Different Data, Different Modalities! Reinforced Data Splitting for Effective Multimodal Information Extraction from Social Media Posts

Bo Xu[1], **Shizhou Huang**[1], **Ming Du**[1,*], **Hongya Wang**[1], **Hui Song**[1],
Chaofeng Sha[2] and **Yanghua Xiao**[2]
[1]School of Computer Science and Technology, Donghua University
[2]Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University
xubo@dhu.edu.cn, 2202408@mail.dhu.edu.cn, {duming, hywang, songhui}@dhu.edu.cn
{cfsha, shawyh}@fudan.edu.cn

Abstract

Recently, multimodal information extraction from social media posts has gained increasing attention in the natural language processing community. Despite their success, current approaches overestimate the significance of images. In this paper, we argue that different social media posts should consider different modalities for multimodal information extraction. Multimodal models cannot always outperform unimodal models. Some posts are more suitable for the multimodal model, while others are more suitable for the unimodal model. Therefore, we propose a general data splitting strategy to divide the social media posts into two sets so that these two sets can achieve better performance under the information extraction models of the corresponding modalities. Specifically, for an information extraction task, we first propose a data discriminator that divides social media posts into a multimodal and a unimodal set. Then we feed these sets into the corresponding models. Finally, we combine the results of these two models to obtain the final extraction results. Due to the lack of explicit knowledge, we use reinforcement learning to train the data discriminator. Experiments on two different multimodal information extraction tasks demonstrate the effectiveness of our method. The source code of this paper can be found in https://github.com/xubodhu/RDS.

1 Introduction

Social media, with its wealth of user-generated posts, provides a rich platform for understanding events, opinions and preferences of groups and individuals (Moon et al., 2018). Information extraction, such as named entity recognition (Yu et al., 2020), relation extraction (Zheng et al., 2021a) and sentiment detection (Yang et al., 2021), is a critical step in uncovering these hidden insights in social media posts.

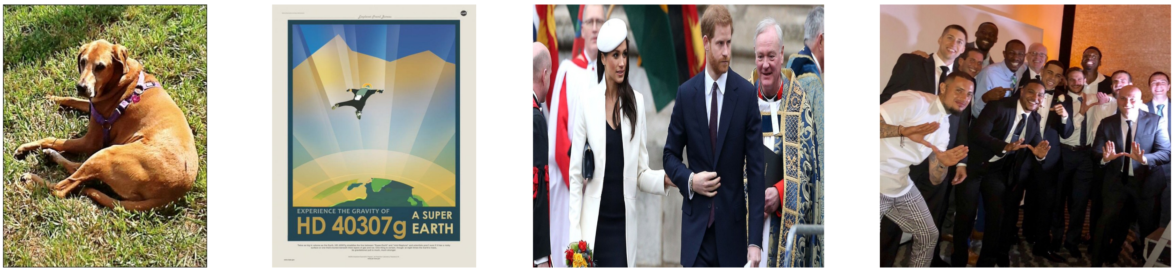
In social media scenarios, information is expressed not only through textual modality, but through multiple modalities (e.g., text, image, etc.). Considering only text modality may lead to inaccurate information extraction. For example in Fig. 1(a) and Fig. 1(c), the text-based named entity recognition model cannot recognize `Kolo` as a dog, and the text-based relation extraction model cannot determine that `Meghan Markle` and `Prince Harry` are couples.

Therefore, many multimodal models for information extraction have been proposed which are using visual modality to complement text modality. They mainly focus on using a multimodal interaction mechanism to combine text representation with image representation. For example, Zhang et al. (2018) propose an adaptive co-attention network to control the combination of text representation and image representation dynamically for multimodal named entity recognition. Zheng et al. (2021a) propose a multimodal relation extraction neural network with an effective alignment strategy for textual and visual graphs to find the correlations between visual objects and textual entities. Yang et al. (2021) propose multi-channel graph neural networks for multimodal sentiment detection.

Despite their success, current approaches overestimate the significance of images. In fact, images are not always needed to understand information on social media posts. It is possible to get comparable performance using only text modality, while using *mismatched* visual modality can hinder performance. The *mismatched* phenomenon is very common in social media posts. As reported in Vempala and Preoţiuc-Pietro (2019), about 33.8% of tweets had textual content that was not reflected in the images, and the images did not add additional con-

[*] Corresponding author. This paper was supported by the National Natural Science Foundation of China (No. 61906035), Information Development Project of Shanghai Economic and Information Commission (No. 202002009), Natural Science Foundation of Shanghai (No. 22ZR1402000) and Shanghai Science and Technology Innovation Action Plan (No. 19511120400)

Proceedings of the 29th International Conference on Computational Linguistics, pages 1855–1864
October 12–17, 2022.



(a) [**Kolo MISC**] loves the sun and is so pretty

(b) [**Nasa ORG**] produces vintage travel posters for newly discovered planets

(c) **Meghan Markle** and **Prince Harry** announce their first official royal tour <Meghan Markle, Prince Harry, couple>

(d) Congrats to **Angel** and **Jesenia Rodriguez** on their marriage last night <Angel, Jesenia Rodriguez, couple>

Figure 1: Four Examples of Multimodal Information Extraction Tasks. (a) and (b) are examples of multimodal named entity recognition, (c) and (d) are examples of multimodal relation extraction. The named entity and its type are highlighted in brackets. The entities and relations between entities are in angle brackets.

tent. For example, in Fig. 1(b) and Fig. 1(d), with the *mismatched* images, the multimodal named entity recognition model may mistakenly associate the `person` in the image with `Nasa` and make a wrong prediction, and the multimodal relation extraction model may incorrectly classify the relation between `Angel` and `Jesenia Rodriguez` as `colleague`.

Based on the above observations, we argue that different social media posts should consider different modalities for multimodal information extraction. Multimodal models cannot always outperform unimodal models [1]. Some posts are more suitable for the multimodal model, while others are more suitable for the unimodal model. This is consistent with human behavior in accomplishing multimodal information extraction tasks. When someone reads a post on social media, he first determines whether the image will help him complete the task; if not, he goes directly to the text, and if so, he views both the image and the text. According to a research on the multimodal named entity recognition (*NER*) benchmark dataset `TWITTER-2017` by Yu et al. (2020), about 22% of entities were incorrectly predicted by the state-of-the-art text-based *NER* model, but correctly predicted by the state-of-the-art multimodal *NER* model; and about 12% of entities were correctly predicted by the text-based *NER* model, but incorrectly predicted by the multimodal *NER* model.

In this paper, we propose a general data splitting strategy to divide the social media posts into two sets so that these two sets can achieve better performance under the information extraction models of the corresponding modalities. Specifically, for an information extraction task, we first propose a data discriminator that divides social media posts into a multimodal and a unimodal set, which is used to determine whether the posts are more suitable for the multimodal or unimodal model. Then we feed these sets into the corresponding models. Finally, we combine the results of these two models to obtain the final extraction results. The core component is the data discriminator. Due to the lack of explicit knowledge about which data are more suitable for multimodal models and which data are more suitable for unimodal models, we use reinforcement learning to train the data discriminator. The reward based on the performances of the multimodal model and the unimodal model on both the multimodal set and the unimodal set will be used as a reinforcement signal to update the parameters of the data discriminator.

Our main contributions are summarized as follows:

- First, to the best of our knowledge, we are the first to argue that different social media posts should consider different modalities to accomplish the multimodal information extraction tasks.

- Secondly, we propose a general data splitting strategy for multimodal information extraction and implement this strategy by reinforcement learning.

- Finally, experiments conducted on two widely used multimodal named entity recognition datasets and a multimodal relation extraction dataset show that our method can effectively

[1] In this paper, unimodal model refers to text-based model.

divide the social media posts and achieves the new state-of-the-art performance.

2 OVERVIEW

In this section, we first formulate our problem, and then introduce our framework.

2.1 Problem Formulation

Let $X = \{(T_i, V_i)\}_{i=1}^{N}$ be the set of text-image posts from social media, where T_i is the text modality and V_i is the corresponding visual information, N represents the number of posts. Our aim is to divide X into two disjoint sets: multimodal set X^M and unimodal set X^U, where X^M contains N^M posts and X^U contains N^U posts with $N^M + N^U = N$, so that these two sets can achieve better performance under the information extraction models of the corresponding modalities (i.e., multimodal and unimodal model).

2.2 Framework

The reinforcement learning framework for training the data discriminator is shown in the left side of Fig. 2, which consists of three main components: a data discriminator, a multimodal model and a unimodal model, where the multimodal and unimodal models can be any existing models. The training process is as follows:

- **STEP 1: Training Set Splitting**. Given the training set $\mathcal{D} = \{(T_j, V_j, Y_j)\}_{j=1}^{G}$, where Y_j is the label for the j-th post. We first randomly divide it into two disjoint sets, namely $\mathcal{D}_{model}$ and $\mathcal{D}_{split}$.

- **STEP 2: Multimodal/Unimodal Model Training**. Then we use $\mathcal{D}_{model}$ to train the multimodal and unimodal models and freeze their parameters.

- **STEP 3: Data Discriminator Training**. Finally, we use reinforcement learning to train the data discriminator with $\mathcal{D}_{split}$. In the data discriminator, each data in $\mathcal{D}_{split}$ has a corresponding action a_i to determine whether to use a multimodal or unimodal model for information extraction. The $\mathcal{D}_{split}$ can be divided into the multimodal set $\mathcal{D}_{split}^{M}$ and the unimodal set $\mathcal{D}_{split}^{U}$ based on the data discriminator. After that, we calculate the reward based on the performances of the multimodal and the unimodal models on both $\mathcal{D}_{split}^{M}$ and $\mathcal{D}_{split}^{U}$. The reward will be used as a reinforcement signal to update the parameters of the data discriminator.

3 METHOD

In this section, we first introduce our data discriminator, then we show how to train it using reinforcement learning, which consists of a reward function and a training algorithm.

3.1 Data Discriminator

The data discriminator is used to determine whether a data should use a multimodal or unimodal model, and the main idea is based on the similarity between text and images. As shown in the right side of Fig. 2, the data discriminator consists of a *CLIP* (Radford et al., 2021) and a multilayer perceptron (*MLP*) with one hidden layer. *CLIP* is the state-of-the-art multimodal vision and language model, which consists of a *CLIPTextModel*, a *CLIPVisionModel* and a projection layer.

Specifically, the *CLIPTextModel* layer is used to encode the text. For the input text T, we first tokenize it by using the byte pair encoding (BPE) (Sennrich et al., 2016) and obtain a token sequence $(t_1, t_2, ..., t_n)$, where n is the length of the token sequence. Then the token sequence is bracketed by [SOS] and [EOS] tokens as $([SOS], t_1, t_2, ..., t_n, [EOS])$. The activation at the [EOS] token in the last layer of *CLIPTextModel* is treated as the representation of the entire text $\boldsymbol{T_s} \in \mathbb{R}^{d_t}$.

The *CLIPVisionModel* layer is used to encode the images. For the input image V, we first resize the image to 224×224 pixels, then the image is split into a sequence of $7 \times 7 = 49$ non-overlapping patches with a pixel size of 32×32, which are then linearly embedded to get each patch representation $(v_1, v_2, ..., v_{49})$, and finally add a [CLS] token with the same dimensions as patch at the beginning of V to get $([CLS], v_1, v_2, ..., v_{49})$, where the activation at the [CLS] token in the last layer of *CLIPVisionModel* as the representation of the image $\boldsymbol{V_g} \in \mathbb{R}^{d_v}$.

The projection layer is used to project the representations of text and images into a latent space with identical dimension. The final representations of text $\boldsymbol{T_c} \in \mathbb{R}^{d_c}$ and image $\boldsymbol{V_c} \in \mathbb{R}^{d_c}$ is obtained by projecting $\boldsymbol{T_s}$ and $\boldsymbol{V_g}$ into the same latent space.

Finally, we perform an element-wise product of $\boldsymbol{T_c}$ and $\boldsymbol{V_c}$ and feed it into an *MLP* layer with one hidden layer to obtain the probability p that the data

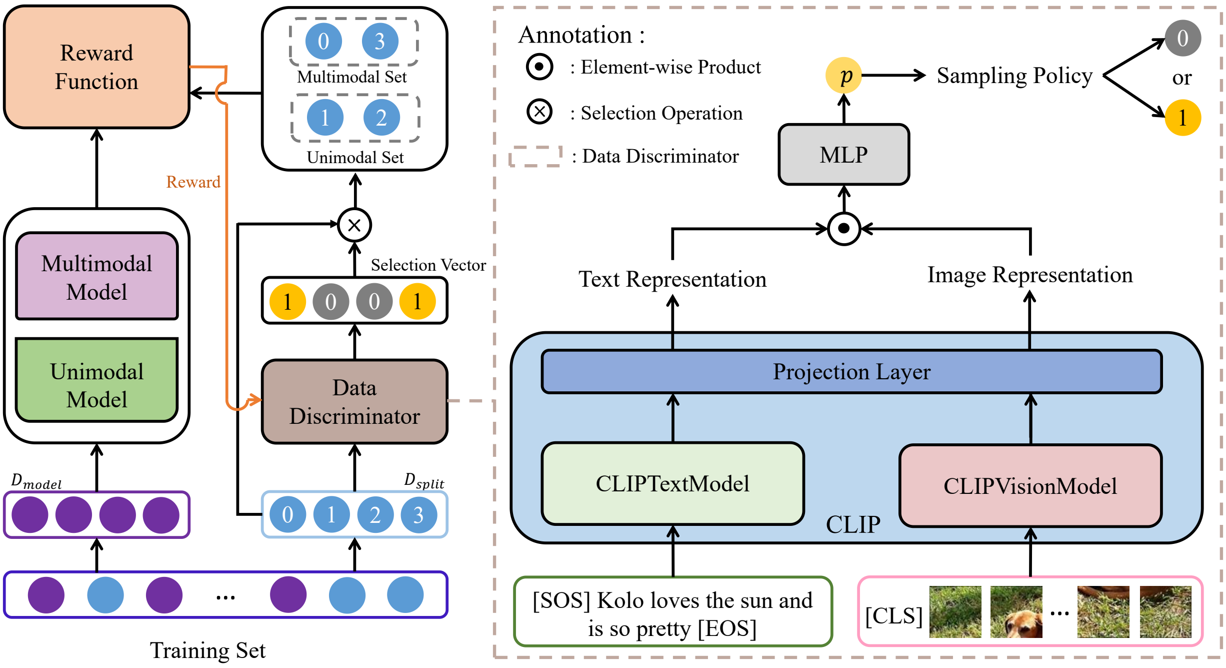


Figure 2: The Reinforcement Learning Framework to Train the Data Discriminator.

is more suitable for the multimodal *IE* model (less suitable for the unimodal *IE* model) as follows:

$$p = sigmoid(\boldsymbol{W_2}\ relu(\boldsymbol{W_1}\ (\boldsymbol{T_c} \odot \boldsymbol{V_c}))) \quad (1)$$

, where $sigmoid$ and $relu$ are the activation functions, $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are the weight matrices.

The sampling policy is used to decide the action of the data discriminator based on the probabilities provided by the data discriminator. In this paper, we propose two sampling policies. In the training phase of the data discriminator, in order to encourage exploration based on the uncertainty of the exponentially-large selection space (Yoon et al., 2020), we use *Bernoulli* sampling (Deshmukh1, 1991) to sample the data. Each data will be sampled according to the probability provided by the data discriminator. While in the prediction phase, the data discriminator puts data with probability greater than 0.5 into the multimodal set and data with probability less than or equal to 0.5 into the unimodal set.

3.2 Reward Function

Due to the lack of supervised data, we use reinforcement learning to train the data discriminator. The core component is to design the reward function, which is used to evaluate the quality of the action of data discriminator and used as a reinforcement signal to adjust the parameters of the data discriminator.

Intuitively, the multimodal *IE* model performs better than the unimodal *IE* model on the $\mathcal{D}^M_{split}$, while the unimodal *IE* model performs better than the multimodal *IE* model on the $\mathcal{D}^U_{split}$. In our task here, we use the performance gaps between the multimodal *IE* model and the unimodal *IE* model on both sets as the reward. For example, the micro F1 score is used to evaluate the performance of the multimodal named entity recognition (*MNER*) models and multimodal relation extraction (*MRE*) models on both sets (Yu et al., 2020; Zheng et al., 2021a).

Specifically, we denote the $\mathcal{D}^M_{split} = \{(T_k, V_k, Y_k)\}^{S_1}_{k=1}$ and $\mathcal{D}^U_{split} = \{(T_l, V_l, Y_l)\}^{S_2}_{l=1}$, respectively. The multimodal *IE* model is denoted as f_m and the unimodal *IE* model is denoted as f_u. The performances of models on both sets are defined as follows:

$$v^k_m = Score(\{Y_k, f_m(T_k, V_k)\}^{S_1}_{k=1}) \quad (2)$$

$$v^l_m = Score(\{Y_l, f_m(T_l, V_l)\}^{S_2}_{l=1}) \quad (3)$$

$$v^k_u = Score(\{Y_k, f_u(T_k)\}^{S_1}_{k=1}) \quad (4)$$

$$v^l_u = Score(\{Y_l, f_u(T_l)\}^{S_2}_{l=1}) \quad (5)$$

, where v^k_m and v^l_m are the performances of multimodal *IE* model performed on $\mathcal{D}^M_{split}$ and $\mathcal{D}^U_{split}$, respectively. v^k_u and v^l_u are the performances of unimodal *IE* model performed on $\mathcal{D}^M_{split}$ and $\mathcal{D}^U_{split}$, respectively. Finally, the reward R is calculated as follows:

$$R = \alpha * (v_m^k - v_u^k) + (1 - \alpha) * (v_u^l - v_m^l) \quad (6)$$

where $\alpha \in (0, 1)$ is the hyperparameter.

3.3 Training Algorithm

Finally, we introduce how to train the data discriminator. Inspired by (Yoon et al., 2020), the training process of the data discriminator is shown in Algorithm 1.

Algorithm 1 The Training Algorithm for the Data Discriminator.

```
Inputs: Training set $\mathcal{D}$, batch size $B_s$, hyperparameter of reward function $\alpha$, learning rate of data discriminator $\eta$
Output: The data discriminator $g_\phi$.
Shuffle $\mathcal{D}$ and divide it into $\mathcal{D}_{model}$ and $\mathcal{D}_{split}$
Train multimodal model and unimodal model using $\mathcal{D}_{model}$ and freeze their parameters
Initialize parameters $\phi$ for the data discriminator $g_\phi$
while until convergence do
    Randomly sample a batch of data $\mathcal{D}_{split}^B = \{(T_i, V_i, Y_i)\}_{i=1}^{B_s}$ from $\mathcal{D}_{split}$
    for $i = 1$ to $B_s$ do
        Calculate probability $p_i = g_\phi(T_i, V_i)$ according to Equation 1
    end for
    Obtain $\mathcal{D}_{split}^M$, $\mathcal{D}_{split}^U$ by Bernoulli sampling
    Calculate reward according to Equation 6
    Update $\phi$ according to Equation 7
end while
```

We first shuffle the training set $\mathcal{D}$ and divide it into two parts. One is the $\mathcal{D}_{model}$, which is used for training the multimodal *IE* model and the unimodal *IE* model, and the other is $\mathcal{D}_{split}$, which is used for training the data discriminator. Then we train the multimodal *IE* model and the unimodal *IE* model by using the $\mathcal{D}_{model}$ and freeze their parameters.

After that, we initialize the parameters ϕ of the data discriminator g_ϕ. For each iteration, we randomly select a batch of data $\mathcal{D}_{split}^B$ from the $\mathcal{D}_{split}$, and use the data discriminator to predict the probability that each data is more suitable for the multimodal *IE* models p_i. Based on the probabilities, we divide the $\mathcal{D}_{split}$ into the $\mathcal{D}_{split}^M$ and $\mathcal{D}_{split}^U$ by using the *Bernoulli* sampling.

Then, we calculate the reward R according to Equation 6, and update the parameters of our data discriminator as follows:

$$\phi \leftarrow \phi + \eta \cdot R \cdot \bigtriangledown_\phi \log \pi_\phi(\mathcal{D}_2^B, (d_1, ..., d_{B_s})) \quad (7)$$

$$\pi_\phi(\mathcal{D}_2^B, (d_1, ..., d_{B_s})) = \prod_{j=1}^{B_s} (p_j)^{d_j} \cdot (1 - p_j)^{1-d_j} \quad (8)$$

, where η is learning rate, $\pi_\phi(\mathcal{D}_2^B, (d_1, ..., d_{B_s}))$ is the probability that the selection vector $(d_1, ..., d_{B_s})$ is selected based on g_ϕ and $d_j = \{0, 1\}$ is an indicator variable, where 1 represents to put the data into the multimodal set $\mathcal{D}_{split}^M$, while 0 represents to put the data into the unimodal set $\mathcal{D}_{split}^U$.

4 Experiment

To validate the effectiveness of our data splitting strategy, we conducted experiments on two different multimodal information extraction tasks, namely multimodal named entity recognition (*MNER*) and multimodal relation extraction (*MRE*).

4.1 Dataset

4.1.1 MNER Datasets

For the *MNER* task, we use two widely used datasets, `Twitter2015` (Zhang et al., 2018) and `Twitter2017` (Lu et al., 2018), which are collected from Twitter. Each tweet contains a text-image pair and the text may contain zero or more named entities. There are four types of entities: Person (*PER*), Organization (*ORG*), Location (*LOC*) and others (*MISC*). We use the pre-processed datasets provided by Yu et al. (2020) [2]. In total, there are 4,000/1,000/3,357 and 3,373/723/723 sentences in train/development/test set contained in `Twitter2015` and `Twitter2017`, respectively.

4.1.2 MRE Datasets

For the *MRE* task, we use the `MNRE` [3] dataset (Zheng et al., 2021a), which is also collected from Twitter. 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 train/development/test set, respectively.

4.2 Metrics

We use the micro precision (**P**), recall (**R**) and F1 score (**F1**) to evaluate the performance of both the

[2] https://github.com/jefferyYu/UMT

[3] https://github.com/thecharm/Mega

MNER models and the *MRE* models, which are widely used in recent *MNER* (Moon et al., 2018; Zhang et al., 2018; Lu et al., 2018; Yu et al., 2020; Chen et al., 2021) and *MRE* (Zheng et al., 2021a,b) works.

4.3 Parameter Settings

We conduct all the experiments on `NVIDIA GTX 2080 Ti` GPUs with PyTorch 1.7.1. The parameter settings of our framework are as follows:

- We randomly split the training data into $\mathcal{D}_{model}$ (80%) and $\mathcal{D}_{split}$ (20%).

- For the *MNER* task, we use *UMT-BERT-CRF* (Yu et al., 2020) and *MAF* (Xu et al., 2022) as the multimodal model, respectively. And use *BERT-CRF* as the unimodal model, which consists of *BERT* (Devlin et al., 2018) and *CRF* (John D. Lafferty, 2001). We use the same hyperparameters provided by Yu et al. (2020) to train both the *UMT-BERT-CRF* model and the *BERT-CRF* model.

- For the *MRE* task, we use *MEGA* (Zheng et al., 2021a) as the multimodal model and *MTB* (Soares et al., 2019) as the unimodal model. We use the same hyperparameters provided by Zheng et al. (2021a) to train the *MEGA* model and follow OpenNRE [4] to train the *MTB* model.

- For the data discriminator, we use $CLIP_{32}$ [5] to obtain the representations of text and images in the same latent space.

- For training data discriminator, we use grid search in the development set to find the learning rate of data discriminator η within [$1e^{-5}$, $1e^{-4}$], the batch size B_s within [128, 512], and the hyperparameter of reward function α within [0.1, 0.9] in Algorithm 1.

- All models use mini-batch backpropagation for training and use adam optimizer (Kingma and Ba, 2015) for optimization.

4.4 Evaluation

We first train the multimodal and unimodal models on the full training set and then evaluate the performance of different models on three test sets, which consists of a *Unimodal test set* and a *Multimodal test set* and the *Full test set*. Specifically, the unimodal and multimodal test sets are obtained by using our data discriminator on the *Full test set*. To evaluate our method on the *Full test set*, as mentioned above, we combine the predictions of the unimodal model on the *Unimodal test set* and the predictions of the multimodal model on the *Multimodal test set* as the predictions of our method.

4.5 Performance Comparison

We compare the performance on both the *MRE* and *MNER* tasks, the comparison results are shown in Table 1, Table 2 and Table 3, respectively.

Table 1: Performance Comparison on *MRE* Task.

Model	Test Set	P	R	F1
MTB	**Unimodal**	**60.60**	**71.43**	**65.57**
	Multimodal	60.68	64.64	62.60
	Full	60.65	66.72	63.54
MEGA	Unimodal	64.88	55.61	59.89
	Multimodal	**70.45**	**62.84**	**66.43**
	Full	68.79	60.63	64.45
Ours	**Full**	**66.83**	**65.47**	**66.14**

Specifically, we first compare the performance of each model on different test sets in the *MRE* task. As shown in Table 1, the unimodal relation extraction model *MTB*, performs the best on the *Unimodal test set* and the worst on the *Multimodal test set*. For the multimodal relation extraction model *MEGA*, it performs the best on the *Multimodal test set* and the worst on the *Unimodal test set*. We perform *MTB* on *Unimodal test set* and *MEGA* on *Multimodal test set* as our *MRE* method on *Full test set*. Compared to the performance (**F1**) of the other two models on the *Full test set*, our method achieves the best performance, beating the state-of-the-art *MRE* model (*MEGA*) by 1.69 points. This shows that our data discriminator can effectively split the data, where the *Unimodal test set* is indeed more suitable for unimodal models and the *Multimodal test set* is indeed more suitable for multimodal models and the prediction results of combining the two models on their suitable data can have better performance.

Then, we compare the performance of each model on different test sets in the *MNER* task. As shown in Table 2 and Table 3, we obtain the

[4] https://github.com/thunlp/OpenNRE

[5] https://huggingface.co/openai/clip-vit-base-patch32

Table 2: Performance Comparison on *MNER* Task (UMT-BERT-CRF as the multimodal model).

Model	**Test Set**	`Twitter2015`			`Twitter2017`		
		P	**R**	**F1**	**P**	**R**	**F1**
BERT-CRF	**Unimodal**	**71.35**	**75.32**	**73.28**	**85.58**	**85.13**	**85.36**
	Multimodal	71.23	73.99	72.58	83.68	81.56	82.60
	Full	71.29	74.63	72.92	84.76	83.57	84.16
UMT-BERT-CRF	Unimodal	70.50	75.12	72.73	84.21	83.55	83.88
	Multimodal	**72.50**	**74.95**	**73.71**	**84.92**	**85.79**	**85.35**
	Full	71.50	74.96	73.19	84.74	84.68	84.71
Ours	**Full**	**71.94**	**75.13**	**73.50**	**85.29**	**85.42**	**85.36**

Table 3: Performance Comparison on *MNER* Task (MAF as the multimodal model).

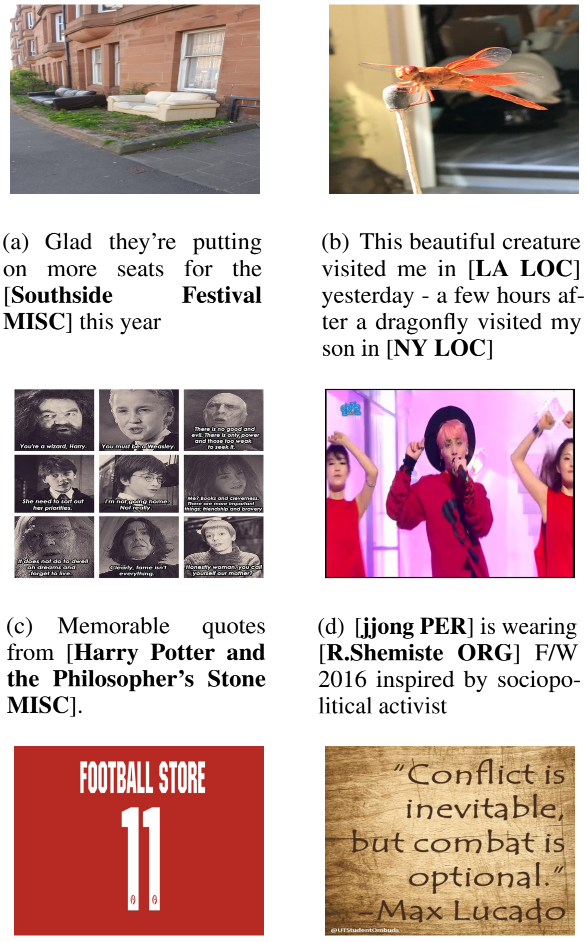
Model	**Test Set**	`Twitter2015`			`Twitter2017`		
		P	**R**	**F1**	**P**	**R**	**F1**
BERT-CRF	**Unimodal**	**71.35**	**75.31**	**73.28**	**85.71**	**84.03**	**84.87**
	Multimodal	71.24	74.00	72.60	84.42	83.40	83.91
	Full	71.29	74.63	72.92	84.76	83.57	84.16
MAF	Unimodal	71.13	75.39	73.20	84.99	84.03	84.51
	Multimodal	**72.53**	**74.70**	**73.60**	**86.44**	**87.22**	**86.83**
	Full	71.85	75.04	73.41	86.06	86.38	86.22
Ours	**Full**	**71.96**	**75.00**	**73.44**	**86.25**	**86.38**	**86.32**

same conclusions as for the *MRE* task, i.e., the unimodal named entity recognition model *BERT-CRF* performs the best on the *Unimodal test set* and the worst on the *Multimodal test set*, and the multimodal named entity recognition model *UMT-BERT-CRF* and *MAF* perform the best on the *Multimodal test set* and the worst on the *Unimodal test set*. We also perform *BERT-CRF* on *Unimodal test set* and *UMT-BERT-CRF* or *MAF* on *Multimodal test set* as our *MNER* method on *Full test set*. When the multimodal model is *UMT-BERT-CRF*, our method outperforms it by 0.31 and 0.65 points on `Twitter2015` and `Twitter2017`, respectively. Our method also outperforms it when the multimodal model is *MAF*, which illustrates that the multimodal model in our method can be replaced by any existing multimodal model, including the one that already considers the mismatched image problem. But the improvement is smaller compared to using *UMT-BERT-CRF* as a multimodal model because (1) we only try a few different sets of hyperparameters on *MAF* and (2) *MAF* considers the problem of mismatched image and has a strong robustness to the mismatched image.

4.6 Case Study

To show the effectiveness of our data discriminator more intuitively, we perform the data discriminator on the test set of `TWITTER-2017` and select six samples for analysis based on the probability predicted by the data discriminator. Note that the data discriminator puts data with probability greater than 0.5 into the multimodal set and data with probability less than or equal to 0.5 into the unimodal set. Specifically, Fig. 3(a) and Fig. 3(b) show the two samples with the lowest probability, Fig. 3(c) and Fig. 3(d) show the two samples with the highest probability, and Fig. 3(e) and Fig. 3(f) show the two samples with the medium probability.

Firstly, for the two samples with the lowest probability, we can observe that there is enough information in their text to recognize the named entities while the images do not provide any useful information to help identify the entities in the text. Specifically, in Fig. 3(a), the unimodal *NER* model can easily recognize from the text that `Southside Festival` is a named entity and the type is *MISC* through the capitalization of the two words and the meaning of the words themselves, and there is no information in the image related to `Southside Festival` to help identify named entities. In Fig. 3(b), the unimodal *NER* model can easily recognize from the text that `LA` and `NY` are named entities and their type is *LOC* through all capitals of the words and the preposition `in`. The main part of the image is a dragonfly, which cannot help identify named entities. In summary, the two samples with the lowest probability are suitable for the unimodal *NER* model, and the use of the multimodal *NER* model will result in poor results due to the introduction of image noise.


You're a wizard, Harry.
You must be a Weasley.
There is no good and evil. There is only power and those too weak to seek it.
She need to sort out her priorities.
I'm not going home. Not really.
Me? Books and cleverness. There are more important things: friendship and bravery
It does not do to dwell on dreams and forget to live.
Clearly, fame isn't everything.
Honestly woman, you call yourself our mother?
FOOTBALL STORE
11
"Conflict is inevitable, but combat is optional."
-Max Lucado

(a) Glad they're putting on more seats for the [**Southside Festival MISC**] this year

(b) This beautiful creature visited me in [**LA LOC**] yesterday - a few hours after a dragonfly visited my son in [**NY LOC**]

(c) Memorable quotes from [**Harry Potter and the Philosopher's Stone MISC**].

(d) [**jjong PER**] is wearing [**R.Shemiste ORG**] F/W 2016 inspired by sociopolitical activist

(e) Take a look at our new look on -line #Football store here:

(f) Work through your conflicts with the student ombuds!

Figure 3: Case Study on `TWITTER-2017`.

Secondly, for the two samples with the highest probability, we can observe that there is not enough information in their text to recognize the named entities. Specifically, in Fig. 3(c), `Harry Potter and the Philosopher's Stone` in the text is the name of a movie and should be classified as *MISC*, but it may also be divided into two parts, `Harry Potter` and `the Philosopher's Stone`, where `Harry Potter` is classified as *PER*. Obviously, there is ambiguity in using only text information, so additional image information is required. In Fig. 3(d), `R.Shemiste` is usually recognized as the name of a person, with the help of the image, we can infer that it is a brand name. In summary, the two samples with the highest probability are suitable for the multimodal *NER* model, and the use of the unimodal *NER* model will result in poor results due to the lack of sufficient information.

Finally, for the two medium probability samples, we find an interesting phenomenon: their images are composed of text and backgrounds. This phenomenon is very common in medium probability samples. The image encoder neither obtains useful information from these images nor introduces noise. Therefore, the multimodal *NER* model degenerates into a unimodal *NER* model. Therefore, it does not matter whether using a multimodal *NER* model or a unimodal model.

5 Related Work

In this section, we review and summarize three multimodal *IE* tasks, namely multimodal named entity recognition, multimodal sentiment detection and multimodal relation extraction.

For the multimodal named entity recognition task, at the earliest, Moon et al. (2018) inputs the whole image into a convolutional neural network (e.g. *ResNet*) to obtain a representation of the whole image to establish the relationship between the text and the image. Since only some regions in the image are useful for recognizing entities, (Lu et al., 2018; Zhang et al., 2018; Yu et al., 2020; Chen et al., 2021; Sun et al., 2021; Xu et al., 2022) divide the image into multiple regions, obtained a representation of the image regions and establish a relationship between the image regions and each word in the text. Next, since the image and text representations come from different encoders, there is a semantic gap that affects the establishment of image and text relationships. To solve the above problem, Xu et al. (2022) proposes an alignment and matching framework to make the text and image representations more consistent by contrastive learning. More directly, (Wu et al., 2020; Wang et al., 2021) extract the semantic information of the image directly to represent the image: Wu et al. (2020) extracts the objects in the image by the object detection model and uses the labels of the objects (e.g., apple, trophy) to represent the image, while (Wang et al., 2021) uses the image captioning model and the OCR model in addition to the object detection model to obtain the overall semantic information of the image and the textual information in the image, respectively. In addition, (Sun et al., 2021; Xu et al., 2022) noticed that the mismatched image can impair the prediction of multimodal models and both propose a method to calculate the image and text similarity scores to filter the image information.

For the multimodal relation extraction task, (Zheng et al., 2021b) first propose this task and demonstrate that previous text-based relation extraction models perform poorly in social media texts, and that incorporating visual information can help improve relation extraction model performance. Subsequently, in order to be able to fully exploit the relationships between objects in the image and to establish the alignment of the image with the text, Zheng et al. (2021a) use graph structure information to align the relations between entities in text and images and then uses image information to supplement the missing semantic information.

For the multimodal sentiment detection task, Xu et al. (2018) obtain the information more important for sentiment by capturing the correlation between text and images. Huang et al. (2018) use an adversarial learning model to learn a joint multimodal representation to combine text and image representations. Yang et al. (2021) use a novel graph neural network based on the global characteristics that encode different modalities to capture hidden representations and learn multimodal representations.

However, current approaches overestimate the significance of images. Although several works (Sun et al., 2021; Xu et al., 2022) have proposed methods to filter image information, they all use the similarity scores of images and text to filter the image information as a whole, more or less retaining some image information and not accurately filtering out the noise in the images. Therefore, we propose a general data splitting technique to process different data using different models (i.e., multimodal model and unimodal model).

6 Conclusion

In this paper, we propose a general data splitting strategy to divide the social media posts into two sets so that these two sets can achieve better performance under the information extraction models of the corresponding modalities. The core component is the data discriminator. Due to the lack of explicit knowledge, we use reinforcement learning to train the data discriminator. Experiments conducted on two widely used multimodal named entity recognition datasets and a multimodal relation extraction dataset show that our data discriminator can effectively split the data and our proposed data splitting strategy for multimodal information extraction achieves the best performance.

References

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen. 2021. Multimodal named entity recognition with image attributes and image knowledge. In *Database Systems for Advanced Applications*, pages 186–201.

Shailaja R Deshmukh1. 1991. Bernoulli sampling. *Australian Journal of Statistics*, 33(2):167–176.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Feiran Huang, Xiaoming Zhang, and Zhoujun Li. 2018. Learning joint multimodal representation with adversarial attention networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1874–1882.

Fernando C. N. Pereira John D. Lafferty, Andrew McCallum. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the 18th Intl. Conf. on Machine Learning (ICML-2001)*, pages 282–289.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1990–1999.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. 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)*, pages 852–860.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International*

Conference on Machine Learning, pages 8748–8763. PMLR.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Livio Baldini Soares, Nicholas Fitzgerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

Lin Sun, Jiquan Wang, Kai Zhang, Yindu Su, and Fangsheng Weng. 2021. Rpbert: a text-image relation propagation-based bert model for multimodal ner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13860–13868.

Alakananda Vempala and Daniel Preoţiuc-Pietro. 2019. Categorizing and inferring the relationship between the text and image of twitter posts. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics*, pages 2830–2840.

Xinyu Wang, Min Gui, Yong Jiang, Zixia Jia, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Kewei Tu. 2021. Ita: Image-text alignments for multi-modal named entity recognition. *arXiv preprint arXiv:2112.06482*.

Zhiwei Wu, Changmeng Zheng, Yi Cai, Junying Chen, Ho-fung Leung, and Qing Li. 2020. 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*, pages 1038–1046.

Bo Xu, Shizhou Huang, Chaofeng Sha, and Hongya Wang. 2022. 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*, pages 1215–1223.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 929–932.

Xiaocui Yang, Shi Feng, Yifei Zhang, and Daling Wang. 2021. Multimodal sentiment detection based on multi-channel graph neural networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 328–339.

Jinsung Yoon, Sercan Arik, and Tomas Pfister. 2020. Data valuation using reinforcement learning. In *International Conference on Machine Learning*, pages 10842–10851. PMLR.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. 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*, pages 3342–3352.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, pages 5674–5681.

Changmeng Zheng, Junhao Feng, Ze Fu, Yi Cai, Qing Li, and Tao Wang. 2021a. Multimodal relation extraction with efficient graph alignment. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 5298–5306.

Changmeng Zheng, Zhiwei Wu, Junhao Feng, Ze Fu, and Yi Cai. 2021b. Mnre: A challenge multimodal dataset for neural relation extraction with visual evidence in social media posts. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.