Variational Hybrid-Attention Framework for Multi-Label Few-Shot Aspect Category Detection

Cheng Peng, Ke Chen [*], Lidan Shou, Gang Chen [*]

The State Key Laboratory of Blockchain and Data Security, Zhejiang University, Hangzhou 310000, China
{chengchng, chenk, should, cg}@zju.edu.cn

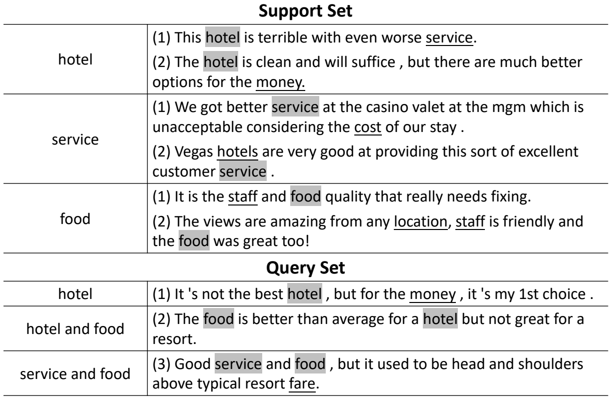
Abstract

Multi-label few-shot aspect category detection (FS-ACD) is a challenging sentiment analysis task, which aims to learn a multi-label learning paradigm with limited training data. The difficulty of this task is how to use limited data to generalize effective discriminative representations for different categories. Nowadays, all advanced FS-ACD works utilize the prototypical network to learn label prototypes to represent different aspects. However, such point-based estimation methods are inherently noise-susceptible and bias-vulnerable. To this end, this paper proposes a novel Variational Hybrid-Attention Framework (VHAF) for the FS-ACD task. Specifically, to alleviate the data noise, we adopt a hybrid-attention mechanism to generate more discriminative aspect-specific embeddings. Then, based on these embeddings, we introduce the variational distribution inference to obtain the aspect-specific distribution as a more robust aspect representation, which can eliminate the scarce data bias for better inference. Moreover, we further leverage an adaptive threshold estimation to help VHAF better identify multiple relevant aspects. Extensive experiments on three datasets demonstrate the effectiveness of our VHAF over other state-of-the-art methods. Code is available at https://github.com/chengzju/VHAF.

Introduction

Aspect category detection (ACD) (Pontiki et al. 2016) is an important task in sentiment analysis, which aims to discern the aspect categories discussed in a given sentence from a predefined set of aspect categories (e.g., price, food). Given that sentences frequently encompass multiple aspects, ACD inherently embodies a multi-label classification challenge. The efficacy of prevalent ACD methodologies (Hu et al. 2019; Li et al. 2020) is contingent upon substantial amounts of supervised data. However, in real scenarios, ACD usually suffers from a lack of training data due to the labor-intensive collection and annotation efforts.

Nowadays, the success of Few-Shot Learning (FSL) provides an effective solution to address the above challenges. FSL is a human-like learning paradigm that can quickly generalize novel classes with a few training data by exploiting prior knowledge. The pioneering work (Hu et al. 2021) is the first to address ACD in the few-shot scenario, which formulates a few-shot aspect category detection problem (FS-ACD). Afterward, many advanced FS-ACD works (Zhao et al. 2022; Liu et al. 2022) have been proposed. FS-ACD follows the meta-learning paradigm (Vinyals et al. 2016) by constructing a set of N-way K-shot meta-tasks. A meta-task aims to infer a query set with a small labeled support set, and an example case is shown in Figure 1. In this way, FS-ACD methods can achieve good generalization ability by leveraging a large number of different meta-tasks.



Support Set	
hotel	(1) This hotel is terrible with even worse service. (2) The hotel is clean and will suffice , but there are much better options for the money.
service	(1) We got better service at the casino valet at the mgm which is unacceptable considering the cost of our stay . (2) Vegas hotels are very good at providing this sort of excellent customer service .
food	(1) It is the staff and food quality that really needs fixing. (2) The views are amazing from any location, staff is friendly and the food was great too!
Query Set	
hotel	(1) It 's not the best hotel , but for the money , it 's my 1st choice .
hotel and food	(2) The food is better than average for a hotel but not great for a resort.
service and food	(3) Good service and food , but it used to be head and shoulders above typical resort fare.

Figure 1: A 3-way 2-shot meta-task example. The words in the gray background represent target aspects, while the words with underscores represent irrelevant aspects.

Despite the promise, the recent FS-ACD research still encounters several thorny challenges. The issues come from two folds. Firstly, since an ACD sentence may contain multiple aspects, the noise from irrelevant aspects will inevitably disturb the learning of the target aspect. As shown in Figure 1, as ”food” is the target aspect, ”staff” will be treated as a noise aspect for the sentence ”It is the staff and food quality that really needs fixing.” Although previous works (Hu et al. 2021; Zhao et al. 2022; Liu et al. 2022) exploit the aspect information to guide an attention mechanism to alleviate this issue, it is hard to ensure that novel aspects can establish accurate attention associations with sentence features during inference. Secondly, existing FS-ACD methods consistently follow the prototypical network (Snell et al.

[*] Ke Chen and Gang Chen are the corresponding authors.

2017), which learns a prototype for each aspect and uses the distance between query samples and prototypes to predict labels. Here, we call these methods as point estimation methods, which utilize a specific point to represent the same class samples. However, scarce support samples often lead to a biased point, which is insufficient for indicating the distribution of the entire class. Therefore, we emphasize that the support samples should be considered as sampling from a high-dimensional distribution, which is more worthwhile to be estimated. In summary, the above issues of FS-ACD should be properly addressed.

To this end, we propose a novel Variational Hybrid-Attention Framework (VHAF) for the FS-ACD task. Unlike the recent prototypical network-based works, VHAF exploits a variational distribution estimation that differs from point estimation to deal with scarce biased data for more robust performance. Specifically, our VHAF framework encapsulates three key components. First, we adopt a hybrid-attention mechanism to learn more discriminative aspect-specific embeddings. The hybrid attention consists of two modules, aspect-wise attention to alleviate the noise of irrelevant aspects and cross-instance attention to highlight highly consistent features among the same aspect instances. Then, based on these embeddings, we introduce a variational distribution inference strategy to represent each aspect with an aspect-specific distribution, which can better eliminate the bias of point estimation for a more robust prediction. Moreover, we further leverage an adaptive threshold estimation to help VHAF better identify positive aspects. The main contributions of this work can be summarized as follows:

- We propose a novel Variational Hybrid-Attention Framework (VHAF), which employs the variational distribution inference to derive more robust estimations from limited data. To the best of our knowledge, we are the first to leverage distribution estimation for FS-ACD.
- To alleviate the noise from irrelevant aspects and highlight highly consistent features among the same aspect instances, we design two effective attention mechanisms, i.e., aspect-wise attention and cross-instance attention.
- We conduct extensive experiments on three benchmark datasets. And results demonstrate that our proposed VHAF method achieves state-of-the-art performance.

Related Work

Aspect Category Detection

Aspect Category Detection (ACD) aims to categorize a sentence into a predefined aspect set. Previous works can be roughly divided into two types: unsupervised and supervised methods. Unsupervised methods use semantic association (Su et al. 2006) or co-occurrence frequency (Hai, Chang, and Kim 2011; Schouten et al. 2018) to extract aspects. Nonetheless, these methodologies demand significant corpus resources and exhibit results that are less remarkable. Supervised methods exploit hand-crafted features (Kiritchenko et al. 2014), representation learning (Zhou, Wan, and Xiao 2015), multi-task learning (Hu et al. 2019), or topic-attention model (Movahedi et al. 2019) to address the ACD task. However, these methods exhibit a strong dependence on substantial quantities of labeled data, prompting the exploration of the ACD task within a few-shot scenario.

Multi-label Few-shot Learning

Few-shot learning (FSL) is a human-like learning paradigm that can quickly generalize unseen classes with limited supervised data by exploiting the prior knowledge learned from seen classes. Many works have successfully applied FSL to computer vision (Liu et al. 2019; Vinyals et al. 2016) and natural language processing (Kumar et al. 2021; Yu et al. 2021). Compared with single-label FSL, the multi-label FSL is less investigated. Previous works focus on image synthesis (Alfassy et al. 2019), signal processing (Cheng, Chou, and Yang 2019), and intent detection (Hou et al. 2021).

Recently, Proto-AWATT (Hu et al. 2021) is the first work that formalizes aspect category detection in the few-shot scenario. It attempts to leverage support-set and query-set attention mechanisms to alleviate the negative effect of noisy aspects. Following Proto-AWATT, some prototypical network-based methods (Zhao et al. 2022; Liu et al. 2022) are proposed, which use label information as auxiliary knowledge to guide the attention mechanism to learn more discriminative aspect prototypes. However, all these methods use point estimation for inference, which are difficult to deal with biased scarce data for robust performance.

To tackle the above issues, inspired by variational FSL (Zhang et al. 2019), VHAF utilizes variational inference to predict the specific distribution for each aspect. Differently, our VHAF improves the distribution calculation process and directly utilizes the similarity between distributions for inference, thus achieving the adaptation to FS-ACD.

Preliminaries

Multi-Label Classification (MLC)

MLC needs to identify multiple labels for each instance. Given the training dataset $\mathcal{D}=\{(\boldsymbol{x}_i,\boldsymbol{y}_i)|\boldsymbol{x}_i\in\mathbb{X},\boldsymbol{y}_i\in\mathbb{Y}\}_{i=1}^{T}$, where $\mathbb{X}$ is the instance space and $\mathbb{Y}$ is the label space with $\mathcal{C}$ predifined classes. Each instance $\boldsymbol{x}_i$ is associated with a multi-hot label vector $\boldsymbol{y}_i=\{0,1\}^{\mathcal{C}}$, where the sign $\boldsymbol{y}_{i,j}=1$ indicates that $\boldsymbol{x}_i$ belongs to the class j, otherwise $\boldsymbol{y}_{i,j}=0$. MLC aims to learn a function $f:\mathbb{X}\times\mathbb{Y}\mapsto\mathbb{R}$ that can predict relevant labels for unseen instances. Specifically, the function f is usually formulated as a real-value function, and the score $f(\boldsymbol{x},c)$ evaluates the relevance between instance $\boldsymbol{x}$ and class $c\in\{1\dots\mathcal{C}\}$. Finally, the predicted relevant labels are derived as $\{c|f(\boldsymbol{x},c)>t\}$, where t is the threshold value.

Few-Shot Learning (FSL)

FSL aims to acquire prior knowledge with few samples to satisfy effective adaption to new tasks. In general, FSL performs a meta-learning paradigm (Vinyals et al. 2016) with two phases: meta-training and meta-testing. Given $\mathcal{C}$ predefined classes, the dataset is divided into base (seen) classes $\mathcal{C}_b$ and novel (unseen) classes $\mathcal{C}_n$ where $\mathcal{C}_b\cup\mathcal{C}_n=\mathcal{C}$ and $\mathcal{C}_b\cap\mathcal{C}_n=\emptyset$. The model is trained with $\mathcal{C}_b$ and then quickly adapted to tasks of $\mathcal{C}_n$. To ease learning, meta-learning constructs a set of meta-tasks on a N-way K-shot setting. In

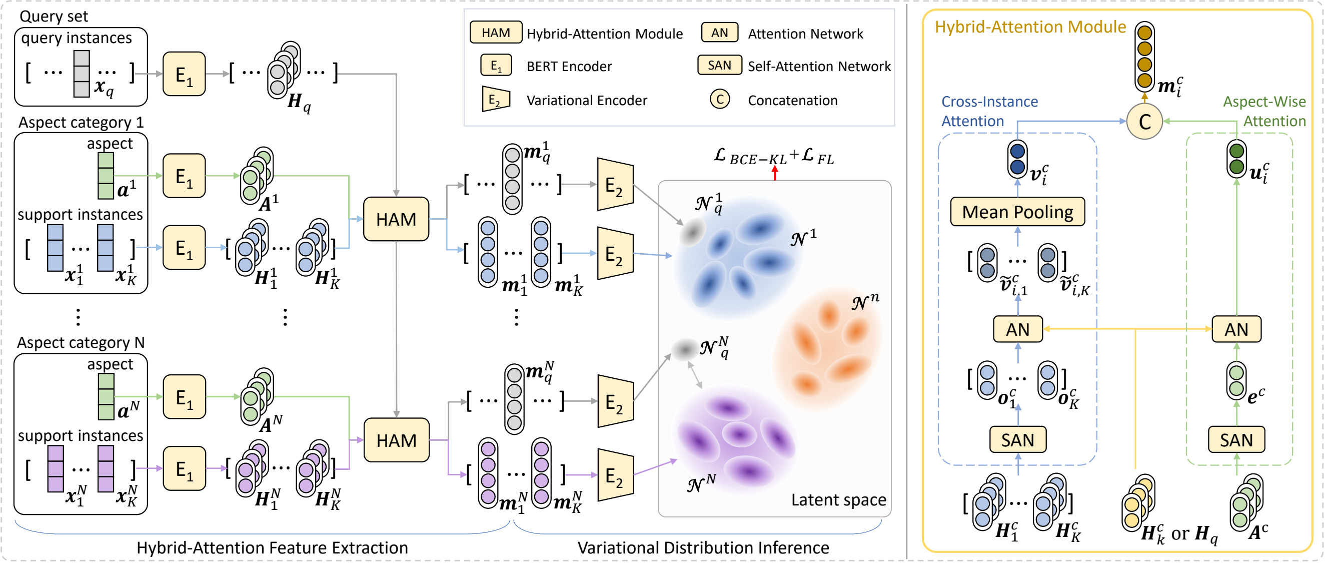


Figure 2: The overview of our VHAF framework. The detailed structure of the Hybrid-Attention Module is on the right.

each meta-task, we randomly select N classes with K instances to form a support set $\mathcal{S} = \{(\boldsymbol{x}_1^c, \dots, \boldsymbol{x}_K^c), \boldsymbol{a}^c\}_{c=1}^N$, where each $\boldsymbol{x}_k^c$ is an instance with the aspect category $\boldsymbol{a}^c$. Meanwhile, a query set is constructed with M instances sampled from the remaining data of the same N classes as $\mathcal{Q} = \{(\boldsymbol{x}_q, \boldsymbol{y}_q)\}_{q=1}^M$, where $\boldsymbol{y}_q$ is a binary label vector.

Methodology

As shown in Figure 2, the VHAF framework contains three components: hybrid-attention feature extraction, variational distribution inference, and adaptive threshold estimation.

Hybrid-attention Feature Extraction (HAFE)

Given a T-length input sentence $\boldsymbol{x}_i = \{w_{i,1}, w_{i,2}, \dots w_{i,T}\}$, we first utilize the pre-trained language model (e.g. BERT model (Devlin et al. 2019)) as the encoder to transform it into an embedding sequence $\boldsymbol{H}_i = [\boldsymbol{h}_{i,1}, \boldsymbol{h}_{i,2}, \dots, \boldsymbol{h}_{i,T}] \in \mathbb{R}^{T \times d}$. Then, to better extract the discriminative information about a specific aspect, we introduce a hybrid attention mechanism to generate more accurate representations that eliminate the distraction of noisy aspects. The hybrid attention consists of two modules, the aspect-wise attention to discover the most relevant semantic features to each aspect and the cross-instance attention to explore the highly similar features among instances within each aspect set.

Aspect-wise Attention (AWA) Considering that each instance is typically associated with multiple aspects, using one identical representation as the support embedding for different aspects can introduce noise information from other aspects. Therefore, we leverage the attention mechanism to discover the aspect-relevant features from $\boldsymbol{H}_i$ and filter out irrelevant semantic information. Specifically, the attentional representation $\boldsymbol{u}_i^c \in \mathbb{R}^d$ of the aspect c is calculated as:

$$\boldsymbol{u}_i^c = \sum_{t=1}^{T} \alpha_{i,t}^c \boldsymbol{h}_{i,t}\,, \;\; \alpha_{i,t}^c = \frac{\exp(\boldsymbol{e}^{c\top} \boldsymbol{h}_{i,t})}{\sum_{t'=1}^{T} \exp(\boldsymbol{e}^{c\top} \boldsymbol{h}_{i,t'})}, \tag{1}$$

where $\boldsymbol{e}^c \in \mathbb{R}^d$ denotes the embedding of the aspect c and $\alpha_{i,t}^c$ is the normalized coefficient of $\boldsymbol{h}_{i,t}$. To obtain the embedding $\boldsymbol{e}^c$, we first encode the label text $\boldsymbol{a}^c$ into an embedding matrix $\boldsymbol{A}^c \in \mathbb{R}^{T_c \times d}$ in the similar manner as $\boldsymbol{x}_i$, and then follow (Lin et al. 2017b) to learn a self-attention embedding as:

$$\begin{aligned} \boldsymbol{e}^c &= \text{squeeze}(\boldsymbol{R}^{c\top} \boldsymbol{A}^c), \\ \boldsymbol{R}^c &= \text{softmax}\left(\tanh\left(\boldsymbol{A}^c \boldsymbol{W}_1\right) \boldsymbol{W}_2\right), \end{aligned} \tag{2}$$

where $\boldsymbol{R}^c \in \mathbb{R}^{T_c \times 1}$ is the self-attention coefficient, $\boldsymbol{W}_1 \in \mathbb{R}^{d \times d_1}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_1 \times 1}$ are trainable parameter weights and are shared in the classes of all meta-tasks, which are learned to be class-agnostic. And squeeze$(\cdot)$ denotes the operation of squeezing dimensions to transform a matrix into a vector.

In this way, different aspects will adaptively focus on different semantic features, thereby the instance $\boldsymbol{x}_i$ can be decoupled into N aspect-specific representations as $\boldsymbol{U}_i = \{\boldsymbol{u}_i^c | c \in \{1, \dots, N\}\} \in \mathbb{R}^{N \times d}$.

Cross-instance Attention (CIA) Intuitively, instances with the same aspect usually contain similar semantic features, which are precisely discriminative information that can also reflect the corresponding aspect. Therefore, we employ the cross attention to explore the feature correlations among instances of the same aspect. Specifically, given the support subset $\mathcal{S}^c = \{\boldsymbol{x}_k^c | k \in \{1, \dots, K\}\}$ of the aspect c, we calculate the cross-attention representations of $\boldsymbol{x}_i$ with respect to all K $\boldsymbol{x}_k^c$. Taking $\boldsymbol{x}_k^c$ as an example, the cross-attention representation $\tilde{\boldsymbol{v}}_{i,k}^c \in \mathbb{R}^d$ is calculated as,

$$\tilde{\boldsymbol{v}}_{i,k}^c = \sum_{t=1}^{T} \beta_{i,t}^{c,k} \boldsymbol{h}_{i,t}\,, \;\; \beta_{i,t}^{c,k} = \frac{\exp(\boldsymbol{o}_k^{c\top} \boldsymbol{h}_{i,t})}{\sum_{t'=1}^{T} \exp(\boldsymbol{o}_k^{c\top} \boldsymbol{h}_{i,t'})}, \tag{3}$$

where $\boldsymbol{o}_k^c \in \mathbb{R}^d$ is the self-attention embedding of $\boldsymbol{x}_k^c$, which is generated similarly as $\boldsymbol{e}^c$ after obtaining its embedding matrix $\boldsymbol{H}_k^c$. In the same way, we can obtain all K representations as $\{\tilde{\boldsymbol{v}}_{i,k}^c | k \in \{1, \ldots, K\}\}$, and then we calculate the final cross-attention representation in an average manner as:

$$\boldsymbol{v}_i^c = \frac{1}{K} \sum_{k=1}^{K} \tilde{\boldsymbol{v}}_{i,k}^c \in \mathbb{R}^d. \tag{4}$$

It is worth noting that the support instance only performs CIA within its corresponding aspect subset to conclude common features, while the query instance performs CIA with all N support subsets separately to extract features that are highly similar to them.

After obtaining $\boldsymbol{u}_i^c$ and $\boldsymbol{v}_i^c$, we concatenate them into a holistic embedding $\boldsymbol{m}_i^c \in \mathbb{R}^{d_2}$ as the final representation to complement each other.

$$\boldsymbol{m}_i^c = [\boldsymbol{u}_i^c || \boldsymbol{v}_i^c] \in \mathbb{R}^{d_2}, \tag{5}$$

where || represents the concatenation operation.

Variational Distribution Inference (VDI)

Prior works (Hu et al. 2021; Zhao et al. 2022; Liu et al. 2022) usually follow (Snell et al. 2017) to aggregate K support instances into one feature vector (prototype), and then measure the distance between the query instance and the prototype to make predictions. However, a single vector is difficult to represent the whole class distribution, making such point estimation susceptible to data noise and bias. To this end, we propose to leverage a distribution-level measure for more robust effect. Therefore, we need to address two issues: how to represent aspect-specific distributions and how to leverage these distributions to estimate the output of query instances.

Given a query instance $\boldsymbol{x}_q$, we aim to infer its output via calculating the confidence score $p(\hat{\boldsymbol{y}}_q | \boldsymbol{z}(\boldsymbol{x}_q), \mathcal{Z})$, where $\boldsymbol{z}(\boldsymbol{x}_q)$ denotes the latent variable of $\boldsymbol{x}_q$ and $\mathcal{Z}$ denotes the entire aspect distribution. We introduce variational inference to model the posterior distribution of the variable $\boldsymbol{z}$. Specifically, we approximate the true posterior distribution $p(\boldsymbol{z}|\boldsymbol{x}_q)$ with another parameterized distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}_q)$ by minimizing the Kullback-Leibler (KL) divergence as:

$$D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_q) \| p(\boldsymbol{z}|\boldsymbol{x}_q)) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x}_q) \log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x}_q)}{p(\boldsymbol{z}|\boldsymbol{x}_q)}, \tag{6}$$

which is equivalent to maximizing the evidence lower bound (ELBO) defined as follows:

$$\mathrm{ELBO} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}_q)}[\log p(\boldsymbol{x}_q|\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_q) \| p(\boldsymbol{z})). \tag{7}$$

Based on the ELBO, we construct a simplified variational objective that allows efficient optimization and effective adaption for multi-label few-shot learning. First, $p(\boldsymbol{z})$ in Eq. 7 denotes the prior distribution, which is always assigned as an identical Gaussian $\mathcal{N}(0, I)$ (Kingma and Welling 2014). However, we emphasize that specifying a consistent and fixed prior hinders the model from generalizing specific distributions of different aspects. Therefore, we replace the prior with $p_\theta(\boldsymbol{z}|\mathcal{S}^c)$ by conditioning it on the aspect-specific support subset. Second, since we aim to develop a distribution estimation method rather than generating reconstructed features from $\boldsymbol{z}$, we focus on the second term in Eq. 7 and elaborate it with an aspect-specific KL divergence $\mathcal{L}_{kl}$ as:

$$\mathcal{L}_{KL} = \sum_{c=1}^{N} D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}_q) \| p_\theta(\boldsymbol{z}|\mathcal{S}^c)). \tag{8}$$

Then, the next step is how to define posterior distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x}_q)$ and prior distribution $p_\theta(\boldsymbol{z}|\mathcal{S}^c)$.

Estimate Prior Distribution After we obtain the support embedding set $\boldsymbol{M}^c = \{\boldsymbol{m}_k^c | k \in \{1, \ldots, K\}\}$ of aspect c, we directly map each embedding $\boldsymbol{m}_k^c$ to an individual Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_k^c, (\boldsymbol{\sigma}_k^c)^2)$ via:

$$\begin{aligned} \boldsymbol{\mu}_k^c &= (\boldsymbol{m}_k^c \boldsymbol{W}_3) \boldsymbol{W}_\mu \in \mathbb{R}^{d_z}, \\ \log((\boldsymbol{\sigma}_k^c)^2) &= (\boldsymbol{m}_k^c \boldsymbol{W}_3) \boldsymbol{W}_\sigma \in \mathbb{R}^{d_z}, \end{aligned} \tag{9}$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{d_2 \times d_3}$ and $\boldsymbol{W}_\mu, \boldsymbol{W}_\sigma \in \mathbb{R}^{d_3 \times d_z}$. Then we aggregate those k priors with the variance-weighted average operation, which produces the overall aggregated distribution $\mathcal{N}(\boldsymbol{\mu}^c, (\boldsymbol{\sigma}^c)^2)$ (abbreviated $\mathcal{N}^c$) for the aspect c as:

$$\boldsymbol{\mu}^c = \frac{\sum_{k=1}^{K} (\boldsymbol{\sigma}_k^c)^{-2} \boldsymbol{\mu}_k^c}{\sum_{k=1}^{K} (\boldsymbol{\sigma}_k^c)^{-2}}, \ (\boldsymbol{\sigma}^c)^2 = \frac{K}{\sum_{k=1}^{K} (\boldsymbol{\sigma}_k^c)^{-2}}. \tag{10}$$

Compared to the equal-weighted average operation, the variance-weighted average operation gives more weight to the distribution with less variance, thereby enhancing the more representative distribution and constraining the less important distribution.

Estimate Posterior Distribution Unlike the prior distribution, the posterior is solely conditional on the query instance $\boldsymbol{x}_q$. After obtaining the c-th aspect representation $\boldsymbol{m}_q^c$ of the query instance $\boldsymbol{x}_q$, we directly use the transformed Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_q^c, (\boldsymbol{\sigma}_q^c)^2)$ (abbreviated $\mathcal{N}_q^c$) as the posterior of $\boldsymbol{x}_q$ with respect to the aspect c.

It is worth noting that $\boldsymbol{x}_q$, as a query instance, will generate the corresponding distribution for each aspect, but this does not mean that is belongs to all aspects. Intuitively, when $\boldsymbol{x}_q$ belongs to the aspect c, $\mathcal{N}_q^c$ should have a low KL divergence with $\mathcal{N}^c$, high KL divergence otherwise. Therefore, given a query instance $\boldsymbol{x}_q$, we can compute the conditional probability $p(\hat{\boldsymbol{y}}_{q,c}|\boldsymbol{x}_q, \mathcal{S})$ (abbreviated $p(\hat{\boldsymbol{y}}_{q,c})$) to predict its aspect labels based on negative KL divergence as:

$$p(\hat{\boldsymbol{y}}_{q,c}|\boldsymbol{x}_q, \mathcal{S}) = \frac{\exp(-D_{\mathrm{KL}}(\mathcal{N}_q^c || \mathcal{N}^c))}{\sum_{c'=1}^{N} \exp(-D_{\mathrm{KL}}(\mathcal{N}_q^{c'} || \mathcal{N}^{c'}))}. \tag{11}$$

Based on the above considerations, we can further formulate Eq. 8 in a binary cross-entropy (BCE) loss form as:

$$\begin{aligned} \mathcal{L}_{BCE-KL} = -\sum_{c=1}^{N} \big[& \boldsymbol{y}_{q,c} \log \left(p(\hat{\boldsymbol{y}}_{q,c}|\boldsymbol{x}_q, \mathcal{S}) \right) + \\ & (1 - \boldsymbol{y}_{q,c}) \log \left(1 - p(\hat{\boldsymbol{y}}_{q,c}|\boldsymbol{x}_q, \mathcal{S}) \right) \big], \end{aligned} \tag{12}$$

where $\boldsymbol{y}_q$ is the ground-truth label of instance $\boldsymbol{x}_q$.

Model	5-way 5-shot		5-way 10-shot		10-way 5-shot		10-way 10-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Matching Network	97.05	81.89	97.49	84.62	96.30	70.95	96.72	73.28
Prototypical Network	96.49	83.30	97.53	86.29	95.97	74.23	96.71	76.83
Relation Network	93.31	75.79	90.86	72.02	91.81	63.78	90.54	61.15
Graph Network	96.54	81.45	97.46	85.04	95.45	70.75	96.97	77.84
Proto-HATT	96.45	83.33	97.62	86.71	95.71	73.42	97.00	77.65
Proto-AWATT	97.56	86.71	97.96	88.54	97.01	80.28	97.55	82.97
LDF	98.29	88.16	98.38	89.32	97.51	81.73	97.96	84.20
LPN	99.29	**94.43**	**99.49**	94.40	99.14	89.40	99.28	90.43
VHAF	**99.40**	94.26	99.43	**94.64**	**99.16**	**89.60**	**99.31**	**90.67**

Table 1: AUC and marco-F1 scores (%) on the FewAsp (single) dataset.

Adaptive Threshold Estimation (ATE)

To effectively determine multiple relevant aspects in a query instance, we further propose an adaptive threshold estimation method. Specifically, the first step is to derive the threhold from the support set. Given a N-way K-shot support set, we can obtain the aggregated distribution $\mathcal{N}^c$ for each aspect c according to Eq. 10. Similarly, we can aggregate the negative support embeddings $\{\boldsymbol{m}_i^c | \boldsymbol{y}_{i,c} = 0, i \in \{1, \dots, N \times K\}\}$ of the aspect c to generate a distribution $\bar{\mathcal{N}}^c$, as the negative distribution opposite to $\mathcal{N}^c$. Then, we calculate the KL divergence between $\mathcal{N}^c$ and $\bar{\mathcal{N}}^c$ as the input of a multi-layer perceptron (MLP) $g_t(\cdot)$ to estimate the threshold t^c for the aspect c:

$$\begin{aligned} t^c &= \mathrm{sigmoid}(\tanh(\boldsymbol{r}^c \boldsymbol{W}_4)\boldsymbol{W}_5)\,, \\ \boldsymbol{r}^c &= D_{\mathrm{KL}}(\bar{\mathcal{N}}^c || \mathcal{N}^c) \in \mathbb{R}^{d_z}, \end{aligned} \tag{13}$$

where $\boldsymbol{W}_4 \in \mathbb{R}^{d_z \times d_4}$ and $\boldsymbol{W}_5 \in \mathbb{R}^{d_4 \times 1}$.

The second step is to fit the threshold t^c to an appropriate value with query instances. Intuitively, an ideal threshold should be somewhere between ground-truth positive and negative predictions, acting as a division. To this end, we first scale the difference between $p(\hat{\boldsymbol{y}}_{q,c})$ and t^c by the Sigmoid activation function to between 0 and 1:

$$s = \mathrm{sigmoid}(p(\hat{\boldsymbol{y}}_{q,c}) - t^c). \tag{14}$$

Then, we introduce a focal loss (Lin et al. 2017a) to fine-tune the MLP $g_t(\cdot)$ to generate the appropriate thresholds as:

$$\mathcal{L}_{FL} = \begin{cases} -\alpha(1-s)^\gamma \log(s), & \text{when } \boldsymbol{y}_{q,c} = 1 \\ -(1-\alpha)s^\gamma \log(1-s), & \text{when } \boldsymbol{y}_{q,c} = 0 \end{cases}, \tag{15}$$

where $\gamma \geq 0$ is the tunable focusing parameter and $\alpha \in [0, 1]$ is a weighting factor to balance the importance of positive/negative examples. To avoid ATE affecting the inference of $p(\hat{\boldsymbol{y}}_{q,c})$, we stop gradient back-propagation so that the loss $\mathcal{L}_{FL}$ is only used to train the MLP $g_t(\cdot)$.

Combining Eq. 12 and 15, the final loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{BCE-KL} + \lambda \mathcal{L}_{FL}, \tag{16}$$

where λ is the trade-off parameter. By minimizing the loss $\mathcal{L}$, we train the entire framework in an end-to-end manner.

Experiments

Datasets

We perform experiments on three benchmark datasets: FewAsp (single), FewAsp (multi) and FewAsp. All datasets are constructed from YelpAspect (Bauman, Liu, and Tuzhilin 2017), which is a large-scale multi-domain dataset for aspect recommendation. Following (Han et al. 2018), we split the 100 aspects without intersection into 64 aspects for training, 16 aspects for validation, and 20 aspects for testing. Specifically, FewAsp (single) consists of single-aspect sentences, FewAsp (multi) consists of multi-aspect sentences, and FewAsp consists of both types of sentences.

Baselines

- **Matching Network** (Vinyals et al. 2016) develops an attention and memory-based method with a distance metric based on the cosine similarity.
- **Prototypical Network** (Snell et al. 2017) measures the distance between query instances and prototypes learned from the support set to achieve classification.
- **Relation Network** (Sung et al. 2018) utilizes a deep neural network to learn the relation between query and support samples instead of fixed metrics.
- **Graph Network** (Garcia and Bruna 2018) novelty casts the few-shot learning as a supervised message passing task via graph neural network.
- **Proto-HATT** (Gao et al. 2019) develops an attention-based prototypical network that addresses the noise with hybrid instance- and feature-level attention mechanisms.
- **Proto-AWATT** (Hu et al. 2021) devises support-set and query-set attention to alleviate the noise and learns a dynamic threshold per instance by a policy network.
- **LDF** (Zhao et al. 2022) introduces label-guided attention strategy and label-weighted contrastive loss to produce denoised prototypes.
- **LPN** (Liu et al. 2022) proposes a label-enhanced prototypical network with contrastive learning to facilitate the learning of discriminative prototypes.

Evaluation Metrics

Following (Hu et al. 2021; Liu et al. 2022), we adopt two metrics, i.e., AUC (Area Under Curve) and macro-F1 score.

Model	5-way 5-shot		5-way 10-shot		10-way 5-shot		10-way 10-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Matching Network	89.54	65.70	91.38	69.02	88.28	50.86	89.94	54.42
Prototypical Network	89.67	67.88	91.60	72.32	88.01	52.72	90.68	58.92
Relation Network	84.91	58.38	86.21	61.37	84.22	43.71	84.72	44.85
Graph Network	87.97	59.25	90.45	64.63	86.05	45.42	88.44	48.49
Proto-HATT	91.10	69.15	93.03	73.91	90.44	55.34	92.38	60.21
Proto-AWATT	91.45	71.72	93.89	77.19	89.80	58.89	92.34	66.76
LDF	92.62	73.38	94.34	78.81	90.87	62.06	92.93	68.23
LPN	95.66	79.48	96.55	82.81	94.51	67.28	95.66	71.87
VHAF	**97.09**	**84.64**	**97.57**	**87.31**	**96.01**	**75.92**	**96.78**	**79.43**

Table 2: AUC and marco-F1 scores (%) on the FewAsp (multi) dataset.

Model	5-way 5-shot		5-way 10-shot		10-way 5-shot		10-way 10-shot	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Matching Network	90.76	67.14	92.39	70.09	88.44	51.27	89.90	54.61
Prototypical Network	88.88	66.96	91.77	73.27	87.35	52.06	90.13	59.03
Relation Network	85.56	59.52	86.98	62.78	84.94	45.62	83.77	44.70
Graph Network	89.48	61.49	92.35	69.89	87.35	47.91	90.19	56.06
Proto-HATT	91.54	70.26	93.43	75.24	90.63	57.26	92.86	61.51
Proto-AWATT	93.35	75.37	95.28	80.16	92.06	65.65	93.42	69.70
LDF	94.65	78.27	95.71	81.87	92.74	67.13	94.29	71.97
LPN	96.45	82.22	97.15	84.90	95.36	71.42	96.55	76.51
VHAF	**97.88**	**87.25**	**98.17**	**89.22**	**97.02**	**79.72**	**97.58**	**82.41**

Table 3: AUC and marco-F1 scores (%) on the FewAsp dataset.

Implementation Details

We adopt the BERT-base model (Devlin et al. 2019) as the encoding backbone. Following (Liu et al. 2022), we freeze the first 6 layers of BERT and fine-tune the final 6 layers. For model parameters, we set $d = d_z = 768$, $d_1 = d_4 = 256$, $d_2 = 1536$ and $d_3 = 1024$. For the loss function, we set $\lambda = 1$, $\gamma = K$ and $\alpha = 1 - 1/N$. We employ the AdamW (Wolf et al. 2019) optimizer with the initial learning rate 1e-5. For better statistical robustness, all experiments are repeated 5 runs to reduce randomness, and the results are averaged over 600 test episodes in each run. We perform experiments with 5/10-way and 5/10-shot settings on all datasets. All experiments are conducted with one NVIDIA RTX A5000 GPU.

Performance Comparison

The overall performance comparisons on datasets FewAsp (single), FewAsp (multi), and FewAsp are shown in Table 1, 2 and 3 respectively. From these tables, we can observe that:

(1) Our VHAF achieves the best results on almost all metrics of three datasets, which demonstrates the effectiveness of VHAF. Specifically, on FewAsp, VHAF significantly outperforms the best baseline approach LPN by 1.02%-1.50% and 4.50%-8.64% in terms of AUC and macro-F1 score respectively. On FewAsp (multi), VHAF leads a performance boost of 0.93%-1.66% and 4.32%-8.30% upon LPN in terms of AUC and macro-F1 score respectively. This indicates that complementary AWA and CIA modules can adequately capture discriminative features for different aspects. Meanwhile, the robust measurement based on aspect-specific distribution between the support set and query instances can address the bias and noise issues caused by scarce data.

Models			FewAsp		FewAsp (multi)	
ID	HAFE	VDI	AUC	F1	AUC	F1
1	w/o HA	w/o	96.54	83.06	93.17	64.46
2	w/ AWA	w/o	97.26	85.82	94.69	72.65
3	w/ CIA	w/o	97.84	87.15	94.32	69.53
4	w/ HA	w/o	97.90	88.38	95.56	74.99
5	w/ HA	w/	**98.17**	**89.22**	**96.01**	**75.92**

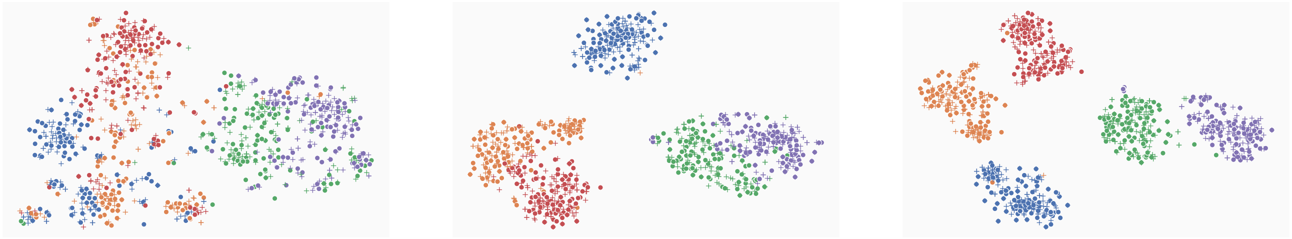
Table 4: Ablation study of the 10-way 5-shot scenario on FewAsp and FewAsp (multi) datasets. All ablated models adopt adaptive threshold estimation for a fair comparison.

(2) For all methods, the performance on FewAsp (multi) is consistently worse than on FewAsp and FewAsp (single). This is because the instances in FewAsp (multi) have more aspects, which increases the complexity of the dataset. Despite this, compared to other methods, our VHAF still maintains optimal performance compared on FewAsp (multi), which further illustrates the robustness of our method in dealing with more severe data noise.

Ablation Study

To further demonstrate the effectiveness of each component, we conduct the ablation study. As shown in Table 4:

(1) *Effect of hybrid-attention feature extraction.* To implement HAFE: *w/o HA* means not using any attention mechanism to generate the aspect-specific representation but using a unified self-attention representation as in Eq. 2. On the contrary, *w/ HA* means using the hybrid-attention mechanism. *w/ AWA* and *w/ CIA* means only adopting aspect-wise attention or cross-instance attention respectively. It can



(a) VHAF w/o HA&VDI (b) VHAF w/o VDI (c) VHAF

Figure 3: Visualization of aspect-specific embeddings and distributions obtained from VHAF w/o HA&VDI, VHAF w/o VDI and VHAF respectively. Circles represent support instances, while plus signs represent query instances.

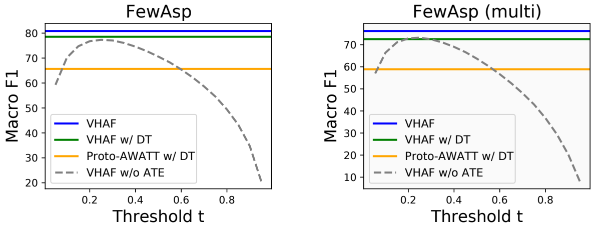


Figure 4: Macro-F1 scores of different thresholds on FewAsp and FewAsp (multi) in the 10-way 5-shot scenario.

be observed that without the HA module, the performance drops a lot, indicating that a unified instance representation cannot eliminate the noise from irrelevant tags. Then we can also see that using AWA or CIA can achieve various degrees of performance improvement, while using both of them achieves the optimal effect, which confirms both attention mechanisms help reduce noise for better representations and also illustrates the complementary effect between them.

(2) *Effect of variational distribution inference. w/o VDI* means that the VHAF model does not introduce the VDI module to generate aspect-specific distribution for inference, but directly aggregates the embeddings extracted from the HAFE module into an aspect prototype to perform point estimation. We can observe that without the VID module, VHAF shows a huge performance degradation, which validates that VDI can help the model obtain a more robust aspect distribution to overcome the scarce data bias.

Effect of Adaptive Thresholds

To verify the effectiveness of the adaptive threshold, we evaluate the impact of different thresholds. As shown in Figure 4, VHAF holds the best performance, which validates the effectiveness of adaptive threshold estimation. However, VHAF *w/o ATE* can only achieve a decent effect when given an appropriate threshold, which is not easily grasped. The DT strategy (Hu et al. 2021) learns a dynamic threshold via the policy network. VHAF *w/ DT* achieves higher macro-F1 than Proto-AWATT *w/ DT*, which indicates that the proposed variational distribution inference with hybrid-attention feature extraction provides a more effective and robust measurement. Moreover, VHAF slightly outperforms VHAF *w/ DT*, which indicates the advantage of adaptive thresholds based on feature distribution. In addition, DT, as a two-stage method, makes the training process more complicated, so our method is more cost-effective.

Visualization

We use t-SNE (Van der Maaten and Hinton 2008) to visualize the extracted feature embeddings and aggregated distributions. We randomly select 5 aspects from the test set of FewAsp (multi) and then sample 20 times of 5-way 5-shot meta-tasks for these aspects. Each meta-task has 5 query instances per aspect. Therefore, with a total of 500 support instances and 500 query instances, Figure 3 shows the visualization result of feature embeddings obtained from *VHAF w/o HA&VDI* and *VHAF w/o VDI*, and distribution obtained from *VHAF*. Data points with the same color represent instances of the same aspect.

In Figure 3 a, without hybrid attention, the extracted embeddings of different aspects are lumped together, which means that the noise from irrelevant aspects cannot be effectively eliminated. In contrast, hybrid attention facilitates *VHAF w/o VDI* (Figure 3 b) to a better embedding separating effect. For ease of presentation, we only use the mean vector to represent the individual distribution of each embedding in Figure 3 c. By using both HA and VDI, *VHAF* produces well-separated and more distinguishable aspect-specific distribution to deal with the bias from scarce data.

Conclusion

In this paper, we propose a variational hybrid-attention framework (VHAF) for multi-label few-shot aspect category detection. Different from the recent point estimation methods based on the prototypical network, our VHAF devises a variational distribution inference approach that can obtain more robust estimates under limited data. To better capture discriminative embeddings to promote distribution approximation, we utilize aspect-wise attention and cross-instance attention to alleviate the noise of irrelevant aspects and highlight highly consistent features. Moreover, we further leverage an adaptive threshold estimation to achieve better multi-label inference. Extensive experimental results on three datasets demonstrate the effectiveness of our VHAF over state-of-the-art methods. In future work, we hope to generalize our VHAF to more multi-label few-shot tasks.

Acknowledgments

This work is supported by the National Key R&D Program of China (No.2022YFB3304100) and the Pioneer R&D Program of Zhejiang "Data Preparation, Optimization, and Augmentation for Domain-Specific Large Models".

References

Alfassy, A.; Karlinsky, L.; Aides, A.; Shtok, J.; Harary, S.; Feris, R.; Giryes, R.; and Bronstein, A. M. 2019. Laso: Label-set operations networks for multi-label few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6548–6557.

Bauman, K.; Liu, B.; and Tuzhilin, A. 2017. Aspect based recommendations: Recommending items with the most valuable aspects based on user reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 717–725.

Cheng, K.-H.; Chou, S.-Y.; and Yang, Y.-H. 2019. Multi-label few-shot learning for sound event recognition. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 1–5. IEEE.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 4171–4186.

Gao, T.; Han, X.; Liu, Z.; and Sun, M. 2019. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI conference on artificial intelligence*, 6407–6414.

Garcia, V.; and Bruna, J. 2018. Few-shot learning with graph neural networks. In *ICLR*.

Hai, Z.; Chang, K.; and Kim, J.-j. 2011. Implicit feature identification via co-occurrence association rule mining. In *Computational Linguistics and Intelligent Text Processing: 12th International Conference, CICLing 2011, Tokyo, Japan, February 20-26, 2011. Proceedings, Part I 12*, 393–404. Springer.

Han, X.; Zhu, H.; Yu, P.; Wang, Z.; Yao, Y.; Liu, Z.; and Sun, M. 2018. Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. In *EMNLP*.

Hou, Y.; Lai, Y.; Wu, Y.; Che, W.; and Liu, T. 2021. Few-shot learning for multi-label intent detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13036–13044.

Hu, M.; Zhao, S.; Guo, H.; Xue, C.; Gao, H.; Gao, T.; Cheng, R.; and Su, Z. 2021. Multi-label few-shot learning for aspect category detection. In *ACL*.

Hu, M.; Zhao, S.; Zhang, L.; Cai, K.; Su, Z.; Cheng, R.; and Shen, X. 2019. CAN: constrained attention networks for multi-aspect sentiment analysis. In *EMNLP*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Kiritchenko, S.; Zhu, X.; Cherry, C.; and Mohammad, S. 2014. NRC-Canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, 437–442.

Kumar, M.; Kumar, V.; Glaude, H.; de Lichy, C.; Alok, A.; and Gupta, R. 2021. Protoda: Efficient transfer learning for few-shot intent classification. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, 966–972. IEEE.

Li, Y.; Yin, C.; Zhong, S.-h.; and Pan, X. 2020. Multi-instance multi-label learning networks for aspect-category sentiment analysis. In *EMNLP*.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017a. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017b. A structured self-attentive sentence embedding. In *ICLR*.

Liu, H.; Zhang, F.; Zhang, X.; Zhao, S.; Sun, J.; Yu, H.; and Zhang, X. 2022. Label-enhanced prototypical network with contrastive learning for multi-label few-shot aspect category detection. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 1079–1087.

Liu, M.-Y.; Huang, X.; Mallya, A.; Karras, T.; Aila, T.; Lehtinen, J.; and Kautz, J. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10551–10560.

Movahedi, S.; Ghadery, E.; Faili, H.; and Shakery, A. 2019. Aspect category detection via topic-attention network. *arXiv preprint arXiv:1901.01183*.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, 19–30. Association for Computational Linguistics.

Schouten, K.; Van Der Weijde, O.; Frasincar, F.; and Dekker, R. 2018. Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data. *IEEE transactions on cybernetics*, 48(4): 1263–1275.

Snell, J.; Swersky, K.; Zemel, R. S.; and O, O. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Su, Q.; Xiang, K.; Wang, H.; Sun, B.; and Yu, S. 2006. Using pointwise mutual information to identify implicit features in customer reviews. In *International Conference on Computer Processing of Oriental Languages*, 22–30. Springer.

Sung, F.; Yang, Y.; Zhang, L.; Xiang, T.; Torr, P. H.; and Hospedales, T. M. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1199–1208.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Yu, D.; He, L.; Zhang, Y.; Du, X.; Pasupat, P.; and Li, Q. 2021. Few-shot intent classification and slot filling with retrieved examples. In *NAACL-HLT*.

Zhang, J.; Zhao, C.; Ni, B.; Xu, M.; and Yang, X. 2019. Variational few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1685–1694.

Zhao, F.; Shen, Y.; Wu, Z.; and Dai, X. 2022. Label-Driven Denoising Framework for Multi-Label Few-Shot Aspect Category Detection. In *EMNLP*.

Zhou, X.; Wan, X.; and Xiao, J. 2015. Representation learning for aspect category detection in online reviews. In *Proceedings of the AAAI conference on artificial intelligence*.