به نام خدا

پدرام رستمی – ۸۱۰۱۰۰۳۵۳

تمرین چهارم درس پردازش زبان طبیعی

سوال ۱ – ParsiNLU dataset classification

۱- خیر. در این بخش برای داده‌های متنی پیش پردازش خاصی صورت نگرفته است. زیرا داده‌ها خودشان به نسبت تمیز هستند و در هر کدام از ستون‌های sent1 و sent2 تنها یک جمله قرار گرفته است. در نتیجه حروف نشانه گذاری هم اگر داشته باشند، بسیار کم و قابل صرف نظر کردن است. همچنین چون جملات به زبان فارسی هستند، بر خلاف انگلیسی نیازی به تبدیل حروف بزرگ به کوچک وجود ندارد.

۲- در این بخش به کمک مدل xlm-RoBERTa سعی می‌شود لیبل جملات تشخیص داده شود. جملات ابتدا به کمک توکنایزر انکود می‌شوند و سپس جملات انکود شده در کنار هم قرار داده می‌شوند (بین دو جمله یک توکن جداکننده گذاشته می‌شود). سپس برای اینکه سایز تمام ورودی‌هایمان یکی باشد، به دنباله‌ی ترکیب جملات توکن‌های ۱ افزوده می‌شوند. هنگام تولید بردار attention_mask وزن خانه‌ی معنادار ۱ و وزن توکن‌های ۱، صفر در نظر گرفته می‌شود. عملکرد این مدل در شرایطی که lr برابر 1e-5 و سایز batch برابر ۸ فرض شده و برای مدل برای ۱۰ ایپاک آموزش داده شده، بررسی می‌شود. دقت مدل بر روی داده‌های تست برابر ۳۶,۴۱٪ بدست آمد.

۳- در این بخش، تمام اتفاقات بخش قبل تکرار می‌شود فقط به جای استفاده از مدل xlm-RoBERTa از مدل ParsBERT استفاده می‌شود. دقت این مدل بر روی داده‌های تست برابر ۴۵,۹۱٪ بدست آمد. همانطور که از مقایسه‌ی دقت این مدل با مدل xlm-RoBERTa مشخص است، عملکرد این مدل بر روی داده‌های متنی فارسی بهتر است.

سوال ۲ – Multilingual classification

۱- در این بخش، به کمک مدل DistilRoBERTa این مجموعه داده را طبقه بندی می‌کنیم. در این بخش برای پیش پردازش متون انگلیسی، حروف آن‌ها را کوچک کرده و حروف نشانه‌گذاری را هم حذف می‌کنیم. این مدل را مطابق پارامترهایی که در صورت سوال آمده، آموزش می‌دهیم. برای به دست آوردن AUC لازم است تا احتمال اختصاص داده شده برای هر داده به هر کلاس به دست آورده شود که برای به دست آوردن احتمالات، بر روی خروجی‌های مدل تابع softmax اعمال می‌شود تا خروجی‌ها به فرم

احتمال در آیند. هم چنین برای استفاده از تابع roc_auc_curve کتابخانه‌ی sklearn، مقادیر خروجی باید به صورت one hot انکود شوند. در جدول ۱ عملکرد این مدل قابل مشاهده است.

جدول ۱ – عملکرد مدل DistilRoBERTa بر روی متن‌های انگلیسی

	Precision	Recall	F1-score	support
Quran	0.96	0.97	0.97	900
Bible	0.96	0.98	0.97	900
Mizan	0.98	0.95	0.97	900
accuracy			0.97	2700
Macro avg	0.97	0.97	0.97	2700
Micro avg	0.97	0.97	0.97	2700
AUC:			0.9965	

۲- در این بخش به کمک مدل ParsBERT و به کمک داده‌های فارسی سعی می‌کنیم داده‌ها را طبقه بندی کنیم. در این بخش برای پیش پردازش، تنها حروف نشانه گذاری را از متن حذف می‌کنیم. برای محاسبه‌ی امتیازات مدل در معیارهای مختلف، همان کارهایی که در بخش قبلی انجام شد را انجام می‌دهیم. در جدول ۲ عملکرد مدل مشخص است. همانطور که مشخص است، طبقه بندی داده‌ها به کمک متون فارسی عملکرد ضعیف‌تری نسبت به انگلیسی داشته است.

جدول ۲ – عملکرد مدل ParsBERT بر روی متن‌های فارسی

	Precision	Recall	F1-score	support
Quran	0.96	0.96	0.96	900
Bible	0.98	0.94	0.96	900
Mizan	0.93	0.97	0.95	900
accuracy			0.96	2700
Macro avg	0.96	0.96	0.96	2700
Micro avg	0.96	0.96	0.96	2700
AUC:			0.9957	

۳- در این بخش به کمک مدل xlm-RoBERTa داده‌ها را به کمک هر دو ستون متون فارسی و انگلیسی طبقه بندی می‌کنیم. برای پیش پردازش هر ستون، از تابع مخصوص به خود استفاده می‌کنیم. (متون فارسی تنها حذف حروف نشانه گذاری و متون انگلیسی هم حذف حروف نشانه گذاری و هم تبدیل کردن تمام حروف به حروف کوچک). در جدول ۳ عملکرد مدل مشخص است. عملکرد این مدل از مدل ParsBERT بهتر بوده است. همچنین با مقایسه‌ی امتیاز AUC آن با نتایج مدل DistilRoBERTa متوجه می‌شویم که عملکرد آن از آن مدل هم بهتر بوده است.

در این مسئله، زمانی که از مدل چند زبانی استفاده می‌کنیم، به دلیل افزایش تعداد فیچرها، عملکردمان در طبقه بندی بهتر شده است ولی در این مسئله، این پیشرفت بسیار نامحسوس است زیرا مدل‌های تک زبانی خودشان به تنهایی عملکردهای بسیار خوبی داشتند (مدل DistilRoBERTa دقت ۹۷٪ و مدل PartBERT دقت ۹۶٪).

جدول ۳ – عملکرد مدل xlm-RoBERTa بر روی متن‌های فارسی و انگلیسی

	Precision	Recall	F1-score	support
Quran	0.95	0.99	0.97	900
Bible	0.99	0.99	0.99	900
Mizan	0.99	0.95	0.97	900
accuracy			0.97	2700
Macro avg	0.97	0.97	0.97	2700
Micro avg	0.97	0.97	0.97	2700
AUC:			0.9984	

سوال ۳ – Cross-lingual zero-shot transfer learning

۱- انتظار ما با توجه به نتایج به دست آمده در سوال ۱، از عملکرد مدل‌های multilingual بر روی داده‌های فارسی، خیلی زیاد نیست. از مقایسه‌ی نتایج عملکرد مدل xlm-RoBERTa با مدل ParsBERT متوجه شدیم که این مدل‌ها عملکردشان بر روی داده‌های متنی فارسی تقریبا قابل قبول ولی نسبت به مدل ParsBERT اختلاف قابل توجهی دارد. همچنین چون مدل با داده‌های انگلیسی قرار است آموزش ببیند ولی با داده‌های فارسی تست شود، انتظار نداریم عملکرد قابل قبولی داشته باشد.

۲- عملکرد مدل بر روی داده‌هایی که ندیده است بسیار قابل قبول است. دلیل این عملکرد این است که مدل بعد از یادگیری طبقه بندی به کمک متون انگلیسی، به کمک transfer learning عملکرد قابل قبولی بر روی داده‌هایی که از قبل ندیده است دارد. در جدول ۴ عملکرد این مدل بر روی داده‌های فارسی آمده است. دلیل این عملکرد دانشی است که مدل با دیدن متن‌های انگلیسی به دست آورده است. زیرا بخشی از همین دانش در زبان فارسی هم استفاده شده است و به همین دلیل، مدل بر روی متن‌های فارسی هم عملکرد قابل قبولی داشته است.

جدول ۳ – عملکرد مدل xlm-RoBERTa در حالتی که با داده‌های انگلیسی آموزش دیده و با داده‌های فارسی تست شده است

	Precision	Recall	F1-score	support
Quran	0.74	0.72	0.73	900
Bible	0.80	0.55	0.67	900
Mizan	0.70	0.97	0.81	900
accuracy			0.75	2700
Macro avg	0.76	0.75	0.74	2700
Micro avg	0.76	0.75	0.74	2700
AUC:			0.9211	

۳- اصلی‌ترین کاربرد روش cross-lingual zero-shot transfer learning زمانی است که بخواهیم مدلی را برای یکی از تسک‌های NLP برای زبانی آموزش دهیم که داده‌هایی از آن زبان نداشته باشیم ولی در زبان‌های دیگر داده‌های مناسبی برای این کار داشته باشیم. در این صورت مدلمان را با داده‌های زبان‌های دیگر آموزش می‌دهیم تا بتوانیم از دانشی که در زبان‌های دیگر به کار رفته است استفاده کنیم (transfer learning) تا روی زبان دیگری (cross-lingual) که آن را تا به حال ندیده‌ایم استفاده کنیم (zero-shot).