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

Bo Xu¹, Shizhou Huang¹, Ming Du¹,*, Hongya Wang¹, Hui Song¹, Chaoefeng Sha² and Yanghua Xiao²

¹School of Computer Science and Technology, Donghua University
²Shanghai 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 Preo˘ tiuc-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-
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 entities and relations between entities are in angle brackets. The entities and relations between entities are in angle brackets.

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. 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
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 $D = \{(T_j, V_j, Y_j)\}_{j=1}^G$, where $Y_j$ is the label for the $j$-th post. We first randomly divide it into two disjoint sets, namely $D^\text{model}$ and $D^\text{split}$.

- **STEP 2: Multimodal/Unimodal Model Training.** Then we use $D^\text{model}$ 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 $D^\text{split}$. In the data discriminator, each data in $D^\text{split}$ has a corresponding action $a_t$ to determine whether to use a multimodal or unimodal model for information extraction. The $D^\text{split}$ can be divided into the multimodal set $D^M$ and the unimodal set $D^U$ based on the data discriminator. After that, we calculate the reward based on the performances of the multimodal and the unimodal models on both $D^M$ and $D^U$. 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 $([\text{SOS}], t_1, t_2, ..., t_n, [\text{EOS}])$. The activation at the [EOS] token in the last layer of CLIPTextModel is treated as the representation of the entire text $T_s \in \mathbb{R}^{d_t}$.

The CLIPVisionModel layer is used to encode the images. For the input image $V$, we first resize the image to $224 \times 224$ pixels, then the image is split into a sequence of $7 \times 7 = 49$ non-overlapping patches with a pixel size of $32 \times 32$, 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 $([\text{CLS}], v_1, v_2, ..., v_{49})$, where the activation at the [CLS] token in the last layer of CLIPVisionModel as the representation of the image $V_g \in \mathbb{R}^{d_v}$.

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 $T_c \in \mathbb{R}^{d_c}$ and image $V_c \in \mathbb{R}^{d_c}$ is obtained by projecting $T_s$ and $V_g$ into the same latent space.

Finally, we perform an element-wise product of $T_c$ and $V_c$ and feed it into an MLP layer with one hidden layer to obtain the probability $p$ that the data
is more suitable for the multimodal IE model (less suitable for the unimodal IE model) as follows:

\[ p = \text{sigmoid}(W_2 \text{ relu}(W_1 (T_c \odot V_c))) \]  \hspace{1cm} (1)

where sigmoid and relu are the activation functions, \( W_1 \) and \( W_2 \) 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 \( D^M_{\text{split}} \), while the unimodal IE model performs better than the multimodal IE model on the \( D^U_{\text{split}} \). 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 \( D^M_{\text{split}} = \{(T_k, V_k, Y_k)\}_{k=1}^{S_1} \) and \( D^U_{\text{split}} = \{(T_l, V_l, Y_l)\}_{l=1}^{S_2} \), 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:

\[ v^k_m = \text{Score}(\{Y_k, f_m(T_k, V_k)\}_{k=1}^{S_1}) \]  \hspace{1cm} (2)

\[ v^l_m = \text{Score}(\{Y_l, f_m(T_l, V_l)\}_{l=1}^{S_2}) \]  \hspace{1cm} (3)

\[ v^k_u = \text{Score}(\{Y_k, f_u(T_k)\}_{k=1}^{S_1}) \]  \hspace{1cm} (4)

\[ v^l_u = \text{Score}(\{Y_l, f_u(T_l)\}_{l=1}^{S_2}) \]  \hspace{1cm} (5)

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

$$R = \alpha \ast (v^k_m - v^k_u) + (1 - \alpha) \ast (v^l_u - v^l_m)$$  \hspace{1cm} (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 $g_\phi$.

1. Shuffle $\mathcal{D}$ and divide it into $\mathcal{D}_{\text{model}}$ and $\mathcal{D}_{\text{split}}$.
2. Train multimodal model and unimodal model using $\mathcal{D}_{\text{model}}$ and freeze their parameters.
3. Initialize parameters $\phi$ for the data discriminator $g_\phi$.
4. while until convergence do
5. Randomly sample a batch of data $\mathcal{D}_\text{split}^B_{\text{split}} = \{(T_i, V_i, Y_i)\}_{i=1}^{B_s}$ from $\mathcal{D}_{\text{split}}$.
6. for $i = 1$ to $B_s$ do
7. Calculate probability $p_i = g_\phi(T_i, V_i)$ according to Equation 1.
8. end for
9. Obtain $\mathcal{D}_\text{split}^M, \mathcal{D}_\text{split}^U$ by Bernoulli sampling.
10. Calculate reward according to Equation 6.
11. Update $\phi$ according to Equation 7.
12. end while

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

After that, we initialize the parameters $\phi$ of the data discriminator $g_\phi$. For each iteration, we randomly select a batch of data $\mathcal{D}_\text{split}^B_{\text{split}}$ from the $\mathcal{D}_{\text{split}}$, 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 $\mathcal{D}_{\text{split}}$ into the $\mathcal{D}_{\text{split}}^M$ and $\mathcal{D}_{\text{split}}^U$ 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:

$$\phi \leftarrow \phi + \eta \ast R \cdot \nabla_\phi \log \pi_{\phi}(D_{\text{split}}^B, (d_1, ..., d_{B_s}))$$  \hspace{1cm} (7)$$

$$\pi_{\phi}(D_{\text{split}}^B, (d_1, ..., d_{B_s})) = \prod_{j=1}^{B_s} (p_j)^{d_j} \cdot (1 - p_j)^{1-d_j}$$  \hspace{1cm} (8)$$

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

### 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 ($\text{PER}$), Organization ($\text{ORG}$), Location ($\text{LOC}$) and others ($\text{MISC}$). 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 ($\text{P}$), recall ($\text{R}$) and F1 score ($\text{F1}$) 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 $D_{model}$ (80%) and $D_{split}$ (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 \(^5\) 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 $\eta$ within $[1e^{-5}, 1e^{-4}]$, the batch size $B_s$ within $[128, 512]$, and the hyperparameter of reward function $\alpha$ 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTB</td>
<td>Unimodal</td>
<td>60.60</td>
<td>71.43</td>
<td>65.57</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>60.68</td>
<td>64.64</td>
<td>62.60</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>60.65</td>
<td>66.72</td>
<td>63.54</td>
</tr>
<tr>
<td>MEGA</td>
<td>Unimodal</td>
<td>64.88</td>
<td>55.61</td>
<td>59.89</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>70.45</td>
<td>62.84</td>
<td>66.43</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>68.79</td>
<td>60.63</td>
<td>64.45</td>
</tr>
<tr>
<td>Ours</td>
<td>Full</td>
<td>66.83</td>
<td>65.47</td>
<td>66.14</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>Twitter2015</th>
<th>Twitter2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>BERT-CRF</td>
<td>Unimodal</td>
<td>71.35</td>
<td>75.32</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>71.23</td>
<td>73.99</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>71.29</td>
<td>74.63</td>
</tr>
<tr>
<td>UMT-BERT-CRF</td>
<td>Unimodal</td>
<td>70.50</td>
<td>75.12</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>72.50</td>
<td>74.95</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>71.50</td>
<td>74.96</td>
</tr>
<tr>
<td>Ours</td>
<td>Full</td>
<td>71.94</td>
<td>75.13</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Set</th>
<th>Twitter2015</th>
<th>Twitter2017</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>BERT-CRF</td>
<td>Unimodal</td>
<td>71.35</td>
<td>75.31</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>71.24</td>
<td>74.00</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>71.29</td>
<td>74.63</td>
</tr>
<tr>
<td>MAF</td>
<td>Unimodal</td>
<td>71.13</td>
<td>75.39</td>
</tr>
<tr>
<td></td>
<td>Multimodal</td>
<td>72.53</td>
<td>74.70</td>
</tr>
<tr>
<td></td>
<td>Full</td>
<td>71.85</td>
<td>75.04</td>
</tr>
<tr>
<td>Ours</td>
<td>Full</td>
<td>71.96</td>
<td>75.00</td>
</tr>
</tbody>
</table>

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.

(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


