Few-shot Multimodal Sentiment Analysis Based on Multimodal Probabilistic Fusion Prompts

Xiaocui Yang*
Northeastern University, China
yangxiaocui@stumail.neu.edu.cn

Shi Feng
Northeastern University, China
fengshi@cse.neu.edu.cn

Daling Wang
Northeastern University, China
wangdaling@cse.neu.edu.cn

Yifei Zhang
Northeastern University, China
zhangyifei@cse.neu.edu.cn

Soujanya Poria
Singapore University of Technology and Design, Singapore
sporia@sutd.edu.sg

ABSTRACT

Multimodal sentiment analysis has gained significant attention due to the proliferation of multimodal content on social media. However, existing studies in this area rely heavily on large-scale supervised data, which is time-consuming and labor-intensive to collect. Thus, there is a need to address the challenge of few-shot multimodal sentiment analysis. To tackle this problem, we propose a novel method called **Multi**modal **Pro**babilistic Fus**i**o**n** Promp**t**s (**MultiPoint**[1]) that leverages diverse cues from different modalities for multimodal sentiment detection in the few-shot scenario. Specifically, we start by introducing a **C**onsistently **D**istributed **S**ampling approach called **CDS**, which ensures that the few-shot dataset has the same category distribution as the full dataset. Unlike previous approaches primarily using prompts based on the text modality, we design unified multimodal prompts to reduce discrepancies between different modalities and dynamically incorporate multimodal demonstrations into the context of each multimodal instance. To enhance the model's robustness, we introduce a probabilistic fusion method to fuse output predictions from multiple diverse prompts for each input. Our extensive experiments on **six** datasets demonstrate the effectiveness of our approach. First, our method outperforms strong baselines in the multimodal few-shot setting. Furthermore, under the same amount of data (1% of the full dataset), our CDS-based experimental results significantly outperform those based on previously sampled datasets constructed from the same number of instances of each class.

CCS CONCEPTS

• **Information systems → Multimedia streaming**; • **Computing methodologies → Natural language processing**.

KEYWORDS

Multimodal sentiment analysis, Multimodal few-shot, Consistently distributed sampling, Unified multimodal prompt, Multimodal demonstrations, Multimodal probabilistic fusion

ACM Reference Format:
Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Soujanya Poria. 2023. Few-shot Multimodal Sentiment Analysis Based on Multimodal Probabilistic Fusion Prompts. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612181

1 INTRODUCTION

With the growing popularity of multimedia platforms, there has been an explosion of data containing multiple modalities such as text, image, video, and etc. Multimodal Sentiment Analysis (MSA) has emerged as a popular research topic due to its wide applications in market prediction, business analysis, and more [1, 11, 35]. In this paper, we specifically focus on the task of multimodal text-image sentiment analysis, which comprises of two subtasks: coarse-grained MSA and fine-grained MSA. Coarse-grained MSA aims to detect the overall sentiment of a text-image pair [13, 23, 26, 27]. On the other hand, fine-grained MSA, also known as Multimodal Aspect-Based Sentiment Classification (MASC), seeks to detect the targeted sentiment for a specific aspect term that is dependent on the corresponding text-image pair [8, 10, 14, 24, 25, 28]. Multimodal sentiment analysis has witnessed significant progress in recent years. Early research primarily focuses on constructing rich and large-scale datasets to facilitate model training [17, 19, 26, 32, 34]. Subsequent studies aim at improving the performance of MSA through the integration of various effective technologies, such as Contrastive Learning [13], Vision-Language Pre-training [14], among others.

One of the limitations of existing multimodal sentiment analysis models is dependency on large-scale annotated datasets, which can be expensive and challenging to obtain. In real-world applications, only a limited amount of labeled data is available, making it more practical to investigate few-shot learning methods that can perform well in low-resource settings. However, in the multimodal few-shot learning setting, it can be challenging to sample diverse and comprehensive few-shot datasets. Existing few-shot classification tasks, such [30, 31], typically sample the same number of instances for each label, without considering the consistency of the category distribution between the full dataset (before sampling)

[*] This paper is completed at Singapore University of Technology and Design with funding from CSC.

[1] Our code and data can be found in the https://github.com/YangXiaocui1215/MultiPoint.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612181

and the few-shot dataset (after sampling). This approach can result in imbalanced and biased few-shot datasets that do not reflect the true distribution of the full dataset. To address this issue, we introduce a novel sampling approach called Consistently Distributed Sampling (**CDS**), which ensures that the few-shot dataset has a category distribution similar to that of the full dataset.

Prompt-based methods have become popular in few-shot learning because they allow pre-trained models to generalize to new tasks with limited or no training data. Despite being widely used for few-shot text tasks, such as LM-BFF [6] and GFSC [7], prompts are rarely utilized in multimodal scenarios. To address this gap, Yu et al. propose a prompt-based vision-aware language modeling (PVLM) approach [30] and a unified pre-training for multimodal prompt-based fine-tuning (UP-MPF) [31] for multimodal sentiment analysis (MSA). PVLM and UP-MPF simply introduce image tokens to a pre-trained language model (PLM) for prompt-based fine-tuning. However, directly feeding image representations into the language model raises the issue of modality discrepancy, as the image encoder is language-agnostic. This can result in suboptimal performance in capturing multimodal cues from multiple modalities. Additionally, it has been observed that different prompts may contain varying amounts of information, and the information conveyed by a single prompt may be insufficient for effective multimodal sentiment analysis. However, previous works on few-shot text tasks [6, 7] and multimodal tasks [30, 31] only apply a single prompt to different models, without considering the fusion of different prompts.

To alleviate the problems raised above, we propose a novel model for Few-shot MSA called **Multi**modal **Pro**babilistic Fus**i**o**n** Promp**t**s, **MultiPoint**, depicted in Figure 1. To begin, we design unified multimodal prompts for our task, as shown in Table 1. For the text modality, we use both manual prompts based on domain knowledge and task-specific requirements, as well as generated prompts that capture diverse and valuable information from pre-trained language models. For the image modality, we generate a textual description of each image and use it as the image prompt to improve compatibility and mitigate discrepancies between the image and text modalities. The text and image prompts are then combined to create a unified multimodal prompt. To improve the robustness of our model, we select the most similar multimodal instances from the training dataset as multimodal demonstrations that are introduced as the multimodal context for each instance. As previously mentioned, the information obtained from a single prompt is limited and different prompts can capture diverse cues from the data. To this end, we propose a novel probabilistic fusion method, based on Bayesian Fusion, which has been shown to be robust in increasingly discrepant sub-posterior scenarios [5]. Our probabilistic fusion approach allows us to incorporate uncertainty in the predictions from different prompts and obtain a more reliable and accurate prediction for each instance. We evaluate our approach on **six** multimodal sentiment datasets through extensive experiments. Our main contributions are summarized as follows:

- We introduce a **C**onsistently **D**istributed **S**ampling approach, called **CDS**, which ensures that the category distribution of the few-shot dataset (only 1% of the full dataset) is similar to that of the full dataset. This approach helps create representative few-shot datasets and enables more accurate evaluation of our model's performance.
- We propose a novel model for Few-shot MSA called **Multi**modal **Pro**babilistic Fus**i**o**n** Promp**t**s (**MultiPoint**). Our model employs unified multimodal prompts with multimodal demonstrations to mitigate the discrepancy between different modalities. Furthermore, probabilistic fusion aggregates predictions from multiple multimodal prompts, enhancing the effectiveness of our model.
- We evaluate MultiPoint and CDS on **six** multimodal sentiment datasets. Our results in the few-shot setting demonstrate that MultiPoint outperforms strong baselines and showcase the benefits of utilizing consistent distribution information.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis (MSA)

MSA encompasses both coarse-grained MSA and fine-grained MSA. **For coarse-grained MSA**, some datasets have been proposed include MVSA-Single and MVSA-Multiple datasets [19], and the TumEmo dataset [26]. Researchers have proposed various methods to tackle the challenges of multimodal sentiment analysis, including co-memory attentional model [23], Multi-channel Graph Neural Networks [27], Contrastive Learning and Multi-Layer Fusion (CLMLF) method [13], and more. **For fine-grained MSA**, there are several datasets for aspect-based sentiment classification, including the Twitter-2015 and Twitter-2017 datasets [17, 32]. Additionally, a large-scale dataset called MASAD (Multimodal Aspect-based Sentiment Analysis Dataset) is built [34] to facilitate research. Several approaches have been proposed to address the challenges of fine-grained MSA. Initially, researchers expand BERT to the multimodal scenario, such as TomBERT [29], EF-CapTrBERT [12]. Recently, external knowledge is introduced to solve fine-grained MSA, e.g., FITE employing facial information [24], KEF with knowledge-enhanced [14], and VLP-MABSA leveraging external pre-training data and multiple pre-traing tasks [14]. Collecting and annotating multimodal data for multimodal sentiment analysis is time-intensive and laborious. To this end, we devote to the multimodal sentiment analysis task in few-shot scenarios.

2.2 Few-shot Learning with PLM

Prompt-based language modeling has emerged as a powerful approach for solving different few-shot tasks using pre-trained language models (PLM) [15]. Prompt-based methods treat the classification task as a masked language modeling (MLM) task, where the model is fine-tuned with a set of prompts to guide its predictions. In the beginning, prompt-based approaches are introduced to handle text few-shot classification task, including LM-BFF [6], LM-SC [9], and so on. Ehsan et al. [7] propose a generative language model (GFSC) that reformulates the task as a language generation problem for text classification. However, the above-mentioned models only handle text-related tasks. Recently, there has been an increasing interest in designing models to handle few-shot multimodal tasks. Existing models for few-shot multimodal tasks, such as Frozen [22], PVLM [30], and UP-MPF [31], primarily rely on introducing image tokens to a pre-trained language model for prompt-based fine-tuning. However, these approaches face the challenge of discrepancy between different modalities since image features are agnostic to language models. To this end, we propose a novel unified multimodal prompt that allows for the joint processing of both text and image modalities in a coherent manner.

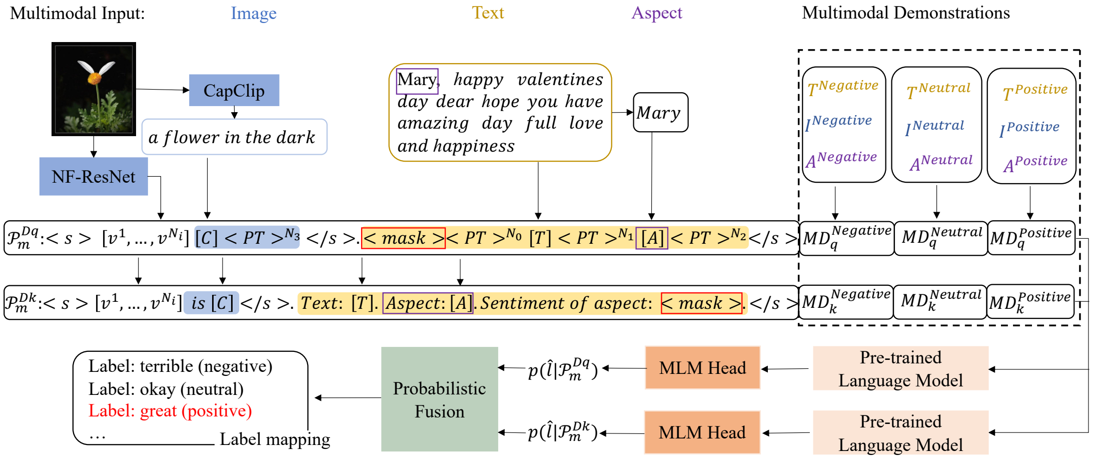

Multimodal Input:
Image
Text
Aspect
Multimodal Demonstrations
CapClip
a flower in the dark
NF-ResNet
Mary, happy valentines day dear hope you have amazing day full love and happiness
Mary
$T^{Negative}$ $I^{Negative}$ $A^{Negative}$
$T^{Neutral}$ $I^{Neutral}$ $A^{Neutral}$
$T^{Positive}$ $I^{Positive}$ $A^{Positive}$
$\mathcal{P}_m^{Dq}$:< s > $[v^1, ..., v^{N_i}]$ [C] < PT >^{N_3} </s >.< mask >< PT >^{N_0} [T] < PT >^{N_1} [A] < PT >^{N_2}</s >
$MD_q^{Negative}$ $MD_q^{Neutral}$ $MD_q^{Positive}$
$\mathcal{P}_m^{Dk}$:< s > $[v^1, ..., v^{N_i}]$ is [C] </s >. Text: [T]. Aspect: [A]. Sentiment of aspect: < mask >.</s >
$MD_k^{Negative}$ $MD_k^{Neutral}$ $MD_k^{Positive}$
Label: terrible (negative)
Label: okay (neutral)
Label: great (positive)
...
Label mapping
Probabilistic Fusion
$p(\hat{l}|\mathcal{P}_m^{Dq})$
$p(\hat{l}|\mathcal{P}_m^{Dk})$
MLM Head
MLM Head
Pre-trained Language Model
Pre-trained Language Model

Figure 1: An illustration of our proposed Multimodal Probabilistic Fusion Prompts (MultiPoint) model for Few-shot Multimodal Sentiment Analysis. We design different unified multimodal prompts with multimodal demonstrations, e.g., $\mathcal{P}_m^{Dq}$ and $\mathcal{P}_m^{Dk}$, here q and k indicate the q-th and k-th multimodal prompt for one instance. A multimodal prompt ($\mathcal{P}_m$) is composed of multiple image slots ($[v^1, ..., v^{N_i}]$), image prompt (blue highlight), and the task-specific text prompt (yellow highlight). C is the image caption from ClipCap, T is the original text, A is an aspect term for fine-grained datasets, which does not exist in coarse-grained datasets. $< mask >, < s >$ and $< /s >$ are special tokens in Pre-trained Language Model. The black dashed boxes represent various demonstrations based on label space, take the $\mathcal{L} = \{Negative, Neutral, Positive\}$ as an example. The multimodal demonstration MD_q for the q-th instance is dynamically selected based on the similarity score with the training dataset for a specific label from $\mathcal{L}$. Given a text-image pair, our model predicts the label $\hat{l}$.

3 CONSISTENTLY DISTRIBUTED SAMPLING

To construct the few-shot dataset for few-shot multimodal sentiment analysis task, it is important to select diverse samples that provide comprehensive coverage. Previous approaches [30, 31] have randomly sampled from the training and development sets to create few-shot datasets with equal amounts of data for each class, without taking into consideration the consistency of the category distribution between the full dataset (before sampling) and the few-shot dataset (after sampling). Additionally, users express emotions with varying proportions on social media, indicating that the distribution of posts with different emotions are differ.

We propose a novel sampling approach called Consistently Distributed Sampling (**CDS**). CDS ensures that the category distribution of the few-shot dataset is similar to that of the full dataset, creating representative few-shot datasets that reflect the real-world sentiment patterns observed on the internet. By constructing few-shot datasets using CDS, we can more accurately evaluate the performance of our model in a few-shot scenario. Specifically, we randomly sample about **1%**[2] of the training dataset based on the sentiment distribution of the full training dataset as the few-shot multimodal training dataset, $\mathcal{D}_{train}$, and construct the development dataset, $\mathcal{D}_{dev}$, with the same sentiment distribution. The MASAD dataset [34] involves 57 aspect categories and 2 sentiments, and our sampled data considers the balance between different aspect categories and sentiments simultaneously. For other datasests, we only consider balance of sentiment categories. The statistics of different datasets are given, as Table 2 and Table 3 show.

4 PROPOSED MODEL

4.1 Task Formulation

We assume access to a pre-trained language model, denoted as $\mathcal{M}$, such as RoBERTa [16]. Our goal is to fine-tune this model for the multimodal sentiment classification task on a specific label space, denoted as $\mathcal{L}$. We construct $\mathcal{D}_{train} = \{(x^j)\}_{j=1}^{K}$ by CDS, where K is the total number of text-image posts. Additionally, we choose the development set, $\mathcal{D}_{dev}$, to be the same size as the few-shot training set, i.e., $|\mathcal{D}_{dev}| = |\mathcal{D}_{train}|$.

In Coarse-grained MSA, $x^j = (t^j, i^j, l^j)$, where t is the text modality, i is the image modality, l is the sentiment label for a text-image pair. The model's objective is to predict the sentiment label l for each text-image pair in an unseen test dataset $(t_{test}, i_{test}, l_{test}) \in \mathcal{D}_{test}$[3].

In Fine-grained MSA, $x^j = (t^j, i^j, a^j, l^j)$, where t is the text modality, i is the image modality, a is the aspect term, l is the sentiment label corresponding to the aspect term a. The objective of the model is to predict the sentiment category l for each aspect term based on the context of both the text and image modalities in the test dataset $(t_{test}, i_{test}, a_{test}, l_{test}) \in \mathcal{D}_{test}$[4].

4.2 Multimodal Prompt-based Fine-tuning

We propose a novel model called MultiPoint, which stands for Multimodal Probabilistic Fusion Prompts. MultiPoint treats multimodal classification as a cloze-filling task, as depicted in Figure 1. We first design separate prompts for different modalities and then create

[2] Following [30, 31], we also randomly sample the 1% data of training dataset as our few-shot training dataset.

[3] For MVSA-Single and MVSA-Multiple, $l \in \{Negative, Neutral, Positive\}$. For TumEmo, $l \in \{Angry, Bored, Calm, Fear, Happy, Love, Sad\}$.

[4] For Twitter-2015 and Twitter-2017, $l \in \{Negative, Neutral, Positive\}$; for MASAD, $l \in \{Negative, Positive\}$.

Table 1: Unified multimodal templates for Few-shot Multimodal Sentiment Analysis. $\mathcal{P}$ is the template for the few-shot sentiment task, where c represents coarse-grained datasets, f represents fine-grained datasets, t represents the text prompt, and m represents the multimodal prompt. T is the original text input, $\tilde{V}$ is image slots from the image input I, and A is the aspect term. The special tokens in the vocabulary of the pre-trained language model are represented as </s>, <mask>, and <PT>. The variable $n^p_{0,\ldots,3}$ represents the number of learned prompt tokens, and for convenience, we set $n^p_0 = n^p_1 = n^p_2 = n^p_3$. Finally, there is a special token, <s>, at the front of each prompt, and the "⊕" symbol denotes concatenation operation.

Dataset	Text Prompts	Unified Multimodal Prompts
Coarse-grained	$\mathcal{P}_t^{c1}(T)$ = <s> [T] </s> It was <mask>.</s> $\mathcal{P}_t^{c2}(T)$ = <s> The sentence "[T]" has <mask> sentiment. </s> $\mathcal{P}_t^{c3}(T)$ = <s> Text: [T]. Sentiment of text: <mask>. </s> $\mathcal{P}_t^{c4}(T)$ = <s> <mask> <PT>$^{n_0^p}$ [T] <PT>$^{n_1^p}$ </s>	$\mathcal{P}_m^{c1}(T,I)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{c1}(T)$ $\mathcal{P}_m^{c2}(T,I)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{c2}(T)$ $\mathcal{P}_m^{c3}(T,I)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{c3}(T)$ $\mathcal{P}_m^{c4}(T,I)$ = <s> $\tilde{V}$ [C] <PT>$^{n_2^p}$ </s> $\oplus$ $\mathcal{P}_t^{c4}(T)$
Fine-grained	$\mathcal{P}_t^{f1}(T,A)$ = <s> [T] [A]</s> It was <mask>.</s> $\mathcal{P}_t^{f2}(T,A)$ = <s> The aspect "[A]" in sentence "[T]" has <mask> sentiment. </s> $\mathcal{P}_t^{f3}(T,A)$ = <s> Text: [T]. Aspect: [A]. Sentiment of aspect: <mask>. </s> $\mathcal{P}_t^{f4}(T,A)$ = <s> <mask> <PT>$^{n_0^p}$ [T] <PT>$^{n_1^p}$ [A] <PT>$^{n_2^p}$ </s>	$\mathcal{P}_m^{f1}(T,I,A)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{f1}(T,A)$ $\mathcal{P}_m^{f2}(T,I,A)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{f2}(T,A)$ $\mathcal{P}_m^{f3}(T,I,A)$ = <s> $\tilde{V}$ is [C] </s> $\oplus$ $\mathcal{P}_t^{f3}(T,A)$ $\mathcal{P}_m^{f4}(T,I,A)$ = <s> $\tilde{V}$ [C] <PT>$^{n_3^p}$ </s> $\oplus$ $\mathcal{P}_t^{f4}(T,A)$

effective multimodal prompts for our task. For the text modality, we manually design several text prompts, including $\mathcal{P}t^{c/f1}$, $\mathcal{P}t^{c/f2}$, and $\mathcal{P}t^{c/f3}$, and use the continuous text prompt, $\mathcal{P}t^{c/f4}$, to extract knowledge from PLMs. We believe that the manual prompts are carefully crafted based on domain knowledge and task-specific requirements, while the generated prompts are automatically generated from pre-trained language models to capture diverse and valuable information. The specific templates for the prompts are presented in Table 1. For the image modality, I, we use ClipCap [18] to generate a textual description of the image and use it as the image prompt, C, to bridge the gap between different modalities.

$$C = ClipCap(I). \tag{1}$$

We further leverage NF-ResNet [2] to extract and project the original image representation into the text feature space.

$$V = W_i Pool(ResNet(I)) + b_i, \tag{2}$$

$$\tilde{V} = reshape(V) = [v^1, ..., v^j, ..., v^{N_i}], v^j \in \mathbb{R}^{d_t}, \tag{3}$$

where $V \in \mathbb{R}^{d_{nt}}$, $W_i \in \mathbb{R}^{d_v \times d_{nt}}$, $b_i \in \mathbb{R}^{d_{nt}}$. $nt = d_t \times N_i$, N_i, a hyperparameter, is the number of slots representing initial image representation in a multimodal prompt, and d_t represents the dimension of text embedding in the pre-trained language model.

Lastly, we design multiple multimodal prompts $\mathcal{P}_m$ based on different text prompts, $\mathcal{P}_t$, and the image prompt. The specific unified multimodal prompts are presented in Table 1. We design three manual multimodal prompts, such as $\mathcal{P}_m^1$, $\mathcal{P}_m^2$, $\mathcal{P}_m^3$, as well as the continuous multimodal prompt $\mathcal{P}_m^4$. We choose to use only three manual prompts for demonstration purposes, as more similar prompts are also capable of handling MSA tasks in our actual experimental process.

4.3 Multimodal Demonstrations

Inspired by recent works, such as GPT-3 [3] and LM-BFF [6], we further design multimodal demonstrations chosen by similarity scores, as shown on the right side of Figure 1. Specifically, we first feed the raw text input t and image prompt c from the image input i, that can be regarded as text description of image, into a pre-trained language model, such as SBERT [21], to obtain embeddings E.

$$E = SBERT([t \oplus a \oplus c]), \tag{4}$$

where $\oplus$ is the concatenation operation. a represents the aspect term and is optional. For the fine-grained task, we combine the text with the aspect term, while for the coarse-grained task, there is no aspect term.

Next, we compute the similarity scores between each query instance $x_{que} = (t_{que}, i_{que}, a_{que})$ and support set with K^l instances for the l-th label category, $D_{sup}^{(l)} = \{(x_{sup}^{(l)})^j\}_{j=1}^{K^l}$. It is worth noting that the support instances are taken from the training dataset $\mathcal{D}_{train}$, both during training and inference stages.

$$Sim(x_{que}, x_{sup}^{(l)}) = cos(E_{que}, E_{sup}^{(l)}). \tag{5}$$

We then select the multimodal support instance with the highest similarity score for each label category l.

$$x_{sup}^{best^{(l)}} = \underset{Label=l,j}{\arg\max}\, Sim(x_{que}, x_{sup}^j)_{j=1}^{K^l}. \tag{6}$$

Finally, we convert the multimodal support instances with the highest similarity scores into $\mathcal{P}_m$ templates, with <mask> tokens replaced by different labels from $\mathcal{L}$. These resulting multimodal prompts are denoted as $\hat{\mathcal{P}}_m$, and we concatenate them with the query instance x_{que}.

$$\mathcal{P}_m^D = \mathcal{P}_m(x_{que}) \oplus \hat{\mathcal{P}}_m(x_{sup}^{best^{(1)}}, l^{(1)}) \oplus ... \oplus \hat{\mathcal{P}}_m(x_{sup}^{best^{(|\mathcal{L}|)}}, l^{(|\mathcal{L}|)}), \tag{7}$$

where $|\mathcal{L}|$ is the number of sentiment categories in each dataset.

4.4 Classification

Let $\boldsymbol{\phi} : \mathcal{L} \to \mathcal{V}$ be a mapping from the task label space to individual words in the vocabulary $\mathcal{V}$ of the pre-trained language model, $\mathcal{M}$. For each text-image pair $x = (t, i)$ for a coarse-grained dataset or $x = (t, i, a)$ for a fine-grained dataset, we input the multimodal prompt from Eq. 7, $\mathcal{P}_m^D$, that contains the $< mask >$ token into the

MLM head. We cast our multimodal classification task as a cloze problem and model the probability of predicting class $\hat{l} \in \mathcal{L}$ as:

$$\begin{aligned} p(\hat{l}|\mathcal{P}_m^D(x)) &= p(<mask>= \boldsymbol{\phi}(\hat{l})|\mathcal{P}_m^D) \\ &= \frac{exp(\mathbf{w}_{\boldsymbol{\phi}(\hat{l})} \cdot \mathbf{h}_{<mask>})}{\sum_{l' \in \mathcal{L}} exp(\mathbf{w}_{\boldsymbol{\phi}(l')} \cdot \mathbf{h}_{<mask>})}, \end{aligned} \tag{8}$$

where $\mathbf{h}_{<mask>}$ is the hidden representation of <mask> token and $\mathbf{w}_v$ indicates the final layer weight of MLM corresponding to $v \in \mathcal{V}$.

4.5 Multimodal Probabilistic Fusion

We find that different prompts contain various amounts of information, and the information conveyed by a single prompt is insufficient. We fuse prediction logits from different multimodal prompts based on Bayes Rule [4, 5] to provide more robust detection than a single prompt. For instance, there are n multimodal prompts $\{\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn}\}$. Crucially, given one instance x that label is classified as $\hat{l}$ by $\mathcal{M}$, we assume that different multimodal prompts are conditionally independent.

$$p(\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn}|\hat{l}) = p(\mathcal{P}_m^{D1}|\hat{l})...p(\mathcal{P}_m^{Dn}|\hat{l}). \tag{9}$$

Therefore, assuming conditional independence between the prediction results of the MLM for different multimodal prompts, we perform multimodal sentiment detection using multiple prompts and propose a novel multimodal probabilistic fusion approach.

$$\begin{aligned} p(\hat{l}|\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn}) &= \frac{p(\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn}|\hat{l})p(\hat{l})}{p(\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn})} \\ &\propto p(\mathcal{P}_m^{D1}, ..., \mathcal{P}_m^{Dn}|\hat{l})p(\hat{l}) \\ &\propto p(\mathcal{P}_m^{D1}|\hat{l})...p(\mathcal{P}_m^{Dn}|\hat{l})p(\hat{l}) \\ &\propto \frac{p(\mathcal{P}_m^{D1}|\hat{l})p(\hat{l})...p(\mathcal{P}_m^{Dn}|\hat{l})p(\hat{l})p(\hat{l})}{p(\hat{l})^n} \\ &\propto \frac{p(\hat{l}|\mathcal{P}_m^{D1})...p(\hat{l}|\mathcal{P}_m^{Dn})}{p(\hat{l})^{n-1}}. \end{aligned} \tag{10}$$

We first train independent classifiers that predict the distributions over the label $\hat{l}$ given each individual multimodal prompt, such as $p(\hat{l}|\mathcal{P}_m^{Dk})$. Then, we obtain the fused distribution of label $\hat{l}$ from the n multimodal prompts based on the probabilistic fusion module.

$$p(\hat{l}|\{\mathcal{P}_m^{Dk}\}_{k=1}^n) \propto \frac{\prod_{k=1}^n p(\hat{l}|\mathcal{P}_m^{Dk})}{p(\hat{l})^{n-1}}, \tag{11}$$

where we set $n = 2$ due to computational resource constraints, which is sufficient to demonstrate the effectiveness of our approach.

5 EXPERIMENTS

5.1 Datasets

We evaluate our proposed model on six multimodal sentiment datasets, including three coarse-grained datasets (MVSA-Single, MVSA-Multiple, and TumEmo) and three fine-grained datasets (Twitter-2015, Twitter-2017, and MASAD), where the label sets $\mathcal{L}$ vary across different datasets. Following [30], we keep the test set unchanged and sample data based on CDS to form few-shot datasets, consisting of about 1% of the training set with $K_{train} = K_{dev}$. The statistics of the different datasets are presented in Tables 2 and 3. The specific method of sampling data is described in Section 3.

5.2 Experimental Setup

In the text prompt, we use the original label set for TumEmo, which has multiple emotion labels. For other datasets, we map the label set {negative, neutral, positive} to {terrible, okay, great}. Our model is constructed using RoBERTa-large with 355M parameters, $\mathcal{M}$. Fine-tuning on small datasets can suffer from instability, and results may change dramatically given a new data split [6, 33]. To account for this, we measure average performance across five randomly sampled $\mathcal{D}_{train}$ and $\mathcal{D}_{dev}$ splits based on different seeds, i.e., *13, 21, 42, 87, 100.* To provide a more reliable measure of performance, we repeat the experiment three times for each split, resulting in a total of 15 (3×5) training runs for each dataset. We report the mean Accuracy (Acc), Weighted-F1 (F1)[5], and the standard deviation over the 15 runs. We set the batch-size to 8. For the number of prompt tokens for $\mathcal{P}^4$ in Table 1, we set $n_0^p = n_1^p = n_2^p = n_3^p = 1$ for Twitter-2017 and MASAD and $n_0^p = n_1^p = n_2^p = n_3^p = 2$ for other datasets. Our model performs best in the Acc metric when $N^i = 1$ in Eq. 3, and we set learning rates of 5e-6/2e-6/1e-5/3e-6 for MVSA-Single/Twitter-2017/MASAD/other datasets. Unless otherwise specified, we use these hyperparameters. MultiPoint has a total of approximately 410M parameters, and all parameters are updated during training. The training time varies depending on the dataset. For example, we train our model up to 1000 training steps in approximately 60 minutes for the MVSA-Single, MVSA-Multiple, Twitter-2015, and Twitter-2017 datasets. For the MASAD/TumEmo dataset, training for 1000 training steps takes around 100/120 minutes.

5.3 Baselines

We compare our model with three groups of baselines[6]. **The first group** consists of previous text-based models, including **RoBERTa** [16], **Prompt Tuning (PT)** only uses a single textual prompt based on the multimodal prompt, such as [<s> [T] It was <mask>. </s>] for coarse-grained datasets and [<s> [T] [A] It was <mask>. </s>] for fine-grained datasets, **LM-BFF** [6] utilizes generated text prompts based on each specific dataset and text demonstrations to solve few-shot text classification tasks, **LM-SC** [9] introduces supervised contrastive learning based on LM-BFF to few-shot text tasks, and **GFSC** [7] converts the classification task into a generation task to solve text classification tasks in the few-shot setting through the pre-trained generation model, i.e., GPT2 [20].

The second group consists of multimodal approaches that are trained in full MSA datasets from published papers. **For the coarse-grained MSA task: Multimodal Fine Tuning (MFN)** is a baseline that doesn't use any designed prompts and employs the representation of the "<s>" token for classification. **CLMLF** [13] is the state-of-the-art model for coarse-grained MSA. **For the fine-grained MSA task: TomBERT** [29] is a multimodal BERT for the fine-grained MSA task. **EF-CapTrBERT** [12] translates images in input space to construct an auxiliary sentence that provides multimodal information to BERT. **KEF** [14] exploits adjective-noun pairs extracted

[5] Since most datasets have highly imbalanced categories, the Weighted-F1 value is a more reasonable metric.

[6] Unless otherwise specified, all baselines are based on RoBERTa-large.

Table 2: Statistics for five datasets, including MVSA-Single, MVSA-Multiple, Twitter-2015, Twitter-2017, and MASAD. For A/B, B represents the number of original data, and A represents the number of few-shot data sampled based on CDS. For all datasets, the few-shot dataset represents approximately 1% of the overall training data. In the few-shot setting, the number of development datasets is equal to the number of training datasets.

Dataset		Train				Test			
		Negative	Neutral	Positive	Total	Negative	Neutral	Positive	Total
Coarse-grained	MVSA-Single	**10**/1004	**4**/345	**20**/1921	**34**/3270	126	37	249	412
	MVSA-Multiple	**20**/1909	**32**/3170	**82**/8166	**134**/13245	217	405	1014	1636
Fine-grained	Twitter-2015	**4**/368	**19**/1883	**10**/928	**33**/3179	113	607	317	1037
	Twitter-2017	**4**/416	**16**/1638	**15**/1508	**35**/3562	168	573	493	1234
	MASAD	**69**/5605	**0**/0	**101**/9263	**170**/14868	1767	0	3168	4935

Table 3: Statistics for the TumEmo dataset that has the same few-shot setting as other datasets.

Dataset	Angry	Bored	Calm	Fear	Happy	Love	Sad	Total
Train	**60**/5879	**108**/10823	**63**/6300	**86**/8625	**222**/22215	**150**/15016	**68**/6829	**757**/75687
Test	736	1354	788	1079	2776	1875	855	9463

from the image for the fine-grained MSA task. **FITE** [24] is the state-of-the-art model for fine-grained MSA, which leverages facial information from the image modality. **VLP-MABSA** [14] designs a unified multimodal encoder-decoder architecture and different pre-training tasks to improve the fine-grained MSA task.

The last group includes multimodal approaches that have been trained for few-shot MSA. **PVLM** [30] directly introduces image features to pre-trained language models to solve the MAS task in a few-shot scenario. **UP-MPF** [31] is the state-of-the-art model in the multimodal few-shot setting for the MSA task. It further employs pre-training data and tasks based on PVLM. **MultiPoint** is our model that introduces multiple multimodal prompts with demonstrations and probabilistic fusion to improve the performance of MSA in a few-shot scenario. Note that we reproduced the LM-BFF, LM-SC, EF-CapTrBERT, FITE, VLP-MABSA, PVLM, and UP-MPF models based on the RoBERTa-large model, while TomBERT and KEF are based on the BERT-base model.

5.4 Experimental Results and Analysis

Following [30, 31], we report the results of our model and baselines on few-shot datasets with 1% training data. We introduce different combinations of multimodal prompts in MultiPoint from Table 1, such as $[\mathcal{P}_m^{c3}, \mathcal{P}_m^{c4}] \rightarrow \mathcal{P}_m^{c[3-4]}$. The performance comparison of our model (MultiPoint) with the baselines is shown in Table 4 for coarse-grained MSA datasets and Table 5 for fine-grained MSA datasets. We make the following observations:

(1) Our model outperforms other robust models, including SOTA multimodal baselines (CLMLF and FITE), text-only prompt tuning models (PT, LM-BFF, LM-SC, and GFSC), and multimodal prompt tuning models (PVLM, UP-MPF). MultiPoint outperforms the existing SOTA few-shot multimodal model, UP-MPF, by more than 3-6% on different datasets, especially for fine-grained datasets. This is due to our use of image prompts to bridge the gap between text and image modalities, introduction of multimodal demonstrations to improve the robustness of our model, and the utilization of probabilistic fusion modules to capture more practical information, including handcrafted prompts and learnable prompts. (2) Our model yields varying results when using different combinations of prompts, and the combination of manual prompts and learnable prompts outperforms using only different manual prompts. (3) Most multimodal models trained on complete datasets outperform text-only models in the few-shot setting, indicating the importance of the image modality for sentiment analysis. However, multimodal models that perform very well on the full dataset perform poorly in the few-shot setting, like CLMLF, VLP-MABSA, and others, mainly due to overfitting on the few-shot data. (4) Similar to previous studies, most prompt-based approaches (denoted with $*$) outperform state-of-the-art multimodal approaches (the second group) by a large margin, even using prompts to tune the model on the text-only modality. (5) Prompt-based generative models for few-shot classification tasks perform poorly compared to cloze-based pre-trained masked language models. There is still much room for exploration using generative models to solve few-shot classification problems.

5.5 Ablation Experiments

We conduct ablation experiments on the MultiPoint model to demonstrate the effectiveness of its different modules, and the results are listed in Table 6. Removing any of these modules affects the model's performance, indicating their significance in few-shot MSA. Here are our specific findings: First, we remove the image modality (w/o Image), including image slots and captions, to verify the effectiveness of image information. The model's performance drops significantly, indicating that image modality is critical in few-shot MSA. Second, we remove the image prompt (w/o Caption) and only apply image slots to the pre-trained language model. The model's performance drops drastically, suggesting that simply introducing image modalities into the pre-trained model fails to capture adequate image information due to the discrepancy of different modalities. Third, we remove the Multimodal Demonstration (w/o MD) to verify the validity of multimodal demonstrations. The model's performance drops, indicating that multimodal demonstrations are effective in few-shot MSA. Fourth, we utilize only one multimodal prompt, such as $\mathcal{P}_m^{1,2,3,4}$, to affirm the usefulness of our proposed multiple multimodal prompts and the probabilistic fusion module (PF). The results drop significantly across all datasets, suggesting that multiple multimodal prompts can furnish more informative

Table 4: Our main results for few-shot experiments on three multimodal coarse-grained datasets, including MVSA-Single, MVSA-Multiple, and TumEmo. The standard deviation is in parentheses. “$*$” indicates baselines with prompt tuning and applies multiple prompts from Table 1. We report the best performance of the baselines applying different prompts. “$\mathcal{P}_m$” means the multimodal prompt, c is coarse-grained. “[q-k]” means combine q-th prompt with k-th prompt.

Modality	Model	MVSA-Single		MVSA-Multiple		TumEmo	
		Acc	F1	Acc	F1	Acc	F1
Text	RoBERTa	61.21 (±2.11)	56.11 (±2.74)	63.40 (±0.86)	61.34 (±1.40)	55.03 (±0.45)	54.87 (±0.57)
	PT*	65.73 (±1.96)	64.13 (±1.77)	65.91 (±1.88)	64.05 (±1.42)	55.97 (±0.30)	55.84 (±0.33)
	LM-BFF*	65.58 (±2.81)	63.41 (±3.00)	66.36 (±0.88)	64.08 (±1.09)	56.03 (±0.66)	55.85 (±0.63)
	LM-SC*	66.51 (±1.09)	64.62 (±0.98)	65.37 (±0.87)	63.63 (±1.56)	55.95 (±0.40)	56.00 (±0.52)
	GFSC*	63.39 (±4.10)	58.72 (±6.52)	64.72 (±1.18)	63.53 (±0.56)	53.28 (±0.50)	52.83 (±0.59)
Text-Image	MFN	64.08 (±2.44)	60.60 (±2.97)	64.04 (±1.97)	61.46 (±1.98)	56.83 (±0.38)	56.82 (±0.40)
	CLMLF	61.19 (±0.65)	51.34 (±3.24)	63.86 (±1.76)	57.95 (±4.32)	42.65 (±9.32)	38.41 (±12.19)
Text-Image	PVLM*	66.94 (±1.20)	63.10 (±2.79)	67.40 (±0.99)	63.67 (±2.56)	55.43 (±0.72)	55.02 (±0.70)
	UP-MPF*	66.84 (±2.05)	64.96 (±1.37)	67.35 (±0.97)	61.00 (±2.23)	54.91 (±0.94)	54.38 (±1.05)
	MultiPoint($\mathcal{P}_m^{c[1-4]}$)	**69.95 (±2.47)**	**68.60 (±1.73)**	68.04 (±0.57)	65.39 (±1.28)	**58.09 (±0.43)**	58.05 (±0.37)
	MultiPoint($\mathcal{P}_m^{c[2-4]}$)	69.66 (±1.48)	67.96 (±1.13)	67.67 (±0.85)	65.15 (±1.47)	57.97 (±0.51)	57.92 (±0.47)
	MultiPoint($\mathcal{P}_m^{c[3-4]}$)	69.76 (±1.08)	68.02 (±1.47)	**68.27 (±1.15)**	**65.34 (±1.87)**	58.05 (±0.53)	**58.06 (±0.50)**
	MultiPoint($\mathcal{P}_m^{c[1-2]}$)	68.11 (±1.40)	67.03 (±1.05)	67.24 (±0.87)	64.83 (±1.34)	57.74 (±0.45)	57.69 (±0.45)
	MultiPoint($\mathcal{P}_m^{c[1-3]}$)	68.59 (±0.59)	67.40 (±0.88)	67.63 (±1.15)	65.28 (±1.32)	57.80 (±0.77)	57.79 (±0.71)
	MultiPoint($\mathcal{P}_m^{c[2-3]}$)	68.15 (±1.98)	66.87 (±1.39)	67.12 (±1.43)	65.04 (±1.29)	57.48 (±0.68)	57.44 (±0.64)

Table 5: Our main results for few-shot experiments on three multimodal fine-grained datasets, including Twitter-2015, Twitter-2017, and MASAD. The standard deviation is in parentheses. f represents fine-grained. “–” indicates no reproducible results on MASAD, as these baselines require external knowledge to model, such as captions, adjective-noun pairs, etc.

Modality	Model	Twitter-2015		Twitter-2017		MASAD	
		Acc	F1	Acc	F1	Acc	F1
Text	RoBERTa	55.58 (±4.13)	52.32 (±2.28)	48.22 (±2.95)	46.37 (±3.17)	68.81 (±1.76)	67.88 (±1.43)
	PT*	61.97 (±3.15)	60.11 (±3.38)	58.77 (±3.70)	57.85 (±3.63)	77.62 (±1.34)	77.60 (±1.37)
	LM-BFF*	60.87 (±3.38)	59.63 (±3.04)	56.84 (±3.51)	55.96 (±3.48)	78.87 (±0.94)	78.35 (±0.77)
	LM-SC*	61.16 (±3.31)	60.99 (±3.28)	54.78 (±1.93)	52.89 (±2.63)	77.94 (±0.97)	77.61 (±0.92)
	GFSC*	52.77 (±0.38)	52.01 (±0.56)	54.426 (±2.47)	53.15 (±2.70)	75.96 (±1.50)	76.14 (±1.32)
Text-Image	MFN	55.86 (±1.66)	52.81 (±1.45)	50.91 (±2.86)	49.20 (±3.05)	78.98 (±1.60)	78.28 (±2.10)
	CLMLF	56.97 (±2.08)	52.04 (±2.35)	49.63 (±2.40)	45.72 (±2.17)	74.33 (±2.85)	72.51 (±1.95)
	TomBERT	55.95 (±5.17)	43.248 (±0.06)	47.47 (±2.26)	36.93 (±5.89)	72.34 (±2.37)	70.55 (±3.04)
	EF-CapTrBERT	57.81 (±1.45)	42.72 (±1.00)	47.41 (±1.01)	33.58 (±3.58)	–	–
	KEF	57.58 (±2.04)	43.09 (±0.25)	45.74 (±0.78)	31.29 (±2.39)	–	–
	FITE	58.42 (±0.18)	43.29 (±0.11)	46.20 (±0.52)	29.97 (±0.70)	–	–
	VLP-MABSA	53.36 (±1.07)	43.23 (±3.75)	55.32 (±3.39)	48.96 (±1.26)	–	–
Text-Image	PVLM*	59.25 (±2.02)	54.45 (±3.33)	54.28 (±3.17)	51.02 (±5.24)	77.94 (±1.25)	77.85 (±1.09)
	UP-MPF*	61.56 (±2.43)	60.16 (±2.54)	54.93 (±2.22)	51.87 (±4.08)	77.75 (±2.14)	77.84 (±1.93)
	MultiPoint($\mathcal{P}_m^{f[1-4]}$)	65.15 (±0.88)	64.34 (±1.02)	60.31 (±1.78)	59.65 (±1.67)	83.72 (±0.84)	83.53 (±0.84)
	MultiPoint($\mathcal{P}_m^{f[2-4]}$)	66.23 (±0.83)	65.59 (±1.09)	60.18 (±1.86)	59.41 (±1.77)	82.73 (±1.05)	82.53 (±1.04)
	MultiPoint($\mathcal{P}_m^{f[3-4]}$)	**67.33 (±1.07)**	**66.61 (±1.36)**	**61.88 (±2.56)**	**61.23 (±2.58)**	**84.05 (±0.77)**	**83.86 (±0.86)**
	MultiPoint($\mathcal{P}_m^{f[1-2]}$)	65.48 (±0.99)	64.99 (±0.90)	56.89 (±1.04)	56.14 (±1.27)	81.55 (±0.89)	81.09 (±0.95)
	MultiPoint($\mathcal{P}_m^{f[1-3]}$)	65.98 (±1.86)	65.65 (±1.55)	58.82 (±1.95)	58.05 (±2.35)	81.90 (±1.47)	81.76 (±1.43)
	MultiPoint($\mathcal{P}_m^{f[2-3]}$)	66.31 (±0.81)	66.06 (±0.84)	58.51 (±2.31)	58.22 (±2.28)	82.05 (±0.99)	81.82 (±0.96)

few-shot sentiment analysis. In the single-prompt setting, different datasets achieve the best results applying different prompts. Note that the learnable prompt $\mathcal{P}_m^4$ achieves the best results on most datasets, such as TumEmo, Twitter-2015 and MASAD, followed by $\mathcal{P}_m^3$. These results show that the amount of information mined by different prompts is distinct, an observation further supported by the results for multiple prompts combinations in Table 4 and Table 5. Finally, we replace the probabilistic fusion module with average fusion (w/ Average Fusion), i.e., averaging multiple logits from the

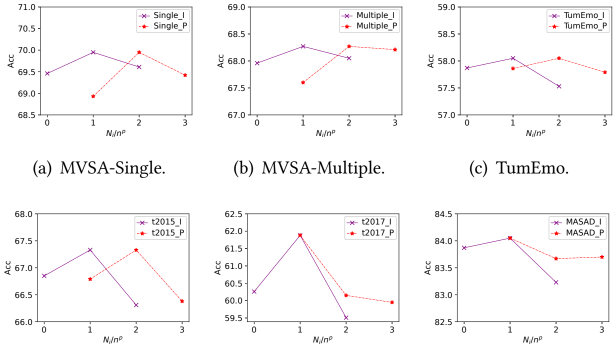
Table 6: Ablation experimental results about on Acc metric on six datasets.

Model	MVSA-Single	MVSA-Multiple	TumEmo	Twitter-2015	Twitter-2017	MASAD
w/o Image	65.77 (±2.21)	66.83 (±1.01)	56.37 (±0.42)	63.22 (±1.50)	60.26 (±2.39)	79.46 (±1.45)
w/o Caption	66.41 (±1.62)	67.55 (±1.09)	56.38 (±0.56)	66.788 (±1.36)	61.12 (±2.48)	79.72 (±1.94)
w/o MD	69.56 (±1.89)	67.86 (±0.97)	57.88 (±0.47)	64.77 (±1.56)	61.28 (±2.72)	82.57 (±1.08)
w/ MultiPoint($\mathcal{P}_m^1$)	67.62 (±1.77)	66.09 (±1.75)	56.94 (±0.99)	63.72 (±1.39)	57.62 (±1.57)	80.37 (±1.26)
w/ MultiPoint($\mathcal{P}_m^2$)	67.52 (±1.88)	*67.25 (±1.63)*	56.75 (±0.56)	65.42 (±1.49)	54.78 (±1.84)	80.13 (±2.32)
w/ MultiPoint($\mathcal{P}_m^3$)	*68.84 (±2.38)*	66.69 (±0.59)	57.06 (±0.70)	*66.25 (±1.05)*	58.56(±1.70)	80.55 (±1.74)
w/ MultiPoint($\mathcal{P}_m^4$)	68.59 (±2.26)	66.65 (±0.97)	*57.19 (±0.59)*	64.22 (±2.96)	*60.52 (±4.11)*	*82.33 (±1.06)*
w/ Average Fusion	69.71 (±1.24)	68.22 (±1.21)	58.04 (±0.55)	67.18 (±0.63)	60.10 (±2.51)	83.76 (±1.29)
MultiPoint	**69.95 (±2.47)**	**68.27 (±1.15)**	**58.05 (±0.53)**	**67.33 (±1.07)**	**61.88 (±2.56)**	**84.05 (±0.77)**

Table 7: Experimental results on Acc metric on few-shot datasets with the same amount of data for each category. The symbol ∇ denotes the decrease in performance compared to our few-shot datasets based on CDS.

Model	MVSA-Single	MVSA-Multiple	TumEmo	Twitter-2015	Twitter-2017	MASAD
PVLM	59.95 (±3.27) ∇**6.99**	59.18 (±3.21) ∇**8.22**	52.67 (±0.95) ∇**2.76**	51.17 (±6.78) ∇**8.08**	51.47 (±0.96) ∇**2.81**	71.96 (±2.44) ∇**5.98**
UP-MPF	61.75 (±3.82) ∇**5.09**	57.32 (±2.76) ∇**10.03**	51.44 (±1.78) ∇**3.47**	54.83 (±8.10) ∇**6.73**	53.21 (±2.46) ∇**1.72**	75.18 (±1.62) ∇**2.57**
MultiPoint	63.11 (±3.96) ∇**6.84**	61.14 (±1.66) ∇**7.13**	55.15 (±0.47) ∇**2.94**	57.92 (±3.53) ∇**9.41**	58.46 (±2.75) ∇**3.42**	81.52 (±1.86) ∇**2.53**

model. The results on all datasets slightly decreased, indicating that the proposed probabilistic fusion module is effective.



(a) MVSA-Single.

(b) MVSA-Multiple.

(c) TumEmo.

(d) Twitter-2015.

(e) Twitter-2017.

(f) MASAD.

Figure 2: Acc comparisons of different Hyperparameters on different datasets, e.g., the number of image tokens, N_i, and the number of prompt tokens, n^p. I means the image token, P means the prompt token.

5.6 Image Tokens and Prompt Tokens Amount

In order to preserve adequate information from the image by NF-ResNet, we conduct experiments on all few-shot datasets under different settings of the hyperparameter N_i in Eq. 3, and the corresponding results are shown by solid purple lines in Figure 2. We obtain the best performance for all datasets when $N_i = 1$. When N_i is smaller, the image information is not fully utilized, while retaining more image features brings redundant information to the model. We also leverage the continuous prompt tokens, $< PT >$ in $\mathcal{P}^4$, to mine knowledge from the pre-trained language model. We conduct hyperparameter experiments on the amount of prompt tokens, n^p, as the red dotted line shows in Figure 2. Our model achieves the best performance on Twitter-2017 and MASAD when $n^p = 1$, and on other datasets when $n^p = 2$.

5.7 Effect of Consistently Distributed Sampling

We design diverse and comprehensive few-shot datasets based on CDS, as shown in Tables 2 and 3. Following the approach of [30, 31], we sample the data to create **f**ew-**s**hot datasets with an **e**qual number of instances for each **s**entiment **c**ategory, **ESCFS**, while keeping the total amount of data consistent with few-shot datasets based on CDS. We reproduce our model, MultiPoint, as well as the PVLM and UP-MPF models for the few-shot multimodal MSA task on these datasets, as reported in Table 7. We observe that the performance of each model on all datasets has decreased by 2.5-10% when trained on ESCFS (indicated by the symbol ∇), indicating the effectiveness of our few-shot datasets with consistent distribution. The CDS approach is particularly beneficial for smaller datasets, such as MVSA-Single, MVSA-Multiple, and Twitter-2015.

6 CONCLUSION

In this paper, we first present a Consistently Distributed Sampling approach called CDS to construct the few-shot dataset with a category distribution similar to that of the full dataset. We further propose a novel approach to the few-shot MSA task, which is comprised of a Multimodal Probabilistic Fusion Prompts model with Multimodal Demonstrations (MultiPoint). Our model leverages a unified multimodal prompt, which combines image prompt and textual prompt, and dynamically selects multimodal demonstrations to improve model robustness. Additionally, we introduce a probabilistic fusion module to fuse multiple predictions from different multimodal prompts. Our extensive experiments on six datasets demonstrate the effectiveness of the CDS and the MultiPoint, outperforming state-of-the-art models on most datasets. In future work, we plan to explore more effective fusion approaches for different prompts to further improve the performance of few-shot multimodal sentiment analysis.

ACKNOWLEDGMENTS

Thanks to all co-authors for their hard work. The work is supported by National Natural Science Foundation of China (62172086, 62272092), Doctoral Research Innovation of Northeastern University (N2216004), and Chinese Scholarship Council. This work is also supported by an AcRF Tier-2 grant from the Ministry of Education, Singapore (Project Number: T2MOE2008, Funder Reference Number: MOE-T2EP20220-0017).

REFERENCES

[1] Sarah A. Abdu, Ahmed H. Yousef, and Ashraf Salem. 2021. Multimodal Video Sentiment Analysis Using Deep Learning Approaches, a Survey. *Inf. Fusion* (2021), 204–226. https://doi.org/10.1016/j.inffus.2021.06.003

[2] Andrew Brock, Soham De, and Samuel L. Smith. 2021. Characterizing signal propagation to close the performance gap in unnormalized ResNets. In *ICLR*. https://openreview.net/forum?id=IX3Nnir2omJ

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, and et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] Yi-Ting Chen, Jinghao Shi, Christoph Mertz, Shu Kong, and Deva Ramanan. 2021. Multimodal object detection via bayesian fusion. *arXiv preprint arXiv:2104.02904* (2021).

[5] Hongsheng Dai, Murray Pollock, and Gareth Roberts. 2021. Bayesian Fusion: Scalable unification of distributed statistical analyses. *arXiv preprint arXiv:2102.02123* (2021).

[6] Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making Pre-trained Language Models Better Few-shot Learners. In *ACL/IJCNLP*. 3816–3830. https://doi.org/10.18653/v1/2021.acl-long.295

[7] Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A Generative Language Model for Few-shot Aspect-Based Sentiment Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz (Eds.). Association for Computational Linguistics, 770–787. https://doi.org/10.18653/v1/2022.findings-naacl.58

[8] Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-Domain Targeted Sentiment Analysis via Span-Based Extraction and Classification. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 537–546. https://doi.org/10.18653/v1/p19-1051

[9] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. 2022. Contrastive Learning for Prompt-based Few-shot Language Learners. In *NAACL*. 5577–5587. https://doi.org/10.18653/v1/2022.naacl-main.408

[10] Xincheng Ju, Dong Zhang, Rong Xiao, Junhui Li, Shoushan Li, Min Zhang, and Guodong Zhou. 2021. Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 4395–4405. https://doi.org/10.18653/v1/2021.emnlp-main.360

[11] Ramandeep Kaur and Sandeep Kautish. 2019. Multimodal Sentiment Analysis: A Survey and Comparison. *Int. J. Serv. Sci. Manag. Eng. Technol.* (2019), 38–58. https://doi.org/10.4018/IJSSMET.2019040103

[12] Zaid Khan and Yun Fu. 2021. Exploiting BERT for Multimodal Target Sentiment Classification through Input Space Translation. In *MM ’21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3034–3042. https://doi.org/10.1145/3474085.3475692

[13] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. *CoRR* abs/2204.05515 (2022). https://doi.org/10.48550/arXiv.2204.05515 arXiv:2204.05515

[14] Yan Ling, Jianfei Yu, and Rui Xia. 2022. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, 2149–2159. https://doi.org/10.18653/v1/2022.acl-long.152

[15] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *CoRR* abs/2107.13586 (2021). arXiv:2107.13586 https://arxiv.org/abs/2107.13586

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[17] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual Attention Model for Name Tagging in Multimodal Social Media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, 1990–1999. https://doi.org/10.18653/v1/P18-1185

[18] Ron Mokady, Amir Hertz, and Amit H. Bermano. 2021. ClipCap: CLIP Prefix for Image Captioning. *CoRR* abs/2111.09734 (2021). arXiv:2111.09734 https://arxiv.org/abs/2111.09734

[19] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. 2016. Sentiment Analysis on Multi-View Social Data. In *MMM*. 15–27. https://doi.org/10.1007/978-3-319-27674-8_2

[20] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).

[21] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. https://doi.org/10.18653/v1/D19-1410

[22] Maria Tsimpoukelli, Jacob Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal Few-Shot Learning with Frozen Language Models. In *NeurIPS*. 200–212. https://proceedings.neurips.cc/paper/2021/hash/01b7575c38dac42f3cfb7d500438b875-Abstract.html

[23] Nan Xu, Wenji Mao, and Guandan Chen. 2018. A Co-Memory Network for Multimodal Sentiment Analysis. In *SIGIR*. 929–932. https://doi.org/10.1145/3209978.3210093

[24] Hao Yang, Yanyan Zhao, and Bing Qin. 2022. Face-Sensitive Image-to-Emotional-Text Cross-modal Translation for Multimodal Aspect-based Sentiment Analysis. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, 3324–3335. https://aclanthology.org/2022.emnlp-main.219

[25] Li Yang, Jin-Cheon Na, and Jianfei Yu. 2022. Cross-Modal Multitask Transformer for End-to-End Multimodal Aspect-Based Sentiment Analysis. *Inf. Process. Manag.* 59, 5 (2022), 103038. https://doi.org/10.1016/j.ipm.2022.103038

[26] Xiaocui Yang, Shi Feng, Daling Wang, and Yifei Zhang. 2021. Image-Text Multimodal Emotion Classification via Multi-View Attentional Network. *IEEE Trans. Multim.* (2021), 4014–4026. https://doi.org/10.1109/TMM.2020.3035277

[27] Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In *ACL/IJCNLP*. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28

[28] Jianfei Yu, Kai Chen, and Rui Xia. 2022. Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing* (2022).

[29] Jianfei Yu and Jing Jiang. 2019. Adapting BERT for Target-Oriented Multimodal Sentiment Classification. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Sarit Kraus (Ed.). ijcai.org, 5408–5414. https://doi.org/10.24963/ijcai.2019/751

[30] Yang Yu and Dong Zhang. 2022. Few-Shot Multi-Modal Sentiment Analysis with Prompt-Based Vision-Aware Language Modeling. In *ICME*. 1–6. https://doi.org/10.1109/ICME52920.2022.9859654

[31] Yang Yu, Dong Zhang, and Shoushan Li. 2022. Unified Multi-modal Pre-training for Few-shot Sentiment Analysis with Prompt-based Learning. In *MM ’22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, João Magalhães, Alberto Del Bimbo, Shin’ichi Satoh, Nicu Sebe, Xavier Alameda-Pineda, Qin Jin, Vincent Oria, and Laura Toni (Eds.). ACM, 189–198. https://doi.org/10.1145/3503161.3548306

[32] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Sheila A. McIlraith and Kilian Q. Weinberger (Eds.). AAAI Press, 5674–5681. http://www.qizhang.info/paper/aaai2017-twitterner.pdf

[33] Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting Few-sample BERT Fine-tuning. In *ICLR*. OpenReview.net. https://openreview.net/forum?id=cO1IH43yUF

[34] Jie Zhou, Jiabao Zhao, Jimmy Xiangji Huang, Qinmin Vivian Hu, and Liang He. 2021. MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing* 455 (2021), 47–58. https://doi.org/10.1016/j.neucom.2021.05.040

[35] Haidong Zhu, Zhaoheng Zheng, Mohammad Soleymani, and Ram Nevatia. 2022. Self-Supervised Learning for Sentiment Analysis via Image-Text Matching. In *ICASSP*. 1710–1714. https://doi.org/10.1109/ICASSP43922.2022.9747819