به نام خدا

امیرمحمد کمیجانی ۴۰۴۱۲۸۳۴

عنوان مقاله: AraGenEval Shared Task: A Transformer-based Framework for Detecting AI-Generated Arabic Text

مقدمه:

باپیشرفت مدل‌های زبانی بزرگ، این مدل‌ها قادر به تولید متونی هستند که به شدت human-like هستند یعنی با نوع نگارش، نوشتار و بیان انسان همخوانی نسبتا زیادی دارند. علیرغم مزایای بسیار زیاد، این مورد میتواند باعث بروز مشکلات امنیتی بشود مثلا در بحث آموزش باعث تقلب بشود به طوری که مصحح توانایی تفکیک متن نوشته شده توسط انسان و هوش مصنوعی را نداشته باشد یا در موارد دیگر مانند social engineering attacks.

در این مقاله به بررسی روش‌هایی پرداخته شده که در آن از مدل‌های transformer-based استفاده میشود تا ما بتوانیم نوشته‌های هوش مصنوعی را از انسان در زبان عربی از هم تشخیص و تفکیک کنیم.

پیش‌زمینه:

مسئله تشخیص متون تولید شده توسط هوش مصنوعی در زبان‌هایی که منابع بسیار زیادی برای آنها موجود است همانند زبان انگلیسی دارای سادگی نسبی در مقایسه با

زبان‌هاییست که منابع برای آنها کم است همانند زبان عربی.
علاوه بر کمبود منابع به زبان عربی، از مشکلات دیگر زبان عربی اعراب(diacritic) ها هستند. طبق مطالعات صورت گرفته اعراب تاثیر به سزایی در کاهش عملکرد مدل‌ها برای تشخیص متون تولید شده توسط هوش مصنوعی دارند.

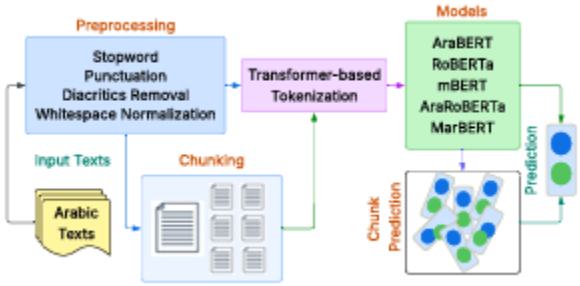
یکی از کارهایی که این مقاله انجام میدهد حذف این اعراب و نورمالایز کردن زبان عربی میباشد و بعد استفاده از مدل‌های transformer-based میباشد.

Description & Dataset:

به طور کلی این مقاله در همایش AraGenEval مورد بحث قرار گرفت و یکی از چالش‌های این همایش بود.
دیتاست این چالش، شامل متونی که توسط انسان نوشته شده و از منابع معتبر جمع آوری شده و همچنین متون تولید شده توسط هوش مصنوعی از مدل‌های زبانی بزرگ همانند GPT4,LLaMA,Mistral استفاده شده است. برای داده‌های قسمت آموزش از ۴۷۹۸ نمونه استفاده شده که شامل ۲۳۹۹ نمونه برای هر کدام از دسته‌ها میباشد. و در قسمت test set به طور کلی ۵۰۰ نمونه داریم.

معماری سیستم:

به طور کلی سیستم ارائه شده در این مقاله به این صورت است که بعد از اعمال preprocessing برروی ورودی، یکبار با chunk کردن و یکبار بدون آن، ورودی را به مدل داده و خروجی را در هر قسمت میسنجیم.



شکل زیر نمایانگر این موضوع است.

Preprocessing : در قسمت پیش‌پردازش اعراب حذف شدند، متون نورمالایز شدند اگر کاراکتری چندبار و بدون معنی تکرار شده بود حذف شدند و همچنین whitespaceها و کاراکترهای خاص نیز حذف شدند.

Transformer-Based Model:

از این نوع مدل‌ها استفاده کردیم چون در فهمیدن متن و اطلاعات موجود در آن به خصوص برای متون طولانی قدرتمند هستند. از چندین مدل همانند Bert, AraBert, RoBerta,AraRoBerta,mBert نیز استفاده کردیم و برای هرکدام از این مدل‌ها از tokenizer مختص به خودشان استفاده کردیم.همچنین با استفاده از padding یا truncating طول دنباله‌ها را به ۵۱۲ توکن رساندیم.

ایده‌ای که در این مقاله پیاده‌سازی شده این است چون ممکن است طول ورودی بیشتر از 512 توکن باشد از chunking استفاده کرده و هر ۴۰۰ توکن را وارد یک chunk کرده و همچنین overlap را برابر ۵۰۰ قرار داده تا دچار به هم ریختگی معنایی نشویم. سپس هر chunk به صورت مستقل توسط مدل بررسی میشود و یک confidence score به آن اختصاص داده میشود و سپس تمام این scoreها میانگین گرفته میشوند. تمام scoreهای مرتبط با یک document میانگین گرفته میشوند.

تمام مدل‌ها برای یک تسک binary classification فاین‌تیون میشوند و هایپرپارامترهای آن به صورت زیر میباشد:

Parameter	Value
Batch Size	16
Epochs	5
Weight Decay	0.001
Learning Rate	2e-5

نتایج :

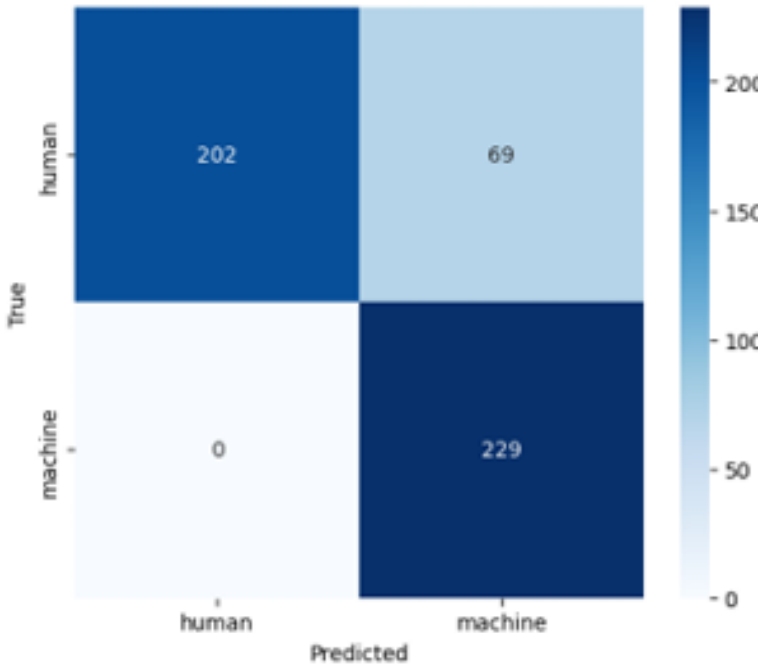
به صورت کلی با اعمال تکنیک chunking تقریبا برای تمامی مدل‌ها در زبان عربی افزایش عملکرد را داشتیم. فقط mbert بود که عملکردمان در آن افزایش نداشته است. همچنین در بخش test-set عملکرد به طور قابل توجهی افزایش داشته است.

در قسمت آموزش نیز افزایش عملکرد را نیز داشته‌ایم؛ اما در بخش test این افزایش چشمگیرتر بوده است. نتایج در جدول زیر میباشد:

Transformer	Approach	Precision	Recall	F1-score	Short	Mid	Long
AraBERT (System Testset)	w/o Chunk	0.47	0.89	0.62	-	-	-
	+ Chunk	0.51	0.97	0.67	-	-	-
	Δ	+0.04	+0.08	+0.05	-	-	-
RoBERTa (System Testset)	w/o Chunk	0.53	0.64	0.58	-	-	-
	+ Chunk	0.47	0.87	0.61	-	-	-
	Δ	+0.06	+0.23	+0.03	-	-	-
AraBERT	w/o Chunk	0.82	0.76	0.79	0.74	0.80	0.73
	+ Chunk	0.88	0.87	0.87	0.89	0.90	0.83
	Δ	+0.06	+0.11	+0.08	+0.15	+0.10	+0.10
RoBERTa	w/o Chunk	0.62	0.54	0.58	0.79	0.78	0.42
	+ Chunk	0.78	0.70	0.73	0.76	0.80	0.84
	Δ	+0.16	+0.16	+0.15	-0.03	+0.02	+0.42
mBERT	w/o Chunk	0.84	0.80	0.81	0.95	0.87	0.64
	+ Chunk	0.77	0.50	0.60	0.37	0.46	0.76
	Δ	-0.07	-0.30	-0.21	-0.58	-0.41	+0.12
Ara-RoBERTa	w/o Chunk	0.23	0.50	0.31	0.64	0.46	0.12
	+ Chunk	0.27	0.52	0.35	0.44	0.53	0.78
	Δ	+0.04	+0.02	+0.04	-0.20	+0.07	+0.66
MARBERT	w/o Chunk	0.83	0.78	0.80	0.87	0.79	0.41
	+ Chunk	0.88	0.86	0.87	0.92	0.86	0.69
	Δ	+0.05	+0.08	+0.07	+0.05	+0.07	+0.28

Error analysis :

برای مدل AraBert که بهترین عملکرد را داشته است confusion matrix زیر را داریم:



باتوجه به این تصویر میدانیم که مدل مجموعا توانسته است 431 مورد را به درستی تشخیص بدهد.

نتیجه‌گیری و کارهای آینده:

به صورت کلی مدل‌های transformer-based با استفاده از chunking افزایش عملکرد دارند به خصوص برای متون بسیار بلند این مورد قابل مشاهده است.

برای کارها و تحقیقات آینده نویسندگان پیشنهاد memory-augmented transformerها و روش‌های بهتر برای chunking و همچنین مدل‌های hierarchial را میدهند.

محدودیت‌ها :

- دیتاست نسبتا کوچک با تنوع موضوعی کم
- مدل‌های بررسی شده میتوانند افزایش کنند
- متدهای chunking پیشرفته‌تری وجود دارند که میتوانند باعث بهبود عملکرد بشوند.