Retrieval-Based Multimodal Data Augmentation for Multimodal Information Extraction in Social Media

Shizhou Huang[1], Bo Xu[2], Yang Yu[1], Changqun Li[1], and Xin Lin[1,3](✉)

[1] School of Computer Science and Technology, East China Normal University,
Shanghai, China
huangshizhou@ica.stc.sh.cn, {52205901014, 52215901009}@stu.ecnu.edu.cn

[2] School of Computer Science and Technology, Donghua University, Shanghai, China
xubo@dhu.edu.cn

[3] Shanghai Key Laboratory of Multidimensional Information Processing
xlin@cs.ecnu.edu.cn

Abstract. Recently, multimodal information extraction (MIE) has attracted increasing attention in social media understanding. The data augmentation methods can effectively address the unique challenges of information extraction on social media, such as data sparsity and insufficient semantics. However, existing data-augmented methods have two weaknesses: (1) existing methods are based on predefined rules or generative models, resulting in the generation of synthetic data that has limited diversity and differs from real-world data; (2) current approaches predominantly focus on text augmentation, overlooking the potential benefits of augmenting image data. To address these issues, we propose a retrieval-based multimodal data augmentation (RMDA) approach by leveraging the social media domain's massive data volumes and high retrievability, which obtains real-world multimodal posts related to the original data as augmented examples through retrieval. We have conducted extensive experiments to demonstrate the effectiveness of our method and demonstrate that it offers significant advantages in both efficiency and performance compared to augmentation methods based on large language models.

Keywords: Multimodal information extraction · Data augmentation · Social media.

1 Introduction

The wealth of user-generated posts on social media carries potential information insights, including opinions and preferences of groups and individuals, and could be used for a variety of purposes to serve humanity more effectively [16, 17]. To achieve this, information extraction serves as an effective approach and a key step in extracting these hidden insights in media posts [17, 27].

Recently, multimodal information extraction (MIE) on social media, such as multimodal named entity recognition (MNER) [15, 16, 29, 31–33], multimodal

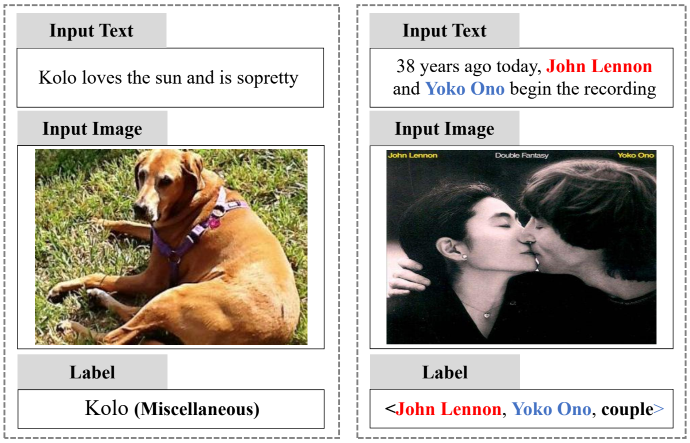

Input Text
Kolo loves the sun and is sopretty
Input Image
Label
Kolo (Miscellaneous)
Input Text
38 years ago today, John Lennon and Yoko Ono begin the recording
Input Image
John Lennon
Double Fantasy
Yoko Ono
Label
<John Lennon, Yoko Ono, couple>

Fig. 1: Examples of MNER (Left) and MRE (Right). Colored texts represent entities for which relations need to be judged, and these entities are given in MRE.

relation extraction (MRE) [3,13,23,25,27,28], has attracted increasing attention. It significantly enhances the traditional text-based methods by incorporating images as additional inputs to obtain more semantic information [3].

As shown in Figure 1, with only text, it is difficult to determine the entity type of `Kolo` (`person` or `miscellaneous`) and similarly challenging to identify the relation between `John Lennon` and `Yoko Ono`. However, with the assistance of their accompanying images, we can easily identify `Kolo` as `miscellaneous` and the relation between `John Lennon` and `Yoko Ono` as `couple`.

In social media, information extraction faces some unique challenges such as data sparsity [19], short text leading to insufficient semantics [16], not following strict syntactic rules [19], and temporal drift by language changes rapidly [2]. To address these issues, some approaches employ data augmentation to increase the size and diversity of the training set and make the model more adaptable to social media data, thereby improving their generalization. For instance, [2] proposes a neural architecture to transform the data representation from the news domain to the social media domain by learning the patterns in the text. [1] utilize the powerful rewriting capabilities of large language models (LLMs) to generate data similar to the original for data augmentation.

Despite their success, existing data augmentation methods face two key limitations: (1) existing methods are based on predefined rules or generative models, resulting in the generation of synthetic data that has limited diversity and differs from real-world data, thereby limiting the generalization ability of the models; (2) current approaches predominantly focus on text augmentation, overlooking the potential benefits of augmenting image data. In contrast, multimodal data augmentation, which combines both text and image augmentation, can generate diverse text-image pairs. This approach enhances the model's ability to understand multimodal content and better capture relationships between different modalities.

However, we note that the social media domain does not only pose challenges for information extraction but also has unique advantages: massive data volumes and high retrievability. Therefore, to address these issues, we propose a retrieval-based multimodal data augmentation (RMDA) approach, which obtains real-world multimodal posts related to the original data as augmented examples through retrieval, ensuring that the augmented examples are consistent with the real-world distribution and is able to perform multimodal data augmentation of both text and images.

Our main contributions can be summarized as follows:

- Firstly, we introduce a novel retrieval-based multimodal data augmentation method that acquires real-world multimodal posts, fully leveraging the advantages of the social media domain, thus enhancing data realism and diversity. To the best of our knowledge, we are the first to propose multimodal data augmentation in the field of information extraction.
- Secondly, our proposed approach is task-independent and can be easily applied to different information extraction tasks, such as named entity recognition and relation extraction.
- Finally, we have conducted experiments on two different multimodal information extraction tasks, namely multimodal named entity recognition and multimodal relation extraction. The results demonstrate that our method effectively improves model performance on all datasets, demonstrating better performance than representative rule-based, generative, and LLM-based data augmentation methods.

2 Related Work

2.1 Multimodal Information Extraction

We review and summarize two crucial MIE tasks in social media, namely multimodal named entity recognition and multimodal relation extraction.

Early methods [15, 16, 31, 33] mainly study how to incorporate text and image representations, mainly based on attention mechanisms. Furthermore, [32] proposes a multimodal graph fusion that explicitly establishes the relationship between text and image.

Subsequently, some methods validate the effectiveness of different image features as image representations. Specifically, [26] uses the objects in the image as the image representation, [24, 28] further propose using the image objects and image caption as the image representation. [35] uses a scene graph model [22] to extract visual objects and their relations as image representations.

Additionally, some methods [20, 27, 29] study the image-text relevance and propose training a classifier to reduce the noise introduced by irrelevant images. Furthermore, [3] regards visual representation as the visual prefix to guide the text representation for addressing error sensitivity. [34] proposes multimodal versions of back-translation and high resource bridging to alleviate the misalignment between modalities, and [8] proposes alignment tasks for entity-object and relation-image pairs using prompts to reduce the modality misalignment. [6] uses an evidence-theory-based method to simulate human decision-making logic and generate decision support degrees for deciding whether image information should participate in the task.

Additionally, [10, 11] model the MNER as a machine reading comprehension task, and [21] proposes a unified information extraction framework that models the information extraction task as a generation problem.

2.2 Data Augmentation for Information Extraction

Data augmentation aims to increase the size and diversity of training data by modifying the copies of already existing data or adding new data from existing data [2]. Compared with sentence-level tasks such as classification and translation, information extraction tasks are more fragile when confronted with data augmentation noise due to the finer granularity of the task (token-level, entity-level), and thus require more careful design [5].

Specifically, [4] proposes a rule-based approach that generates new data by randomly replacing tokens, synonyms, and mentions in the text with tokens of the same label using a binomial distribution, as well as shuffling the order of tokens in the text. [5] concatenates the label corresponding to each entity before the entity, constructing the text and entities within the same text, and trains a generative model to automatically generate entities and their corresponding labels. [36] replaces the entities in the sentence with masks and builds a pre-trained masked language model to predict the tokens corresponding to the masks, enabling entity replacement. [2] proposes a neural architecture to transform the data representation from one domain to another domain by learning the patterns in the text.

With the rise of large language models (LLMs), [1] and [30] proposed using large models to generate similar texts as augmented data. However, these methods rely heavily on predefined rules or generative models, which lead to the creation of synthetic data with limited diversity, hindering the models' ability to generalize effectively. Additionally, these approaches focus mainly on text augmentation, neglecting the potential advantages of augmenting image data to improve model performance. In this paper, we leverage the retrievable nature of social media data and propose using retrieval-based methods for data augmentation to obtain real-world multimodal data.

3 Method

Our proposed RMDA shown in Figure 2, consists of three main steps: (1) construct the query; (2) retrieve the results; (3) filter the results. RMDA expands the training set by retrieving examples similar to those in the training set to perform data augmentation.

Next, we will define our problem and then describe our approach in detail.

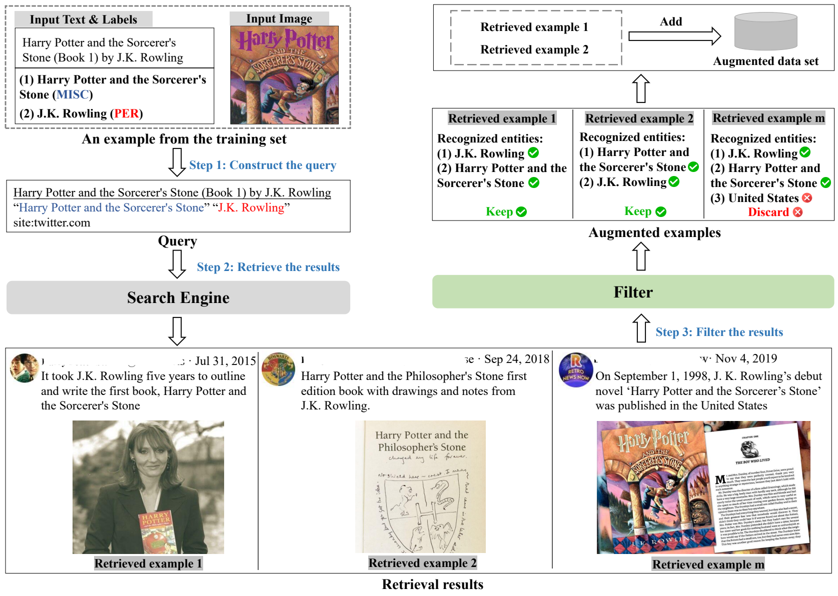

Input Text & Labels
Harry Potter and the Sorcerer's Stone (Book 1) by J.K. Rowling
(1) Harry Potter and the Sorcerer's Stone (MISC)
(2) J.K. Rowling (PER)
Input Image
An example from the training set
Step 1: Construct the query
Harry Potter and the Sorcerer's Stone (Book 1) by J.K. Rowling
"Harry Potter and the Sorcerer's Stone" "J.K. Rowling"
site:twitter.com
Query
Step 2: Retrieve the results
Search Engine
Jul 31, 2015
It took J.K. Rowling five years to outline and write the first book, Harry Potter and the Sorcerer's Stone
Retrieved example 1
Sep 24, 2018
Harry Potter and the Philosopher's Stone first edition book with drawings and notes from J.K. Rowling.
Retrieved example 2
Nov 4, 2019
On September 1, 1998, J. K. Rowling's debut novel 'Harry Potter and the Sorcerer's Stone' was published in the United States
Retrieved example m
Retrieval results
Step 3: Filter the results
Filter
Augmented examples
Retrieved example 1
Recognized entities:
(1) J.K. Rowling ✓
(2) Harry Potter and the Sorcerer's Stone ✓
Keep ✓
Retrieved example 2
Recognized entities:
(1) Harry Potter and the Sorcerer's Stone ✓
(2) J.K. Rowling ✓
Keep ✓
Retrieved example m
Recognized entities:
(1) J.K. Rowling ✓
(2) Harry Potter and the Sorcerer's Stone ✓
(3) United States ✗
Discard ✗
Retrieved example 1
Retrieved example 2
Add
Augmented data set

Fig. 2: Overall architecture of RMDA. The process is illustrated using the MNER task as an example, with user information in the retrieval results being masked.

3.1 Problem Formulation

In this paper, we have conducted experiments on multimodal named entity recognition (MNER) and multimodal relation extraction (MRE), and they are defined as follows.

MNER: Given a text T and its associated images I as input, the task of MNER is to extract a set of named entities from T, and classify each extracted named entity into one of the pre-defined types. Most existing work formulates the task as a sequence labeling problem: let $T = (t_1, t_2, ..., t_n)$ denote a sequence of input words (tokens), and $Y = (y_1, y_2, ..., y_n)$ be the corresponding label sequence, where $y_i \in \mathcal{Y}$ and $\mathcal{Y}$ is the pre-defined label set with the BIO tagging schema.

MRE: Given a text T, its associated images I, the marked head entity e_h, and the tail entity e_t as input, the task of MRE is to classify the corresponding relation tag between e_h and e_t, where e_h and e_t are entities in the text.

3.2 Construct the Query

Given the advanced capabilities and robust performance of text-based search engines, and the labels in MIE tasks typically correspond only to text (e.g., labels in MRE are typically entity categories for each token in the text), we utilize only the text content of the example to construct the query (we will provide a more detailed explanation of the reasons for not using images in the subsequent sections).

Additionally, we strive to ensure that the retrieved posts contain the same entities as those in the original example and that the sources of these posts align with the sources of the original example. This approach helps in obtaining examples with similar distributions. To achieve this, we utilize the exact match and on-site search capabilities provided by the modern search engine's advanced search [1].

Specifically, for examples from the MNER training set, we extract all entities using the provided labels in the example and enclose each entity in quotation marks, resulting in a set of entities represented as $(e_1, e_2, ..., e_u)$, where u denotes the number of entities. This approach utilizes the search engine's advanced search to perform an exact match on the contents within the quotation marks to ensure that these entities are included in the search results. We then concatenate text T with $(e_1, e_2, ..., e_u)$ to construct the query $(T_t, e_1, e_2, ..., e_u)$ for MNER. For examples from the MRE training set, where the task input specifies the two entities (head entity and tail entity) for which the relation needs to be judged, we directly capture these entities, enclosing them in quotation marks to obtain e_h and e_t, respectively. We then concatenate T, e_h, and e_t to construct the query (T_t, e_h, e_t) for MRE.

Finally, to limit the search engine results to specific sources, we use the site search feature of advanced search to append site information prefixed with "site:" at the end of the query (e.g. "site: twitter.com" indicates that only content from "twitter.com" is retrieved). Since current MIE datasets present their data sources, the corresponding websites can be easily obtained. As shown in Figure 2, the text from Twitter containing the entities `Harry Potter and the Sorcerer's Stone` and `J.K. Rowling` will be constructed into the query: `Harry Potter and the Sorcerer's Stone (Book 1) by J.K. Rowling ''Harry Potter and the Sorcerer's Stone'' ''J.K. Rowling'' site:twitter.com`.

3.3 Retrieve the Results

In this paper, we use Google as our search engine and implement the search function using the API provided by Serper[2]. This API allows users to scrape Twitter results from a Google search[3], taking a query as input and returning hyperlinks corresponding to the retrieval results. Additionally, the API supports scraping search results from other social platforms, such as Weibo[4].

Firstly, we input the query into the search engine to get the hyperlinks corresponding to the retrieval results, as shown in the left side of Figure 3. Since we limit our search to a specific social media, the page corresponding to a hyperlink is usually a single post in social media, as shown in the right side of Figure 3.

[1] https://www.google.com/advanced_search

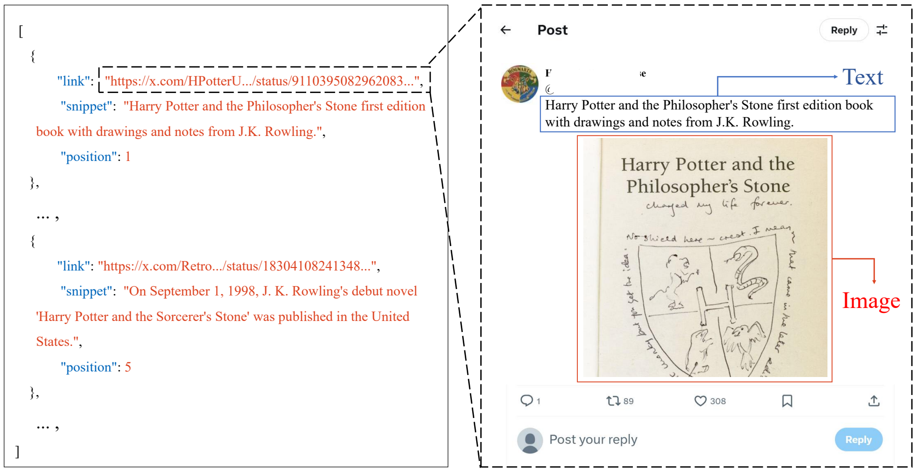
[2] https://serper.dev/

[3] https://serpapi.com/twitter-results

[4] https://serpapi.com/baidu-social-media-results

Then, we visit these links to enter the corresponding HTML page and use the HTML parser (such as Beautiful Soup [5]) to get the text and images of the post on the page. If the post does not contain an image, then we discard the post (it is easy to determine whether the post contains an image through the HTML parser).

Finally, we get retrieval results: $R = (r_1, r_2, ..., r_m)$, where each search retrieval contains the text and its accompanying image (text and its accompanying image are from the same post), and m is the number of retrieval results.



```
[
  {
      "link": "https://x.com/HPotterU.../status/9110395082962083...",
       "snippet":  "Harry Potter and the Philosopher's Stone first edition
   book with drawings and notes from J.K. Rowling.",
       "position": 1
  },
   ... ,
  {
      "link": "https://x.com/Retro.../status/18304108241348...",
       "snippet":  "On September 1, 1998, J. K. Rowling's debut novel
   'Harry Potter and the Sorcerer's Stone' was published in the United
   States.",
       "position": 5
  },
   ... ,
]
```

Harry Potter and the Philosopher's Stone first edition book with drawings and notes from J.K. Rowling.

Post
Reply
Text
Harry Potter and the Philosopher's Stone
changed my life forever.
No shield here ~ crest. I mean
Image
1
89
308
Post your reply
Reply

Search results returned by the search engine

A page accessed via the corresponding link

Fig. 3: Examples of retrieval results and accessing the corresponding pages through the links.

3.4 Filter the Results

To ensure the quality of the obtained examples, further filtering is applied. For MNER, we first fine-tune BERTweet [18] on the training set to obtain an entity boundary detector, which is used to detect entities in the text. We use the detector to recognize entities in the text of the retrieval result, and if any recognized entity is not in $(e_1, e_2, ..., e_u)$, then this retrieval result is discarded.

For MRE, we employ SimSCE [7] to measure the similarity between the text in the retrieval results and the T_t, and any text in retrieval results with a similarity lower than 0.77 (based on our test on 50 instances) are discarded. The remaining results, after deduplication, are kept as augmented examples, and the labels of the original example are assigned to these augmented examples.

[5] https://pypi.org/project/beautifulsoup4/

Finally, we merge the augmented examples with the original training set D_{train} to obtain an expanded training set $D_{aug} = \{(T_i, I_i)\}_{i=1}^{N+M}$, where N is the number of examples in D_{train}, and M is the number of augmented examples.

Note: We do not use images to filter the results for the following reasons: (1) One of the purposes of our multimodal data augmentation is to obtain more diverse images. As shown in Figure 2, many images in the retrieval results are different from the original image. Using images to filter the results would discard these more diverse images; (2) The images and text come from the same post, so regardless of whether the image contributes to the information extraction, it still reflects the real-world co-occurrence relationship between the image and text. This is why we do not use image information for filtering in the retrieval process.

4 Experiments

4.1 Datasets

For the MNER, we use four public datasets in the social media domain, namely Twitter-2015 [33], Twitter-2017 [31], SNAP [15] and MNER-MI [9]. The first three datasets are single-image datasets, where each example is a text-image pair. MNER-MI is a multi-image dataset, where each example contains a text and its corresponding multiple images (up to four). There are four types of entities: Person (PER), Organization (ORG), Location (LOC) and Miscellaneous (MISC). These datasets contain 4,000/1,000/3,357, 3,373/723/723, 4,290/1,432/1,459, and 6,856/860/860 examples in train/development/test set respectively.

For the MRE, we use the MNRE dataset [35], where the data is also collected from social media. It contains 9,201 sentences and 15,485 entity pairs with 23 types of relations. In total, there are 12,247/1,624/1,614 entity pairs in the train/development/test set, respectively.

4.2 Settings

In search, we ask the search engine to return 10 search results (one page). The number of augmented samples for Twitter-2015, Twitter-2017, SNAP, MNER-MI and MNRE is 11,643, 9,681, 11,844, 17,366 and 34,292 respectively. On average, each search yields 2.75 augmented examples.

All models are implemented with PyTorch Framework, and all the experiments are conducted on NVIDIA GTX 4090 GPUs. During training, we use mini-batch backpropagation for training and the AdamW optimizer for optimization. In all experiments, we use the grid search in the development set to find the batch size within [8,64], and the learning rate within $[5e^{-6}, 7e^{-5}]$. All models are trained for 15 epochs and we select the model that performs best on the development set and evaluate it on the test set. Following many recent works [1, 3, 8, 28, 34], we use F1 score (**F1**) as evaluation metrics.

4.3 Baselines

To validate the effectiveness and robustness of our method, we compare the performance of our data augmentation method combined with specific models against other data augmentation methods applied to the same models. Specifically, we use two established text-based information extraction models: (1) BERT [6] [12] and (2) RoBERTa [7] [14], and two representative multimodal information extraction models: (3) UMT [31] and (4) HVPNeT [3].

We compare our method against four baselines in our experiments: (1) **Gold** uses the examples from the original training set without any data augmentation; (2) **LSMS** [4] uses four rule-based methods to transform the text, including label-wise token replacement, synonym replacement, mention replacement, and shuffling of sentence order; (3) **DAGA** [5] trains a generative model to generate both tokens and labels simultaneously; (4) **MELM** [36] trains a RoBERTa model and uses it to replace the entities; (5) **LLM-DA** [30] leverages LLMs to rewrite the original examples and replaces entities in the text with entities of the same type.

The number of augmented examples generated by the above methods exceeds that of our method. For instance, LSMS augments each example four times, while MELM demonstrates in experiments that augmenting three times is a better choice.

4.4 Performance Comparison

As shown in Table 1, our comprehensive evaluation compares our RMDA method with representative data augmentation approaches across multiple models and datasets.

Firstly, we find that, compared to using only the original training set (Gold), all augmentation methods significantly improve model performance. This highlights the effectiveness of data augmentation methods in MIE tasks and underscores data augmentation as a promising strategy for enhancing model generalization.

Secondly, we observe that LLM-DA outperforms rule-based and traditional generative approaches. This is because traditional rule-based and generative methods rely entirely on the training set data, making it difficult for the generated examples to generalize beyond the training set. These methods exhibit low example diversity and struggle to maintain semantic integrity, consequently producing relatively low-quality augmented samples. In contrast, LLM-DA leverages the advanced semantic understanding and generation capabilities of LLMs to produce more diverse and semantically coherent examples. This indicates that the quality of generated examples can significantly impact model performance.

Finally, our proposed RMDA method consistently and significantly outperforms competing approaches across all datasets, with particularly notable im-

[6] https://huggingface.co/google-bert/bert-base-uncased

[7] https://huggingface.co/FacebookAI/roberta-base

Table 1: Performance comparison (F1 score) with different competitive baseline approaches for MNER and MRE. Tw-15: Twitter-2015, Tw-17: Twitter-2017. The marker † refers to significant test p-value < 0.05 when compared with LLM-DA.

Models	**Methods**	**Tw-15**	**Tw-17**	**SNAP**	**MNER-MI**	**MNRE**
BERT	Gold	71.81	83.44	84.61	71.22	60.86
	LSMS	72.32	84.87	85.66	72.13	63.85
	DAGA	72.44	84.99	85.87	72.34	62.97
	MELM	72.65	85.08	86.25	73.02	64.03
	LLM-DA	74.73	86.38	86.92	74.07	65.17
	RMDA (ours)	**75.87**†	**87.25**†	**87.89**†	**75.30**†	**66.05**†
RoBERTa	Gold	73.86	85.72	86.08	73.04	63.96
	LSMS	74.12	86.28	86.54	73.92	65.47
	DAGA	74.25	86.17	86.79	73.48	64.87
	MELM	74.33	86.40	87.12	74.24	65.99
	LLM-DA	74.96	87.50	87.71	74.83	67.27
	RMDA (ours)	**76.32**†	**88.54**†	**89.01**†	**76.38**†	**68.67**†
UMT	Gold	73.41	85.31	85.98	74.13	63.46
	LSMS	74.32	86.87	87.71	74.77	65.15
	DAGA	74.08	86.73	87.98	74.43	65.54
	MELM	74.80	86.71	87.32	74.55	66.03
	LLM-DA	75.56	87.54	88.26	76.74	66.17
	RMDA (ours)	**76.87**†	**89.25**†	**89.63**†	**78.65**†	**68.05**†
HVPNeT	Gold	75.32	86.87	87.73	75.10	81.85
	LSMS	76.10	87.31	88.05	75.86	83.11
	DAGA	75.87	87.03	88.26	75.46	83.20
	MELM	76.02	87.27	88.44	75.78	83.65
	LLM-DA	76.77	87.87	89.13	76.27	84.36
	RMDA (ours)	**78.97**†	**89.94**†	**90.19**†	**78.53**†	**86.25**†

provements in multimodal models. This demonstrates that the augmented examples generated through our method are of superior quality and can more effectively enhance model performance. Moreover, our results emphasize the crucial importance of ensuring the authenticity of augmented examples and validate the significant potential of multimodal data augmentation strategies.

4.5 Ablation Study

To validate the effectiveness of multimodal data augmentation in our proposed RMDA, we conduct an ablation study.

As shown in Table 2, removing text data augmentation, image data augmentation, and filter from RMDA leads to a significant drop in model performance. This demonstrates that both types of augmentation and filter contribute effectively to enhancing the MIE model's performance and that multimodal data augmentation holds advantages over unimodal approaches. Furthermore, we observe that removing text data augmentation results in a more significant performance drop compared to removing image data augmentation. This is because MIE is primarily a text-centered multimodal understanding task, where text data augmentation provides more substantial benefits than image data augmentation.

Table 2: Ablation study of our RMDA. w/o Image DA indicates that image data augmentation is not used, where the augmented text is combined with the original image to form augmented examples. w/o Text DA indicates that text data augmentation is not used, where the augmented image is combined with the original text to form augmented examples. w/o Filter indicates that no filtering is performed after the results are obtained.

Models	**Methods**	**Tw-15**	**Tw-17**	**SNAP**	**MNER-MI**	**MNRE**
UMT	**RMDA**	**76.87**	**89.25**	**89.63**	**78.65**	**68.05**
	w/o Image DA	75.77	88.03	88.61	76.90	66.79
	w/o Text DA	74.38	87.02	87.30	75.16	64.15
	w/o Filter	76.23	88.76	88.73	78.05	67.64
HVPNeT	**RMDA**	**78.97**	**89.94**	**90.19**	**78.53**	**86.25**
	w/o Image DA	77.21	88.76	89.50	77.67	85.17
	w/o Text DA	76.22	87.42	88.91	76.04	83.76
	w/o Filter	78.33	89.26	89.61	78.01	85.99

Table 3: Comparison of efficiency (seconds) and cost (USD) between LLM and our method. The recorded time and cost are for generating 500 augmented examples. The cost refers to the fees required for calling the API.

Dataset	**LLM-DA**		**RMDA**	
	Time	Cost	Time	Cost
Tw-15	945s	$0.6723	**532s**	**$0.0562**
Tw-17	952s	$0.6678	**544s**	**$0.0573**
SNAP	936s	$0.6701	**538s**	**$0.0547**
MNER-MI	977s	$0.7012	**502s**	**$0.0561**
MNRE	967s	$0.6923	**497s**	**$0.0512**

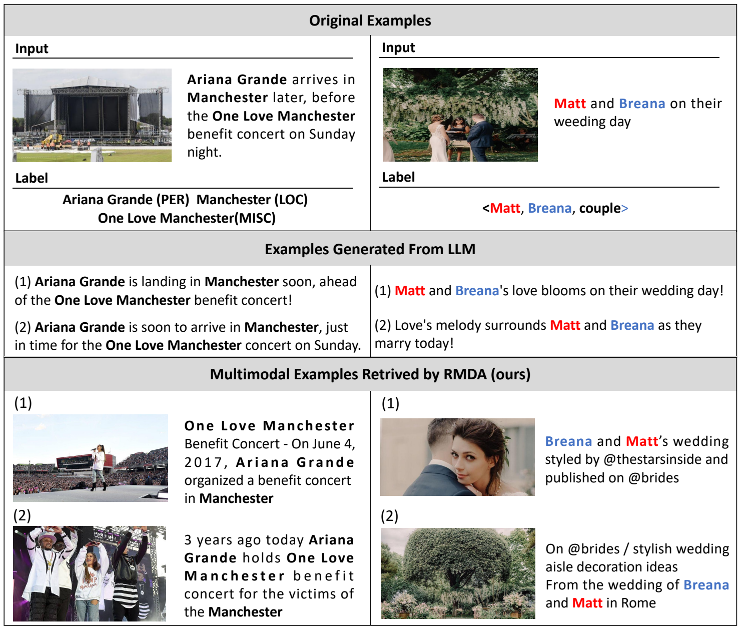
4.6 Efficiency and Cost

As shown in Table 3, we compare the performance of our data augmentation method with LLM-DA (based on gpt3.5-turbo) in terms of efficiency and cost. For convenience, we measure the time and cost required for both methods to generate 500 augmented examples.

We find that, due to the large number of parameters in the LLM, its generation speed is slower, and its usage cost is higher. Our method generates examples at approximately **twice the speed of the LLM and at only one-tenth of the cost**. Specifically, our method takes an average of 1 second to generate one augmented data sample, with a cost of 0.00001 USD, which also makes it more applicable to larger-scale data augmentation. This clearly demonstrates the advantages of our method in both efficiency and cost.

4.7 Comparison of Augmented Examples From LLM and RMDA

To demonstrate the effectiveness of the RMDA module, we compare the augmented examples from the LLM-DA (based on an LLM) and RMDA in Figure 4.



Original Examples

Input

Ariana Grande arrives in **Manchester** later, before the **One Love Manchester** benefit concert on Sunday night.

Label

Ariana Grande (PER) Manchester (LOC)
One Love Manchester(MISC)

Input

Matt and **Breana** on their weeding day

Label

<**Matt**, **Breana**, **couple**>

Examples Generated From LLM

(1) **Ariana Grande** is landing in **Manchester** soon, ahead of the **One Love Manchester** benefit concert!

(2) **Ariana Grande** is soon to arrive in **Manchester**, just in time for the **One Love Manchester** concert on Sunday.

(1) **Matt** and **Breana**'s love blooms on their wedding day!

(2) Love's melody surrounds **Matt** and **Breana** as they marry today!

Multimodal Examples Retrived by RMDA (ours)

(1) **One Love Manchester** Benefit Concert - On June 4, 2017, **Ariana Grande** organized a benefit concert in **Manchester**

(2) 3 years ago today **Ariana Grande** holds **One Love Manchester** benefit concert for the victims of the **Manchester**

(1) **Breana** and **Matt**'s wedding styled by @thestarsinside and published on @brides

(2) On @brides / stylish wedding aisle decoration ideas From the wedding of **Breana** and **Matt** in Rome

Fig. 4: Illustrative comparison of augmented examples from LLM and RMDA. The left and right sides show examples of MNER and MRE, respectively.

First, compared to LLM, the RMDA module performs multimodal data augmentation, which can help the model capture more relationships between text and images. For instance, with only the original MNER image, the model might only capture the correspondence between the One Love Manchester benefit concert and the stage in the image. However, with the multimodal augmented examples, the model might capture the correspondence between Ariana Grande and the singer in the image, as well as the One Love Manchester benefit concert and the details of the stage and the audience.

In addition, the text generated by the LLM varies in expression but maintains consistent semantic content. In contrast, examples retrieved from multiple posts contain different perspectives and diverse information on the same topic. For instance, the retrieval-based augmented examples of MRE exhibit different expressions and expand the semantics, providing additional semantic information such as the location of the wedding. This can help the model understand and process different expressions of the same theme and improve its generalization ability with real-world data.

Lastly, LLMs typically generate formal, well-structured, grammatically correct, and logically coherent sentences. In contrast, social media content is personalized and informal, and retrieved examples can reflect this unique data distribution. For example, in the second retrieval-based augmented example of MNER, the tense is incorrect, and the second retrieval-based augmented example of MRE lacks the coherence of a normal sentence. Examples retrieved by RMDA directly mirror the habits of real users and align with the actual data distribution of social media. Such data augmentation is beneficial for training more accurate and adaptive models, especially for applications related to social media.

5 Conclusion

In this paper, we propose RMDA, which fully leverages the advantages of the social media domain and achieves multimodal data augmentation through retrieval ensuring that the augmented examples are authentic. We conduct extensive experiments to validate the effectiveness of our method across MIE models, including text-based models and multimodal models, achieving better performance with fewer examples compared to other baseline data augmentation methods. Furthermore, our experiments confirm the substantial benefits of multimodal data augmentation compared to unimodal data augmentation. Additionally, our method demonstrates faster speed and lower cost compared to using LLMs, making it more suitable for large-scale data augmentation.

Acknowledgments. This work is supported by National Science and Technology Major Project (2021ZD0111000/2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901,22DZ2229004), the Fundamental Research Funds for the Central Universities 2232023D-19, the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education. Xin Lin is the corresponding author. Xin Lin is also a member of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education.

References

1. Chen, F., Feng, Y.: Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. arXiv preprint arXiv:2306.14122 (2023)

2. Chen, S., Aguilar, G., Neves, L., Solorio, T.: Data augmentation for cross-domain named entity recognition. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 5346–5356. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.434
3. Chen, X., Zhang, N., Li, L., Yao, Y., Deng, S., Tan, C., Huang, F., Si, L., Chen, H.: Good visual guidance make a better extractor: Hierarchical visual prefix for multimodal entity and relation extraction. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1607–1618. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-naacl.121
4. Dai, X., Adel, H.: An analysis of simple data augmentation for named entity recognition. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 3861–3867. International Committee on Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.coling-main.343
5. Ding, B., Liu, L., Bing, L., Kruengkrai, C., Nguyen, T.H., Joty, S., Si, L., Miao, C.: Daga: Data augmentation with a generation approach for low-resource tagging tasks. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6045–6057 (2020)
6. Ding, G., Jiang, T., Zhou, R., Gao, Q.: Aldf: An adaptive logical decision framework for multimodal named entity recognition. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 436–445 (2024)
7. Gao, T., Yao, X., Chen, D.: Simcse: Simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 6894–6910. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.emnlp-main.552
8. Hu, X., Chen, J., Liu, A., Meng, S., Wen, L., Yu, P.S.: Prompt me up: Unleashing the power of alignments for multimodal entity and relation extraction. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 5185–5194. ACM (2023). https://doi.org/10.1145/3581783.3611899
9. Huang, S., Xu, B., Li, C., Ye, J., Lin, X.: Mner-mi: A multi-image dataset for multimodal named entity recognition in social media. In: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 11452–11462 (2024)
10. Jia, M., Shen, L., Shen, X., Liao, L., Chen, M., He, X., Chen, Z., Li, J.: Mner-qg: An end-to-end mrc framework for multimodal named entity recognition with query grounding. In: Proceedings of the AAAI conference on artificial intelligence. vol. 37, pp. 8032–8040 (2023)
11. Jia, M., Shen, X., Shen, L., Pang, J., Liao, L., Song, Y., Chen, M., He, X.: Query prior matters: a mrc framework for multimodal named entity recognition. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3549–3558 (2022)
12. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186 (2019)
13. Liu, X., Hu, C., Zhang, R., Sun, K., Mensah, S., Mao, Y.: Multimodal relation extraction via a mixture of hierarchical visual context learners. In: Proceedings of the ACM on Web Conference 2024. pp. 4283–4294 (2024)
14. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

15. Lu, D., Neves, L., Carvalho, V., Zhang, N., Ji, H.: Visual attention model for name tagging in multimodal social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1990–1999. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/p18-1185
16. Moon, S., Neves, L., Carvalho, V.: Multimodal named entity recognition for short social media posts. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 852–860. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-1078
17. Nasar, Z., Jaffry, S.W., Malik, M.K.: Named entity recognition and relation extraction: State-of-the-art. ACM Computing Surveys (CSUR) **54**(1), 1–39 (2021). https://doi.org/10.1145/3445965
18. Nguyen, D.Q., Vu, T., Nguyen, A.T.: Bertweet: A pre-trained language model for english tweets. EMNLP 2020 p. 9 (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.2
19. Nie, Y., Tian, Y., Wan, X., Song, Y., Dai, B.: Named entity recognition for social media texts with semantic augmentation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1383–1391. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.107
20. Sun, L., Wang, J., Zhang, K., Su, Y., Weng, F.: Rpbert: a text-image relation propagation-based bert model for multimodal ner. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 13860–13868. Association for the Advancement of Artificial Intelligence (AAAI) (2021). https://doi.org/10.1609/aaai.v35i15.17633
21. Sun, L., Zhang, K., Li, Q., Lou, R.: Umie: Unified multimodal information extraction with instruction tuning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 19062–19070 (2024)
22. Tang, K., Niu, Y., Huang, J., Shi, J., Zhang, H.: Unbiased scene graph generation from biased training. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3716–3725 (2020). https://doi.org/10.1109/CVPR42600.2020.00377
23. Wang, X., Cai, J., Jiang, Y., Xie, P., Tu, K., Lu, W.: Named entity and relation extraction with multi-modal retrieval. In: Findings of the Association for Computational Linguistics: EMNLP 2022. pp. 5925–5936 (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.437
24. Wang, X., Gui, M., Jiang, Y., Jia, Z., Bach, N., Wang, T., Huang, Z., Tu, K.: Ita: Image-text alignments for multi-modal named entity recognition. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3176–3189. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.naacl-main.232
25. Wei, P., Huang, Z., Ouyang, H., Hu, Q., Zeng, B., Feng, G.: Cgi-mre: A comprehensive genetic-inspired model for multimodal relation extraction. In: Proceedings of the 2024 International Conference on Multimedia Retrieval. pp. 524–532 (2024)
26. Wu, Z., Zheng, C., Cai, Y., Chen, J., Leung, H.f., Li, Q.: Multimodal representation with embedded visual guiding objects for named entity recognition in social media posts. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 1038–1046. ACM (2020). https://doi.org/10.1145/3394171.3413650

27. Xu, B., Huang, S., Du, M., Wang, H., Song, H., Sha, C., Xiao, Y.: Different data, different modalities! reinforced data splitting for effective multimodal information extraction from social media posts. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 1855–1864. International Committee on Computational Linguistics (2022), https://aclanthology.org/2022.coling-1.160
28. Xu, B., Huang, S., Du, M., Wang, H., Song, H., Xiao, Y., Lin, X.: A unified visual prompt tuning framework with mixture-of-experts for multimodal information extraction. In: International Conference on Database Systems for Advanced Applications. pp. 544–554. Springer, Springer Nature Switzerland (2023). https://doi.org/10.1007/978-3-031-30675-4_40
29. Xu, B., Huang, S., Sha, C., Wang, H.: Maf: a general matching and alignment framework for multimodal named entity recognition. In: Proceedings of the fifteenth ACM international conference on web search and data mining. pp. 1215–1223. ACM (2022). https://doi.org/10.1145/3488560.3498475
30. Ye, J., Xu, N., Wang, Y., Zhou, J., Zhang, Q., Gui, T., Huang, X.: Llm-da: Data augmentation via large language models for few-shot named entity recognition. arXiv preprint arXiv:2402.14568 (2024)
31. Yu, J., Jiang, J., Yang, L., Xia, R.: Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 3342–3352. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.306
32. Zhang, D., Wei, S., Li, S., Wu, H., Zhu, Q., Zhou, G.: Multi-modal graph fusion for named entity recognition with targeted visual guidance. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 14347–14355. Association for the Advancement of Artificial Intelligence (AAAI) (2021). https://doi.org/10.1609/aaai.v35i16.17687
33. Zhang, Q., Fu, J., Liu, X., Huang, X.: Adaptive co-attention network for named entity recognition in tweets. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018). https://doi.org/10.1609/aaai.v32i1.11962
34. Zheng, C., Feng, J., Cai, Y., Wei, X., Li, Q.: Rethinking multimodal entity and relation extraction from a translation point of view. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 6810–6824. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.acl-long.376
35. Zheng, C., Feng, J., Fu, Z., Cai, Y., Li, Q., Wang, T.: Multimodal relation extraction with efficient graph alignment. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 5298–5306. ACM (2021). https://doi.org/10.1145/3474085.3476968
36. Zhou, R., Li, X., He, R., Bing, L., Cambria, E., Si, L., Miao, C.: Melm: Data augmentation with masked entity language modeling for low-resource ner. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2251–2262 (2022)