.31**	**41.78**	**45.71**	**48.20**	**50.10**

Table 9. Comparison of the computational time on Clinc→Banking under 10-way-1-shot. The best and second best results are emphasized using bold fonts and underlines, respectively.

Model	**Total Time**	**Time for Meta-Training**	**Time for Meta-Testing**
ProtoNet	1530 s	1209 s	321 s
HATT	1753 s	1448 s	305 s
DS-FSL	**896 s**	**772 s**	124 s
MLADA	1458 s	1219 s	239 s
DualAN	1651 s	1381 s	270 s
MultiDS	1051 s	938 s	**113 s**

6. Discussion

Here we discuss the potential application scenarios of our proposed method. MultiDS mainly utilizes distributional signatures of multi-domain data to empower few-shot model under domain shift. Consequently, MultiDS can be used in scenarios where data come from multiple domains and is few-labeled, such as fake news detection. Fake news detection involves information from multiple domains. At the same time, in real-world applications, fake news often breaks out from some novel domains with few annotations. To solve this problem, we take the existing fake news detection dataset as the meta-training dataset, and train our MultiDS model according to Algorithm 1. After meta-training, we use the emerging news to be detected as the meta-testing dataset and sample a large number of N-way-K-shot tasks from it. Well-trained MultiDS model can identify fake news from multiple domains using only a small amount of labeled data.

7. Conclusions

In this study, we propose a multi-level distributional signatures based model, MultiDS, to solve DG-FSTC problem. Firstly, we propose a hierarchical attention generator to translate multi-level distributional signatures into high-quality word-level attentions. We utilize a large news corpus to derive domain-agnostic attention. We extract domain-specific attention and class-specific attention from domain-related and class-related information using ridge regression and attention translator. Secondly, we propose a domain calibrator that uses domain features and domain-specific attention to generate domain calibration vectors. Experimental results show that our method achieves the best average performance on four testing datasets. The effectiveness of each module is verified by ablation experiments. In model stability validation, our model can exceed the baseline method in most cases. In the future, we will try to design better ways to use distributional signatures, not just word frequency.

Author Contributions: Conceptualization, X.W.; Methodology, X.W. and D.C.; Software, X.W. and X.L.; Validation, X.W. and D.C.; Formal analysis, Y.D. and X.L.; Investigation, Y.F.; Resources, Y.D.;

Data curation, X.C.; Writing—review and editing, Y.L.; Supervision, C.X. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation (Grant Nos. 61872298, 61802316, and 61902324), the Sichuan Regional Innovation Cooperation Project (Grant No. 2021YFQ0008) and Innovation Fund of Postgraduate, Xihua University (Grant No.YCJJ2021030).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study is available at https://github.com/Willy1005/DG-FSTC-and-DualAN accessed on 11 October 2022.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chakraborty, S.; Singh, A. Active Sampling for Text Classification with Subinstance Level Queries. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, Online Event, 22 February–1 March 2022; pp. 6150–6158.
2. Choi, S.; Jeong, M.; Han, H.; Hwang, S. C2L: Causally Contrastive Learning for Robust Text Classification. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, Online Event, 22 February–1 March 2022; pp. 10526–10534.
3. Finn, C.; Abbeel, P.; Levine, S. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017; pp. 1126–1135.
4. Snell, J.; Swersky, K.; Zemel, R.S. Prototypical Networks for Few-shot Learning. In Proceedings of the Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, Long Beach, CA, USA, 4–9 December 2017; pp. 4077–4087.
5. Geng, R.; Li, B.; Li, Y.; Zhu, X.; Jian, P.; Sun, J. Induction Networks for Few-Shot Text Classification. 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 2019, Hong Kong, China, 3–7 November 2019; pp. 3902–3911.
6. Ye, Z.; Ling, Z. Multi-Level Matching and Aggregation Network for Few-Shot Relation Classification. In Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, 28 July–2 August 2019; pp. 2872–2881.
7. Ohashi, S.; Takayama, J.; Kajiwara, T.; Arase, Y. Distinct Label Representations for Few-Shot Text Classification. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, Online Event, 1–6 Augest 2021; pp. 831–836.
8. Sun, P.; Ouyang, Y.; Zhang, W.; Dai, X. MEDA: Meta-Learning with Data Augmentation for Few-Shot Text Classification. In Proceedings of the 30th International Joint Conference on Artificial Intelligence, IJCAI 2021, Online Event, 19–26 August 2021; pp. 3929–3935.
9. Bao, Y.; Wu, M.; Chang, S.; Barzilay, R. Few-shot Text Classification with Distributional Signatures. In Proceedings of the 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, 26–30 April 2020.
10. Zhang, C.; Song, D. A Simple Baseline for Cross-Domain Few-Shot Text Classification. In Proceedings of the Natural Language Processing and Chinese Computing—10th CCF International Conference, NLPCC 2021, Qingdao, China, 13–17 October 2021; pp. 700–708.
11. Wang, X.; Du, Y.; Chen, D.; Li, X.; Chen, X.; Fan, Y.; Xie, C.; li Lee, Y.; Liu, J.; Li, H. Dual Adversarial Network with Meta Learning for Domain-Generalized Few-Shot Text Classification. *SSRN* **2022**. [CrossRef]
12. Larson, S.; Mahendran, A.; Peper, J.J.; Clarke, C.; Lee, A.; Hill, P.; Kummerfeld, J.K.; Leach, K.; Laurenzano, M.A.; Tang, L.; et al. 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 2019, Hong Kong, China, 3–7 November 2019; pp. 1311–1316.
13. Arora, S.; Liang, Y.; Ma, T. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In Proceedings of the 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 24–26 April 2017.
14. Karimi Mahabadi, R.; Zettlemoyer, L.; Henderson, J.; Mathias, L.; Saeidi, M.; Stoyanov, V.; Yazdani, M. Prompt-free and Efficient Few-shot Learning with Language Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 3638–3652.
15. Gu, Y.; Han, X.; Liu, Z.; Huang, M. PPT: Pre-trained Prompt Tuning for Few-shot Learning. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 8410–8423.
16. Mueller, A.; Krone, J.; Romeo, S.; Mansour, S.; Mansimov, E.; Zhang, Y.; Roth, D. Label Semantic Aware Pre-training for Few-shot Text Classification. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022, Dublin, Ireland, 22–27 May 2022; pp. 8318–8334.
17. Wang, C.; Wang, J.; Qiu, M.; Huang, J.; Gao, M. TransPrompt: Towards an Automatic Transferable Prompting Framework for Few-shot Text Classification. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online Event, 7–11 November 2021; pp. 2792–2802.

18. Zhang, J.; Bui, T.; Yoon, S.; Chen, X.; Liu, Z.; Xia, C.; Tran, Q.H.; Chang, W.; Yu, P.S. Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Online Event, 7–11 November 2021; pp. 1906–1912.
19. Zhang, H.; Zhang, Y.; Zhan, L.; Chen, J.; Shi, G.; Wu, X.; Lam, A.Y.S. Effectiveness of Pre-training for Few-shot Intent Classification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021, Online Event, 7–11 November 2021; pp. 1114–1120.
20. Chen, J.; Zhang, R.; Mao, Y.; Xu, J. ContrastNet: A Contrastive Learning Framework for Few-Shot Text Classification. In Proceedings of the 36th AAAI Conference on Artificial Intelligence, AAAI 2022, Online Event, 22 February–1 March 2022; pp. 10492–10500.
21. Luo, Q.; Liu, L.; Lin, Y.; Zhang, W. Don't Miss the Labels: Label-semantic Augmented Meta-Learner for Few-Shot Text Classification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, 1–6 August 2021; pp. 2773–2782.
22. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
23. Lang, K. NewsWeeder: Learning to Filter Netnews. In Proceedings of the 12th International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 331–339.
24. Bertinetto, L.; Henriques, J.F.; Torr, P.H.S.; Vedaldi, A. Meta-learning with differentiable closed-form solvers. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019.
25. Casanueva, I.; Temčinas, T.; Gerz, D.; Henderson, M.; Vulić, I. Efficient Intent Detection with Dual Sentence Encoders. In Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, Online, 9 July 2020; pp. 38–45.
26. Misra, R.; Grover, J. *Sculpting Data for ML: The First Act of Machine Learning*; University of California San Diego: La Jolla, CA, USA, 2021.
27. Liu, X.; Eshghi, A.; Swietojanski, P.; Rieser, V. Benchmarking Natural Language Understanding Services for Building Conversational Agents. In Proceedings of the Increasing Naturalness and Flexibility in Spoken Dialogue Interaction—10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24–26 April 2019; pp. 165–183.
28. Gao, T.; Han, X.; Liu, Z.; Sun, M. Hybrid Attention-Based Prototypical Networks for Noisy Few-Shot Relation Classification. In Proceedings of the 33rd AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, HI, USA, 27 January–1 February 2019; pp. 6407–6414.
29. Han, C.; Fan, Z.; Zhang, D.; Qiu, M.; Gao, M.; Zhou, A. Meta-Learning Adversarial Domain Adaptation Network for Few-Shot Text Classification. In Proceedings of the Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, 1–6 August 2021; pp. 1664–1673.
30. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.