CSE4078 Introduction to Natural Language Processing

NLP Project Delivery #3
NLP Task and Dataset Report

Group Number 10

150119880 Abdullah Enes Dizer

150119919 Ayberk Türksoy

150119805 Mert Efe Karaköse

150119751 Musa Meriç

150119715 Yağmur Koçoğlu

NLP (Natural Language Processing) is a sub-branch of artificial intelligence used for computers to understand and process human language. It usually works on textual data and is used to make sense of, classify or infer this data. NLP has many application areas and one of them is classification/categorization.

Classification/categorization is a sub-branch of NLP and generally involves the process of assigning text data to specific categories or classes. This process aims to analyze the content of a document or text and place it into the most appropriate category or class. This process is supported by machine learning and deep learning techniques.

We performed model training with multiple r values and alpha values. We trained 7 models in total and tested these models with the same test sentences and compared the results.

1. Train Model:

r=16 / lora_alpha=32

https://huggingface.co/yagmurkocoglu/model_16_32

Time:63 minutes

max_steps:1500

2. Train Model:

r=32 / lora_alpha=64

https://huggingface.co/AbdullahEnesDizer/model_32_64

Time:75 minutes

max_steps:1500

3. Train Model:

r=64 / lora_alpha=64

https://huggingface.co/MKarakose/model_64_64

Time:82 minutes

max_steps:1500

4. Train Model:

r=64 / lora_alpha=128

https://huggingface.co/MKarakose/model_64_128

Time:91 minutes

max_steps:1500

5. Train Model:

r=128 / lora_alpha=128

https://huggingface.co/turksoyayberk/model_128_128

Time:105 minutes

max_steps:1500

6. Train Model:

r=256 / lora_alpha=256

https://huggingface.co/musamrc/model_256_256

Time:117 minutes

max_steps:2000

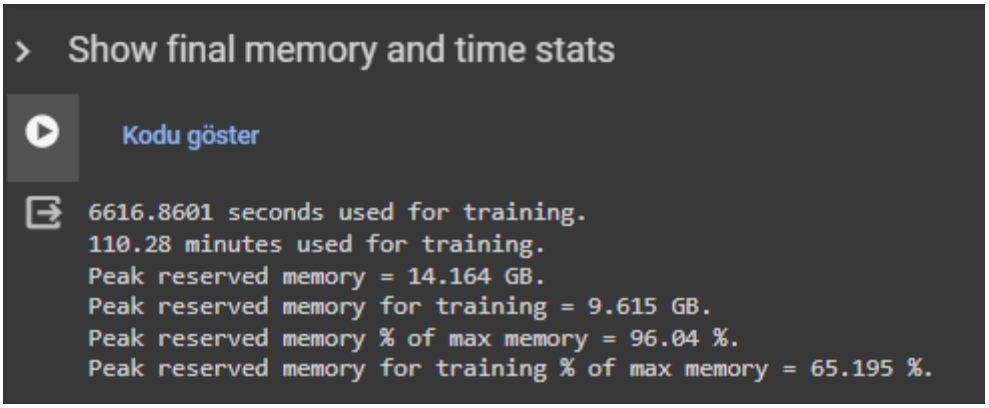
7. Train Model:

r=512 / lora_alpha=512

https://huggingface.co/MKarakose/model_512_512

Time:125 minutes

max_steps:1500



Category/r-alpha	16-32	32-64	64-64	64-128	128-128	256-256	512-512
Dizi-Film	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Kültür	8/10	10/10	9/10	9/10	10/10	8/10	9/10
Sağlık	4/10	5/10	6/10	6/10	5/10	8/10	8/10
Teknoloji	7/10	8/10	7/10	8/10	7/10	7/10	8/10
Ekonomi	6/10	7/10	8/10	7/10	9/10	10/10	9/10
Dünya	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Şikayet/Müşteri Geri Bildirim	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Magazin	9/10	10/10	10/10	10/10	10/10	10/10	10/10
Spor	10/10	10/10	10/10	10/10	10/10	9/10	10/10
Ürün Yorumu	10/10	10/10	10/10	10/10	10/10	10/10	10/10
Kitap	10/10	10/10	10/10	10/10	10/10	10/10	10/10
TOTAL Correct	94/110	100/110	100/110	100/110	101/110	102/110	104/110
TOTAL False	16/110	10/110	10/110	10/110	9/110	8/110	6/110

After training 6 different models with gemma-2b, we obtained multiple results and compared them. According to the results, the best model was r=512 / LoRa alpha=512. This model gave the closest results to the real outputs. We uploaded our models to hugging face.

Each time we increased the max_steps value and r values, we saw that the model gave better results. Thus, we trained the model in the most accurate way.

We trained the model using 366k rows. We tested the model using 110 test sentences.

Example of Excel:

	t Instance	Instruction	Input	Output	Gemma-2b Base Model Output	Fine-Tuned Model Output	ROUGE1	ROUGE2	ROUGE-L	BLEU	BERTScore
	26	Aşağıdaki cümlenin Dünya,	şok iddia hastalıklı	Bu cümle 'Sağlık' kategorisine aittir.	Şok iddia hastalıklı et yiyoruz ! a	Bu cümle 'Ekonomi' kategorisine aittir.	0,8	0,5	0,8	0	0.9452255964279175
	27	Aşağıdaki cümlenin Dünya,	hücre spreyi muciz	Bu cümle 'Sağlık' kategorisine aittir.	Dünyanın çeşitli cümlerin Dünya,	Bu cümle 'Sağlık' kategorisine aittir.	1	1	1	0.1778279410038923	1.0
	28	Aşağıdaki cümlenin Dünya,	beyazlatayım derk	Bu cümle 'Sağlık' kategorisine aittir.	Dünyanın çeşitli cümlerin Dünya,	Bu cümle 'Sağlık' kategorisine aittir.	1	1	1	0.1778279410038923	1.0
	29	Aşağıdaki cümlenin Dünya,	yatalak hastalara iy	Bu cümle 'Sağlık' kategorisine aittir.	The context provides informatio	Bu cümle 'Sağlık' kategorisine aittir.	1	1	1	0.1778279410038923	1.0
	30	Aşağıdaki cümlenin Dünya,	sağlıkçılar çekti gsk	Bu cümle 'Sağlık' kategorisine aittir.	Sağlıkçılar çekti gsk ödüllendirdi	Bu cümle 'Kültür' kategorisine aittir.	0,8	0,5	0,8	0	0.9429526925086975
	31	Aşağıdaki cümlenin Dünya,	megaupload değil	Bu cümle 'Teknoloji' kategorisine aittir.	The context does not provide an	Bu cümle 'Dünya' kategorisine aittir.	0,8	0,5	0,8	0	0.9321044087409973
	32	Aşağıdaki cümlenin Dünya,	mozilla ilk akıllı tel	Bu cümle 'Teknoloji' kategorisine aittir.	Mozilla ilk akıllı telefonlarını tanı	Bu cümle 'Teknoloji' kategorisine aittir.	1	1	1	0.1778279410038923	1.0
	33	Aşağıdaki cümlenin Dünya,	bağdat_caddesi nd	Bu cümle 'Teknoloji' kategorisine aittir.	The context does not provide an	Bu cümle 'Teknoloji' kategorisine aittir.	1	1	1	0.1778279410038923	1.0
	34	Aşağıdaki cümlenin Dünya,	whatsapp a büyük	Bu cümle 'Teknoloji' kategorisine aittir.	Dünya, Ekonomi, Kültür, Sağlık, S	Bu cümle 'Dünya' kategorisine aittir.	0,8	0,5	0,8	0	0.9321044087409973