Information Processing and Management 61 (2024) 103724

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

An empirical study of Multimodal Entity-Based Sentiment Analysis with ChatGPT: Improving in-context learning via entity-aware contrastive learning

Li Yang [a,*], Zengzhi Wang [b], Ziyan Li [b], Jin-Cheon Na [a], Jianfei Yu [b]

[a] *Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore, Singapore*
[b] *School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China*

ARTICLE INFO

Keywords:
In-context learning
Multimodal sentiment analysis
Large language models
Entity-based sentiment analysis

ABSTRACT

Multimodal Entity-Based Sentiment Analysis (MEBSA) is an emerging task within sentiment analysis, with the objective of simultaneously detecting entity, sentiment, and entity category from multimodal inputs. Despite achieving promising results, most existing MEBSA studies requires a substantial quantity of annotated data. The acquisition of such data is both costly and time-intensive in practical applications. To alleviate the reliance on annotated data, this work explores the potential of in-context learning (ICL) with a representative large language model, ChatGPT, for the MEBSA task. Specifically, we develop a general ICL framework with task instructions for zero-shot learning, followed by extending it to few-shot learning by incorporating a few demonstration samples in the prompt. To enhance the performance of the ICL framework in the few-shot learning setting, we further develop an Entity-Aware Contrastive Learning model to effectively retrieve demonstration samples that are similar to each test sample. Experiments demonstrate that our developed ICL framework exhibits superior performance over other baseline ICL methods, and is comparable to or even outperforms many existing fine-tuned methods on four MEBSA subtasks.

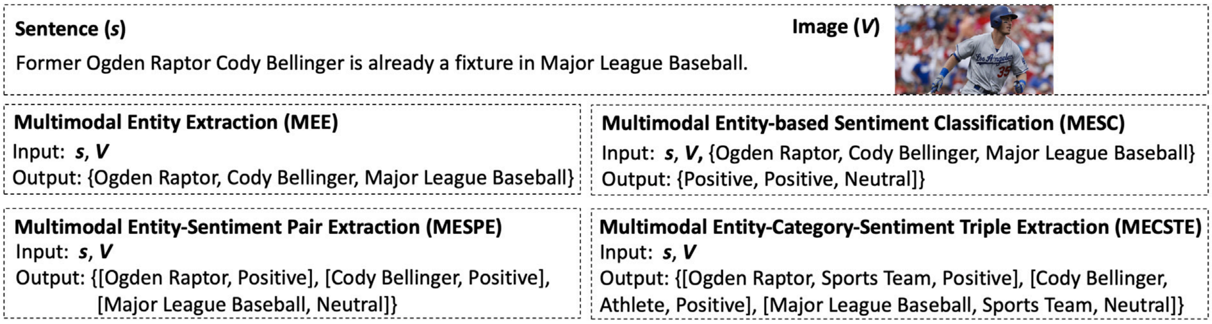
1. Introduction

Due to the prevalence of social media, a growing number of users are utilizing social platforms (e.g., X (formerly Twitter) and Facebook) to share their views and feelings through various modalities, including text, images, and videos. Gaining an understanding of the sentiment expressed in these massive multimodal posts can provide valuable insights for companies and governments, hence enhancing decision-making processes. Multimodal Sentiment Analysis (MSA) has garnered increasing interest recently, owing to its wide range of practical applications. MSA focuses on incorporating modalities beyond text, such as visual and audio, to better gauge sentiment and emotion (Kumar & Garg, 2019; Zhang, Li, Deng, Bing, & Lam, 2022; Zhang, Zhang, Li, Zhu, & Zhou, 2020). This paper explores an emerging fine-grained MSA task known as Multimodal Entity-based Sentiment Analysis (MEBSA), which is initially introduced by Yu, Jiang and Xia (2020). With a pair of text and image as input, MEBSA aims to identify all the entities, corresponding categories, and sentiments (Yang, Wang, Na & Yu, 2023).

Many studies have been exploring various subtasks of MEBSA. Earlier works primarily focused on single-element extraction tasks, including Multimodal Entity Extraction (MEE) (Wu et al., 2019) and Multimodal Entity-based Sentiment Classification (MESC) (Yu, Chen, & Xia, 2023; Yu, Wang, Xia, & Li, 2022; Zhou, Zhao, Huang, Hu, & He, 2021). MEE is to identify the entities that appeared

[*] Corresponding author.
E-mail address: yang0666@e.ntu.edu.sg (L. Yang).

https://doi.org/10.1016/j.ipm.2024.103724
Received 27 October 2023; Received in revised form 16 March 2024; Accepted 18 March 2024



Sentence (*s*)
Former Ogden Raptor Cody Bellinger is already a fixture in Major League Baseball.

Image (*V*)

Multimodal Entity Extraction (MEE)
Input: ***s*, *V***
Output: {Ogden Raptor, Cody Bellinger, Major League Baseball}

Multimodal Entity-based Sentiment Classification (MESC)
Input: ***s*, *V*,** {Ogden Raptor, Cody Bellinger, Major League Baseball}
Output: {Positive, Positive, Neutral]}

Multimodal Entity-Sentiment Pair Extraction (MESPE)
Input: ***s*, *V***
Output: {[Ogden Raptor, Positive], [Cody Bellinger, Positive], [Major League Baseball, Neutral]}

Multimodal Entity-Category-Sentiment Triple Extraction (MECSTE)
Input: ***s*, *V***
Output: {[Ogden Raptor, Sports Team, Positive], [Cody Bellinger, Athlete, Positive], [Major League Baseball, Sports Team, Neutral]}

Fig. 1. An example illustrating the outputs of different MEBSA subtasks.

in multimodal inputs, while MESC is to identify the sentiment towards each given entity. As three components of MEBSA (entity, category, and sentiment) are highly related with each other, some recent studies have explored two multi-element extraction tasks, namely Multimodal Entity-Sentiment Pair Extraction (MESPE) (Xiao et al., 2023) and Multimodal Entity-Category-Sentiment Triple Extraction (MECSTE) (Yang, Wang et al., 2023). The objective of MESPE is to identity entities and the associated sentiments. MECSTE is an extension of the MESPE task, aiming to identity entities along with the associated entity categories and sentiments. Fig. 1 gives an example illustrating the outputs of different MEBSA subtasks.

However, the major limitation of these aforementioned studies is their heavy reliance on a considerable quantity of annotated data to train specialized models. As a result of the labor-intensive nature of human annotation, acquiring fine-grained labels for multimodal social media posts can be a resource-intensive and costly endeavor (Wu et al., 2019). Moreover, the diverse topics and continuously evolving content on social platforms pose a challenge for specialized models to maintain strong generalization performance. Recently, many large language models (LLMs), including PaLM and ChatGPT (Chowdhery et al., 2022; OpenAI, 2023a), have demonstrated their effectiveness and generalization abilities in various NLP tasks (Zhao et al., 2023). In contrast to prior pre-trained language models, these LLMs often contain a substantial amount of parameters, ranging from tens to hundreds of billions. They are provided as a service with a prompting interface, such as ChatGPT API.[1] Therefore, fine-tuning LLMs for downstream tasks is a challenge. Instead, a typical solution to leverage LLMs is through the implementation of in-context learning (ICL). This involves the development of task-specific prompts (zero-shot learning) and conditioning LLMs on some demonstration examples (few-shot learning). By employing these techniques, the robust reasoning abilities of LLMs can be elicited to generate predictions for downstream tasks. While LLM-based ICL has demonstrated remarkable success in various NLP tasks (Dong et al., 2022; Ouyang et al., 2022), it is not devoid of challenges. First, the performance of ICL varies by tasks and exhibits instability when presented with different prompts (Zhao, Wallace, Feng, Klein, & Singh, 2021a). ICL's performance is susceptible to changes in the format and order of demonstration examples (Liu et al., 2022; Lu, Bartolo, Moore, Riedel, & Stenetorp, 2022). This requires human effort in selecting demonstration examples to ensure optimal model performance. Second, although LLMs have demonstrated impressive performance in various NLP tasks, they still display a performance disparity compared to completely fine-tuned models (Wang et al., 2022). Third, LLM-based ICL has not been investigated in MEBSA, and the adaptability of LLMs to various MEBSA subtasks remains uncertain. Hence, our objective is to investigate the potential of utilizing LLMs on the MEBSA. We intend to develop an efficient ICL framework that can address all four subtasks of MEBSA, improving performance of ICL to either narrow the gap or surpass the current state-of-the-art fine-tuned models.

Therefore, this work focuses on investigating the potential of ChatGPT, one of the representative LLMs, in zero-shot and few-shot settings for MEBSA. Specifically, we apply a vision-language model named BLIP (Li, Li, Xiong & Hoi, 2022) to generate image captions and detecting visual entities and sentiments, and the input images are then translated into auxiliary sentences. Following this, we design a set of instructions that utilize both the auxiliary visual text input and the original text input to conduct zero-shot learning. Additionally, we concatenate these task instructions and a few randomly selected demonstration samples to perform few-shot learning. Although the few-shot learning strategy achieves reasonable performance, it lags from existing fine-tuned multimodal approaches in some tasks.

To further enhance the efficacy of few-shot learning, we develop a demonstration exemplar selection framework to retrieve demonstration samples that are similar to each test sample in the prompt. Concretely, a scoring function is developed to quantify the degree of similarity among samples by taking into account multiple factors, including the textual input of each sample (referred to as the original sentence), auxiliary sentences derived from the vision modality, and the corresponding labels. Based on the similarity among samples, we construct positive and negative instance pairs for training a retriever with a contrastive learning model. With this retriever, exemplars that are similar to the test sample we can retrieved, in terms of the input text, the input image (using the auxiliary sentence as a proxy), and the output to be predicted.

The contributions made by this work are mainly described as follows:

[1] https://platform.openai.com/docs/guides/gpt.

- This paper is the first to investigate in-context learning for Multimodal Entity-Based Sentiment Analysis (MEBSA). Unlike most traditional methods using a large amount of training samples for fine-tuning, our in-context learning framework leverages a small number of training samples, which significantly diminishes the computational resources and lessens the demand for data annotation.
- We develop a novel in-context learning framework based on an entity-aware contrastive learning model to retrieve demonstration samples that are similar to each test sample. We also devise an entity-aware similarity scoring function that comprehensively considers various components of a sample to construct positive and negative instance pairs for training the entity-aware contrastive learning model.
- The experimental results demonstrate that incorporating a few randomly selected samples in the prompt significantly boosts in-context learning performance in comparison to the zero-shot learning approach. Moreover, the exemplar samples selected by our contrastive learning-based retriever further enhances the performance of in-context learning. Compared to many representative fine-tuned methods, our contrastive learning-based framework exhibits indistinguishable or superior performance across all the four MEBSA subtasks.

2. Research objectives

This paper aims to delineate three research objectives as outlined below:

- **RO1**: Develop a general ICL framework that can effectively address all four subtasks of the MEBSA task with a small sample size. (Section 4.3)
- **RO2**: Develop a similarity scoring function that comprehensively considers various components of a sample to construct positive and negative instance pairs for training the entity-aware contrastive learning model. (Section 4.4.1)
- **RO3**: Develop an effective demonstration exemplar retriever that can select relevant demonstration samples for each test sample. (Sections 4.4.2 and 4.4.3)

3. Related work

3.1. Entity-based sentiment analysis in text

Sentiment analysis in text focuses on identifying the sentiment or emotional content present in a given text. Sentence-level sentiment analysis offered an overall sentiment polarity for the sentence but often misses detailed opinions about specific entities (Liu, 2012; Zhang et al., 2022). This gap can be ascribed to the emergence of Entity-based Sentiment Analysis (EBSA) in text, which recognized both the entities and their sentiments from the text (Pontiki et al., 2016; Schouten & Frasincar, 2016). Initial studies on EBSA concentrated on Named Entity Recognition (NER) and Entity-based Sentiment Classification (EBSC). NER is to detect entities from a given text and categorize them into predetermined categories. The approaches of NER has evolved from traditional machine learning models (Li, Sun, Han & Li, 2022; Li, Sun, Weng, & He, 2014) to various deep learning techniques, including CNN (Shang & Ran, 2022) and LSTM (Katiyar & Cardie, 2018). EBSC is to detect the sentiment regarding entities that mentioned in the text. Deep learning models are commonly used for this task. For instance, Meškelė and Frasincar (2020) and Zhang, Zhang, and Vo (2016) both utilized gated neural networks to collect syntactic and semantic information in order to facilitate sentiment identification. Given the underling connection between NER and EBSC, Zhang, Zhang, and Vo (2015) proposed Entity-Sentiment Pair Extraction to simultaneously extract both entity and sentiment. Recent studies shifted towards more complex compound EBSA tasks, which aim to simultaneously detect multiple elements. For example, Barnes, Kurtz, Oepen, Øvrelid, and Velldal (2021) and Barnes et al. (2022) included linguistic structure into the sentiment graphs and proposed a dependency graph parsing approach. This approach can simultaneously detect opinion holder, target entity, opinion, and sentiment, which produced a more comprehensive sentiment profile. However, all these studies focused on textual inputs. Different from the aforementioned studies, our work extends beyond mere textual analysis, delving into multimodal entity-based sentiment analysis, which combines both visual and textual modalities.

3.2. Multimodal entity-based sentiment analysis

Related studies on MEBSA can be categorized into four main categories according to the tasks they focus on: MEE, MESC, MESPE and MECSTE. In the early stages of exploring the MEE task, deep learning models were utilized to establish the relationship between visual and textual data. For example, Zhang, Fu, Liu, and Huang (2018) applied an attention mechanism with a co-attention layer based on the LSTM model to model the connection between visual object embeddings and textural features. Additionally, the visual attention mechanism-based model developed by Lu, Neves, Carvalho, Zhang, and Ji (2018) achieved good performance on multiple datasets by using the visual information that most relevant to the textual information and ignoring the irrelevant one. With the recent development of self-attention mechanisms, researchers have recently applied transformer-based models to the MEE task. For instance, the transformer-based model developed by Yu, Jiang, Yang and Xia (2020) utilized unimodal feature representations and a number of multimodal transformers to obtain the inter-modal dynamics between textual and visual inputs. Similarly, research on the MESC subtask exhibits a similar trend to studies on the MEE. Previous works utilized deep learning approaches to model the inter-modal interactions among entity, image and text (Kumar, Srinivasan, Cheng, & Zomaya, 2020; Truong & Lauw, 2017, 2019; Zadeh, Chen, Poria, Cambria, & Morency, 2017). Given the prevalence of the attention-based models, an attention-based co-memory network (Xu, Mao, & Chen, 2018) was developed to capture the pairwise interaction among aspect, text, and image. Following that, Yu and Jiang (2019) added target-aware transformer layers based on top of the attention model to effectively integrate textual and visual modalities.

Given the interrelation and relevance between the MEE and MESC tasks, Ju et al. (2021) introduced the MESPE task, which aims to identify entities and sentences simultaneously. They proposed an auxiliary module to identify the relatedness between text and image, which alleviated the noise brought by image features. Based on this, Yang, Na and Yu (2022) devised a model that could generate sentiment-aware and entity-aware representations from text and images. More recently, Yang, Wang et al. (2023) introduced a new MECSTE task, and proposed a generative multimodal approach, which utilized a pretrained sequence-to-sequence model to produce paraphrases containing entity-category-sentiment triples. However, the aforementioned studies mainly focus on one or two subtasks of MEBSA. In contrast, our study is intended to develop a general approach that covers all the four subtasks of MEBSA. Moreover, all these aforementioned models require fine-tuning, which involves updating the parameters for each downstream task. These fine-tuning models necessitate a substantial quantity of annotated datasets and computational resources. Unlike the aforementioned methods, our study employs LLMs, eliminating the requirement for extensive annotated data. By employing ICL, we enhance the performance of LLMs with a limited number of sample demonstrations, significantly reducing the need for manual annotation of datasets and enhancing the practical applicability.

3.3. Large language models and in-context learning

Recent months have witnessed a notable increase in the level of attention and focus dedicated to the advancement and expansion of neural large language models (LLMs). The LLMs, including GPT-3, Chinchilla and PaLM (Brown, Mann, Ryder, Subbiah et al., 2020; Chowdhery, Narang, Devlin, Bosma et al., 2022; Hoffmann et al., 2022), showed impressive performance in NLP tasks. Moreover, ChatGPT (OpenAI, 2023a) and GPT-4 (OpenAI, 2023b) exhibited extraordinary instruction following ability, exceeding that of many other conversational language models. This has garnered considerable interest from both the academic community and the general population. Fine-tuning these LLMs is becoming increasingly impractical due to the substantial computational resources required. A more common approach now is to design prompts that leverage the models' ability to follow instructions to accomplish specific tasks. When prompted with a few demonstration examples, the output generated by these LLMs became more controllable, namely in-context learning (ICL) (Brown, Mann, Ryder, Subbiah et al., 2020; Dong, Li et al., 2022), a new learning paradigm that can adopt LLMs to new tasks without any parameters updating. However, its performance was not always stable, since it was affected by a number of factors, such as demonstration example selection, the formatting of demonstration, and the order of these examples (Lu et al., 2022; Wei et al., 2022; Zhao, Wallace, Feng, Klein, & Singh, 2021b).

To construct a richer context, some studies employ various methods to select demonstration examples, including both supervised and unsupervised approaches. For unsupervised ones, Liu et al. (2022) showcased that choosing nearest neighbors (based on L2 distance or cosine similarity) as demonstration examples can improve ICL performance. Gonen, Iyer, Blevins, Smith, and Zettlemoyer (2022) chose examples with low perplexity. For supervised ones, Rubin, Herzig, and Berant (2022) developed a two-stage retrieval strategy for example selection. Li et al. (2023) introduced a unified demonstration retriever, streamlining the process of selecting for a variety of tasks. Ye, Wu, Feng, Yu and Kong (2023) proposed a determinantal point processes model to calculate the probability of the subset of demonstration examples, thereby enhancing the effectiveness of example selection. Although there has been some preliminary assessments of ChatGPT's performance in textual sentiment analysis tasks, revealing generally good performance, it still falls short in certain tasks compared to fine-tuned models (Wang, Xie, Ding, Feng, & Xia, 2023). Beyond textual sentiment analysis, we consider multimodal sentiment analysis, specifically, how we can adopt LLMs like ChatGPT to MEBSA. In this work, we explore translating images into captions with rich semantic information (also known as modal translation (Khan & Fu, 2021; Vinyals, Toshev, Bengio, & Erhan, 2015)) and leveraging the ICL capabilities of LLMs to track it (without any gradient updates to the LLMs). An alternative approach to this is to modularistically train LLMs with multimodal instruction data for multimodal abilities, as has been done with mPLUG-Owl (Ye et al., 2023), PandaGPT (Su et al., 2023) and etc. We leave this for our future work.

4. Methodology

This section delineates the tasks and subsequently provides a comprehensive breakdown of each component within our ICL framework. Our ICL framework comprises three modules: input construction (Section 4.2), in-context learning (Section 4.3) and demonstration exemplar retriever (Section 4.4), as depicted in Fig. 2. The notations and definitions utilized in this section are listed in the Table 1. This framework is applicable to all four subtasks of MEBSA, namely MEE, MESC, MESPE and MECSTE.

In the input construction module, the visual input is converted to text through the generation of image captions and the extraction of visual entities, entity categories and visual sentiments from images. Following this, the transformed visual text is integrated with the original textual input. Additionally, in the demonstration exemplar retriever module, a contrastive learning-based sample embedding model is developed to retrieve top-K similar samples for the test sample. To train the entity-aware contrastive learning model, we design a scoring function that measures the similarity among samples, hence constructing positive and negative pairs according to the similarity. Furthermore, in the in-context learning module, we develop various task-specific prompts to instruct the model to be aware of each MEBSA subtask.

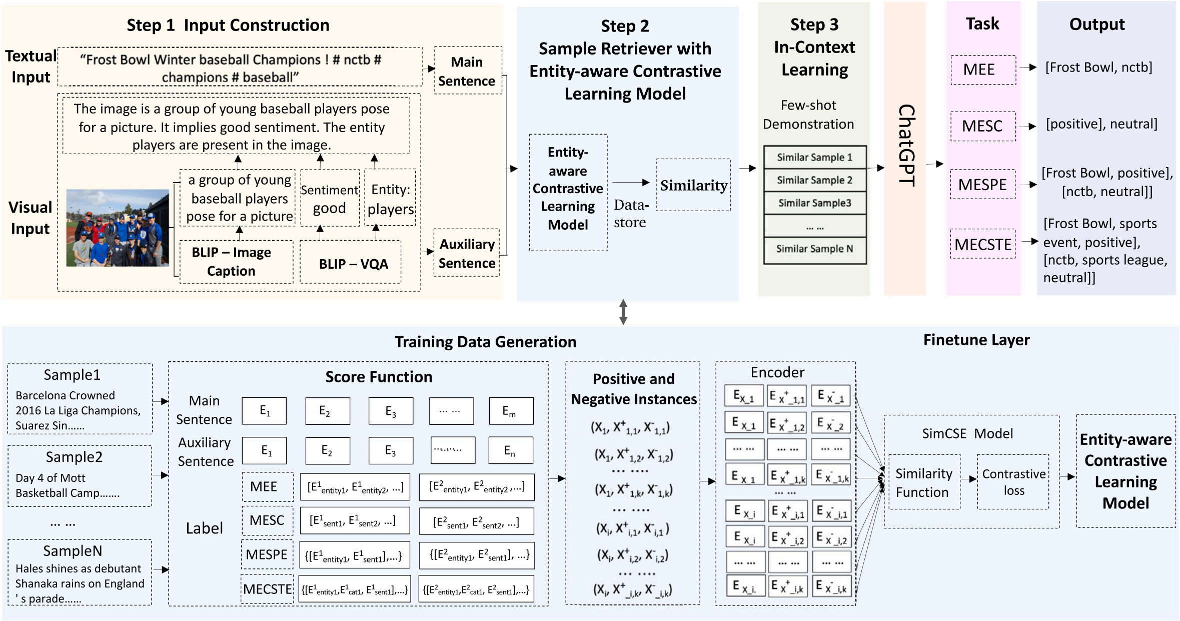

Step 1 Input Construction
Textual Input
"Frost Bowl Winter baseball Champions ! # nctb # champions # baseball"
Main Sentence
Visual Input
The image is a group of young baseball players pose for a picture. It implies good sentiment. The entity players are present in the image.
a group of young baseball players pose for a picture
Sentiment: good
Entity: players
BLIP – Image Caption
BLIP – VQA
Auxiliary Sentence
Step 2 Sample Retriever with Entity-aware Contrastive Learning Model
Entity-aware Contrastive Learning Model
Data-store
Similarity
Step 3 In-Context Learning
Few-shot Demonstration
Similar Sample 1
Similar Sample 2
Similar Sample3
… …
Similar Sample N
ChatGPT
Task
MEE
MESC
MESPE
MECSTE
Output
[Frost Bowl, nctb]
[positive], neutral]
[Frost Bowl, positive], [nctb, neutral]]
[Frost Bowl, sports event, positive], [nctb, sports league, neutral]]
Training Data Generation
Finetune Layer
Sample1
Barcelona Crowned 2016 La Liga Champions, Suarez Sin……
Sample2
Day 4 of Mott Basketball Camp…….
… …
SampleN
Hales shines as debutant Shanaka rains on England ' s parade……
Score Function
Main Sentence
Auxiliary Sentence
Label
E1 E2 E3 … … Em
E1 E2 E3 …… En
MEE [E1entity1, E1entity2, …] [E2entity1, E2entity2 ,…]
MESC [E1sent1, E1sent2, …] [E2sent1, E2sent2, …]
MESPE {[E1entity1, E1sent1],…} {[E2entity1, E2sent1], …}
MECSTE {[E1entity1,E1cat1, E1sent1],…} {[E2entity1,E2cat1, E2sent1],…}
Positive and Negative Instances
(X1, X+1,1, X-1,1)
(X1, X+1,2, X-1,2)
(X1, X+1,k, X-1,k)
(Xi, X+i,1, X-i,1)
(Xi, X+i,2, X-i,2)
(Xi, X+i,k, X-i,k)
Encoder
SimCSE Model
Similarity Function
Contrastive loss
Entity-aware Contrastive Learning Model

Fig. 2. The proposed in-context learning framework based on the entity-aware demonstration exemplar retriever for multimodal entity-based sentiment analysis.

Table 1
Notations and their corresponding definitions.

Notations	Definitions	Notations	Definitions
$\cdot$	Dot product	$*$	Elementwise multiplication
s	The original textual input of the sample	x	The input tokens in s
$\boldsymbol{V}$	The visual input	$\boldsymbol{h}_{\text{main}}$	The contextual embedding of main sentences
$\boldsymbol{h}_{\text{auxiliary}}$	The contextual embedding of auxiliary sentences	$\boldsymbol{h}_{\text{label}}$	The contextual embedding of label
$(\boldsymbol{a}, \boldsymbol{b})$	Sample pair, a^{test} is the test sample b^{train} is the training sample	$\boldsymbol{x}_i$	The sample in training set
$\boldsymbol{x}^+_{i,k}$	The kth positive sample	$\boldsymbol{x}^-_{i,k}$	The kth negative sample
w_i	A trade-off parameter and it is in [0,1]	$\neg w_i$	Equals to $1 - w_1$
$\boldsymbol{z}^{train}$	The hidden representation of the train sample	$\boldsymbol{z}^{test}$	The hidden representation of the test sample

4.1. Task formulation

Given a multimodal tweet, the text comprising n words is denoted as $\boldsymbol{s} = [x_1, \ldots, x_n]$, and the associated image is represented as $\boldsymbol{V}$. The objective is to generate the output for each MEBSA subtask as follows:

- Multimodal Entity Extraction (MEE): The input is $\boldsymbol{s} = [x_1, x_2, \ldots, x_n]$ and $\boldsymbol{V}$, and the output is $\boldsymbol{y}_{\text{MEE}} = \{e_1, e_2, \ldots, e_m\}$, where e_m represents the mth entity to be extracted. The task objective is to identify every entity that appears in the multimodal inputs (Wu, Cheng, Wang, Li & Chi, 2020).
- Multimodal Entity-based Sentiment Classification (MESC): The input is $\boldsymbol{s} = [x_1, x_2, \ldots, x_n]$, $\boldsymbol{V}$, and a set of given entities $\{e_1, e_2, \ldots, e_m\}$, and the output is $\boldsymbol{y}_{\text{MESC}} = \{s_1, s_2, \ldots, s_m\}$, where s_m is the sentiment towards the mth given entity. This task focuses on determining the sentiment towards each entity present in the multimodal inputs, when the entity is provided (Xu, Mao, & Chen, 2019; Yu & Jiang, 2019).
- Multimodal Entity-Sentiment Pair Extraction (MESPE): The input is $\boldsymbol{s} = [x_1, x_2, \ldots, x_n]$ and $\boldsymbol{V}$, and the output is $\boldsymbol{y}_{\text{MESPE}} = \left\{[e_1, s_1], [e_2, s_2], \ldots, [e_m, s_m]\right\}$, where m denotes the total amount of extracted entities, e_m represents the mth entity to the extracted, and s_m is the sentiment towards the mth entity. The aim of this task is recognizing all the entities and associated sentiments (Ju et al., 2021).
- Multimodal Entity-Category-Sentiment Triple Extraction (MECSTE): The input is $\boldsymbol{s} = [x_1, x_2, \ldots, x_n]$ and $\boldsymbol{V}$, and the output is $\boldsymbol{y}_{\text{MECSTE}} = \left\{[e_1, c_1, s_1], [e_2, c_2, s_2], \ldots, [e_m, c_m, s_m]\right\}$, where m denotes the total number of entity-category-sentiment triples, $[e_m, c_m, s_m]$ represents the mth triple of entity-category-sentiment, e_m is the mth entity to the extracted, and c_m and s_m refer to the fine-grained category and sentiment, respectively. The task objective is to identify all the triples of {entity, category, sentiment} from the multimodal inputs (Yang, Wang et al., 2023).

4.2. Input construction

Our work leverages ChatGPT (OpenAI, 2023a), a representative large language model, to conduct in-context learning for MEBSA. Since ChatGPT is currently limited to processing the textual input and cannot deal with the visual input, we transform the visual input into textual counterparts, which are regarded as auxiliary sentences. These auxiliary sentences are subsequently concatenated with the original text input and fed into ChatGPT to generate the output for downstream tasks.

To extract semantic information from visual inputs, we employ a widely-used image captioning model named BLIP (Li, Li et al., 2022), which transforms images into eloquent image captions. For instance, the visual input in Fig. 2 is translated into the text of "a group of young baseball players pose for a picture". Secondly, given the importance of entity detection MEE, MESPE, and MECSTE tasks, we address entity detection from images as a Visual Question Answering (VQA) task, and employ BLIP to extract entity-related information from visual inputs, thereby translating visual entity information into textual representations. Moreover, since sentiment plays an important factor in tasks of MESC, MESPE, and MECSTE, we also utilize BLIP to identify the overall sentiment expressed in the input image based on the VQA task. The details are described below.

4.2.1. Visual input construction

Image Captioning. We apply a vision-language model BLIP (Li, Li et al., 2022) to produce descriptions (captions) for the given image $\boldsymbol{V}$: $\boldsymbol{c}_{\text{caption}} = \text{Caption}_{\text{BLIP}}(\boldsymbol{V})$, where $\boldsymbol{c}_{\text{caption}}$ represents the generated image caption, and $\boldsymbol{V}$ represents the input image.

Visual Entity Detection. To detect the entities contained in the image, we address the entity detection task as a VQA task, and apply BLIP to detect the entities. Specifically, given an image $\boldsymbol{V}$ and a textual question $\boldsymbol{q}_{\text{entity}}$ = "What are the entities in the image?", the BLIP model generates a textual answer about the entities appeared in the input image $\boldsymbol{V}$: $\boldsymbol{a}_{\text{entity}} = \text{VQA}_{\text{entity}}(\boldsymbol{V})$.

Visual Sentiment Detection. Similarly, we address the detection of visual sentiment as a VQA task, and use BLIP to identify the visual sentiment. Given an image $\boldsymbol{V}$ and a textual question $\boldsymbol{q}_{\text{entity}}$ = "What is the general sentiment conveyed by the image?", the BLIP VQA model generates a textual response containing a subjective word such as *good*, *bad*, and *happy*: $\boldsymbol{a}_{\text{sentiment}} = \text{VQA}_{\text{sentiment}}(\boldsymbol{V})$.

Translating the Input Image to Auxiliary Sentences. Once the captions, entities, and sentiment are extracted from the image, an auxiliary sentence is constructed by encompassing the image caption $\boldsymbol{x}_{\text{caption}}$ = "The image isabout $\boldsymbol{c}_{\text{caption}}$", the visual entities $\boldsymbol{x}_{\text{entity}}$ = "The entity of the $\boldsymbol{a}_{\text{entity}}$ is present in the image", as well as the visual sentiment $\boldsymbol{x}_{\text{sentiment}}$ = "The image implies $\boldsymbol{a}_{\text{sentiment}}$ sentiment", as illustrated in Fig. 2

4.2.2. Multimodal input construction

After translating the image into three auxiliary sentences, these auxiliary sentences are combined with the original textual input to form the multimodal input $\boldsymbol{x}$:

$$\boldsymbol{x} = [\boldsymbol{s}, \boldsymbol{x}_{\text{caption}}, \boldsymbol{x}_{\text{entity}}, \boldsymbol{x}_{\text{sentiment}}]. \tag{1}$$

As previously mentioned, it is challenging to fine-tune LLMs because of the considerable quantity of model parameters. To adapt LLMs to downstream tasks, a common practice is to leverage one of the prompt engineering methods known as in-context learning. This involves providing the model with task instructions and sample demonstrations as the prompt. Therefore, this work aims to develop effective prompts to facilitate LLMs' comprehension of each MEBSA subtask.

4.3. In-context learning

Task Instruction. Task instruction provides a description of the task, including task objectives, input components, learning instructions, and output format. Fig. 3 presents the task description and the output format for each MEBSA subtask in the upper section. Moreover, we clarify in the prompt that each input sample consists of a main sentence and three auxiliary sentences, as specified in Section 4.2.1.

Zero-shot Instruction. One straightforward solution to address the given task instruction is feeding the test sample directly into ChatGPT to generate the output for each MEBSA subtask. As shown in the middle of Fig. 3, the zero-shot instruction is "*generate output for the following sentence: Test Sample*". *Test Sample* consists of a main sentence and three auxiliary sentences.

Few-shot Demonstration with Random Selection. Previous research has indicated that zero-shot learning of LLMs is usually incapable of achieving satisfactory performance for relation extraction task and multimodal NER task (Li, Li, Pan & Pan, 2023; Wan et al., 2023). A straightforward method for enhancing LLMs' efficacy on information extraction tasks is randomly selecting a few training samples as the few-shot demonstration for the prompt. With these few-shot samples, LLMs can better understand the task instruction and output format, consequently leading to improved predictions. However, this straightforward method disregards the semantic relationship between few-shot samples and the test sample, potentially introducing noise to the predictions.

Few-shot Demonstration with Entity-aware Exemplar Retriever. To address the issue of random selection, we develop the Entity-aware Exemplar Retriever to effectively retrieve samples for few-shot demonstration. The details of the Entity-aware Exemplar Retriever is described in Section 4.4.3, in which an Entity-aware Contrastive Learning model is trained with the objective of reducing the divergence between similar samples and maximizing the dissimilarity between dissimilar ones. Subsequently, the trained Entity-aware Exemplar Retriever is used to select samples that exhibit semantic similarity to the test sample, thereby increasing the efficacy of few-shot learning.

Inference Based on the Output from ChatGPT. After obtaining the results from ChatGPT, we perform post-processing on the outputs that do not align with the format of labels, such as adjusting the letter case of sentiment polarities. In cases where the

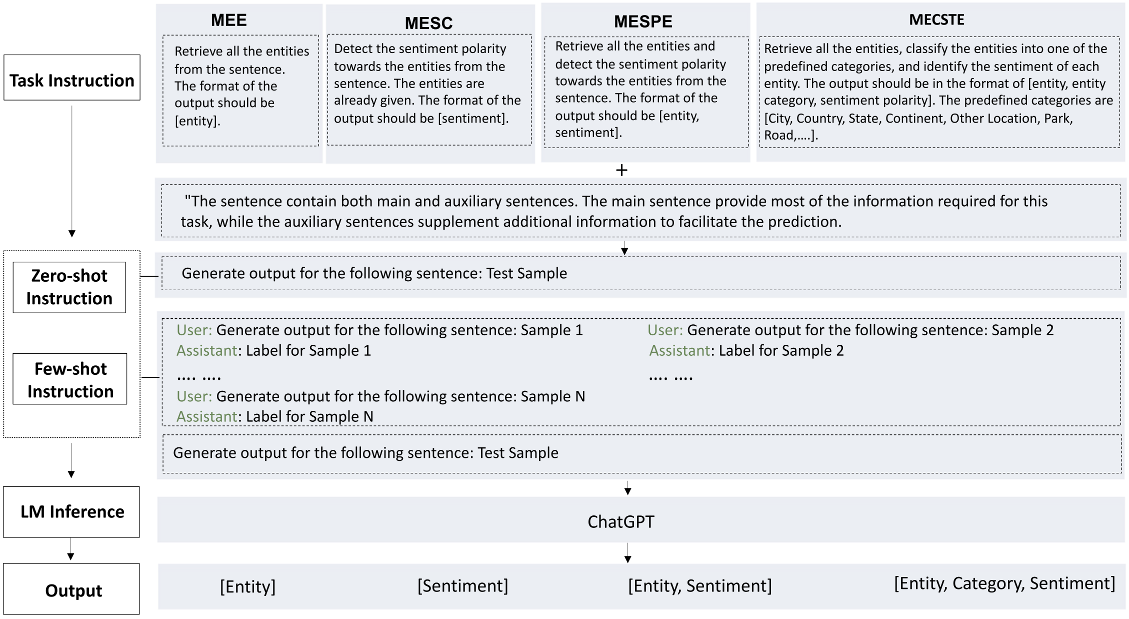

Task Instruction
MEE
Retrieve all the entities from the sentence. The format of the output should be [entity].
MESC
Detect the sentiment polarity towards the entities from the sentence. The entities are already given. The format of the output should be [sentiment].
MESPE
Retrieve all the entities and detect the sentiment polarity towards the entities from the sentence. The format of the output should be [entity, sentiment].
MECSTE
Retrieve all the entities, classify the entities into one of the predefined categories, and identify the sentiment of each entity. The output should be in the format of [entity, entity category, sentiment polarity]. The predefined categories are [City, Country, State, Continent, Other Location, Park, Road,….].
+
"The sentence contain both main and auxiliary sentences. The main sentence provide most of the information required for this task, while the auxiliary sentences supplement additional information to facilitate the prediction.
Zero-shot Instruction
Generate output for the following sentence: Test Sample
Few-shot Instruction
User: Generate output for the following sentence: Sample 1
Assistant: Label for Sample 1
…. ….
User: Generate output for the following sentence: Sample N
Assistant: Label for Sample N
User: Generate output for the following sentence: Sample 2
Assistant: Label for Sample 2
…. ….
Generate output for the following sentence: Test Sample
LM Inference
ChatGPT
Output
[Entity]
[Sentiment]
[Entity, Sentiment]
[Entity, Category, Sentiment]

Fig. 3. The proposed in-context learning schema for different MEBSA subtasks.

generated entity categories do not belong to the 52 fine-grained categories, BERT is used to obtain the embeddings of these generated entity categories (Devlin, Chang, Lee, & Toutanova, 2019). Subsequently, we compute the similarity of these aberrant categories to each of the 52 categories, and finally substitute them with the most similar entity category. Furthermore, we add a post-processing step to the outputs generated by ChatGPT for the MEE task. Specifically, ChatGPT generates a list of entities, accompanied by their corresponding confidence scores, for the test sample. The list of entities is organized in descending order by confidence scores, and top-N entities are chosen as the final predicted entity list of the test sample. Notably, the framework that we have developed places greater emphasis on retrieving demonstration examples using an entity-aware contrastive learning model, hence facilitating ICL. ChatGPT can be replaced with other LLMs flexibly. Our framework has applicability and adaptability to different LLMs.

4.4. Demonstration exemplar retriever based on entity-aware contrastive learning

This subsection provides a comprehensive explanation of our ICL framework with an Entity-Aware Demonstration Exemplar Retriever that is depicted in Fig. 2. A score function is developed to evaluate the pairwise similarity between samples in the training dataset. Based on the similarity score, a set of positive and negative instances is constructed as the training dataset, as explained in Section 4.4.1. Subsequently, the Entity-aware Contrastive Learning model is trained to acquire knowledge about word embedding and the sample distribution for four subtasks, as described in Section 4.4.2. By utilizing the trained Entity-ware contrastive learning model, similar samples are retrieved for the purpose of conducting a few-shot demonstration in the in-context learning module, as outlined in Section 4.4.3. We also delineate the algorithm of each module of our developed framework in Algorithms 1, 2, 3 to facilitate the understanding. Compared to few-shot demonstration with random selection, the sample retriever with the Entity-Aware Contrastive Learning model better captures the semantic similarities of input sentences and labels between demonstration samples and the test sample.

4.4.1. Positive and negative instance construction

A crucial consideration for utilizing the Supervised SimCSE framework (Gao, Yao & Chen, 2021) for the MEBSA task is the construction of effective positive instances $\boldsymbol{x}_i^+$ and negative instances $\boldsymbol{x}_i^-$. Many data augmentation techniques, including word deletion, substitution, and reordering have been employed to generate positive instances in NLP (Meng et al., 2021; Wu et al., 2020). Nevertheless, the instances generated by these data augmentation methods may damage the semantic meaning of the source sentence and exhibit deficiencies in terms of fluency and coherence as a result of the discrete nature of words. Additionally, one commonly employed approach for negative instance generation is to randomly sample the remaining instances within a mini-batch. However, this approach inevitably brings much noise to the contrastive learning process.

To address these two issues, we devise a scoring function that effectively assesses the degree of similarity among each pair of samples, thereby constructing useful positive and negative instances for training the contrastive learning model. The scoring function considers three factors, i.e., semantic similarities of main sentences (the original text input), auxiliary sentences generated from the visual input, and the labels.

Specifically, RoBERTa (Liu et al., 2019) is utilized for the tokenization of main sentences, auxiliary sentences and the label of each sample. The sentences and labels are converted into the corresponding representations as follows:

$$\boldsymbol{h}_{\text{main}} = \text{RoBERTa}(\boldsymbol{s}), \quad \boldsymbol{h}_{\text{auxiliary}} = \text{RoBERTa}([\boldsymbol{x}_{\text{caption}}, \boldsymbol{x}_{\text{entity}}, \boldsymbol{x}_{\text{sentiment}}]), \quad \boldsymbol{h}_{\text{label}} = \text{RoBERTa}(\boldsymbol{x}_{\text{label}}), \tag{2}$$

where $\boldsymbol{h}_{\text{main}}$ and $\boldsymbol{h}_{\text{auxiliary}}$ denote the contextual embedding of main sentences and auxiliary sentences, s denotes the original textual input of the sample, x_{caption}, x_{entity}, and $x_{\text{sentiment}}$ are sentences that describe the image captioning, entity and sentiment of the image, as described in Section 4.2.1, and $\boldsymbol{h}_{\text{label}}$ represents the embedding of the label, which can be either entity, category, or sentiment depending on the task.

Similarity Between Main Sentences. We apply the cosine similarity to determine the semantic similarity of main sentences for each pair of samples as follows:

$$\text{sim-main}(a, b) = \text{cosine}(\boldsymbol{h}^a_{\text{main}}, \boldsymbol{h}^b_{\text{main}}) = \frac{\boldsymbol{h}^a_{\text{main}} \cdot \boldsymbol{h}^b_{\text{main}}}{\|\boldsymbol{h}^a_{\text{main}}\| \cdot \|\boldsymbol{h}^b_{\text{main}}\|}, \tag{3}$$

where sim-main($\boldsymbol{a}, \boldsymbol{b}$) is the cosine similarity of input sentences in sample pair (a, b).

Similarity Between Auxiliary Sentences. The measurement of semantic similarity between the auxiliary sentences in sample pair (a, b) is as follows:

$$\text{sim-auxiliary}(a, b) = \text{cosine}(\boldsymbol{h}^a_{\text{auxiliary}}, \boldsymbol{h}^b_{\text{auxiliary}}) = \frac{\boldsymbol{h}^a_{\text{auxiliary}} \cdot \boldsymbol{h}^b_{\text{auxiliary}}}{\|\boldsymbol{h}^a_{\text{auxiliary}}\| \cdot \|\boldsymbol{h}^b_{\text{auxiliary}}\|}, \tag{4}$$

where sim-auxiliary($\boldsymbol{a}, \boldsymbol{b}$) is the cosine similarity of auxiliary sentences in sample pair (a, b).

Similarity Between Labels. Regarding the semantic similarity between labels, there are two challenges. First, since the labels of different subtasks are disparate, it requires designing specific similarity computation strategy for each subtask. For example, MEE focuses on entity detection, MESC and MESPE focus on entity-related sentiment detection, while MECSTE focuses on entity, category, and sentiment detection. Second, since the label of each subtask is often a set of tuples, it is necessary to consider the similarities between two sets of tuples.[2]

To tackle the first challenge, we measure the similarity of the same element in a pair of tuples via cosine similarity . Let us assume $\textbf{label}^a = \{label^a_1, \dots, label^a_i, \dots, label^a_I\}$ is the label set for sample a, and $\textbf{label}^b = \{label^b_1, \dots, label^b_j, \dots, label^b_J\}$ is the label set for sample b, where I and J are the count of tuples in two samples. We then assign appropriate weights to the entity, category, and sentiment for evaluating their respective contributions to the similarity between $label^a_i$ and $label^b_j$:

$$\text{sim}(label^a_i, label^b_j)_{\text{MEE}} = \text{cosine}(\boldsymbol{h}^a_{\text{entity}}, \boldsymbol{h}^b_{\text{entity}}), \tag{5}$$

$$\text{sim}(label^a_i, label^b_j)_{\text{MESC,MESPE}} = w_1 * \text{cosine}(\boldsymbol{h}^a_{\text{entity}}, \boldsymbol{h}^b_{\text{entity}}) + \neg w_1 * \text{cosine}(\boldsymbol{h}^a_{\text{sentiment}}, \boldsymbol{h}^b_{\text{sentiment}}), \tag{6}$$

$$\begin{aligned} \text{sim}(label^a_i, label^b_j)_{\text{MESCTE}} = {} & w_2 * \text{cosine}(\boldsymbol{h}^a_{\text{entity}}, \boldsymbol{h}^b_{\text{entity}}) + w_3 * \text{cosine}(\boldsymbol{h}^a_{\text{category}}, \boldsymbol{h}^b_{\text{category}}) \\ & + \neg(w_2 + w_3) * \text{cosine}(\boldsymbol{h}^a_{\text{sentiment}}, \boldsymbol{h}^b_{\text{sentiment}}), \end{aligned} \tag{7}$$

where $w_1 \in [0, 1]$ in Eq. (6) is a trade-off parameter, and $\neg w_1$ equals to $1 - w_1$. The same holds true for Eq. (7).

To tackle the second challenge, given any tuple in $\textbf{label}^a$, we calculate its similarity score with all the tuples in $\textbf{label}^b$, and regard the maximum similarity score as the similarity score between this tuple and $\textbf{label}^b$. After obtaining the similarity scores of all the tuples in $\textbf{label}^a$, the final similarity score between $\textbf{label}^a$ and $\textbf{label}^b$ is then determined by averaging the similarity scores, as shown below:

$$\text{sim-label}(a, b) = \frac{1}{I} \sum_{i=1}^{I} \max_j \left(\text{sim}(label^a_i, label^b_j) \right), \tag{8}$$

Lastly, we consider the contributions of main sentences, auxiliary sentences and labels to obtain the similarity score between a sample pair (a, b) as follows:

$$\text{sim-sample}(a, b) = \alpha_0 * \text{sim-main}(a, b) + \alpha_1 * \text{sim-auxiliary}(a, b) + \alpha_2 * \text{sim-label}(a, b), \tag{9}$$

where α_0, α_1, and α_2 are three trade-off parameters.

Positive and Negative Instance Pairs. Given an input sample $\boldsymbol{x}_i$, we calculate its similarity with all the samples in training set. Subsequently, top-K samples with highest similarity scores are chosen as positive samples, whereas the bottom-K samples with the lowest similarity scores are selected as negative samples, as shown in Fig. 2. Therefore, we can obtain K positive and negative instance pairs as follows:

$$\mathcal{D}_i = \{(\boldsymbol{x}_i, \boldsymbol{x}^+_{i,1}, \boldsymbol{x}^-_{i,1}), (\boldsymbol{x}_i, \boldsymbol{x}^+_{i,2}, \boldsymbol{x}^-_{i,2}) \dots, (\boldsymbol{x}_i, \boldsymbol{x}^+_{i,K}, \boldsymbol{x}^-_{i,K})\}, \tag{10}$$

where $\boldsymbol{x}_i$ denotes the sample in training set, and $\boldsymbol{x}^+_{i,k}$ and $\boldsymbol{x}^-_{i,k}$ signify the kth positive sample and negative sample, respectively.

In order to facilitate a better understanding of the construction procedure of positive and negative instances, we show the pseudo-algorithm in Algorithm 1 to clearly articulate each step.

[2] Sample labels for the MEE, MESC, MESPE, and MECSTE subtasks are as follows: $[entity_1, entity_2, \dots, entity_n]$, $[sentiment_1,\ sentiment_2, \dots, sentiment_n]$, $\{[entity_1, sentiment_1], [entity_2, sentiment_2], \dots, [entity_n, sentiment_n]\}$, and $\{[entity_1, category_1, sentiment_1]$, $[entity_2, category_2, sentiment_2], \dots, [entity_n, category_n, sentiment_n]\}$, respectively.

Algorithm 1 Positive and Negative Instance Construction

```
Given training set size $M$, training set $\{x_k\}_{k=1}^{M}$, encoder $\mu_\theta(\cdot)$,
$D = []$
for all $i \in \{1, \cdots, M\}$ do
    $\boldsymbol{h}_i = \mu_\theta(x_i)$
    simlist = []
    for all $j \in \{1, \cdots, M\}$ do
        $\boldsymbol{h}_j = \mu_\theta(x_j)$
        sim-sample($\boldsymbol{h}_i, \boldsymbol{h}_j$) = $a_0 \cdot$ sim-main($\boldsymbol{h}_i, \boldsymbol{h}_j$) + $a_1 \cdot$ sim-auxiliary($\boldsymbol{h}_i, \boldsymbol{h}_j$) + $a_2 \cdot$ sim-label($\boldsymbol{h}_i, \boldsymbol{h}_j$)
        simlist.append(sim-sample($\boldsymbol{h}_i, \boldsymbol{h}_j$))
    end for
    POS = Argsort(simlist, reverse = True)[1 : $K + 1$]
    NEG = Argsort(simlist, reverse = True)[$-K$ :]
    for all $k \in \{1, \cdots, K\}$ do
        $D$.append(($x_i$, $x_{\mathrm{POS}[k]}$, $x_{\mathrm{NEG}[k]}$))
    end for
end for
Return Positive and negative instance set $D = \{(x_i, x_i^+, x_i^-)\}_{i=1}^{M*K}$
```

Algorithm 2 Entity-aware Contrastive Learning

```
Given positive and negative instance set $D$, encoder $f(\cdot)$, neural network projection layer $g(\cdot)$, batch size $N$
for sampled minibatch $\{(x_k, x_k^+, x_k^-)\}_{k=1}^{N}$ do
    for all $i \in \{1, \ldots, N\}$ do
        $\boldsymbol{h}_i = f(x_i)$, $\boldsymbol{h}_i^+ = f(x_i^+)$, $\boldsymbol{h}_i^- = f(x_i^-)$
        $\boldsymbol{z}_i = g(\boldsymbol{h}_i)$, $\boldsymbol{z}_i^+ = g(\boldsymbol{h}_i^+)$, $\boldsymbol{z}_i^- = g(\boldsymbol{h}_i^-)$
        $\mathrm{sim}(x_i, x_i^+) = \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_i^+}{\|\boldsymbol{z}_i\|\|\boldsymbol{z}_i^+\|}$, $\mathrm{sim}(x_i, x_i^-) = \frac{\boldsymbol{z}_i^\top \boldsymbol{z}_i^-}{\|\boldsymbol{z}_i\|\|\boldsymbol{z}_i^-\|}$
    end for
end for
Define $\mathcal{L} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_i^+)/\tau}}{\sum_{j=1}^{N} (e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^+)/\tau} + e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^-)/\tau})}$
Update $f(\cdot)$ and $g(\cdot)$ to minimize $\mathcal{L}$
Return $f(\cdot)$ and $g(\cdot)$
```

4.4.2. Entity-aware contrastive learning

Contrastive learning aims to obtain proficient representations by grouping similar samples and distinguishing dissimilar samples (Hadsell, Chopra, & LeCun, 2006), which is extensively utilized in many NLP tasks such as question answering and text summarization (Dong, Lu, Wang and Caverlee, 2022; Xu, Zhang, Wu, & Wei, 2022). In this work, we utilize a widely-used contrastive learning model named Supervised SimCSE (Gao, Yao et al., 2021) as the backbone and train it to obtain our Entity-aware Contrastive Learning model. The training procedure is outlined in Algorithm 2.

Specifically, the training data is the collection of positive and negative instances $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{x}_i^+, \boldsymbol{x}_i^-)\}$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^+$ are semantically similar while $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^-$ are semantically different. The primary goal is to enhance the similarity between the sample pair $(\boldsymbol{x}_i,\ \boldsymbol{x}_i^+)$ while minimizing the similarity between the sample pair $(\boldsymbol{x}_i,\ \boldsymbol{x}_i^-)$. The representation of sample pairs is obtained using RoBERTa (Liu et al., 2019). Next, a neural network projection layer is used to further transform the retrieved representations into a standardized embedding space, which is subsequently used to calculate the contrastive loss for a group of positive and negative pairs. Following this, we employ the normalized cross-entropy loss with an adjustable temperature within each mini-batch to optimize the parameters. Specifically, the cosine similarity function is employed to determine the distance between each pair of positive and negative instances, aiming to reduce the distance between $(\boldsymbol{x}_i,\ \boldsymbol{x}_i^+)$ and maximizing the distance between $(\boldsymbol{x}_i,\ \boldsymbol{x}_i^-)$ via the following cross-entropy objective function:

$$\mathcal{L} = -\log \frac{e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_i^+)/\tau}}{\sum_{j=1}^{N} (e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^+)/\tau} + e^{\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^-)/\tau})} \tag{11}$$

where τ is a temperature hyperparameter, N means a mini-batch of N pairs, and $\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^+)$ and $\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j^-)$ are the cosine similarity of positive and negative pairs, respectively.

Algorithm 3 Retrieving Similar Samples with the Entity-aware Contrastive Learning Model

```
# Retrieving Similar Samples with the Entity-aware Contrastive Learning Model
Given a test sample $x^{\text{test}}$, training set $\{x_k\}_{k=1}^{M}$, the Entity-aware Contrastive Learning encoder $f(\cdot)$, the neural network projection layer $g(\cdot)$, the number of in-context examples $n$.
$z^{\text{test}}_{\text{main}} = g(f(x^{\text{test}}_{\text{main}}))$, $z^{\text{test}}_{\text{auxiliary}} = g(f(x^{\text{test}}_{\text{auxiliary}}))$
for all $i \in \{1, \cdots, M\}$ do
    $z_{i,\text{main}} = g(f(x_{i,\text{main}}))$, $z_{i,\text{auxiliary}} = g(f(x_{i,\text{auxiliary}}))$
    sim-test$(x^{\text{test}}, x_i) = \alpha_0 \cdot \text{cosine}(z^{\text{test}}_{\text{main}}, z_{i,\text{main}}) + \alpha_1 \cdot \text{cosine}(z^{\text{test}}_{\text{auxiliary}}, z_{i,\text{auxiliary}})$
end for
Select the top-$n$ samples with the highest similarity score (in descending order)
```

4.4.3. Retrieving similar samples with the entity-aware contrastive learning model

As shown in Fig. 2, to retrieve similar samples from a small-scale training set for a given test sample during the inference process, we apply the well-trained Entity-Aware Contrastive Learning model to obtain the hidden representations of each training sample and the test sample. Note that the training samples are from the training set in Twitter-15 and Twitter-17, as described in Section 5.1. Following this, cosine similarity is computed between the test sample and each training sample in a small-scale training set. Additionally, since the ground-truth label for the test sample is lacked, similarities between two sample pairs are computed only based on the similarity between their main sentences and auxiliary sentences as shown below:

$$\text{sim-test}(a^{\text{test}}, b^{\text{train}}) = \alpha_0 \cdot \text{cosine}(z^{a^{test}}_{\text{main}}, z^{b^{train}}_{\text{main}}) + \alpha_1 \cdot \text{cosine}(z^{a^{test}}_{\text{auxiliary}}, z^{b^{train}}_{\text{auxiliary}}) \tag{12}$$

where a^{test} is the test sample, b^{train} is the training sample, sim-test$(a^{\text{test}}, b^{\text{train}})$ is the similarity between them, $z^{a^{test}}_{\text{main}}$, $z^{b^{train}}_{\text{main}}$, $z^{a^{test}}_{\text{auxiliary}}$ and $z^{b^{train}}_{\text{auxiliary}}$ are the hidden representations derived from the trained Entity-Aware Contrastive Learning model, and cosine$(z^{a^{test}}_{\text{main}}, z^{b^{train}}_{\text{main}})$ and cosine$(z^{a^{test}}_{\text{auxiliary}}, z^{b^{train}}_{\text{auxiliary}})$ are similarity of their main sentences and auxiliary sentences, respectively. α_0 and α_1 are trade-off parameters.

Ultimately, the N samples exhibiting the largest similarity are chosen as the few-shot demonstration used in the in-context learning module. The detailed retrieval process is shown in Algorithm 2.

Furthermore, to obtain deeper comprehension of the N samples retrieved by our model, we show the comparison between the retrieved samples from our model and those from the BERT model (Devlin et al., 2019) on a representative test sample. For fair comparison, both approaches utilize the cosine similarity to rank the pairwise similarity between the main and auxiliary sentences of the training and test samples. As shown in Table 2, we can make the following observations: (1) all the top-3 samples retrieved by our model are semantically close to the query sample in terms of the topic (i.e., soccer), whereas the last two samples retrieved by the BERT model are semantically different from the query sample; (2) The labels obtained by our model are more similar to the labels of the query test sample in terms of category and sentiment.

5. Experiments

In this section, we first provide an explanation of the experimental setting in Section 5.1, and then describe the comparison systems in Section 5.2. Next, we report the experimental results in Section 5.3, which includes the results compared with ICL baseline systems (Section 5.3.1) and the results compared with fine-tuned models on each subtask (Section 5.3.2). Moreover, an in-depth analysis is performed in Section 5.4 to evaluate the merits and limitations of the framework we have developed. This analysis includes examining the impact of the quantity of few-shot samples in Section 5.4.1, conducting an ablation study in Section 5.4.2, presenting a case study in Section 5.4.3, performing the error analysis in Section 5.4.4, and demonstrating the effectiveness of the few-shot samples in Section 5.4.5.

5.1. Experimental settings

Dataset. We follow a recent related study (Yang et al., 2023) by selecting a small subset of training and development sets from the Twitter-15 and Twitter-17 (Ju et al., 2021) as our training and development sets. Specifically, we employ the data split method in Yang, Feng et al. (2023) by selecting three subsets from the training and development datasets of both datasets. The subset selection for Twitter-15 and Twitter-17 is based on three different seeds: [13, 42, 100] and [42, 87, 100], respectively.

In Table 3, we show the average statistics across three data splits For Twitter-15, there are 138 training samples selected from the original 2101 training samples, and another 138 validation samples selected from the original 727 development samples. Similarly, for Twitter-17, 132 samples are chosen from the original 1746 training samples, and another 132 samples are selected from the original 577 development samples. The whole test set are employed as our test sets for both datasets. In the two sampled datasets, the sentiment class with the highest proportion of samples is *Neutral*, while *Negative* has the lowest proportion. There are a total of eight coarse-grained entity categories. Most entities belong to *Person*, *Location*, *Organization*, and *Event* categories, whereas a relatively small amount of entities belong to *Art*, *Building*, *Product*, and *Other* categories.

Table 2
Comparison between the few-shot samples retrieved by our Entity-aware Contrastive Learning model and the BERT model.

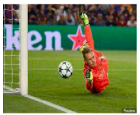
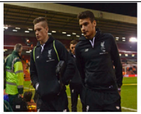
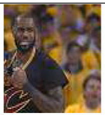
	Few-shot Samples Retrieved by Our Model	Few-shot Samples Retrieved by BERT
Query	Barcelona Crowned 2016 La Liga Champions, Suarez Sink...	
Label	[Barcelona, **Sports Team**, **Positive**], [La Liga, **Sports League**, **Neutral**], [Suarez, **Athlete**, **Negative**]	
Sample 1	 "Ter Stegen saved 56 shots from outside the box in La Liga 16/17 the highest number among La Liga goalkeepers"	Oscar Pareja concedes FC Dallas didn't have any #MLS #MajorLeagueSoccer #bettingtips.
Label	[Ter Stegen, **Athlete**, **Positive**], [La Liga, **Sports League**, **Neutral**], [La Liga, **Sports League**, **Neutral**]	[Oscar Pareja, **Athlete**, **Neutral**], [FC Dallas, **Sports League**, **Negative**], [MLS, **Sports Event**, **Neutral**]
Sample 2	"Sergio Ramos almost got DECAPITATED during Real Madrid ' s victory parade #RMCF #uclfinal."	Harry Styles has taken the most #GOALS photo of all time
Label	[Sergio Ramos, **Athlete**, **Negative**], [Real Madrid, **Sports Team**, **Neutral**], [RMCF, **Sports Team**, **Neutral**]	[Harry Styles, **Musician**, **Positive**]
Sample 3	 premierleague: Jurgen Klopp makes 8 changes to the LFC side that drew with Newcastle	 NBA Finals: Is LeBron James facing his own basketball mortality after falling
Label	[premierleague, **Sports League**, **Neutral**], [Jurgen Klopp, **Athlete**, **Neutral**], [Newcastle, **Sports Team**, **Neutral**]	[NBA, **Sports League**, **Neutral**], [LeBron James, **Athlete**, **Negative**]

Table 3
A statistical analysis of the Twitter-15 and Twitter-17 datasets utilized in our work. Note that **Sample** denotes the quantity of samples in the dataset. **Entities** represent the quantity of entities in MEE and MESC tasks, the entity-sentiment pair in MESPE task, or the entity-category-sentiment triplet in MECSTE task. As the category in MECSTE is comprised of 51 fine-trained categories that are excessively numerous to be displayed in their entirety here, we combine the fine-trained categories and present them on coarse-grained entity categories.

Dataset		Samples	Entities	Sentiments			Coarse-grained entity categories							
				Positive	Neutral	Negative	Person	Location	Organization	Event	PROD	Art	BLD	Other
Twitter-15	Train	**138**/2101	**209**/3173	**62**/928	**116**/1877	**31**/368	**77**/1172	**53**/725	**38**/495	**9**/121	**12**/129	**4**/99	**6**/124	**8**/108
	Dev	**138**/727	**221**/1114	**67**/301	**123**/665	**31**/148	**81**/423	**37**/232	**60**/175	**13**/35	**9**/46	**7**/29	**5**/44	**7**/39
	Test	674	1028	314	601	113	380	299	154	43	48	31	40	33
Twitter-17	Train	**132**/1746	**268**/3561	**89**/1508	**135**/1637	**44**/416	**128**/1663	**26**/342	**78**/1032	**20**/37	**1**/28	**8**/135	**4**/51	**4**/49
	Dev	**132**/577	**265**/1176	**94**/515	**123**/517	**48**/144	**131**/527	**21**/105	**77**/379	**14**/75	**4**/10	**9**/40	**5**/21	**4**/19
	Test	587	1235	494	573	168	545	113	395	104	7	34	19	18

Training Details. To develop the Entity-Aware Contrastive Learning model, we utilizes the Supervised SimCSE model (Gao, Yao et al., 2021) as the backbone model with the *princeton-nlp/sup-simcse-roberta-large* version.[3] The model is trained with three epochs. The optimal parameters for fine-tuning are identified through the implementation of a small-scale grid search. The learning rate is set to 1e−5, and the sample size is set to 8. The sentence representation is [CLS] representation with a multi-layer processing layer superimposed on top of it. The model is executed on an NVIDIA RTX3090 GPU, and all experiments are conducted utilizing PyTorch.

In addition, to train the Entity-Aware Contrastive Learning model, we construct positive and negative instances using a score function and set the following values for the score function's trade-off parameters. The trade-off parameter w_1 in Eq. (6) is configured to 0.5, while the two trade-off parameters w_2 and w_3 in Eq. (7) are both set to 1/3. α_0, α_1, and α_2 in Eq. (9) are assigned as 1/3, whereas α_0 and α_1 in Eq. (12) are assigned as 60% and 40%, respectively. When determining the number of positive and negative instance pairs K as defined in Eq. (10), we also conduct a small-scale grid search over 3, 5, and 7 on the development set, and select 5 as the optimal value of K. Furthermore, the ChatGPT API with its *gpt-3.5-turbo* version (Brown et al., 2020)[4] is used for all the experiments. Note that since each dataset has three different data splits, we report the average performance for each method.

Evaluation Metrics. To make comparison between different approaches, we use Micro-F1 score (F1), Recall (Rec.), and Precision (Prec.) to assess their performance on MEE, MESPE and MECSTE subtasks.

$$Prec. = \frac{\#correct}{\#predict}, \quad Rec. = \frac{\#correct}{\#gold}, \quad F1 = \frac{2 \times Prec. \times Rec.}{Prec. + Rec.}, \tag{13}$$

[3] The model is available from both https://huggingface.co/princeton-nlp/sup-simcse-roberta-large and https://github.com/princeton-nlp/SimCSE.
[4] https://platform.openai.com/docs/guides/gpt.

where $\#correct$ represents the count of predicted outputs that align with the golden output, and $\#gold$ and $\#predict$ represent the counts of golden outputs and predicted outputs, respectively.

$$\text{correct}_{MEE} = \begin{cases} 0 & \text{if } p_{\text{entity}} \neq g_{\text{entity}}, \\ 1 & \text{otherwise} \end{cases}$$

$$\text{correct}_{MESPE} = \begin{cases} 0 & \text{if } (p_{\text{entity}} \neq g_{\text{entity}}) \text{ or } (p_{\text{sentiment}} \neq g_{\text{sentiment}}), \\ 1 & \text{otherwise} \end{cases}$$

$$\text{correct}_{MECSTE} = \begin{cases} 0 & \text{if } (p_{\text{entity}} \neq g_{\text{entity}}) \text{ or } (p_{\text{category}} \neq g_{\text{category}}) \text{ or } (p_{\text{sentiment}} \neq g_{\text{sentiment}}), \\ 1 & \text{otherwise} \end{cases}$$

where $p_{\text{sentiment}}$, p_{category}, and p_{entity} are the predicted sentiment, category and entity, and $g_{\text{sentiment}}$, g_{category}, and g_{entity} are the golden sentiment, category, and entity, respectively. Since the MESC subtask is a classification task, the performance of various methods is assessed using Accuracy.

5.2. Comparison systems

Since this work focuses on four different MEBSA subtasks, we consider three types of approaches for comparison.

First, we examine the subsequent in-context learning approaches, all of which are based upon ChatGPT and adapted to the four subtasks with task-specific instructions. Due to the exploratory nature of this study, there are no existing learning methods within the MEBSA setting that can be used as baseline models. Therefore, we follow the approaches from a related study that works on the text generation task (Liu et al., 2022), and include three ICL methods as the comparative ICL approaches.

- **0-shot learning** and **10-shot learning with Random Selection** are two in-context learning approaches introduced in Section 4.3. The former only shows the task instruction in the prompt, whereas the latter provides 10 samples which picked at random from the training set.
- **10-shot with RoBERTa-large** (Liu et al., 2022) involves utilizing the RoBERTa-large model (Liu et al., 2019) as the encoder to retrieve similar samples. Subsequently, the top 10 similar samples are selected as demonstration examples to be incorporated into the prompt, and then fed into ChatGPT to obtain the final prediction.
- **10-shot with RoBERTa-large-nli-mean-tokens** (Liu et al., 2022) utilizes RoBERTa-large as the backbone model, and it is fine-tuned with two natural language inference datasets: SNLI (Bowman, Angeli, Potts, & Manning, 0000) and MultiNLI (Williams, Nangia, & Bowman, 2017). The fine-tuned model is utilized as the encoder to obtain most similar samples, which are subsequently inputted into ChatGPT to aid in the detection.
- **10-shot with RoBERTa-large-nli-stsb-mean-tokens** is similar with **10-shot with RoBERTa-large-nli-mean-tokens**. It undergoes the first fine-tuning on the SNLI and MultiNLI datasets, followed by a second round of fine-tuning on the STS-B dataset (Cer, Diab, Agirre, Lopez-Gazpio, & Specia, 2017). STS-B dataset comprises sentences pairs and their corresponding similarity ratings. This dataset is widely used for various NLP tasks, including information retrieval, machine translation, and text summarization.
- **10-shot learning with Entity-Aware Exemplar Retriever** is our developed in-context learning method based on Entity-aware Contrastive Learning, as introduced in Section 4.4.

Second, we utilize the following comparison systems for the three extraction subtasks (MEE, MESPE, and MECSTE), among which the first four methods are text-only, whereas the remaining approaches are multimodal:

- **BARTNER** (Yan et al., 2021) generates indexes of the entities together with their respective entity categories and sentiments by using BART (Lewis et al., 2019) as the foundation.
- **SpanABSA** (Hu, Peng, Huang, Li, & Lv, 2019) extracts multiple entities from the sentence based on entity span boundaries, followed by classifying the sentiment polarity using the span representations of those entities.
- **D-GCN** (Chen, Tian, & Song, 2020) encodes syntactic information from multimodal inputs and performs end-to-end identification of entity and sentiment.
- **T5-Paraphrase** (Zhang et al., 2021) is a text-only approach that formulates information extraction tasks as generation problems and generates the tuples via the template-based natural language.
- **UMT** (Yu, Jiang, Yang et al., 2020) is a representative multimodal method for multimodal NER. It employs a stack of cross-modal transformer layers to effectively catch the connections between textual and visual representations.
- **CMMT** (Yang, Na et al., 2022) is a transformer-based cross-modal framework designed for the MESPE task. It comprises two intra-modal modules, which are responsible for processing images and text, and an interaction module that regulates the influence of images on the prediction.
- **MM-BARTNER** (Yang, Wang et al., 2023) is a BART extension in which the encoder is fed with a concatenation of textual inputs and extracted features, followed by the decode to generate the target index sequence from the decoder.
- **JML** (Ju et al., 2021) is an extension of **SpanABSA**, which supports an end-to-end extraction of entities and sentiments for multiple modalities. It incorporates a detection module for alleviating the noise from the visual information.
- **NVLP** (Yang, Feng et al., 2023) is a framework that builds upon a vision-language model VLP (Ling, Yu, & Xia, 2022), which undergoes training with a Twitter corpus. For fair comparison with other methods, **NVLP** excludes the pre-training stage and only trains the BART-based generative model on the datasets for downstream tasks.

Table 4
Comparison between different in-context learning approaches on each MEBSA subtask.

Dataset	Samples	MEE			MESC	MESPE			MECSTE		
		Prec.	Rec.	F1	Accuracy	Prec.	Rec.	F1	Prec.	Rec.	F1
Twitter-15	0-shot	46.16	61.95	52.90	47.18	30.38	36.16	33.02	18.57	28.20	22.39
	10-shot w. Random Selection	62.60	74.45	68.00	56.42	40.29	47.69	43.68	27.18	33.37	29.96
	10-shot w. RoBERTa-large	60.95	73.68	66.71	51.43	41.06	47.31	43.95	28.34	32.49	30.27
	10-shot w. RoBERTa-large-nli-mean-tokens	59.42	75.95	66.67	53.72	41.12	50.22	45.21	28.62	34.40	31.25
	10-shot w. RoBERTa-large-nli-stsb-mean-tokens	58.88	75.59	66.19	55.33	40.65	50.03	44.84	28.43	34.20	31.05
	10-shot w. Entity-Aware Examplar Retriever (Ours)	**62.97**	**77.22**	**69.37**	**62.35**	**42.33**	**49.61**	**45.66**	**28.92**	**34.08**	**31.28**
Twitter-17	0-shot	61.01	61.77	61.39	54.78	36.92	31.76	34.15	28.36	29.98	29.15
	10-shot w. Random Selection	76.20	79.70	77.90	61.70	48.18	48.66	48.41	43.01	44.31	43.65
	10-shot w. RoBERTa-large	73.93	**81.78**	77.66	59.73	48.72	51.84	50.23	42.85	44.93	43.87
	10-shot w. RoBERTa-large-nli-mean-tokens	74.71	80.89	77.68	59.97	46.95	50.87	48.83	42.96	43.97	43.46
	10-shot w. RoBERTa-large-nli-stsb-mean-tokens	74.54	80.16	77.25	60.67	48.45	52.14	50.20	43.03	44.61	43.81
	10-shot w. Entity-Aware Examplar Retriever (Ours)	**75.87**	81.56	**78.61**	**64.44**	**51.38**	**54.06**	**52.69**	**43.97**	**45.15**	**44.56**

- **GMP** (Yang, Feng et al., 2023) is a BART-based multimodal prompt approach for the MESPE task in few-shot setting, comprising a multimodal encoder and an n-stream decoder.
- **MM-Paraphrase** (Yang, Wang et al., 2023) addresses the extraction of entities and sentiments by treating it as a paraphrase generation task, and uses a T5 model to generates the target sentence.

Lastly, for the MESC subtask, we consider several additional representative approaches for comparison, in addition to the variants of the aforementioned methods:

- **LM-BEF** (Gao, Fisch & Chen, 2021) and **GFSC** (Hosseini-Asl, Liu, & Xiong, 2022) are two text-only few-shot learning approaches for different text classification tasks. The former approach designs different dataset-specific prompts and task-specific demonstrations, whereas the latter approach addresses the classification task as a generation problem using GPT-2 (Radford, Narasimhan, Salimans, Sutskever, et al., 2018).
- **TomBERT** (Yu & Jiang, 2019) adapts BERT to model the intra-modal with inter-modal interactions in order to obtain entity-aware multimodal representations for sentiment classification.
- **CapTrBERT** (Khan & Fu, 2021) transforms images into supplementary sentences using image caption, and concatenates the auxiliary sentences with the original sentences for sentiment classification.
- **FITE** (Yang, Zhao & Qin, 2022), an extension of **CapTrBERT**, incorporates the facial information generated from visual inputs alongside the image captions to facilite the prediction.

5.3. Main results

The results generated from our ICL framework is compared to other ICL baseline methods and representative fine-tuned methods in this section. All these fine-tuned models are trained with the dataset explained in Section 5.1.

5.3.1. Comparison with ICL baseline systems

Table 4 illustrates the results of various in-context learning approaches on four subtasks. First, the performance of 0-shot learning is relatively limited on all the tasks. In particular, its performance on multi-element extraction tasks including MESPE and MECSTE is poor, and the F1 scores on these tasks are consistently lower than 0.35. This indicates the challenge of applying LLMs to complex MEBSA subtasks with zero-shot learning setting.

Second, few-shot learning offers ChatGPT with more relevant and valuable samples than zero-shot learning. This enables ChatGPT to enhance its comprehension of various downstream tasks and make more accurate predictions for each task. Therefore, a notable improvement in performance can be observed in the 10-shot learning setting across all the four subtasks. Furthermore, among the three adapted ICL methods, *10-shot with RoBERTa-large-nli-stsb-mean-tokens* achieves the best performance on average across the four subtasks, followed by *10-shot with RoBERTa-large-nli-mean-tokens* and *10-shot with RoBERTa-large*. Moreover, we can observe that the three adapted ICL methods generally achieve better performance than *10-shot with Random Selection* on the multi-element extraction tasks.

Finally, the results of Twitter-15 demonstrate that our developed ICL approach with entity-aware contrastive learning outperforms the 10-shot learning with random selection method by 1.37%, 5.93%, 1.98%, and 1.32% on the MEE, MESC, MESPE, and MECSTE subtasks, respectively. The performance trend is almost the same on Twitter-17. When compared to the three ICL comparison systems, our ICL framework also outperforms them. For example, the ICL framework that we develop demonstrates superior performance compared to RoBERTa-large-nli-stsb-mean-tokens by 1.36%, 3.77%, 2.49%, and 0.75% on the MEE, MESC, MESPE, and MECSTE subtasks of Twitter-17. A similar trend is observed on Twitter-15. These results demonstrate that our retriever can indeed identify samples that are similar to the test sample, and thus helps ChatGPT predict more accurately. The aforementioned observations provide evidence of the efficacy of our developed Entity-Aware Exemplar Retriever.

Table 5
Comparison between different approaches on the MECSTE subtask. The comparison systems are categorized into fine-tuning models and in-context learning approaches. The modality is categorized as a purely textual model and a model capable of handling both image and text inputs.

Model type	Modality	Methods	Twitter-15			Twitter-17		
			Prec.	Rec.	F1	Prec.	Rec.	F1
Fine-tuning	T	BARTNER (Yan et al., 2021)	11.52	14.83	15.02	29.63	26.90	28.20
		T5-Paraphrase (Zhang et al., 2021)	27.26	25.00	26.06	42.07	40.87	41.45
	T+V	MM-BARTNER (Yan et al., 2021)	21.24	18.54	19.80	34.03	31.18	32.52
		UMT (Yu, Jiang, Yang et al., 2020)	20.81	25.14	22.76	33.50	38.00	35.63
		MM-Paraphrase (Yang, Wang et al., 2023)	29.39	26.25	27.71	41.86	40.27	41.04
In-context learning	T+V	10-shot w. Random Selection	27.18	33.37	29.96	43.01	44.31	43.65
		10-shot w. Entity-Aware Examplar Retriever (Ours)	**28.92**	**34.08**	**31.28**	**43.97**	**45.15**	**44.56**

Table 6
Comparison between different approaches on the MESPE subtask. The comparison systems are categorized into fine-tuning models and in-context learning approaches. The modality is categorized as a purely textual model and a model capable of handling both image and text inputs.

Model type	Modality	Methods	Twitter-15			Twitter-17		
			Prec.	Rec.	F1	Prec.	Rec.	F1
Fine-tuning	T	BARTNER (Lewis et al., 2019)	47.03	41.90	44.28	48.59	44.97	46.70
		D-GCN (Chen et al., 2020)	42.03	40.07	40.85	45.66	45.81	44.89
		SpanABSA (Hu et al., 2019)	48.52	39.80	43.71	51.67	48.44	49.98
	T+V	NVLP (Ling et al., 2022)	46.04	42.40	44.14	50.66	45.92	48.16
		JML (Ju et al., 2021)	48.51	41.59	44.77	50.13	48.65	49.38
		GMP (Yang, Feng et al., 2023)	**51.67**	47.19	**49.33**	**54.28**	53.31	**53.79**
In-context learning	T+V	10-shot w. Random Selection	40.29	47.69	43.68	48.18	48.66	48.41
		10-shot w. Entity-Aware Examplar Retriever (Ours)	**42.33**	**49.61**	**45.66**	**51.38**	**54.06**	**52.69**

5.3.2. Comparison with fine-tuned models on each subtask

Results analysis on MECSTE. The results of different approaches to the MECSTE subtask are presented in Table 5. First, it is evident that the 10-shot learning with random selection method outperforms all fine-tuned models on both Twitter-15 and Twitter-17. Moreover, our ICL framework enhances the performance by effectively retrieving similar few-shot samples. It is noteworthy to mention that our framework only uses a restricted amount of samples. One possible rationale for the superior performance of our framework is that due to the complexity of the MECSTE subtask, fine-tuned models necessitates a larger amount of training data to acquire a deeper comprehension of the underlying patterns. In contrast, despite the limited number of demonstration samples, LLMs can understand the complexities of the intricate task due to their extensive internal knowledge and strong reasoning capabilities.

Results analysis on MESPE. Table 6 shows that our ICL framework exhibits superior performance over the majority of fine-tuned models, albeit slightly inferior to the GMP model. Notably, our framework attains the highest performance in the recall metric in comparison to all the fine-tuned models. In order to examine the comparatively higher recall and lower precision of our framework, we conduct an error analysis in Section 5.4.4. The investigation reveals that the ChatGPT-based ICL approach has a tendency to generate more entity-sentiment pairs than golden labels. Approximately 45% of these predictions can be ascribed to the fact that our framework extracts more entities than anticipated, while the remaining 55% can be attributed to ambiguous annotation. Consequently, this leads to a high recall but a low precision.

Results analysis on MEE and MESC. The results of various methods for single element extraction tasks, MESC and MEE, are illustrated in Tables 7 and 8. Our ICL framework generally exhibits lower competitiveness compared to fine-tuned models. We conjecture the reasons as follows: (1) the aforementioned fine-tuned models are fine-tuned on task-specific datasets. In contrast, our ICL framework does not require fine-tuning of downstream tasks; (2) Due to the relative simplicity of single element extraction subtasks, fine-tuned models can effectively capture the underlying patterns of each subtask using a small set of training samples.

Although ICL approaches usually do not attain comparable performance to supervised fine-tuning models (Bhatia, Narayan, De Sa, & Ré, 2023), the advantages of ICL methods over fine-tuned models are evident. In particular, the process of fine-tuning models requires tailoring models to downstream tasks and parameter optimization on task-specific datasets. This fine-tuning process has a negative impact on model's adaptability to unseen tasks. However, our ICL framework does not necessitate retraining the model for each downstream subtask. Instead, a single model can be effectively applied to multiple subtasks simultaneously.

To sum up, our proposed ICL framework has indistinguishable or even better performance than most fine-tuned models in multi-element extraction subtasks such as MESPE and MECSTE. For single element extraction subtasks including MEE and MESC, the performance of our ICL framework is slightly worse than that of some fine-tuned models. However, our ICL framework exhibits advantages in the generalization ability and the reliance on large-scale training samples.

Table 7
Comparison between different approaches on the MESC subtask. The comparison systems can be categorized into fine-tuning models and in-context learning approaches. The modality is categorized as a purely textual model and a model capable of handling both image and text inputs.

Model type	Modality	Methods	Twitter-15 accuracy	Twitter-17 accuracy
Fine-tuning	T	BART (Lewis et al., 2019)	65.57	63.12
		LM-BFF (Gao, Fisch et al., 2021)	64.87	52.08
		GFSC (Hosseini-Asl et al., 2022)	60.75	61.72
	T+V	TomBERT (Yu & Jiang, 2019)	61.78	59.97
		CapTrBERT (Khan & Fu, 2021)	58.76	56.48
		JML (Ju et al., 2021)	60.36	61.62
		FITE (Yang, Yu, Zhang, & Na, 2021)	63.11	60.89
		NVLP (Ling et al., 2022)	63.84	62.72
		GMP (Yang, Feng et al., 2023)	**67.06**	**66.20**
In-context learning	T+V	10-shot w. Random Selection	56.42	61.70
		10-shot w. Entity-Aware Examplar Retriever (Ours)	**62.35**	**64.44**

Table 8
Comparison results on the MEE subtask. The comparison systems can be categorized into fine-tuning models and in-context learning approaches. The modality is categorized as a purely textual model and a model capable of handling both image and text inputs.

Model type	Modality	Models	Twitter-15 F1	Twitter-17 F1
Fine-tuning	T	BARTNER (Lewis et al., 2019)	66.67	70.12
	T+V	NVLP (Ling et al., 2022)	65.95	71.52
		GMP (Yang, Feng et al., 2023)	**73.65**	79.95
		JML (Ju et al., 2021)	71.95	82.14
		CMMT (Yang, Na et al., 2022)	73.19	**82.50**
In-context learning	T+V	10-shot w. Random Selection	68.00	77.90
		10-shot w. Entity-Aware Examplar Retriever (Ours)	**69.37**	**78.61**

5.4. In-depth analysis

5.4.1. The impact of the quantity of few-shot samples

To examine the influence of the quantity of demonstration samples on few-shot learning, we select different number of examples to conduct experiments on all the four subtasks, and subsequently analyze the variations in performance.

Table 9 shows that the performance tends to be subpar when the number of examples is limited, such as 2 or 5. A rise in performance is observed as the number of shots increases, reaching its peak value after 10 shots. However, the performance decreases when the count of shots increases to 15 or 20. It can be inferred that the utilization of demonstration samples can assist the LLM better understand the pattern of datasets and improve model's adaptability to downstream tasks. Nevertheless, an excessive quantity of samples introduces noise, hence diminishing the precision of prediction. Our datasets for selecting similar samples are small, consisting of 138 samples for Twitter-15 and 136 samples for Twitter-17. As the number of shots increase, more irrelevant samples are selected from a small sample pool. Therefore, 10 shots are sufficient for our ICL framework to attain competitive performance, which significantly reduces the computation resources and data annotation effort.

5.4.2. Ablation study

To assess the efficiency of different components within the ICL framework, ablation experiments are performed on the MEPSE subtask by deleting each component as follows:

- **Without Image Captioning**: the image caption description explained in Section 4.2.1 is removed.
- **Without Entity Detection with VQA**: the main entities detected by BLIP-VQA explained in Section 4.2.1 is removed.
- **Without Sentiment Detection with VQA**: the subjective words indicating the sentiment detected by BLIP-VQA described in Section 4.2.1 is removed.
- **Without training Entity-aware Contrastive Learning model**: extracting few-shot samples directly using the backbone model SimCSE model, without training it with the constructed positive and negative instances described in Section 4.4.

Table 10 demonstrates that there is a discernible decrease in performance upon the removal of Image Captioning. This finding illustrates that the inclusion of image captions provides valuable insights into entities and categories, hence enhancing the performance of MEBSA tasks. Furthermore, excluding sentiment information identified by BLIP-VQA leads to a drop in F1 scores for both datasets. The results indicate that subjective terms generated by BLIP-VQA effectively capture the subjective information from images, hence assisting in the identification of sentiment. In addition, removing entity information from the image results in a fall in F1 scores for both datasets. The detection of these primary entities from images contributes directly to the prediction of entities, thereby enhancing the overall performance. Moreover, the results indicate that removing our designed score function and the constructed positive–negative instances led to a decrease in performance. It demonstrates that the score function that we have

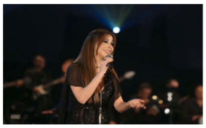
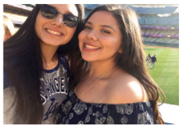
Table 9
Impact of the number of demonstration samples in our proposed method. The experiments encompass all four subtasks of MEBSA, namely MEE, MESC, MESPE, and MECSTC. The experimental shots are 2, 5, 10, 15, and 20.

Dataset	Samples	MEE			MESC	MESPE			MECSTE		
		Prec.	Rec.	F1	Accuracy	Prec.	Rec.	F1	Prec.	Rec.	F1
Twitter-15	2-shot	54.60	76.60	63.76	51.02	34.11	45.15	38.83	23.30	30.29	26.34
	5-shot	62.23	76.31	68.56	51.85	40.77	48.48	44.25	24.45	30.49	27.14
	10-shot	**62.97**	**77.22**	**69.37**	**62.35**	**42.33**	**49.61**	**45.66**	**28.92**	**34.08**	**31.28**
	15-shot	62.81	75.44	68.55	55.23	41.78	49.19	45.17	25.39	33.49	28.89
	20-shot	63.24	76.02	69.05	57.88	40.77	40.77	40.77	24.55	30.39	27.16
Twitter-17	2-shot	75.67	63.73	69.19	61.16	40.56	46.48	43.32	39.04	41.53	40.25
	5-shot	72.73	81.35	76.80	59.40	46.46	50.23	48.27	38.7	41.46	40.03
	10-shot	**75.87**	**81.56**	**78.61**	**64.44**	**51.38**	**54.06**	**52.69**	**43.97**	**45.15**	**44.89**
	15-shot	76.35	80.38	78.32	61.48	48.46	50.47	49.44	40.12	42.91	41.47
	20-shot	74.62	78.86	76.68	61.51	49.93	51.07	50.49	38.67	41.86	40.20

Table 10
Ablation study results on the MESPE subtask. There are a total of four ablation experiments presented. To assess the efficacy of Visual Input Construction module, the following components are eliminated: image captioning, entity detection with VQA, and sentiment detection with VQA. Entity-aware Contrastive Learning module is also excluded to evaluate its effectiveness.

Ablation models	Twitter-15			Twitter-17		
	Prec.	Rec.	F1	Prec.	Rec.	F1
Our method	**42.33**	**49.61**	**45.66**	**51.38**	**54.06**	**52.69**
- w/o Image Captioning	41.28	50.55	45.41	47.53	53.20	50.20
- w/o Visual Entity Detection	38.56	46.80	42.23	48.52	52.58	50.47
- w/o Visual Sentiment Detection	38.30	43.92	40.88	47.80	50.66	49.19
- w/o Entity-aware Contrastive Learning module	39.88	50.49	44.56	47.85	50.36	49.07

Table 11
Inaccurate predictions by compared systems in contrast to accurate predictions by our ICL framework. The image caption, VQA sentiment and VQA entity that are detected by the visual inputs construction module are shown. The comparison systems consist of three competitive fine-tuned models, including MM-Paraphrase, UMT-RoBERTa, and MM-BARTNER.

Text	 a. "Nancy Ajram during the Beirut Cultural Festival; beautiful as always."	 b. "Some of that Dodger baseball @ alyssajacinto"
Image Caption	a woman in a black dress and a microphone	two girls in the stands at the stadium
VQA Sentiment	Positive	Positive
VQA Entity	sound	friends
Human Label	(Nancy Ajram, Musician, Positive) (Beirut Cultural Festival, Festival, Neutral)	(Dodger, Sports Team, Positive) (alyssajacinto, Common Person, Positive)
MM-Paraphrase	(Nancy Ajram, Musician, Positive) ✓ (Beirut Cultural Festival, Festival, Positive) ×	(Dodger, Sports Team, Positive) ✓ (alyssajacinto, Athlete, Positive) ×
UMT-RoBERTa	(Nancy, Athlete, Neutral) × (Ajram, Athlete, Neutral) × (Beirut, City, Neutral) × (Cultural Festival, Festival, Positive) ×	(Dodger, Sports Team, Neutral) × (alyssajacinto, Sports Team, Neutral) ×
MM-BARTNER	(Nancy Ajram, Musician, Neutral) × (Beirut Cultural Festival, Festival, Positive) ×	(Dodger, Sports Team, Positive) ✓ (alyssajacinto, Athlete, Neutral) ×
Our Model	(Nancy Ajram, Musician, Positive) ✓ (Beirut Cultural Festival, Festival, Neutral) ✓	(Dodger, Sports Team, Positive) ✓ (alyssajacinto, Common Person, Positive) ✓

developed efficiently computes semantic similarity among samples, thereby facilitating the construction of positive and negative instances. Based on this, we train an efficient Entity-aware Contrastive learning module to acquire useful few-shot samples. The aforementioned findings collectively demonstrate the effectiveness of the three modules.

5.4.3. *Case study*

A comparative analysis is conducted on the results of our ICL framework in comparison to three baseline approaches: MM-BARTNER, MM-Paraphrase, and UMT-RoBERTa. Table 11 presents two test examples to illustrate the advantages of our ICL framework. Table 11.a shows that all these three baseline models show deficiencies in their predictions of sentiment polarities. Both MM-BARTNER and UMT-RoBERTa make an inaccurate prediction by categorizing the sentiment towards *Nancy Ajram* as *Neutral*.

Table 12
Few-shot sample demonstration similarity scores for MESPE and MECSTE subtasks on Twitter-15 and Twitter-17.

Task	Element	Twitter-15			Twitter-17		
		Match count	Few-shot sample count	Similarity score	Match count	Few-shot sample count	Similarity score
MESPE	Sentiment	7883	16 590	47.12%	12 573	28 683	43.70%
MECSTE	Category	1998	16 890	11.86%	6676	27 363	24.75%
	Sentiment	7318	16 890	43.03%	11 961	27 363	43.61%

However, our ICL framework efficiently utilizes the information from Image Captioning and VQA, and makes correct predictions. Additionally, UMT-RoBERTa fails to correctly extract *Nancy Ajram* as an entity and erroneously categorized her as *Athlete*. However, our ICL framework effectively extracts the entity and classifies the entity as *Musician*. Moreover, as illustrated in Table 11.b, none of MM-BARTNER, MM-Paraphrase, and UMT-RoBERTa accurately predict the category for the entity *alyssajcino*. By contrast, our ICL framework accurately classifies the entity as *Common Person* by utilizing the information obtained from Image Captioning and VQA.

5.4.4. Error analysis

An error analysis is conducted to explore incorrect predictions made by our ICL framework in this subsection. A total of one hundred errors samples were manually evaluated for the MESPE task, resulting in the following observations: in 30% of the error instances, our framework generated more predictions compared to the golden labels; in 25% of the error cases, our framework inaccurately predicted the sentiment polarities towards entities. For another 25% of the error cases, our framework generating more words than the golden reference when handling entities composed of multiple words, leading to a lack of precise matches. The remaining 10% of errors were primarily attributed to a lower number of predicted entity-sentiment pairs compared to the labels.

In a majority (55%) of instances, where errors occurred in predicting more entity-sentiment pairs than the golden reference, it was observed that there were missing annotations. In the instance "London enjoys eating puffs while watching @OSUBaseball! #OrangeFriday #GoPokes", our ICL framework made predictions of (London, *Neutral*) and (OSUBaseball, *Positive*) as outputs, which did not align with the label of (OSUBaseball, *Positive*). However, it can be observed that (London, *Neutral*) was not labeled. In the remaining 45% of cases, our ICL framework identified extraneous entity-sentiment pairs that should be excluded. As an illustration, within the given instance "Barack Obama is extremely concerned about the impact that carbon will have on our planet", our ICL framework predicted (Barack Obama, *Negative*), (carbon, *Negative*) and (planet, *Positive*) as the outputs. However, it is inappropriate to include (carbon, *Negative*) and (planet, *Positive*).

In cases when entities were composed of multiple words, the error typically arose from the inclusion or exclusion of words throughout the entirety of the entity. In the instance of "Harry Potter and The Order of the Phoenix", the predictions generated were (Harry Potter, *Positive*) and (The Order of the Phoenix, *Positive*), while the label was (Harry Potter and The Order of the Phoenix, *Neutral*).

Finally, we observed frequent occurrences of sentiment prediction errors when sentences contained different sentiment polarities. In the given instance of "I loved seeing @VerticalHorizon on Saturday at @CelebrateFFX. Brilliant as expected. #music #summer #concert", the anticipated outputs were (VerticalHorizon, *Positive*) and (CelebrateFFX, *Neutral*), while predictions were ((VerticalHorizon, *Positive*) and (CelebrateFFX, *Positive*). The presence of different sentiments towards different entities inside a single sentence posed a challenge for accurate predictions.

5.4.5. Few-shot demonstration

To enhance comprehension regarding the effectiveness of our ICL framework for sample selection, as expounded upon in Section 4.4, we provide several illustrative instances in Table 13. While our framework only computes the sentence similarity between the test sample and samples in training set, it is apparent that the selected demonstration samples indeed exhibit similarities to the test samples with regard to their labels, particularly entity categories and sentiments. This finding demonstrates that our ICL framework can effectively identify samples with similar labels with the test sample, thereby improving the effectiveness of few-shot learning.

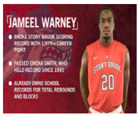
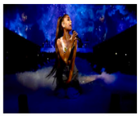
For example, the sample of Table 13.a is related to sports and athletes. The anticipated entities for extraction pertain to *Athlete*, *Sports Teams* and *Sports Leagues*, and the sentiment polarities are mainly *Positive* and *Neutral*. From the few-shot samples, it can be observed that all of these samples pertain to sports. The entities mentioned in these samples are primarily related to *Athlete*, *Sports Teams* and *Sports Leagues*, with sentiments predominantly being *Positive* and *Neutral*. Furthermore, in the sample of Table 13.b, the entities are related to *Music* and *Festival*, with a *Positive* sentiment. The labels of selected few-shot samples are also related to *Music* and *Festival*, with most sentiments being *Positive*. Therefore, our ICL framework can select samples that closely resemble the test sample, which effectively assists the model in generating precise predictions.

Furthermore, we conduct a quantitative analysis to assess label similarity between few-shot samples and the test sample. The similarity score of each entity category is defined as follows:

$$\mathbf{SimilarityScore}_{\text{sentiment/category}} = \mathbf{C}_{\text{match}} / \mathbf{C}_{\text{few-shot sample}} \tag{14}$$

where $\mathbf{C}_{\text{match}}$ is the count of sentiment or category labels that match the test sample in all the few-shot samples and $\mathbf{C}_{\text{few-shot sample}}$ represents the count of sentiment or category labels in few-shot samples.

Table 13
The demonstration of few-shot samples that our ICL framework selects for the MECSTE subtask on Twitter-17.

Sample		"Excited to talk w / @ Fullcoursemeelz later tonight about his career at Stony Brook and his future in the NBA!"		"pitchfork : Tune into ArianaGrande's 'One Love Manchester' benefit concert live stream, which has begun..."
Label	[Fullcoursemeelz, **Athlete**, **Positive**], [Stony Brook, **Sports Team**, **Neutral**], [NBA, **Sports League**, **Neutral**]		[ArianaGrande, **Musician**, **Positive**], [One Love Manchester, **Festival**, **Positive**]	
Few-shot Sample 1		"MLB Alshon Jeffery sings 'Take Me Out To The Ballgame' at #Cubs game"		"Justin Bieber and David Guetta's new collaboration #2 U drops Friday."
Label	[MLB, **Sports League**, **Neutral**], [Alshon Jeffery, **Athlete**, **Positive**]		[Justin Bieber, **Musician**, Neutral], [David Guetta, **Musician**, Neutral]	
Few-shot Sample 2		"Join us in wishing wilsonchandler of the nuggets a HAPPY 29th BIRTHDAY! #NBABDAY."		"Justin Bieber, Rich the Kid and Diplo have a banger on the way"
Label	[wilsonchandler, **Athlete**, **Positive**], [nuggets, **Sports Team**, **Neutral**]		[Justin Bieber, **Musician**, Neutral], [David Guetta, **Musician**, Neutral]	
Few-shot Sample 3		"Moorhead announces Bormann as new boys basketball coach #Spuds @ inforum."		"Katy Perry and Lady Gaga appreciation tweet"
Label	[Moorhead, **Sports Team**, **Positive**], [Spuds, **Sports Team**, **Positive**], [Bormann, Coach, **Neutral**]		[Katy Perry, **Musician**, **Positive**], [Lady Gaga, **Musician**, **Positive**]	
Few-shot Sample 4		"Paul Pierce has high confidence in Kevin Durant. (via @ ESPNNBA)"		"Taylor with her "Taylor Swift Award" she is such a legend... your fav could never"
Label	[Paul Pierce, **Athlete**, **Positive**], [Kevin Durant, **Athlete**, **Positive**], [ESPNNBA, News Agency, **Neutral**]		[Taylor, **Musician**, **Positive**], [Taylor Swift Award, Award, Neutral]	
Few-shot Sample 5		"Congrats to UTSA's first-ever NFL draft pick! #WeAreUTSA-."		"Buy Your Beyonce 2016 Formation World Tour Concert Tickets Here"
Label	[UTSA, **Sports Team**, **Positive**], [NFL, **Sports League**, **Positive**]		[Beyonce, **Musician**, Neutral]	

As indicated in Table 12, the sentiment similarity between few-shot samples and test samples is relatively high. Compared to sentiment, the level of similarity in the category is comparatively lower. The reason is that there are 52 fine-grained categories, whereas there are only three sentiment polarities. However, the alignment of sentiment and category labels between the few-shot samples and the test sample improve the effectiveness of few-shot learning and facilitate the model's predictions.

6. Conclusion

This work explores the potential of ICL with ChatGPT for Multimodal Entity-Based Sentiment Analysis (MEBSA). Specifically, we design a set of task-specific instructions to perform zero-shot learning and few-shot learning on four MEBSA subtasks. To improve the performance of ICL, we develop an entity-aware contrastive learning model, which can retrieve demonstration samples that are similar to each test sample. Extensive experiments show that our developed ICL framework outperforms other baseline ICL approaches, and is comparable to or even surpasses that of many existing fine-tuned approaches. Furthermore, the in-depth analysis demonstrates the efficiency of different modules of our developed ICL framework.

Although the framework that we have developed has been tested on Twitter-15 and Twitter-17, it can also be applied to analyze multimodal posts from various social media sites and multimodal product reviews from E-commerce platforms. Our framework allows users to efficiently analyze numerous opinion aspects, such as entities, categories, and feelings, due to its promising performance in all four MEBSA subtasks. Moreover, compared to fine-tuned models, the framework that we have developed achieves competitive performance while utilizing a significantly smaller sample size. This reduction in sample size greatly reduces the computational resources required and also diminishes the need for dataset labeling effort. Hence, with this framework, large-scale multimodal data can be analyzed both efficiently and effectively, facilitating a comprehensive comprehension of trending topics and user opinions across various platforms.

This study is subject to some limitations as it is the first exploration of ICL for MEBSA. First, the ICL framework exhibits a relatively limited capability for single element extraction tasks when compared to fine-tuned methods. Second, we utilize ChatGPT, a representative LLM, as the backbone model in our developed framework. However, ChatGPT model has certain limitations, and in recent months there has been a significant surge in the development of open-source LLMs and Multimodal Large Language Models (MLLMs). We hope this work can inspire more researchers to investigate the potential of LLMs or MLLMs for MEBSA.

CRediT authorship contribution statement

Li Yang: Writing – original draft, Visualization, Software, Methodology, Investigation, Conceptualization. **Zengzhi Wang:** Methodology, Investigation. **Ziyan Li:** Visualization, Software. **Jin-Cheon Na:** Writing – review & editing, Supervision. **Jianfei Yu:** Writing – review & editing, Methodology, Funding acquisition.

Data availability

Data will be made available on request.

References

Barnes, J., Kurtz, R., Oepen, S., Øvrelid, L., & Velldal, E. (2021). Structured sentiment analysis as dependency graph parsing. In *Proceedings of ACL-IJCNLP* (pp. 3387–3402).

Barnes, J., Oberlaender, L., Troiano, E., Kutuzov, A., Buchmann, J., Agerri, R., et al. (2022). SemEval 2022 task 10: Structured sentiment analysis. In *Proceedings of the 16th international workshop on semantic evaluation* (pp. 1280–1295).

Bhatia, K., Narayan, A., De Sa, C., & Ré, C. (2023). TART: A plug-and-play Transformer module for task-agnostic reasoning. arXiv preprint arXiv:2306.07536.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. A large annotated corpus for learning natural language inference.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., et al. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in neural information processing systems 33: annual conference on neural information processing systems 2020, neurIPS 2020, December 6-12, 2020, virtual*.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. arXiv preprint arXiv:1708.00055.

Chen, G., Tian, Y., & Song, Y. (2020). Joint aspect extraction and sentiment analysis with directional graph convolutional networks. In *Proceedings of COLING* (pp. 272–279).

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., et al. (2022). Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., et al. (2022). PaLM: Scaling language modeling with pathways. CoRR abs/2204.02311. arXiv:2204.02311.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL* (pp. 4171–4186).

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., et al. (2022). A survey for in-context learning. arXiv preprint arXiv:2301.00234.

Dong, X., Lu, J., Wang, J., & Caverlee, J. (2022). Closed-book question generation via contrastive learning. arXiv preprint arXiv:2210.06781.

Gao, T., Fisch, A., & Chen, D. (2021). Making pre-trained language models better few-shot learners. In *Joint conference of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing, ACL-IJCNLP 2021* (pp. 3816–3830). Association for Computational Linguistics (ACL).

Gao, T., Yao, X., & Chen, D. (2021). Simcse: Simple contrastive learning of sentence embeddings. arXiv preprint arXiv:2104.08821.

Gonen, H., Iyer, S., Blevins, T., Smith, N. A., & Zettlemoyer, L. (2022). Demystifying prompts in language models via perplexity estimation. CoRR abs/2212.04037. arXiv:2212.04037.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition* (pp. 1735–1742).

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., et al. (2022). Training compute-optimal large language models. CoRR abs/2203.15556. arXiv:2203.15556.

Hosseini-Asl, E., Liu, W., & Xiong, C. (2022). A generative language model for few-shot aspect-based sentiment analysis. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 770–787).

Hu, M., Peng, Y., Huang, Z., Li, D., & Lv, Y. (2019). Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 537–546).

Ju, X., Zhang, D., Xiao, R., Li, J., Li, S., Zhang, M., et al. (2021). Joint multi-modal aspect-sentiment analysis with auxiliary cross-modal relation detection. In *Proceedings of EMNLP* (pp. 4395–4405).

Katiyar, A., & Cardie, C. (2018). Nested named entity recognition revisited. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies*.

Khan, Z., & Fu, Y. (2021). Exploiting BERT for multimodal target sentiment classification through input space translation. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 3034–3042).

Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications*, *78*(17), 24103–24119.

Kumar, A., Srinivasan, K., Cheng, W.-H., & Zomaya, A. Y. (2020). Hybrid context enriched deep learning model for fine-grained sentiment analysis in textual and visual semiotic modality social data. *Information Processing & Management*, *57*(1), Article 102141.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461.

Li, J., Li, H., Pan, Z., & Pan, G. (2023). Prompt ChatGPT In MNER: Improved multimodal named entity recognition method based on auxiliary refining knowledge from ChatGPT. arXiv preprint arXiv:2305.12212.

Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888–12900). PMLR.

Li, X., Lv, K., Yan, H., Lin, T., Zhu, W., Ni, Y., et al. (2023). Unified demonstration retriever for in-context learning. In A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 4644–4668).

Li, J., Sun, A., Han, J., & Li, C. (2022). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge & Data Engineering*, *34*(01), 50–70.

Li, C., Sun, A., Weng, J., & He, Q. (2014). Tweet segmentation and its application to named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, *27*(2), 558–570.

Ling, Y., Yu, J., & Xia, R. (2022). Vision-language pre-training for multimodal aspect-based sentiment analysis. In *Proceedings of ACL* (pp. 2149–2159).

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, *5*(1), 1–167.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Liu, J., Shen, D., Zhang, Y., Dolan, W. B., Carin, L., & Chen, W. (2022). What makes good in-context examples for GPT-3? In *Proceedings of deep learning inside out (deeLIO 2022): the 3rd workshop on knowledge extraction and integration for deep learning architectures* (pp. 100–114).

Lu, Y., Bartolo, M., Moore, A., Riedel, S., & Stenetorp, P. (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)* (pp. 8086–8098).

Lu, D., Neves, L., Carvalho, V., Zhang, N., & Ji, H. (2018). Visual attention model for name tagging in multimodal social media. In *Proceedings of ACL* (pp. 1990–1999).

Meng, Y., Xiong, C., Bajaj, P., Bennett, P., Han, J., Song, X., et al. (2021). Coco-lm: Correcting and contrasting text sequences for language model pretraining. *Advances in Neural Information Processing Systems*, *34*, 23102–23114.

Meškelė, D., & Frasincar, F. (2020). ALDONAr: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, *57*(3), Article 102211.

OpenAI (2023a). ChatGPT. https://openai.com/blog/chatgpt.

OpenAI (2023b). GPT-4 technical report. CoRR abs/2303.08774. arXiv:2303.08774.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., AL-Smadi, M., et al. (2016). SemEval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th international workshop on semantic evaluation* (pp. 19–30).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Rubin, O., Herzig, J., & Berant, J. (2022). Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies* (pp. 2655–2671).

Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830.

Shang, F., & Ran, C. (2022). An entity recognition model based on deep learning fusion of text feature. *Information Processing & Management*, *59*(2), Article 102841.

Su, Y., Lan, T., Li, H., Xu, J., Wang, Y., & Cai, D. (2023). PandaGPT: One model to instruction-follow them all. CoRR abs/2305.16355. arXiv:2305.16355.

Truong, Q.-T., & Lauw, H. W. (2017). Visual sentiment analysis for review images with item-oriented and user-oriented CNN. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 1274–1282).

Truong, Q.-T., & Lauw, H. W. (2019). Vistanet: Visual aspect attention network for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 305–312).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015* (pp. 3156–3164). IEEE Computer Society, http://dx.doi.org/10.1109/CVPR.2015.7298935.

Wan, Z., Cheng, F., Mao, Z., Liu, Q., Song, H., Li, J., et al. (2023). GPT-RE: In-context learning for relation extraction using large language models. arXiv e-prints, arXiv–2305.

Wang, Y., Mukherjee, S., Liu, X., Gao, J., Awadallah, A., & Gao, J. (2022). LiST: Lite prompted self-training makes parameter-efficient few-shot learners. In *Findings of the association for computational linguistics: NAACL 2022* (pp. 2262–2281).

Wang, Z., Xie, Q., Ding, Z., Feng, Y., & Xia, R. (2023). Is ChatGPT a good sentiment analyzer? A preliminary study. CoRR abs/2304.04339. arXiv:2304.04339.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv:1704.05426.

Wu, H., Cheng, S., Wang, J., Li, S., & Chi, L. (2020). Multimodal aspect extraction with region-aware alignment network. In *Proceedings of NLPCC* (pp. 145–156).

Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., & Ma, H. (2020). Clear: Contrastive learning for sentence representation. arXiv preprint arXiv:2012.15466.

Wu, S., Xu, Y., Wu, F., Yuan, Z., Huang, Y., & Li, X. (2019). Aspect-based sentiment analysis via fusing multiple sources of textual knowledge. *Knowledge-Based Systems*, *183*, Article 104868.

Xiao, L., Wu, X., Yang, S., Xu, J., Zhou, J., & He, L. (2023). Cross-modal fine-grained alignment and fusion network for multimodal aspect-based sentiment analysis. *Information Processing & Management*, *60*(6), Article 103508.

Xu, N., Mao, W., & Chen, G. (2018). A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval* (pp. 929–932).

Xu, N., Mao, W., & Chen, G. (2019). Multi-interactive memory network for aspect based multimodal sentiment analysis. In *Proceedings of AAAI* (pp. 371–378).

Xu, S., Zhang, X., Wu, Y., & Wei, F. (2022). Sequence level contrastive learning for text summarization. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 11556–11565).

Yan, H., Gui, T., Dai, J., Guo, Q., Zhang, Z., & Qiu, X. (2021). A unified generative framework for various NER subtasks. In *Proceedings of ACL-IJCNLP* (pp. 5808–5822).

Yang, X., Feng, S., Wang, D., Sun, Q., Wu, W., Zhang, Y., et al. (2023). Few-shot joint multimodal aspect-sentiment analysis based on generative multimodal prompt. In *Findings of the association for computational linguistics: ACL 2023* (pp. 11575–11589).

Yang, L., Na, J.-C., & Yu, J. (2022). Cross-modal multitask transformer for end-to-end multimodal aspect-based sentiment analysis. *Information Processing & Management*, *59*(5), Article 103038.

Yang, L., Wang, J., Na, J.-C., & Yu, J. (2023). Generating paraphrase sentences for multimodal entity-category-sentiment triple extraction. *Knowledge-Based Systems*, *278*, Article 110823.

Yang, L., Yu, J., Zhang, C., & Na, J.-C. (2021). Fine-grained sentiment analysis of political tweets with entity-aware multimodal network. In *International conference on information* (pp. 411–420).

Yang, H., Zhao, Y., & Qin, B. (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* (pp. 3324–3335).

Ye, J., Wu, Z., Feng, J., Yu, T., & Kong, L. (2023). Compositional exemplars for in-context learning. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of machine learning research*: *vol. 202*, *International conference on machine learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA* (pp. 39818–39833). PMLR, URL: https://proceedings.mlr.press/v202/ye23c.html.

Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., et al. (2023). mPLUG-owl: Modularization empowers large language models with multimodality. CoRR abs/2304.14178. arXiv:2304.14178.

Yu, J., Chen, K., & Xia, R. (2023). Hierarchical interactive multimodal transformer for aspect-based multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, *14*(03), 1966–1978.

Yu, J., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. In *Proceedings of IJCAI* (pp. 5408–5414).

Yu, J., Jiang, J., & Xia, R. (2020). Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 429–439.

Yu, J., Jiang, J., Yang, L., & Xia, R. (2020). Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of ACL* (pp. 3342–3352).

Yu, J., Wang, J., Xia, R., & Li, J. (2022). Targeted multimodal sentiment classification based on coarse-to-fine grained image-target matching. In *Proceedings of IJCAI* (pp. 4482–4488).

Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L.-P. (2017). Tensor fusion network for multimodal sentiment analysis. In *Proceedings of EMNLP* (pp. 1103–1114).

Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of EMNLP* (pp. 9209–9219).

Zhang, Q., Fu, J., Liu, X., & Huang, X. (2018). Adaptive co-attention network for named entity recognition in tweets. In *Thirty-second AAAI conference on artificial intelligence* (pp. 5674–5681).

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2022). A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge & Data Engineering*, (01), 1–20.

Zhang, D., Zhang, W., Li, S., Zhu, Q., & Zhou, G. (2020). Modeling both intra-and inter-modal influence for real-time emotion detection in conversations. In *Proceedings of ACM MM* (pp. 503–511).

Zhang, M., Zhang, Y., & Vo, D.-T. (2015). Neural networks for open domain targeted sentiment. In *Proceedings of EMNLP* (pp. 612–621).

Zhang, M., Zhang, Y., & Vo, D.-T. (2016). Gated neural networks for targeted sentiment analysis. In *Proceedings of AAAI* (pp. 3087–3093).

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021a). Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning* (pp. 12697–12706). PMLR.

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021b). Calibrate before use: Improving few-shot performance of language models. In M. Meila, & T. Zhang (Eds.), *Proceedings of the 38th international conference on machine learning, ICML 2021, 18-24 July 2021, virtual event* (pp. 12697–12706).

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. arXiv preprint arXiv:2303.18223.

Zhou, J., Zhao, J., Huang, J. X., Hu, Q. V., & He, L. (2021). MASAD: A large-scale dataset for multimodal aspect-based sentiment analysis. *Neurocomputing*, *455*, 47–58.