

# Supplementary Material: Generative Multimodal Data Augmentation for Low-Resource Multimodal Named Entity Recognition

Anonymous Authors

## A DATA STATISTICS

In this section, we report the total number of tweets and entities in the original benchmark datasets and the synthetic datasets generated from GMDA as well as the synthetic datasets generated from mixGen in both low-resource and full-supervision settings.

As shown in Table 1, we can make the following observations. Firstly, the synthetic tweets generated from mixGen and those generated from GMDA are all less than the number of the original training data. The reasons are as follows. For mixGen, we delete tweets that are longer than 50 tokens or that contain more than 10 entities to be more consistent with the distribution of the original dataset. For GMDA, due to the synthetic data filtering strategy, the number of remaining synthetic tweets is much less than the number of training data. Secondly, as the number of samples in the original training set gradually increases, the proportion of retained synthetic tweets also tends to increase, probably because the performance of the base MNER or GMNER model gradually improves. Meanwhile, the Twitter-FMNERG dataset has the fewest retained synthetic tweets, mainly due to the fact that generating accurate fine-grained

**Table 1: Statistics of the three benchmark datasets and synthetic datasets generated from GMDA as well as those generated from mixGen in low-resource and full-supervision settings.**

Dataset	Prop.	The Number of Tweets			
		# Train	# Dev	# mixGen	# GMDA
Twitter-2015	10%	400	100	393	216
	20%	800	200	792	583
	40%	1,600	400	1,571	1,221
	100%	4,000	1,000	3,923	3,093
Twitter-GMNER	10%	700	100	700	367
	20%	1,400	200	1,400	716
	40%	2,800	400	2,800	1,592
	100%	7,000	1,000	7,000	4,084
Twitter-FMNERG	10%	700	100	700	374
	20%	1,400	200	1,400	538
	40%	2,800	400	2,800	1,577
	100%	7,000	1,000	7,000	2,284

  

Dataset	Prop.	The Number of Entities			
		# Train	# Dev	# mixGen	# GMDA
Twitter-2015	10%	624	146	1,637	240
	20%	1,242	304	3,292	846
	40%	2,501	617	6,707	1,800
	100%	6,176	1,546	16,651	4,555
Twitter-GMNER	10%	1,186	244	3,341	572
	20%	2,375	484	6,305	1,117
	40%	4,735	973	12,581	2,507
	100%	11,779	2,450	31,731	6,355
Twitter-FMNERG	10%	1,186	244	3,272	596
	20%	2,375	484	6,433	847
	40%	4,735	973	12,788	2,527
	100%	11,779	2,450	32,128	3,330

entity labels is much more challenging than generating coarse-grained labels. Lastly, due to the simple concatenation of samples in the mixGen method, the number of entities per tweet significantly increases compared to the original samples. In contrast, the number of entities per tweet generated by GMDA appears closer to the number of entities in the original samples.

## B CASE STUDY AND ERROR ANALYSIS

To better understand the distinction between the synthetic data generated from data augmentation methods and the samples in the original training set, we select several representative synthetic samples generated from GMDA for analysis, as shown in Table 2.

### B.1 Case Study

In case (a), it can be observed that GMDA integrates the *Notch* character from the original image with the *Minecraft* reference from the synthetic sentence, resulting in a cartoon depiction of *Notch* with a rich *Minecraft* style. This fusion effectively conveys the semantic information of both two entities. In case (b), the entity *Oscars*, along with its label, guided the GMDA framework to generate a synthetic sentence related to *Oscars* but different from the original sentence in the semantic content, and the synthetic image conveying the iconic visual characteristics associated with the Academy Awards. Based on GMDA, we obtained a synthetic text-image pair that only shares the same named entities with the original tweet but is semantically different.

### B.2 Error Analysis

Additionally, we have observed that some synthetic data contains errors, which may introduce noise to the model training in downstream tasks. Most of these error cases can be generally categorized into the following two groups:

- **The loss of semantic information.** For example, in case (c), based on the synthetic text, we can infer that the Multimodal Text Generation stage in GMDA correctly identifies *Pluto* as a cartoon character from Disney. However, during the Multimodal Image Generation stage, the model did not accurately generate the associated image of *Pluto*; instead, it generated a cartoon dog based on the synthetic sentence’s description.
- **Fabricating non-existent entities.** For example, in case (d), the pyramid of *Egypt* in the original tweet is an analogy for the *Jim Beam American Stillhouse*, which is a stillhouse associated with the American brand Jim Beam. However, the synthetic image incorrectly combines *Egypt* and the *Jim Beam American Stillhouse*, fabricating a structure incorporating features of both entities.

**Table 2: Comparison between the original samples and the synthetic samples generated from GMDA.**

	Original Samples	Synthetic Samples	Original Samples	Synthetic Samples	
(a)			(b)		
	RT @WiredUK : [Minecraft, OTHER] creator [Notch, PER] says his billions have made him miserable <a href="http://t.co/wMnoZp2hLv">http://t.co/wMnoZp2hLv</a> <a href="http://t.co/iUogGT6d3N">http://t.co/iUogGT6d3N</a>	RT @Joy vs Suffer : [Minecraft, OTHER] icon [Notch, PER] is the king of the world! <a href="http://t.co/jj7Bjtwj1g">http://t.co/jj7Bjtwj1g</a>	RT @CBSunday : What treasures worth at least \$ 125K will be gifted to the losers at the [Oscars, OTHER] ? <a href="http://t.co/xlb10CnwQZ">http://t.co/xlb10CnwQZ</a> <a href="http://t.co/WhilRub240">http://t.co/WhilRub240</a>	RT @adeslnor : We are #Handbag-Bad at the [Oscars, OTHER] booth. <a href="http://t.co/LXR4mJjQew">http://t.co/LXR4mJjQew</a>	
(c)			(d)		
	RT @steventurous : This [Pluto, OTHER] image is really stunning . <a href="http://t.co/P3bjN0v46h">http://t.co/P3bjN0v46h</a>	RT @DisneyLikesPluto : My [Pluto, OTHER] is a dog. <a href="http://t.co/kUyNa2syvkD">http://t.co/kUyNa2syvkD</a>	RT @JimBeam : [Egypt, LOC] has pyramids , we have rackhouses . Visit the [Jim Beam American Stillhouse, LOC] <a href="http://t.co/2tvkUNKPji">http://t.co/2tvkUNKPji</a> <a href="http://t.co/dOro4MislS">http://t.co/dOro4MislS</a>	RT @bourbonstories : From [Egypt, LOC] to [Jim Beam American Stillhouse, LOC] : <a href="http://t.co/ogfopQhF9V">http://t.co/ogfopQhF9V</a>	

117  
118  
119  
120  
121  
122  
123  
124  
125  
126  
127  
128  
129  
130  
131  
132  
133  
134  
135  
136  
137  
138  
139  
140  
141  
142  
143  
144  
145  
146  
147  
148  
149  
150  
151  
152  
153  
154  
155  
156  
157  
158  
159  
160  
161  
162  
163  
164  
165  
166  
167  
168  
169  
170  
171  
172  
173  
174

175  
176  
177  
178  
179  
180  
181  
182  
183  
184  
185  
186  
187  
188  
189  
190  
191  
192  
193  
194  
195  
196  
197  
198  
199  
200  
201  
202  
203  
204  
205  
206  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232