خلاصه‌سازی اخبار

نفیسه نیک اقبال[۱]، آرش عسگری[۲]، سید بهداد عبدالهی مقدم[۳]، الیاس اسماعیلی[۴]
هادی حاجی حسینی[۵]
[۱]دانشجوی کارشناسی ارشد مهندسی کامپیوتر، نرم افزار، دانشگاه صنعتی شریف، تهران، nafise.nikeghbal32@gmail.com
[۲]دانشجوی کارشناسی ارشد مهندسی کامپیوتر، نرم افزار، دانشگاه صنعتی شریف، تهران، ArashAsgari1378@gmail.com
[۳]دانشجوی کارشناسی ارشد مهندسی کامپیوتر، نرم افزار، دانشگاه صنعتی شریف، تهران، behdad.a.moghadam@gmail.com
[۴]دانشجوی کارشناسی ارشد مهندسی کامپیوتر، نرم افزار، دانشگاه صنعتی شریف، تهران، elyas.esmaeili1@gmail.com
[۴]دانشجوی کارشناسی ارشد مهندسی کامپیوتر، نرم افزار، دانشگاه صنعتی شریف، تهران، m.hadi.hajihosseini@gmail.com

چکیده

خلاصه‌سازی انتزاعی متن یکی از حوزه‌هایی است که تحت تأثیر پیدایش مدل‌های زبانی از پیش آموزش‌دیده قرار گرفته است. کارهای از پیش آموزش دیده کنونی در خلاصه‌سازی انتزاعی به خلاصه‌هایی که حاوی کلمات مشترک بیشتری با متن اصلی باشند امتیاز بیشتری می‌دهند و کمتر به شباهت معنایی جملات تولید شده با سند اصلی توجه می‌کنند.

ما در این پروژه از مدل‌های مختلف استفاده کردیم و هر کدام از این مدل‌ها معماری مخصوص به خودشان را دارند که روی داده‌های خبری آموزش دیده‌اند. ما مدل‌های پیشنهادی خود را توسط دو معیار ROUGE و BERTScore بر روی داده‌های مختلف خبری ارزیابی کردیم و توانستیم با توجه به محدودیت‌های موجود به عملکردی خوبی برسیم.

کلمات کلیدی

خلاصه‌سازی انتزاعی، مدل‌های زبانی، ROUGE، BERTScore

۱- مقدمه

خلاصه کردن یکی از چالش‌های مهم در زبان طبیعی است و هدف تولید یک نمایش فشرده از یک متن ورودی است که شامل معنای اصلی متن ورودی است. اکثر سیستم‌های خلاصه‌سازی از روش‌های استخراجی[۱] استفاده می‌کنند که بخش‌هایی از متن اصلی را بر اساس اهمیتی که دارند استخراج می‌کنند و به هم متصل می‌کنند تا یک نسخه فشرده تولید شود. در مقابل وظیفه خلاصه‌سازی انتزاعی[۲] متن، تولید متنی کوتاه، روان و مختصر است که حاوی کلمات و عبارات بدیع غیر از سند اصلی باشد و موضوعات اصلی سند را حفظ کند.

با معرفی ترنسفرمرها (Vaswani et al, 2017) و تاثیر مثبتی که مدل‌های از پیش آموزش دیده[۳] روی کارهای مختلفی که در زمینه پردازش زبان‌های طبیعی انجام شده است گذاشتند باعث state-of-the-art خیلی از روش‌ها با استفاده از مدل‌های از پیش آموزش دیده و ترنسفرمرها شدند. در خلاصه‌سازی انتزاعی نیز به نتایج بسیار خوبی با استفاده از مدل‌های از پیش آموزش دیده و ترنسفرمرها رسیدند که بهتر از روش های قبلی بود (Liu and Lapata, 2019; Zhang et al., 2020; Qi et al., 2020) اما کارهای از پیش آموزش دیده به خلاصه‌هایی که کلمات مشترک بیشتری با متن اصلی داشته باشند امتیاز بیشتری می‌دهد.

زبان فارسی یکی از ۲۵ زبان برتر دنیا است ولی تعداد خیلی محدودی ریسرچ در زمینه خلاصه‌سازی متن فارسی وجود دارد و اکثر آن‌ها خلاصه‌سازی به شیوه استخراجی هستند و ما در این پروژه سعی کردیم که مدل‌های مختلفی که وجود دارد را برای خلاصه‌سازی انتزاعی اخبار تنظیم-دقیق[۴] کرده و مورد ارزیابی قرار داده تا بتوانیم مقایسه بین روش‌های مختلف انجام داده به یک جمع‌بندی جامع برسیم. در نهایت در هر مدل به عملکرد خوبی در خلاصه‌سازی متن اخبار رسیدیم و این مدل‌ها را با استفاده از دو معیار ROUGE (Lin, 2004) و BERTScore (Zhang et al., 2020b) بر روی مجموعه دادگان مختلف ارزیابی کردیم.

ما مشارکت‌های زیر را در این پروژه انجام دادیم:

- ابتدا یک مجموعه دادگان جامع از خبرگزاری‌های مختلف و معروف فارسی به همراه خلاصه آن‌ها جمع-آوری کرده و پیش پردازش‌های لازم را انجام دادیم.

- تعدادی مدل با معماری‌های مختلف روی داده‌های خبرگزاری‌های مختلف آموزش داده و تنظیم دقیق کردیم و به دقت و عملکرد خوبی رسیدیم.
- مدل‌های آموزش دیده به همراه دیتاست جمع آوری شده را با استفاده از Docker و Django و سایت HuggingFace در دسترس عموم قرار دادیم.

در این گزارش در بخش ۲ ابتدا با کارهای گذشته که در زمینه خلاصه‌سازی متن اخبار انجام شده است آشنا می‌شویم و سپس در بخش ۳ روند اصلی کار را ذکر کردیم که شامل زیر بخش های مختلفی است که در زیر بخش داده‌های پروژه مشخص می‌کنیم از چه داده‌هایی استفاده نمودیم و داده چگونه جمع‌آوری شده است، در زیر بخش تحلیل کاوشگرانه داده در رابطه با داده‌ها اطلاعات آماری را مشخص می‌کنیم، در زیر بخش تمیز کردن داده‌ها با پیش پردازش-هایی که جهت تمیز کردن داده‌ها انجام شده است آشنا می‌شویم، در زیربخش مدل سازی با معماری و عملکرد مدل‌هایی که در پروژه استفاده نمودیم آشنا می‌شویم، در زیر بخش ارزیابی مدل‌ها را ارزیابی کرده و نتایج ارزیابی را مشخص نمودیم، در زیر بخش کارهای آینده کارهایی که می‌توان در آینده برای بهبود پروژه انجام داد معرفی شده است، در بخش نتیجه‌گیری هم یک نتیجه کلی از پروژه را ذکر کردیم و در آخر هم مراجع و مقالاتی که برای این پروژه استفاده و مطالعه شده‌اند را ذکر نمودیم.

۲- کارهای گذشته

پس از معرفی ترنسفورمرها و تولید مدل زبانی ماسک شده‌ی Bert[۶] (Delvin et al., 2019) ، بسیاری از مسائل حوزه پردازش زبان طبیعی، با بهره‌گیری از این روش‌ها و معماری‌های از پیش یادگرفته شده، به پیشرفت گسترده‌ای دست پیدا کردند. به دنبال رویکرد Bert ، مدل‌های زبانی بسیاری (Liu et al., 2019; Joshi et al., 2020) با اندازه‌ی دادگان متفاوت جهت استفاده در پیش-یادگیری و برخی بهینه‌سازی‌ها روی روش پیش-یادگیری Bert، یادگرفته شدند. به علاوه، مدل‌های رمزگذار-رمزگشای[۷] با تلفیق مسائل پیش-یادگیری، یادگرفته شدند، از جمله؛ Bart (Lewis et al., 2020) ، T5 (Raffel et al., 2020) و MT5 (Linting et al., 2021) که نسخه‌ی چند زبانه‌ی T5 است و دادگانی که روی آن پیش-یادگیری صورت گرفته، شامل ۱۰۱ زبان می‌شود و زبان فارسی نیز مشمول آن می‌شود.

در زبان فارسی، روش‌های خلاصه‌سازی استخراجی کمی وجود ندارند (Khademi et al., 2018; Rezaei et al., 2019; Kermani and Fakhredanesh, 2020) ولی در زمینه‌ی خلاصه‌سازی انتزاعی، تلاش زیادی صورت نگرفته است. Farahani et al. (2020b) از ParsBert (Farahani et al., 2020a) به کمک روش Rothe et al. (2020) و وزن‌های از پیش-یادگیری شده برای بخش رمزگذاری-رمزگشای، جهت یادگیری مدل جدید توالی به توالی[۸] استفاده کرد. از دیگر کارها در این راستا، می‌توان به ARMAN (Salemi et al., 2021) اشاره کرد. ARMAN یک مدل رمزگذار-رمزگشای مبتنی بر ترنسفورمر است که از مدل‌های از پیش-یادگیری شده، مانند PEGASUS (Zhang et al., 2020a) که مخصوص خلاصه‌سازی دادگان متون خبری بزرگ، یادگرفته شده‌است. مدل از پیش-یادگیری شده‌ی دیگر، STEP (Zou et al., 2020) است که با اهداف مدل زبانی ماسک شده، تولید جمله‌ی پسین[۹] و ضبط جمله ناد[۱۰] یادگرفته شده‌است.

۳- روند اصلی

۳-۱- داده‌های پروژه

- داده BBC

سایت فارسی BBC، زیرمجموعه‌ای از سرویس جهانی BBC است که به زبان فارسی از طریق رادیو، اینترنت و تلویزیون، فعالیت خبری و رسانه‌ای می‌کند سایت خبری BBC یکی از سایت‌های بسیار معروف اخبار است که از سال ۱۳۷۹ راه‌اندازی شده است و ما از دادگان خبری BBC برای مدل خود استفاده کردیم و این داده‌ها از این لینک قابل دسترس است و می‌توان آن‌ها دانلود نموده و استفاده کرد که در این داده‌ها ما فقط به متن خبر و خلاصه آن نیاز داریم. مدل mT5_multilingual_XLSum که به نتایجی خوبی در زمینه خلاصه‌سازی خبر دست یافته است از این دادگان برای آموزش مدل خود استفاده کرده است.

- داده تبیان

مؤسسه فرهنگی تبیان وابسته به «سازمان تبلیغات اسلامی» یکی از بزرگ‌ترین و شناخته‌شده‌ترین مؤسسات فرهنگی ایران است که در زمینه‌های مختلف برای حمایت از جشنواره‌های فرهنگی و پخش با سایر مؤسسات فرهنگی همکاری داشته است و روزانه یک میلیون و ۶۰۰ هزار کاربر از وب سایت آن بازدید می کنند و ما این مجموعه داده از این لینک دریافت کردیم که در اینجا نیز ما فقط به متن خبر و خلاصه آن نیاز داریم. مدل ARMAN که نتایج خوبی را در زمینه خلاصه‌سازی متن خبر فارسی دریافت کرده است از این مجموعه داده برای بخشی از آموزش مدل خود استفاده نموده است.

- دادگان pn-summary

این مجموعه داده توسط شرکت هوشواره، از شرکت‌های دانش‌بنیان ایرانی، تهیه و تنظیم شده است. این مجموعه

داده از اخبار موجود در وبسایت خبرگزاری‌های فارسی متفاوتی نظیر تحلیل بازار، ایمنا، شانا، مهر، ایرنا وخبر آنلاین بدست آمده است. همچنین اخبار موجود در این مجموعه داده شامل اخبار اقتصادی، راه‌سازی، بانکداری، کشاورزی، بین‌المللی، انرژی، صنعتی، حمل و نقل، تکنولوژی، محلی، ورزشی، سیاسی، فرهنگی، اجتماعی، سلامت، دانشگاه، توریستی و پژوهشی می‌شود. در موارد متعددی مدل‌های خلاصه‌سازی فارسی با عملکرد خوبی با استفاده از این داده آموزش دیده اند. یکی از این موارد می‌توان به استفاده از مدل ParsBert به عنوان کدکننده[11] و کدگشا[12]برای ایجاد یک مدل خلاصه‌کننده اشاره کرد. (Farahani et al., 2021)

- داده تسنیم

خبرگزاری تسنیم یک خبرگزاری خصوصی است که در حوزه‌های گوناگون اعم از سیاسی، فرهنگی، اجتماعی، اقتصادی، ورزشی، بین‌الملل، رسانه و ... فعالیت دارد. ما این دادگان را از طریق خزش بر روی بخش‌های مختلف سایت بدست آوردیم. از هر حوزه ۱۰۰ صفحه را که شامل متن خبر، تیتر خبر و خلاصه خبر بود را خزش کردیم، که ما فقط به متن خبر و خلاصه آن احتیاج داریم. برای خزش از زبان Python و کتابخانه‌ی Scrapy استفاده کردیم.

- داده عصر ایران

عصر ایران یک خبرگزاری بدون جناح خاص است و به عنوان یک خبرگزاری مردمی اقدام به فعالیت کرده است.بخشی از داده های ما برای آموزش دادن مدل از این سایت جمع آوری شده اند.

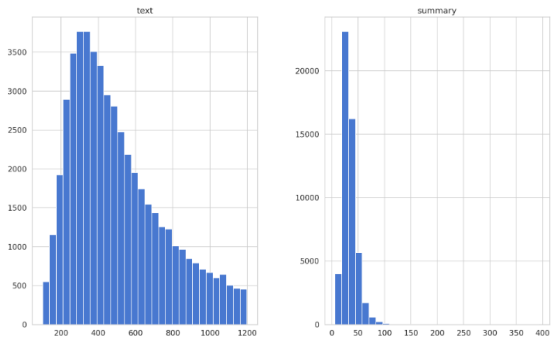
در اینجا ذخیره سازی در یک فایل csv انجام شده است. برای انجام این کار از فریمورک قدرتمند scrapy که تحت زبان پایتون می باشد استفاده شده است. این فریمورک بسیار قدرتمند است و برای انتخاب آن برای این پروژه تحقیق انجام شده است. بین استفاده از beautiful soup و selenium و همینطور scrapy بررسی انجام شد و به علت asynchronous بودن فرایند های scrapy و همینطور ساختار مشخص آن، این فریمورک برای این کار انتخاب شد. البته یک بار این خزشگر را با beautiful soup نیز پیاده کردیم ولی برای اطمینان از بازدهی بالاتر، در نهایت از فریمورک scrapy استفاده کردیم و تفاوت چشمگیری را مشاهده کردیم.

۲-۳- تحلیل کاوشگرانه داده‌ها

برای درک بهتر داده‌ها و آماده کردن آن‌ها برای مدل خود روی آن‌ها بررسی‌هایی انجام دادیم و اطلاعات آماری از دادگان خود استخراج نمودیم. در این بخش کارهای زیادی می‌توان انجام داد اما فقط به طور محدود کاوش‌های مختلفی را روی دادگان جمع‌آوری شده انجام داده‌ایم.

- دادگان BBC

ما برای تعداد کلمات متن خبر و خلاصه خبر نمودار رسم کردیم که نمودار رسم شده در شکل (۱) آورده شده است و این شکل به ما کمک می‌کند تا بدانیمه اخبار و خلاصه خبر حدودا از چه تعداد کلمه تشکیل شده‌اند.



شکل(۱) – تعداد کلمات خلاصه و متن خبر BBC

در مدل ما نیاز داریم که برای متن اخبار و متن خلاصه‌های خود بیشینه طول تعیین کنیم که برای اینکه مدل عملکرد بهتری داشته و دقت خوبی به ما بدهد در اینجا بررسی می‌کنیم که ۹۰ درصد دادگان چه طولی دارند و سپس بر همان اساس بیشینه طول را در مدل خود تعیین می‌کنیم.
در شکل (۲) نشان دادیم که ۹۰ درصد از دادگان BBC چه طولی دارند.

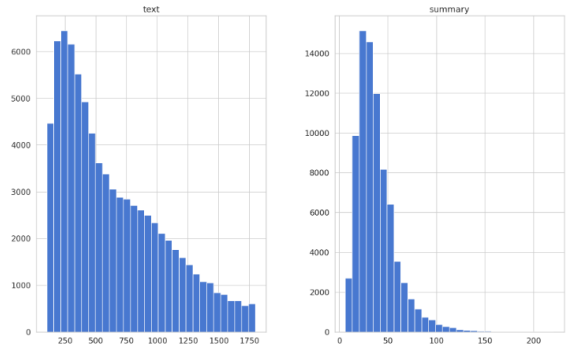
```
percentile 90 of length of news: 1286.0
longest sentence:  26573

percentile 90 of length of summaries: 54.0
longest sentence:  394
```

شکل(۲) – طول ۹۰ درصد از دادگان خبر و خلاصه BBC

ما در دادگان بعدی نیز دقیقا همین کاوش‌ها را انجام داده‌ایم.

- دادگان تبیان



شکل(۳)– تعداد کلمات خلاصه و متن خبر تبیان

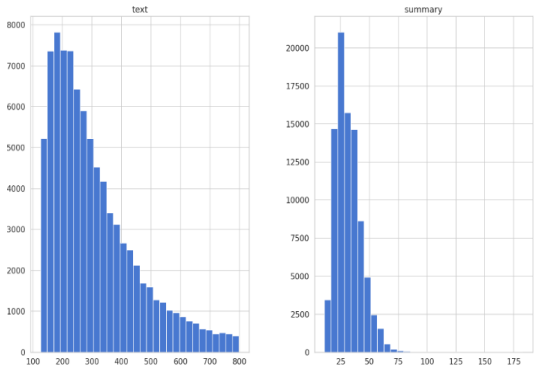
```
percentile 90 of length of news: 1706.0
longest sentence:  50148

percentile 90 of length of summaries: 65.0
longest sentence:  242
```

شکل(۴) – طول ۹۰ درصد از دادگان خبر و خلاصه تبیان

- دادگان pn-summary

مشابه مجموعه داده‌های قبل به منظور تعیین طول بیشینه بردارهای توکن نیاز به بررسی نمودار پراکندگی طول کلمات در متن خبر و خلاصه آن قابل مشاهده است. همانطور که قابل مشاهده است ۹۰ درصد اخبار موجود در این مجموعه داده دارای حداکثر طول ۵۸۳ کلمه هستند که این مجوعه داده را تبدیل به یکی از بهترین مجموعه دادگان موجود برای خلاصه‌سازی ماشینی می‌کند.



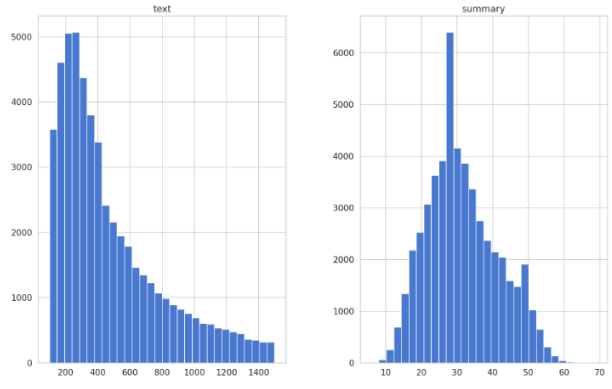
شکل(۵) – تعداد کلمات خلاصه و متن خبر pn-summary

```
percentile 90 of length of news: 621.0
longest sentence:  5449

percentile 90 of length of summaries: 48.0
longest sentence:  212
```

شکل(۶) – تعداد کلمات خلاصه و متن خبر pn-summary

- داده تسنیم



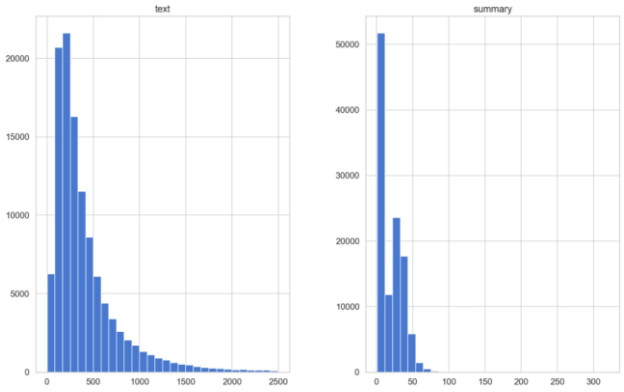
شکل(۷) – تعداد کلمات خلاصه و متن خبر تسنیم

```
percentile 90 of length of news: 1316.0
longest sentence:  19446

percentile 90 of length of summaries: 47.0
longest sentence:  69
```

شکل(۸) – تعداد کلمات خلاصه و متن خبر تسنیم

- داده عصر ایران



شکل(۹) – تعداد کلمات خلاصه و متن خبر عصر ایران

```
percentile 90 of length of news: 1015.0
longest sentence:  19069

percentile 90 of length of summaries: 52.0
longest sentence:  682
```

شکل(۱۰) – طول ۹۰ درصد از دادگان خبر و خلاصه عصر ایران

۳-۳- تمیز کردن داده ها

ما می‌دانیم که هر داده‌ای قبل از اینکه به مدل داده شود نیاز به پیش پردازش دارد که وابسته به پروژه‌ای که داریم این پیش پردازش‌ها می‌تواند متفاوت باشد.

ما ابتدا برای تمام دادگان فقط ستون متن اصلی خبر و ستون خلاصه خبر را نگه می‌داریم چون به بقیه ستون‌ها نیازی نیست و سپس سطرهایی که مقدار خالی یا تکراری دارند را حذف می‌کنیم و چون می‌خواهیم داده‌ها یکپارچگی بیشتری داشته باشند نام تمام ستون‌ها در تمام دادگان را یکسان می‌کنیم.

در این پروژه نمی‌توان علائم نگارشی را حذف نمود زیرا پایان جملات و علائم نگارشی دیگر مانند نقل قول، پرانتزها و... برای آموزش مدل‌ما اهمیت دارد و نمی‌توان آن‌ها حذف نمود ولی برای یکسان‌سازی تمامی حروف کلمات، آن دسته از حروفی که شکل نوشتاری آن‌ها به شکل عربی می‌باشد را یکسان و به شکل فارسی آن تبدیل می‌کنیم تا متن از نظر نوشتاری یکپارچه شود.

تعدادی کاراکتر مانند \u200e، \xad و... در متن وجود دارند که در پیش پردازش آن‌ها حذف می‌کنیم و هم چنین اگر فاصله و نیم فاصله‌ای اضافه در متن وجود دارد نیز باید آن‌ها را حذف نماییم تا در هنگام آموزش مدل مشکلی ایجاد نشود و دقت مدل پایین نیاید. مجدد پس از پیش پردازش چون ممکن است بعضی از سطرها خالی شده باشند مجدد سطرهایی که مقادیر خالی دارند را از داده‌های خود حذف می‌کنیم.

در نهایت داده‌های تمیز و پیش پردازش شده را در یک مجموعه دادگان جدید ذخیره می‌کنیم تا از آن‌ها در مدل خود استفاده نماییم.

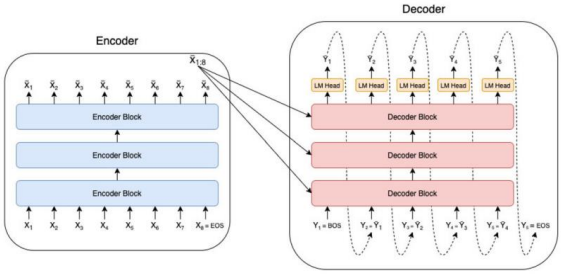
۴-۳- مدل‌سازی

BERT[۲]BERT -3-4-1

BERT یک شبکه عصبی ترنسفورمر دوسویه است که با در نظر گرفتن دو هدف مدل‌سازی زبانی نقاب‌دار[۴]و پیش‌بینی جمله بعدی، بر روی مجموعه وسیعی از داده‌های متنی آموزش داده شده است. مدلهای از پیش آموزش داده شده چند زبانه متعددی منتشر شده‌اند که از زبانهای مختلفی از جمله زبان فارسی پشتیبانی می‌کنند. علاوه بر مدلهای چند زبانه، مدلهای تک زبانه‌ای نیز وجود دارند که به طور خاص بر روی زبان فارسی آموزش داده شده‌اند.

برخلاف مدلهای دنباله به دنباله[۵]که از دو بخش رمزنگار[۶]و رمزگشا[۷]تشکیل شده‌اند، BERT به عنوان یک مدل فقط رمزنگار عمل می‌کند. شکل ۱۱ شبکه مبتی بر ترنسفورمر رمزنگار-رمزگشا دنباله به دنباله را در سطح بالا نشان می‌دهد. لایه‌های ترنسفورمر رمزنگار معمولا شامل اتصالات دوسویه‌ای[۸]هستند که شباهت زیادی به مدل BERT دارند. در حالیکه لایه‌های رمزگشا فقط شامل اتصالات یک طرفه (چپ به راست) می‌باشند. اگر چه BERT یک مدل فقط رمزنگار است، اما امکان این وجود دارد که از وزن مدلهای از پیش آموزش داده شده استفاده کرد تا یک مدل دنباله به دنباله ساخته شود که بخش‌های رمزنگار و رمزگشای آن توسط وزن‌های مدل از پیش آموزش داده شده مقداردهی می‌شود. برای دستیابی به این هدف با کمک BERT (۱) یک لایه توجه متقاطع[۹]که به صورت تصادفی مقدار دهی اولیه شده است بین لایه‌های خود-توجهی[۱۰]و لایه‌های پیشخور[۱۱]اضافه شده است (۲) لایه‌های خود-توجهی BERT در رمزگشا به لایه‌های خود-توجهی یک طرفه تبدیل شده‌اند (۳) یک لایه مدل زبانی به بالای لایه رمزگشا اضافه شده است تا یک تابع توزیع احتمال شرطی برای تولید خروجی تعریف شود. در نتیجه، وزن‌های از پیش آموزش داده شده به طور مستقیم به مدل رمزنگار-رمزگشای ساخته شده منتقل می‌شوند که تنها استثنا لایه‌های توجه متقاطع اضافی‌ای هستند که به طور تصادفی مقداردهی اولیه شده‌اند.

در این پژوهش از مدل BERT تک زبانه ParsBERT به عنوان مدل از پیش آموزش دیده شده برای آموزش یک مدل دنباله به دنباله جدید استفاده شده است که این مدل را ParsBERT-sum می‌نامیم.



شکل(۱۱) – یک شبکه رمزنگار-رمزگشا مبتنی بر ترنسفرمر

WikiBert2WikiBert -3-4-2

یکی از مدل‌های دیگری که مشابه معماری قبل می‌توان پیاده‌سازی کرد، مدل WikiBert است. این مدل برروی مجموعه دادگان ویکی‌پدیا آموزش دیده است. در این مدل از وزن‌های مدل آموزش دیده‌ی ParsBert برای معماری رمزنگار و رمزگشا استفاده کردیم و آن را برروی مجموعه دادگان pn-summary و BBC فارسی آموزش دادیم. این مدل را WikiBert2WikiBert می‌نامیم.

علاوه بر مدل WikiBert مدل‌های ترنسفرمر چندزبانه دیگری هستند که می‌توانند در معماری فوق به کار روند از جمله پرآوازه‌ترین این مدل‌ها مدل mbert است(Pyysalo et al., 2020). پس از پیشنهاد مدل WikiBert عملکرد این مدل با mbert[۱۲] مقایسه شد. برپایه مقا آن‌ها مدل WikiBert در زبان فارسی در معیار LAS امتیاز ۸۸,۶۰ را به خود اختصاص داده است که دو درصد بالاتر از مدل mbert با امتیاز ۸۶,۶۰ است. به همین دلیل تصمیم برآن شد تا از آن برای آموزش مدل خلاصه‌گر خود بهره ببریم.

در این میان پارامترهایی هستند که جهت آموزش مدل به صورت بهینه برروی مموری محدود کولب درنظر گرفته شده است و در جدول ۱ آورده شده است.

جدول ۱ - مقادیر پارامترهای مدل WikiBert2WikiBer

پارامتر	مقدار
Num_train_epochs	5

500	Warm_up_steps
1	Per_device_eval_batch_size
1	Per_device_train_bath_size
500	Eval_steps
16	Gradient_accumulation_steps

نتایج آموزش مدل در بخش ارزیابی ضمیمه شده است.

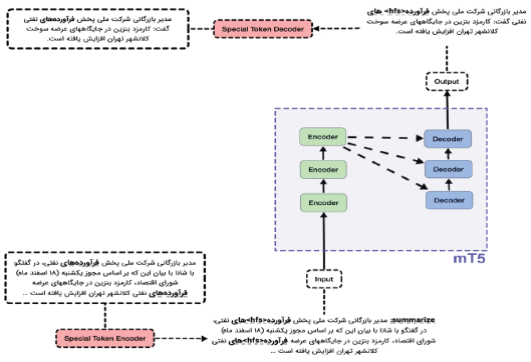
MT5 -3-4-3

یکی از مدل‌های بسیار معروفی که در زمینه متن به متن [۲۳] می‌توان استفاده نمود MT5 است که بر پایه T5 است ولی با این تفاوت که مدل MT5 از ۱۰۱ زبان پشتیبانی می‌کند که فارسی یکی از این زبان‌ها است. MT5 سعی کرده کمترین تغییر را در معماری T5 بدهد در نتیجه تمام مزایای آن را به ارث می‌برد.

با استفاده از MT5 کارهای مختلفی مانند ترجمه ماشینی، طبقه‌بندی، خلاصه‌سازی و... می‌توان انجام داد که یکی از معروف‌ترین مدل‌های ترنسفرمر که برای خلاصه-سازی استفاده می‌شود MT5 است.

MT5 خود سه نوع دارد که شامل MT5-small، MT5-base و MT5-large است. که در این پروژه به دلیل محدودیتی که در حافظه داشتیم فقط از MT5-small می‌توانستیم استفاده کنیم. MT5 گوگل روی مجموعه mc4 آموزش داده شده است. این مدل حتما قبل از اینکه برای یک کار مانند خلاصه‌سازی یا ترجمه یا... استفاده شود باید تنظیم-دقیق شود.

معماری MT5 مانند T5 است و یک ترنسفرمر رمزگذار-رمزگشا است و 3 هدف اصلی را دنبال می‌کند: ۱- مدلسازی برای پیش‌بینی کلمه بعدی، ۲- ترکیب کردن [۲۴]برای رسیدن به متن اصلی، ۳-پیش‌ینی کلمات پوشش[۲۵] داده شده و شکل(۱۲) معماری یک MT5 را نشان می-دهد(Farahani et al., 2020) در این مدل ابتدا متن ورودی با استفاده از یک توکن ساز رمزگذار داده می‌شود و سپس به یک رمزگذار سه لایه داده شده و حاصل آن مجدد به یک رمزگشا سه لایه داده می‌شود و در نهایت خروجی تولید شده با استفاده از یک توکن‌ساز رمزگشا تولید شده و به عنوان خروجی نهایی برگردانده می‌شود.



شکل(12) – معماری MT5

مدل MT5 نیز مانند بسیاری از مدل‌ها هایپر پارامترهای مخصوص به خودش را دارد که درجدول۲ هایپرپارامترهایی که در مدل خود استفاده کردیم به همراه مقادیر آن‌ها مشخص شده است:

جدول۲ - مقادیر پارامترهای مدل MT5

پارامتر	مقدار
#Beams	**2**
Length Penalty	**2.0**
Early Stopping status	**ACTIVE**
Repetition Penalty	**1.0**

در این مدل از دادگان pn-summary و بی بی سی برای آموزش استفاده شده است و دو دوره[۲۶]روی ترکیب این دو داده تنظیم دقیق شده است و سپس هر کدام از دادگان را دو دوره به صورت جداگانه به مدل داده شده است تا مجدد تنظیم دقیق شود.

۵-۳- کارهای آینده

از جمله کارهایی که می‌توان در آینده انجام داد؛ جمع‌آوری دادگان برای گویش‌ها و زبان‌های محلی ایران است و مسئله خلاصه‌سازی را با استفاده از آن‌ها انجام داد. همچنین می‌توان تنها به دادگان خبری بسنده نکرد و از دادگان ادبی و داستانی نیز برای پوشش بیشتر متون استفاده کرد. از طرفی با جمع‌آوری بیشتر دادگان، مدل‌ها دقت بیشتری بدست می‌آورند.

۶-۳- ارزیابی

جهت درنظر گرفتن امتیاز معنایی[۲۷]در محاسبه‌ی شباهت دو جلمه، از معیار BERTScore استفاده کردیم. BERTScore امتیاز شباهت را برای هر توکن در جمله‌ی کاندید با هر توکن در جمله‌ی مرجع، با استفاده از امبدینگ محتوایی[۲۸] محاسبه می‌کند. (Zhang et al., 2020b). برای یک جمله‌ی مرجع x و جمله کاندید x`، مقادیر یادآوری، دقت و امتیاز F1 به صورت زیر محاسبه می‌شوند.

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad ,$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \quad ,$$

$$F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad .$$

معیار دیگری که برای محاسبه شباهت بین جمله‌ی کاندید و دسته‌ای از جملات مرجع استفاده شد، ROUGE است. ROUGE شباهت را برا اساس N-gramهای روی هم قرار گرفته محاسبه می‌کند (Lin, 2004). هر چه امتیاز ROUGE بین دو متن بیشتر باشد، شباهت آن‌ها بیشتر است. ROUGE-N به صورت زیر محاسبه می‌شود.

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

در اینجا، n مخفف طول n-gram است، $gram_n$ و $Count_{match}(gram_n)$ به ترتیب بیشینه تعداد رخداد n-gram در یک خلاصه‌ی کاندید و یک مجموعه از خلاصه‌های مرجع است.

در معیار ROUGE-L، L مخفف طولانی‌ترین زیرتوالی مشترک[۱] است. یک توالی $Z = [z_1, z_2, \ldots, z_n]$ یک زیرتوالی از یک توالی دیگر $X = [x_1, x_2, \ldots, x_n]$ است، اگر یک توالی اکیدا صعودی از اندیس‌های X به صورت $[i_1, i_2, \ldots, i_n]$ که برای هر $j = 1, 2, \ldots, k$ داشته باشیم، $x_{ij} = z_j$ (Cormen et al., 1989).

ParsBERT-sum -3-6-1

ارزیابی کمی:

برای ارزیابی کمی مدل‌های معرفی شده از معیار ROUGE استفاده شده است که یک معیار ارزیابی رایج در خلاصه سازی متن می‌باشد. ROGUE1، ROUGE2 و ROUGE-L برای مدل‌ها گزارش شده است. معیار ROUGE-n میزان غنی بودن خلاصه تولید شده توسط مدل را با شماردن تعداد n-gramهای مشترک میان خلاصه‌ی تولید شده و خلاصه مرجع محاسبه می‌کند. ROUGE-L تعداد n-gramهای همپوشان را براساس طولانی‌ترین زیردنباله‌های مشترک محاسبه می‌کند و میزان روان بودن خلاصه‌های تولید شده را نشان می‌دهد.

نتایج حاصل از ارزیابی مدل ParsBERT-sum بر روی خبرهایی که به صورت تصادفی از مجموعه دادگان خبرگزاری تسنیم و عصر ایران به دست آمده‌اند در جدول ۳ و ۴ مشاهده می‌شود.

جدول ۳ - نتایج ارزیابی مدل ParsBERT-sum

Model	R1	R2	RL
ParsBERT-sum	۳۲/۸۵	۱۷/۴۵	۲۸/۵

جدول۴- نتایج ارزیابی مدل ParsBERT-sum توسط معیار BERTScore

Model	Precision	Recall	F1
ParsBERT-sum	۰/۷۴	۰/۷۲	۰/۷۳

ارزیابی کیفی:

در این بخش به ارزیابی کیفی مدل‌های آموزش داده شده پرداخته شده است. برای ارزیابی میزان غنی و روان بودن خلاصه‌های تولید شده توسط مدل‌های مختلف تعدادی نمونه به عنوان ورودی به مدل‌ها داده شده است و خروجی آنها با چکیده اصلی خبر مقایسه شده است. این ارزیابی این امکان را فراهم می‌کند که بتوانیم میزان کارایی مدل را بر روی داده‌هایی که از قبل دیده نشده‌اند بررسی کنیم.

متن خبر
اداره املاک و مستغلات دوبی از جریمه ۵۰ هزار درهمی برای درج آگهی اجاره واحدهای مسکونی و تجاری در شیخ نشین دوبی خبر داد. به گزارش ایسنا، اداره املاک و مستغلات دوبی در اطلاعیه‌ای اعلام کرد که انتشار آگهی اجاره واحدهای مسکونی و تجاری در روزنامه‌ها و نشریات و حتی در سطح شهر باید با مجوز این اداره صورت گیرد و در غیر این صورت فرد و یا افراد اگهی دهنده به پرداخت ۵۰ هزار درهم جریمه نقدی محکوم خواهند شد. بر اساس این اطلاعیه، اداره املاک و مستغلات دوبی تمامی قیمت اجاره‌های مسکونی و تجاری به منظور ثبات و کنترل قیمت اجاره‌ها را مورد بررسی قرار می‌دهد و از افزایش بی‌رویه اجاره‌ها جلوگیری خواهد کرد.
چکیده
اداره املاک و مستغلات دوبی تمامی قیمت اجاره‌های مسکونی و تجاری به منظور ثبات و کنترل قیمت اجاره‌ها را مورد بررسی قرار می‌دهد و از افزایش بی‌رویه اجاره‌ها جلوگیری خواهد کرد.
خلاصه تولید شده توسط مدل ParsBERT-sum
اداره املاک و مستغلات امارات تمامی قیمت اجاره واحدهای مسکونی و تجاری به منظور ثبات و کنترل قیمتها را مورد بررسی قرار میدهد.

WikiBert2WikiBert -3-6-2

ارزیابی کمی:

پس از آموزش مدل با استفاده از دو مجموعه داده pn-summary و bbc بخشی از داده که به عنوان داده تست توسط مدل دیده نشده است را به عنوان ورودی به مدل داده و مدل را آزمودیم. داده تست شامل ۵۰۰ داده از مجموعه داده pn-summary و ۵۰۰ داده از مجموعه داده bbc می‌شود. نتایج این بررسی در جداول ۵ و ۶ موجود است.

جدول۵– نتایج ارزیابی معیار rouge در مدل WikiBert2WikiBert

	Rough-1	Rough-2	Rough-l
Pn-summary	35.51	15.65	30.91
BBC	32.75	9.72	27.80

جدول۶– نتایج ارزیابی معیار BERTScore در مدل mt5

	precision	recall	F1
Pn-summary	72.86	78.46	75.50

BBC	71.17	74.98	72.98

ارزیابی کیفی:

متن خبر
به گزارش شانا، علی کاردر امروز (۲۷ دی ماه) در مراسم تودیع محسن قمصری، مدیر سابق امور بین الملل شرکت ملی نفت ایران و معارفه سعید خوشرو، مدیر جدید امور بین الملل این شرکت، گفت: مدیریت امور بین الملل به عنوان یکی از تاثیرگذارترین مدیریت های شرکت ملی نفت ایران در دوران تحریم های ظالمانه غرب علیه کشورمان بسیار هوشمندانه عمل کرد و ما توانستیم به خوبی از عهده تحریم ها برآییم. وی افزود: مجموعه امور بین الملل در همه دوران ها با سختی ها و مشکلات بسیاری مواجه بوده است، به ویژه در دوره اخیر به دلیل مسایل پیرامون تحریم وظیفه سنگینی بر عهده داشت که با تدبیر مدیریت خوب این مجموعه سربلند از آن بیرون آمد. کاردر با قدردانی از زحمات محسن قمصری، به سلامت مدیریت امور بین الملل این شرکت اشاره کرد و افزود: محوریت کار مدیریت اموربین الملل سلامت مالی بوده است. وی بر ضرورت نهادینه سازی جوانگرایی در مدیریت شرکت ملی نفت ایران تاکید کرد و گفت: مدیریت امور بین الملل در پرورش نیروهای زبده و کارآزموده آنچنان قوی عملکرده است که برای انتخاب مدیر جدید مشکلی وجود نداشت. کاردر، حرفه ای گری و کار استاندارد را از ویژگی های مدیران این مدیریت برشمرد و گفت: نگاه جامع، خلاقیت و نوآوری و بکارگیری نیروهای جوان باید همچنان مد نظر مدیریت جدید امور بین الملل شرکت ملی نفت ایران باشد.
چکیده
مدیرعامل شرکت ملی نفت، عملکرد مدیریت امور بین الملل این شرکت را در دوران تحریم بسیار هوشمندانه خواند و گفت: امور بین الملل در دوران پس از تحریم ها نیز می تواند نقش بزرگی در تسریع روند توسعه داشته باشد.
خلاصه تولید شده توسط مدل WikiBert2WikiBert
مدیرعامل شرکت ملی نفت ایران گفت : مدیریت امور بین الملل در پرورش نیروهای جوان و ماهر بسیار هوشمندانه عمل میکند و باید به خوبی از عهده این مدیریت براید و بکارگیری نیروهای جوان در این مدیریت را در دستور کار قرار دهد و به خوبی عمل کرد. وی گفت : رویکرد جامع ، خلاقیت و نواوری در مدیریت امور بینالمللی سلامت مالی بوده است.

mt5 -3-6-3

ارزیابی کمی:

در بخش ارزیابی مدل mt5 ما به ۳ صورت نتایج را ارزیابی کردیم: ابتدا مدل را روی ۵۰۰ داده تصادفی از دادگان بی بی سی مورد ارزیابی قرار دادیم، سپس روی ۵۰۰ داده تصادفی از دادگان pn-summary و یکبار هم روی ۱۰۰۰ داده تصادفی از ترکیب این دو داده ارزیابی کردیم که نتایج را در جدول۷ و ۸ گزارش شده است.

جدول۷– نتایج ارزیابی معیار rough در مدل mt5

	Rough-1	Rough-2	Rough-l
Pn-summary	31.61	14.64	26.27
BBC	22.98	6.17	18.48
Merged datasets	۲۵,۵۳	۸,۶۶	۲۰,۳۸

جدول۸– نتایج ارزیابی معیار BERTScore در مدل mt5

	precision	recall	F1
Pn-summary	75.53	72.21	73.76
BBC	72.76	69.53	71.06
Merged datasets	۷۳,۷۷	۷۰,۰۴	۷۱,۷۹

ارزیابی کیفی:

متن خبر
به گزارش شانا، علی کاردر امروز (۲۷ دی ماه) در مراسم تودیع محسن قمصری، مدیر سابق امور بین الملل شرکت ملی نفت ایران و معارفه سعید خوشرو، مدیر جدید امور بین الملل این شرکت، گفت: مدیریت امور بین الملل به عنوان یکی از تاثیرگذارترین مدیریت های شرکت ملی نفت ایران در دوران تحریم های ظالمانه غرب علیه کشورمان بسیار هوشمندانه عمل کرد و ما توانستیم به خوبی از عهده تحریم ها برآییم. وی افزود: مجموعه امور بین الملل در همه دوران ها با سختی ها و مشکلات بسیاری مواجه بوده است، به ویژه در دوره اخیر به دلیل مسایل پیرامون تحریم وظیفه سنگینی بر عهده داشت که با تدبیر مدیریت خوب این مجموعه سربلند از آن بیرون آمد. کاردر با قدردانی از زحمات محسن قمصری، به سلامت مدیریت امور بین الملل این شرکت اشاره کرد و افزود: محوریت کار مدیریت اموربین الملل سلامت مالی بوده است. وی بر ضرورت نهادینه سازی جوانگرایی در مدیریت شرکت ملی نفت ایران تاکید کرد و گفت: مدیریت امور بین الملل در پرورش نیروهای زبده و کارآزموده آنچنان قوی عملکرده است که برای انتخاب مدیر جدید مشکلی وجود نداشت. کاردر، حرفه ای گری و کار استاندارد را از ویژگی های مدیران این مدیریت برشمرد و گفت: نگاه جامع، خلاقیت و نوآوری و بکارگیری نیروهای جوان باید همچنان مد نظر مدیریت جدید امور بین الملل شرکت ملی نفت ایران باشد.
چکیده
مدیرعامل شرکت ملی نفت، عملکرد مدیریت امور بین الملل این شرکت را در دوران تحریم بسیار هوشمندانه خواند و گفت: امور بین الملل در دوران پس از تحریم ها نیز می تواند نقش بزرگی در تسریع روند توسعه داشته باشد.
خلاصه تولید شده توسط مدل MT5
مدیریت امور بین الملل در همه دوران ها با سختی ها و مشکلات بسیاری مواجه بوده است، به ویژه در دوره اخیر به دلیل مسایل پیرامون تحریم وظیفه سنگینی بر عهده داشت.

۴- رابط گرافیکی و نسخه دمو

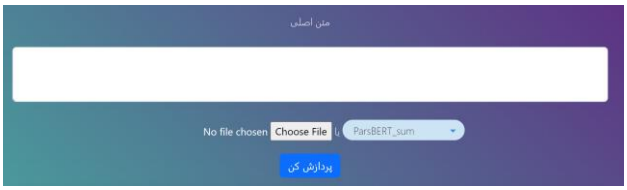
نرم افزار تحت وب تولید شده، برای پردازش مستقیم داده های متنی توسط مدل های train شده یادگیری ماشین برای تولید خلاصه اخبار ورودی می باشد، به این صورت که ورودی متنی را از کاربر دریافت می کند و پردازش هایی را بر روی آن انجام داده و سپس خروجی را به کاربر تحویل می‌دهد.

برای ساخت این برنامه تحت وب، از Django استفاده شده است. Django فریمورکی قدرتمند مبتنی بر زبان برنامه نویسی محبوب Python می‌باشد که امنیت و سرعت توسعه با آن و مقیاس پذیری محصولات نرم افزاری تولید شده با آن زبانزد دنیای اهل فن است. لازم به ذکر است که هر سایتی برای ذخیره سازی داده های خود نیاز به

یک پایگاه داده دارد، در این پروژه از پایگاه داده‌ی محبوب و قدرتمند PostgreSql استفاده شده است.

۴-۱- استفاده از دموی وبسایت برای پردازش

بدیهی است که برای دریافت اطلاعات از کاربر نیاز به یک فرم داریم. فرم استفاده شده در این پروژه را در شکل (۱۲) مشاهده می کنیم.



شکل (۱۲) — فرم دموی وب

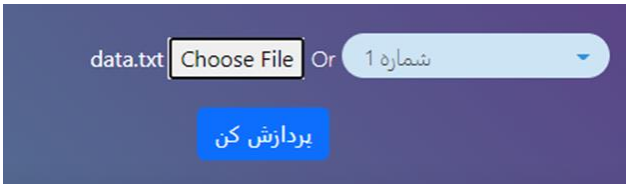
فرم ذکر شده می تواند اطلاعات متنی را هم از طریق کادر متن و هم از طریق انتخاب فایل متنی دریافت کند، منوی کشویی نیز برای انتخاب مدل یادگیری ماشین مورد نظر جهت پردازش متن تعبیه شده است و دکمه ای هم جهت ارسال اطلاعات جهت پردازش قرار داده شده است.

در ادامه به چند نمونه از ورودی و خروجی می پردازیم. شکل (۱۳) نمونه ای از ورودی برای پردازش با مدل انتخابی است. شکل (۱۴) خروجی آن پردازش است که می‌بینیم جمله خلاصه تولید شده به همراه زمان پردازش آن را نشان می‌دهد.

شکل (۱۳) — نمونه ورودی از طریق جعبه دریافت متن

شکل (۱۴) — نمونه خروجی

همچنین برای ایجاد امکانات بیشتر برای سایت، امکان ورودی دادن متن به کمک فایل نیز مهیا شده است به صورتی که اگر فایلی را با پسوند txt. که شامل متن ورودی است به فرم بدهیم متن آن را به مدل داده و پردازش لازم را انجام می‌دهد. شکل (۱۵) نمونه ورودی با فایل و شکل (۱۶) متن داخل فایل را نشان می‌دهد. بدیهی است که خروجی این روش نیز مانند شکل (۱۴) است.



شکل (۱۵) — ارسال متن ورودی از طریق فایل

همچنین زیر فرم، جدولی جهت نمایش اطلاعات پردازش های گذشته تدارک دیده شده تا از طریق آن اطلاعات مربوط به پردازش های گذشته را بتوان مشاهده کرد و تاریخچه ای از پردازش ها را در اختیار داشت. نمونه جدول نمایش تاریخچه متون ورودی و خروجی را در شکل (۱۷) مشاهده میکنید.

متن اصلی	شماره مدل nlp	خلاصه	Processing time
به گزارش شانا، علی کاردر امروز (۲۷ دی ماه) در مراسم تودیع محسن قمصری، مدیر سابق امور بین الملل شرکت ملی نفت ایران و معا...	1	مدیر سابق امور بین الملل و بین الملل شرکت ملی نفت ایران در دوران تحریم به دلیل پیرامون تحریم وظیفه اصلی این شرکت ملی در...	2.05793

شکل (۱۷) — جدول نمایش تاریخچه

همانطور که در شکل (۱۷) مشخص است علاوه بر ورودی ها و خروجی ها، مدت زمان پردازش ها نیز ذخیره می شوند که بتوان در آینده مقایسه کرد. این جدول به صورت کامل اطلاعات را نگه می‌دارد تا بتوان ورودی ها و خروجی های مختلف را به مرور زمان بررسی کرد.

۴-۲- امکانات دیگر سایت

امکاناتی نظیر دو زبانه بودن سایت و همینطور احراز هویت و ذخیره تاریخچه نیز برای سایت آماده شده اند تا بتوان به بهترین شکل از سایت استفاده کرد.

۵- تهدیدات علیه اعتبار

در این مدل ما دادگان خود را محدود کردیم زیرا امکان آموزش روی کل دادگان نبود زیرا کمبود زمان و منابع سخت افزاری این اجازه را به ما نمی‌داد که مدل خود را روی تمامی دادگان آموزش دهیم و اگر مدلی باشد که بتواند کل این اخبار را پوشش دهد احتمال می‌رود که مدل با دقت بهتری به دست آید.

همچنین ما به دلیل کمبود منابع سخت افزاری و محدودیت زمانی تعداد محدودی epoch روی دادگان خود زدیم و اگر این محدودیت‌ها وجود نداشت با تعداد دوره بیشتر امکان داشت بتوانیم به دقت‌های بهتری روی دادگان خود برسیم.

ما برای تنظیم دقیق روی دادگان خود به دلیل کمبود حافظه مجبور بودیم از مدل‌های کوچک‌تر استفاده کنیم و اگر با کمبود حافظه

روبه رو نبودیم می‌توانستیم از مدل‌های قوی‌تری کـه عملکـرد بهتـری دارند استفاده نماییم.

۶- نتیجه‌گیری

ما در این پروژه، از سایت‌های خبری معتبر ایران، مجموعه‌ای از دادگان خبری و خلاصه‌هایشان را جمع آوری کردیم و تمام این دادگان را کـه شامل متن خبر، خلاصه و طول خبر و خلاصه است، در دسترس عموم قرار دادیم. سپس مدل‌های بروزی که در زمینه‌ی خلاصه‌سـازی وجـود دارنـــد، از جملـــه؛ mt5، Bert2Bert و WikiBert2WikiBert را تنظیم-دقیق کرده و بر روی دادگان جمع‌آوری شده، فرایند یـادگیری را انجام داده و نتایجشان را با هم مقایسه کرده و عملکرد هـر مـدل را ثبت کردیم.

این مدل‌ها را در بستر وب قرار داده شده و کد آن‌ها به صورت عمومی قابل دسترس است و افراد می‌توانند روند بهبود کد و مجموعه دادگـان را ادامه دهند.

مراجع

[1] Ashish Vaswani, Noam Shazeer, Niki Parmer, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

[2] Yang Liu and Mirella Lapata. 2019, Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730-3740, Hong Kong, China. Association for Computational Linguistics.

[3] Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401-2410, Online, Association for Computational Linguistics.

[4] Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization Branches Out*, pages 74-81, Barcelona, Spain. Association for Computational Linguistics.

[5] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30*, 2020. OpenReview.net.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long and Short Papers)*, pages 4171-4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[7] Y. Liu, Myle Ott, Naman Goyal Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692

[8] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Impeoving pre-training by representing and predicting spans. *Transactions of the Association for Computational LinguisticsI*, 8:64-77.

[9] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Chazvininejad, Abdelrahman Mohaed, Omer Levy, Veseling Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871-7880, Online. Association for Computational Linguistics.

[10] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Hournal of Machine Learning Research*, 21(140):1-67.

[11] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami, Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *NAACL*.

[12] Mohammad Ebrahim Khademi, Mohammad Fakhredanesh, and Seyed Mojtaba Hoseini. 2018. Conceptual text summarizer: A new model in continuous vector space.

[13] Hosein Rezaei, Seyed Amid Moeinzadeh, A. Shahgholian, and M. Saraee. 2019. Features in extractive supervised single-document summarization: Case of Persian news. *Arxiv*, abs/1909.02776.

[14] Mehrdad Farahani, Mohammad Gharachorloo, and Mohammad Manthouri. 2020b. Leveraging parsbert and pretrained mt5 for Persian abstractive text sumemarization.

[15] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2020a. Parsbert: Transformer-based model for Persian language understanding.

[16] Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging Pre-trained Chechpoints for Sequence Generation Tasks. *Transactions of the Association for Computational Linguistics*, 8:264-280.

[17] Alireza Salemi, Emad Kebriaei, Ghazal Neisi Minaei, Azadeh Shakery. 2021. ARMAN: Pre-training with Semantically Selecting and Reordering of Sentences for Persian Abstractive Summarization. *ArXiv*, abs/2109.04098.

[18] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference of Machine Learning Research*, pages 11328-11339. PMLR.

[19] Yanyan Zou, Xingxing Zhang, Wei Lu, Furu Wei, and Ming Zhou. 2020. Pre-training for abstractive document summarization by reinstating source text. In *Proceedings of the 2020 Conference on Epmpirical Methods in Natural Lanquage Processinq (EMNLP)*, pages 3646-3660, Online. Association for Computational Linguistics.

[20] Pyysalo, S., Kanerva, J., Virtanen, A., & Ginter, F. (2020). Wikibert models: deep transfer learning for many languages. *arXiv preprint arXiv:2006.01538*.

[21] Farahani, M., Gharachorloo, M., & Manthouri, M. (2021, March). Leveraging ParsBERT and Pretrained mT5 for Persian Abstractive Text Summarization. In 2021 26th International Computer Conference, Computer Society of Iran (CSICC) (pp. 1-6). IEEE.

زیرنویس‌ها

۱ extractive
۲ abstractive
۳ transformer
٤ Pre-training
٥ Fine-tune
٦ Masked Language Modeling
۷ Encoder-Decoder
۸ Sequence to Sequence
۹ Next Sentence Generation (NSG)
۱۰ Sentence Recording (SR)
۱۱ Encoder
۱۲ Decoder
۱۳ nan
۱٤ Masked language modeling
۱٥ Sequence to sequence
۱٦ encoder
۱۷ decoder
۱۸ bidirectional
۱۹ Cross attention layer
۲۰ Self-attention
۲۱ Feed-forward
۲۲ Multilingual Bert
۲۳ text to text
۲٤ De-shuffling
۲٥ mask
۲٦ epoch
۲۷ Semantic Score
۲۸ Contextual Embedding
۲۹ Longest Common Subsequence