



جوانہ

چالش سلامت

امیر محمد علی آبادی

روش پیشنهادی

از آنجایی که این چالش به صورت یک چالش واقعی و کاربردی مطرح شده است برای حل آن سعی شده است که راه حل نیز کاملاً کاربردی باشد.

یکی از راه های رسیدن به این نتیجه، بررسی رویکرد انسان در مواجهه با این چالش است.

مغز انسان برای درک توهین بودن یا نبودن یک جمله تنها به ساختار جمله و الفاظی که در جمله به کار رفته است نگاه میکند. همانطور که ناظرها آن را انجام میدهند.

به همین دلیل رویکردی که برای حل این چالش انتخاب کردم، استفاده از مدل های زبانی بود. به طوری که در نهایت مدل هوش مصنوعی ما همانند انسان تنها با خواندن جمله بتواند تشخیص دهد که آن جمله توهین بوده است یا خیر. و دیگر ویژگی های موجود در دیتاست را مدنظر قرار نخواهد داد.

تا امروز مدل های زبانی و روشهای گوناگونی برای فهم جمله و متن ارائه شده است که جدیدترین و به روزترین متدهای در حال حاضر دنیا برای متن کاوی، ترنسفورمر لرنینگ (Transformers learning) می باشد.

به همین سبب من هم مدل زبانی BERT که بر پایه ترنسفورمرها هستند را انتخاب کردم و مدل خود را بر پایه و اساس آن پیش برده ام که در ادامه به توضیح دقیق و نتایج به دست آمده از آن می پردازیم.

در ابتدا توضیح مختصری راجع به برت و نحوه عملکرد آن خواهیم داد:

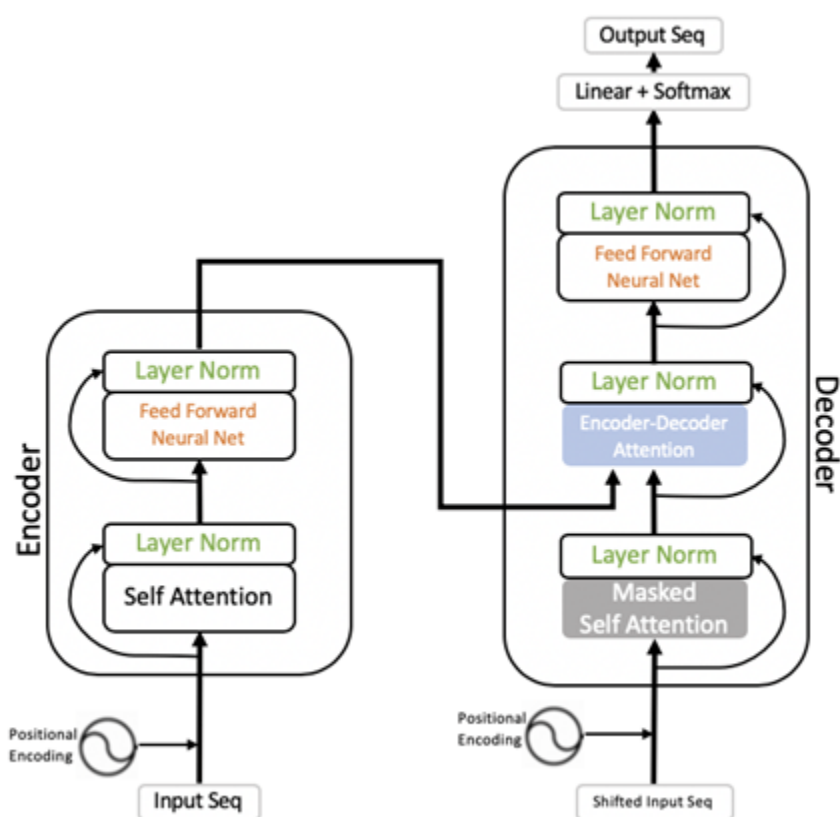
BERT چیست؟

Bert مخفف عبارت Bidirectional Encoder Representations from Transformers است. این الگوریتم به دنبال درک بهتر زبان طبیعی و مفهوم جستجوی ما و تطابق آن با محتواهای مختلف است.

اگر بخواهیم الگوریتم برت را به زبان ساده تعریف کنیم، باید بگوییم این الگوریتم مدلی از زبان محاوره و روزمره است که به باقی الگوریتم های گوگل کمک میکند تا حتی کوچک ترین بخش های زبان محاوره مثل حروف اضافه را مانند یک انسان به خوبی درک کند.

ترانسفورمر

ترانسفورمر جز سازنده BERT است. ترانسفورمر یک توپولوژی شبکه است که با استفاده از شبکه های کاملاً متصل مبتنی بر توجه رمزگذار و رمزگشا را پیاده سازی و اجرا می کند.



ورودی ها تعبیه کلمات به همراه رمزگذاری مکانی است. ترانسفورمر با بهره گیری از خودتوجه تمامی کلمات را با در نظر گرفتن سایر کلمات حاضر در دنباله رمزگذاری می کند. در مرحله بعد، ترانسفورمر از طریق نرمال سازی لایه ورودی ها در سرتاسر ویژگی ها نرمال سازی می کند (توجه داشته باشید که کلمات به فضای

بردارای چندبعدی نگاشت شده‌اند). در رمزگشا ورودی را با استفاده از یک کلمه تغییر می‌دهیم، در نتیجه می‌توانیم ترانسفورمر را آموزش دهیم که چگونه می‌تواند کلمه بعدی در جمله را پیش‌بینی کند. برای پیش‌بینی کلمه، ورودی رمزگشا باید پنهان شود (در نتیجه کلماتی که باید پیش‌بینی کند را نمی‌بیند)، پس رمزگشا از یک خودتوجه پوشیده برای ورودی‌های خود استفاده می‌کند. توجه رمزگذار و رمزگشا پیش از فرایند پیش‌بینی، اطلاعات بافت را ترکیب می‌کند. در آخر، رمزگشا یک لایه نرمال‌سازی دیگر خواهد داشت و خروجی‌ها نرمال‌سازی می‌شوند و در نتیجه نرمال‌سازی خروجی‌ها چندین ترانسفورمر می‌توانند با یکدیگر پشته شوند. ترانسفورمرها جز سازنده شبکه‌های مدرن NLU هستند چرا که تمامی مزایا و نکات مثبت سه توپولوژی شبکه را با یکدیگر در خود جمع کرده‌اند. ترانسفورمرهای مبتنی بر شبکه توانمند اما هزینه‌بر هستند؛ این شبکه‌ها معمولاً فضای زیادی از حافظه را اشغال می‌کنند و توان محاسباتی زیادی برای آموزش و همچنین استنتاج استفاده می‌کنند.

مدل های پیاده سازی شده:

1. مدل دپ لرنینگ با count vectorizing:

در این مدل ابتدا لیستی از تعداد کلمات واحد و تعداد تکرار آن ساخته شد. استاپ ورد های هر جمله حذف شد و در نهایت بردار جمله آن ساخته شد. در ادامه با توجه به دیتاست 7 ویژگی به بردار کلمات اضافه شد و در نهایت بردار وارد یک شبکه کانوولوشن با دو لایه کانوولوشن و 3 لایه شبکه عصبی تمام متصل آموزش دیده شد. 7 ویژگی اضافه شده از دیتاست به شرح زیر می باشد:

1. آی دی دسته بندی آگهی
2. تعداد کلمات تیتل آگهی
3. تعداد کلمات متن آگهی
4. تعداد عکس استفاده شده در آگهی
5. مخفی بودن یا نبودن شماره تلفن
6. تعداد جملات رد و بدل شده در چت
7. گزارش شده یا نشده بودن آگهی

نتایج به دست آمده از این روش در انتهای گزارش در جدول نمایش داده شده است.

2. مدل زبانی BERT:

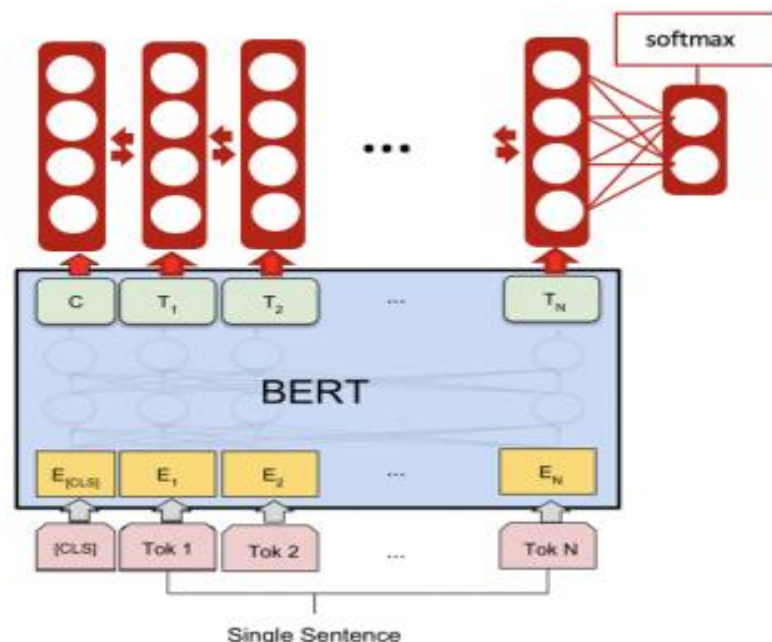
برای پیاده سازی این مدل از مدل های pretrained استفاده شده است. تنها ورودی این مدل مجموع چت های صورت گرفته می باشد. کاراکترهای اضافی از متن پاک شده و برای کاهش خطا در تشخیص کلمات تمامی حروفی که احتمال تایپ شدن آن با کیبورد عربی بود با حروف فارسی جابه جا شده است.

بهترین مدل pretrain فارسی مربوط به مجموعه هوشاره می باشد که در سه نسخه منتشر شده است. که هر سه نسخه آن مورد آزمایش قرار گرفته و نتایج آن ثبت شده است که بهترین آن مربوط به نسخه 2 می باشد.

3. تغییر لایه آخر مدل BERT:

در حالت معمولی خروجی ترانسفورمر به یک لایه شبکه عصبی تماما متصل وصل شده و نتیجه مشخص میشود.

در این مدل با اضافه کردن یک لایه کانولوشن و سه لایه شبکه عصبی تماما متصل؛ مدل جزئیات بیشتری را یاد بگیرد و دقت آن افزایش یابد که همانطور هم شد.



جدول نتایج :

ردیف	مدل	دقت (F1) در آموزش	دقت (AUC) در کوئرا
1	Count vectorizing	72.46	74.72
2	BERT-parsbert V1	86.61	88.11
3	BERT-parsbert V2	89.43	90.37
4	BERT-parsbert V3	88.59	89.5
5	BERT + CNN	90.21	91.71

نکته ای که باید به آن اشاره شود میزان سخت افزار مورد نیاز برای آموزش شبکه برت می باشد. با توجه به اینکه این پردازش ها بسیار زیاد می باشد نیازمند سخت افزار خوبی هم است. این نتایج با یک لبتاپ با مشخصات رم 8 گیگ و یک GPU 4 گیگ بدست آمده است که از حداکثر توان آن استفاده شده است.

حداکثر میزان batch که پاسخگوی سخت افزار من بود 5 بود.

لذا برای آموزش و تست هریک از مدل های بالا زمان زیادی صرف شده است. و طبیعتا هرچه سخت افزار قوی تری موجود باشد نتایج بدست آمده بهبود پیدا خواهند کرد.

*سخن آخر:

درست است که نتایج بدست آمده در جدول امتیازات بالاتر از این مقدار هم وجود دارد اما همانطور که ابتدای این گزارش بیان شد؛ هوش مصنوعی آمده است تا شبیه ساز مغز و رفتار انسان باشد و در این حوزه می بایست بر لبه تکنولوژی حرکت کرد. پس روی عقیده خودم ماندم و از متودولوژی ترنسفورمر ها استفاده کردم؛ تا مدل همانند مغز انسان تنها با خواندن جمله تشخیص دهد که جمله توهین آمیز است یا خیر.