مقاله پژوهشی

درون‌سازی معنایی واژه‌ها با استفاده از BERT روی وب فارسی

شکوفه بستان، علی‌محمد زارع‌بیدکی و محمدرضا پژوهان

چکیده: استفاده از بافت و ترتیب واژگان در یک عبارت از مواردی است که می‌تواند به فهم بهتر آن عبارت منجر گردد. در سال‌های اخیر، مدل‌های زبانی از پیش‌آموزش‌یافته، پیشرفت شگرفی در زمینه پردازش زبان طبیعی به وجود آورده‌اند. در این راستا مدل‌های مبتنی بر ترنسفورمر مانند الگوریتم BERT از محبوبیت فزاینده‌ای برخوردار گردیده‌اند. این مسئله در زبان فارسی کمتر مورد بررسی قرار گرفته و به‌عنوان یک چالش در حوزه وب فارسی مطرح می‌گردد. بنابراین در این مقاله، درون‌سازی واژگان فارسی با استفاده از این الگوریتم مورد بررسی قرار می‌گیرد که به درک معنایی هر واژه بر مبنای بافت متن می‌پردازد. در رویکرد پیشنهادی، مدل ایجادشده بر روی مجموعه دادگان وب فارسی مورد پیش‌آموزش قرار می‌گیرد و پس از طی دو مرحله تنظیم دقیق با معماری‌های متفاوت، مدل نهایی تولید می‌شود. در نهایت ویژگی‌های مدل استخراج می‌گردد و در رتبه‌بندی اسناد وب فارسی مورد ارزیابی قرار می‌گیرد. نتایج حاصل از این مدل، بهبود خوبی نسبت به سایر مدل‌های مورد بررسی دارد و دقت را نسبت به مدل برت چندزبانه تا حداقل یک درصد افزایش می‌دهد. همچنین اعمال فرایند تنظیم دقیق با ساختار پیشنهادی بر روی سایر مدل‌های موجود توانسته به بهبود مدل و دقت درون‌سازی بعد از هر فرایند تنظیم دقیق منجر گردد. نتایج رتبه‌بندی بر مبنای مدل‌های نهایی، بیانگر بهبود دقت رتبه‌بندی وب فارسی نسبت به مدل‌های پایه مورد ارزیابی با افزایش حدود ۵ درصدی دقت در بهترین حالت است.

***کلیدواژه:* بردار معنایی، درون‌سازی واژه، رتبه‌بندی، یادگیری عمیق.**

۱- مقدمه

پردازش زبان طبیعی، یکی از زیرشاخه‌های یادگیری ماشین است که اغلب با پردازش متن سروکار دارد. با توجه به اینکه متن از واحدهای کوچک‌تری همچون واژه تشکیل می‌شود، نمایش عددی واژه‌ها و متون در استفاده به‌عنوان ورودی الگوریتم‌های حوزه یادگیری و یا دسته‌بندی لغات و اسناد اهمیت می‌یابد. یکی از چالش‌هایی که اغلب موتورهای جستجو با آن مواجه هستند، دستیابی به روشی مؤثر در درک بهتر منظور کاربر از پرس‌وجوی وارد شده و ارائه نتایج مرتبط به نیاز اوست. از روش‌های نوین در نمایش متون، استفاده از نمایش برداری واژه‌ها و جملات است که اخیراً در بازیابی اطلاعات نیز مورد توجه قرار گرفته است [۱]. لذا مهم‌ترین گام در این راستا دستیابی به نمایش برداری مناسب و بامعنی از پرس‌وجو و سند است تا منجر به درک عمیق معنای متن و نمایش مناسب آن گردد. تا کنون روش‌های مختلفی بر پایه الگوریتم‌های یادگیری عمیق و با هدف پردازش جملات و یادگیری ارتباط واژگان در آنها مطرح گردیده است. می‌توان از درون‌سازی واژگان[1] و متون به‌عنوان تکنیک‌های موفق در این مسیر نام برد. بنابراین هدف از این پژوهش، ارائه راهکاری در راستای درک بهتر مفهوم واقعی یک عبارت با استفاده از درون‌سازی معنایی است.

در روش‌های مرسوم از درون‌سازی معنایی در گسترش پرس‌وجو استفاده می‌شود، اما به‌عنوان اولین نوآوری در این مقاله از درون‌سازی معنایی متون به‌صورت مستقیم در رتبه‌بندی اسناد استفاده گردیده است. در واقع به‌جای استفاده از نمایش برداری متون و استفاده از آن در گسترش پرس‌وجو، از بردارهای معنایی در فضای چندبعدی و به‌صورت مستقیم در رتبه‌بندی استفاده گردیده است. به عبارت دیگر تمام محاسبات و سنجش میزان شباهت و ارتباط پرس‌وجو و اسناد، کاملاً مبتنی بر بردار معنایی متون در همان فضای چندبعدی است و از این رو در ابتدا به بررسی الگوریتم‌ها و معماری‌های موفق درون‌سازی پرداخته شده است. در سال‌های اخیر، الگوریتم BERT[2] از محبوبیت فراوانی برخوردار گردیده و در حال حاضر توسط موتور جستجوی گوگل در حال استفاده است. اما همان گونه که انتظار می‌رود، درون‌سازی واژگان در زبان فارسی کمتر مورد بررسی قرار گرفته است. همچنین الگوریتم BERT در حوزه زبان فارسی به‌صورت محدود استفاده گردیده و بنابراین الگوریتم BERT می‌تواند انتخاب مناسبی برای دستیابی به درون‌یابی واژگان و متون باشد. نوآوری دوم در استفاده از مجموعه دادگان وب فارسی به‌صورت یک مجموعه مستقل جهت آموزش مدل و استفاده از آن در رتبه‌بندی اسناد وب بر مبنای پرس‌وجوی کاربر است. فرایند پیش‌آموزش BERT از ابتدا، بسیار هزینه‌بر و پیاده‌سازی آن با سیستم‌های معمولی کار دشواری است.

در این مقاله با تغییر یک سری از پارامترها، پیچیدگی زمانی مدل کاهش خواهد یافت. سپس با تنظیم‌های دقیق متوالی به صورت نوآورانه، یک مدل برت سفارشی ارائه خواهد گردید که در رتبه‌بندی معنایی اسناد وب فارسی و اولویت‌بخشیدن به اسناد مرتبط، مؤثر خواهد بود. در نهایت به ارزیابی مدل‌ها و رتبه‌بندی اسناد بر مبنای بردارهای معنایی حاصل از درون‌سازی پیشنهادی واژگان پرداخته خواهد شد.

ساختار مقاله به این ترتیب است که در بخش دوم، پژوهش‌های پیشین بیان می‌گردد. بخش سوم به چهارچوب پیشنهادی مبتنی بر درون‌سازی BERT می‌پردازد. در بخش چهارم، پیاده‌سازی‌های صورت‌گرفته و نحوه آموزش مدل مطرح می‌گردد و در بخش پنجم، فرایند تنظیم دقیق طی دو معماری متفاوت تشریح می‌شود. در بخش ششم، ارزیابی مدل‌های پیشنهادی مورد بررسی قرار می‌گیرد و نهایتاً در بخش هفتم، جمع‌بندی و

این مقاله در تاریخ ۸ مهر ماه ۱۴۰۱ دریافت و در تاریخ ۲۸ دی ماه ۱۴۰۱ بازنگری شد.

شکوفه بستان، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: sbostan@stu.yazd.ac.ir).

علی‌محمد زارع‌بیدکی (نویسنده مسئول)، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: alizareh@yazd.ac.ir).

محمدرضا پژوهان، دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران، (email: pajoohan@yazd.ac.ir).

1. Word Embedding
2. Bidirectional Encoder Representations from Transformers

نتیجه‌گیری نهایی مقاله بیان می‌گردد.

۲- پژوهش‌های پیشین

پژوهش‌های پیشین در قالب چهار دسته بیان می‌گردند. در دسته اول رویکردهای نمایش برداری واژگان مورد بررسی قرار می‌گیرد. دسته دوم و سوم به رویکردهای درون‌سازی ایستا و پویای واژگان می‌پردازند. نهایتاً در دسته چهارم، درون‌سازی واژگان فارسی و کارهای صورت‌پذیرفته در راستای الگوریتم BERT بیان می‌گردند.

۲-۱ رویکردهای نمایش برداری واژگان

در روش‌های برداری سنتی همچون BoW[1] [۲] و TF-IDF[2] [۳] از وزن‌دهی به هر واژه بر مبنای معیارهای مختلف در متن استفاده می‌شود، اما این نوع نمایش عددی، محدودیت و کاستی‌های فراوانی دارد که علاوه بر عدم دستیابی به ارتباط مفهومی بین واژه‌ها، نمی‌تواند به نمایش برداری بامعنی از واژه‌ها به‌صورت مجزا دست یابد. روش BoW نمایش ساده‌ای از واژه‌هاست که در پردازش زبان‌های طبیعی و بازیابی اطلاعات مورد استفاده قرار می‌گیرد. در این مدل، یک متن که می‌تواند یک جمله یا سند باشد، بر مبنای تعداد تکرار واژه‌ها در آن و بدون درنظرگرفتن دستور زبان و معنی و حتی نظم واژه‌ها به نمایش درمی‌آید. همچنین TF-IDF که توسط آقای سالتون در سال ۱۹۸۸ معرفی گردید [۳] از تکرار واژه‌های پرس‌وجو و سند برای محاسبه وزن واژه‌ها استفاده می‌کرد. TF بیانگر تعداد تکرار واژه است و در صورتی که یک واژه چندین بار در متن تکرار شود می‌تواند بیانگر توصیفی از متن مورد نظر باشد. همچنین IDF بیانگر عکس تکرار یک واژه در کل اسناد است و به بیان اهمیت واژه بر مبنای تکرار آن در سایر اسناد می‌پردازد. با مقایسه دو روش برداری BoW و TF-IDF درمی‌یابیم که هر دو روش بسیار ساده و سریع هستند. با این تفاوت که میزان اهمیت واژه در TF-IDF تا حدی در نظر گرفته می‌شود اما در BoW کاملاً نادیده گرفته می‌شود. نکته دیگر در مورد این دو روش عدم وجود ارتباط معنایی بین واژه‌هاست.

۲-۲ رویکردهای درون‌سازی ایستای واژگان

بنجیو[3] و همکارانش [۴] در سال ۲۰۰۳ مدلی را معرفی کردند که از یک شبکه عصبی با یک لایه مخفی تشکیل شده بود و به پیش‌بینی واژه بعدی در متن می‌پرداخت. این مفهوم بعد از مدتی با نام درون‌سازی واژه‌ها مطرح گردید. درون‌سازی واژه، نمایش برداری آن واژه در فضای n بعدی است که تلاش می‌کند معنای لغت و محتوای آن را بر اساس میزان نزدیکی به واژه‌های مشابه، محاسبه کند و به‌صورت عددی در فضای n بعدی به نمایش درآورد. در سال ۲۰۱۳، الگوریتم Word۲vec گوگل [۵]، توسط میکلوو[4] و همکارانش معرفی گردید. در این روش از جملات به‌عنوان ورودی مدل استفاده می‌شود و بردارهای درون‌ساز واژه‌ها به‌عنوان خروجی ارائه می‌گردند. Word۲vec دارای یک معماری مطلوب جهت نمایش بامعنی واژه‌ها است. در Word۲vec، نمایش درون‌سازی واژه‌های هم‌معنی به هم نزدیک و بیانگر ارتباط آنها با یکدیگر است. به‌منظور پیمایش جملات در این مدل از یک پنجره لغزان استفاده می‌شود که روی متون حرکت می‌کند. هدف از این کار، مشاهده واژه‌ها و ارتباط آنها با همسایگان خود و یافتن رابطه معنایی واژه‌ها با یکدیگر است. در این صورت واژه‌هایی که به همدیگر مربوط هستند و معمولاً در یک جمله در کنار هم یا با فاصله کم از یکدیگر قرار دارند، در فضای برداری نزدیک و با بار معنایی مشابه در نظر گرفته می‌شوند. معماری داخلی Word۲vec از نوع شبکه کاملاً متصل[5] است؛ به این صورت که نورون‌ها در هر لایه به‌صورت کاملاً متصل به نورون‌های لایه بعدی که تحت عنوان لایه متراکم هم شناخته می‌شود اتصال دارند [۶]. این الگوریتم از دو روش CBOW[6] و Skip-gram برای یادگیری مدل استفاده می‌کند.

مدل بردار سراسری یا GloVe[7] [۷] در سال ۲۰۱۴ توسط پنینگتون[8] و و همکاران در دانشگاه استنفورد مطرح گردید. در روش‌های یادگیری مبتنی بر پنجره لغزان، احتمال ضعف یادگیری روی مجموعه دادگان بزرگ بسیار محتمل است؛ لذا ایده الگوریتم GloVe برخلاف Word۲vec مبتنی بر احتمال هم‌زمانی واژه‌ها در یک مجموعه متن است. به عبارت دیگر، GloVe بررسی می‌کند که چگونه واژه j در محتوای متن شامل واژه i در تمام اسناد مجموعه دادگان ظاهر می‌شود و بنابراین می‌توان GloVe را مبتنی بر تعداد رخداد واژه‌ها در نظر گرفت.

در سال ۲۰۱۴، الگوریتم FastText [۸] توسط فیسبوک مطرح گردید. در این الگوریتم از مدل Skip-gram الگوریتم Word۲vec ایده گرفته شده، اما در آن از تابع وزن‌دهی متفاوتی استفاده گردیده است. در این روش هر واژه به‌صورت کیفی از واژه‌ها به‌صورت n-gram در نظر گرفته می‌شود و از یک سری نشانه در آغاز و پایان هر واژه استفاده شده است. سپس به ازای تمام n-gramهای هر واژه، بردارهای عددی به شیوه مشابه با الگوریتم Word۲vec به دست می‌آید و نهایتاً بردار هر واژه از مجموع تمامی بردارهای n-gram آن واژه حاصل می‌شود.

۲-۳ رویکردهای درون‌سازی پویای واژگان

الگوریتم Word۲vec نمونه ساده‌ای از یادگیری انتقالی[9] [۹] است که تنها از یک لایه وزن‌دار تحت عنوان درون‌ساز واژه‌ها استفاده می‌کند. اما یک شبکه عصبی می‌تواند شامل لایه‌های فراوانی باشد که قدرت شبکه و در عین حال پیچیدگی آن را افزایش می‌دهد. خروجی Word۲vec، بردار واژه‌هاست که شباهت معنایی بین واژه‌ها را نشان می‌دهد و این شباهت از طریق همسایگان آن واژه در مجموعه دادگان به دست می‌آید. یکی از محدودیت‌های این الگوریتم، اختصاص یک بردار ثابت درون‌ساز برای هر واژه است؛ یعنی فرض بر این است که معنای یک واژه در تمام جملات یکسان باشد. اما در واقعیت چنین نیست و هر واژه می‌تواند معانی مختلفی داشته باشد که از معنای سایر واژه‌ها در جمله برداشت می‌شود. همچنین عدم توجه به ترتیب واژه‌ها و محل قرارگیری آنها در متن و تشخیص واژه‌های کلیدی و بااهمیت، از دیگر ضعف‌های مدل‌های فوق است. در ادامه، مدل‌های زبانی ELMo[10] [۱۰]، BERT [۱۱] و GPT[11] [۱۲] ارائه گردیدند که دارای ماهیت پویا در درون‌سازی واژگان هستند.

روش ELMo نوع جدیدی از نمایش واژه‌هاست که در سال ۲۰۱۸

1. Bag of Words
2. Term Frequency-Inverse Document Frequency
3. Bengio
4. Mikolov
5. Fully Connected Network
6. Continuous Bag-of-Words
7. Global Vectors
8. Pennington
9. Transfer Learning
10. Embeddings from Language Models
11. Generative Pre-Trained Transformer

معرفی گردید و به فهم عمیق معنایی و نحوی واژه‌ها می‌پردازد [۱۰]. برخلاف روش‌های درون‌سازی Word۲vec و GloVe، مدل ELMo به ارائه درون‌سازی‌های متفاوتی از یک واژه می‌پردازد. به عبارت دیگر، بردار نمایش یک واژه در دو جمله متفاوت، یکسان نیست و بر اساس مفهوم آن واژه در جمله، نمایش برداری متفاوتی در سایر جملات به دست می‌آورد. در واقع ELMo از پیش‌بینی واژه بعدی در دنباله واژه‌های مبتنی بر مدل زبانی استفاده می‌کند اما به جای اختصاص یک درون‌ساز ثابت به هر واژه، به کل جمله نگاه می‌کند و با ارائه درون‌سازهای مختلف به نمایش بردار معنایی هر واژه، متناسب با جمله‌ای که واژه مورد نظر در آن ظاهر شده است می‌پردازد. در معماری ELMo از LSTM[۱] [۱۳] استفاده شده که یک نوع RNN[۲] [۱۴] است و به‌خوبی می‌تواند به‌عنوان یک مدل زبانی در نظر گرفته شود. مدل ELMo از شبکه‌های LSTM دوطرفه برای مدل‌کردن واژه‌ها استفاده می‌کند؛ به این صورت که واژه‌های قبلی و بعدی آن واژه در جمله در نظر گرفته می‌شوند. دنباله‌ای از واژه‌ها به‌صورت واژه به واژه وارد LSTM می‌شوند و واژه قبلی به همراه وضعیت داخلی LSTM برای پیش‌بینی احتمال واژه بعدی استفاده می‌شود.

دولین[۳] و همکارانش در سال ۲۰۱۸ الگوریتم BERT [۱۱] را معرفی نمودند. BERT مبتنی بر ترنسفورمر و از نظر منطقی به ELMo شبیه است. در واقع BERT یک مدل زبانی دوطرفه از نمایش رمزگذاری‌شده ترنسفورمرها است که ترکیبی از معماری شبکه عصبی بازگشتی دوجهته [۱۵] و شبکه عصبی عمیق بازگشتی است. این معماری با عنوان معماری شبکه عصبی عمیق بازگشتی دوجهته شناخته می‌شود که به معماری BERT نزدیک است؛ با این تفاوت که در معماری BERT از هیچ RNNی استفاده نمی‌شود و مبتنی بر شبکه‌های کاملاً متصل در بدنه ترنسفورمرهاست که از دقت بالایی برخوردار می‌باشد. از دلایل موفقیت معماری BERT می‌توان به وجود دو مکانیزم معروف و بااهمیت دیگر با عناوین مکانیزم توجه[۴] و رمزگذاری مکانی[۵] در بدنه ترنسفورمر اشاره نمود [۱۶]. الگوریتم‌های ALBERT[۶] [۱۷]، RoBERTa[۷] [۱۸] و DistilBERT[۸] [۱۹] به‌عنوان الگوریتم‌های گسترش‌یافته BERT در سال ۲۰۱۹ ارائه گردیدند.

همچنین در سال ۲۰۱۸، مدل GPT [۱۲] توسط openAI منتشر شد. این مدل به‌صورت یک ترنسفورمر مبتنی بر رمزگشای چندلایه معرفی می‌گردد. GPT در نسخه نخست خود از ۱۲ لایه رمزگشا که روی هم پشته[۹] شده‌اند استفاده می‌نماید و هر رمزگشا متن را مورد پردازش قرار می‌دهد و قسمت‌های مهم آن را جست‌وجو می‌کند. سپس بر اساس میزان ارتباط هر واژه با سایر واژگان جمله دریافتی به درون‌یابی واژگان و متون می‌پردازد. به همین منظور، هر لایه از دو زیرلایه که دربرگیرنده مکانیزم خودتوجهی چندسر[۱۰] و یک شبکه کاملاً متصل هستند تشکیل شده است. تعداد پارامترها و داده‌های مورد آموزش در این مدل، بسیار بزرگ‌تر از مدل‌های قبلی در نظر گرفته شده است.

1. Long Short-Term Memory
2. Recurrent Neural Network
3. Devlin
4. Attention
5. Position Encoding
6. A Lite BERT
7. A Robustly Optimized BERT Pretraining Approach
8. A Distilled Version of BERT
9. Stack
10. Multi Head Self Attention

۲-٤ رویکردهای درون‌سازی واژگان فارسی بر اساس الگوریتم BERT

در سال‌های اخیر، بهره‌گیری از درون‌سازی واژگان و کشف روابط معنایی آنها مورد توجه بسیاری از پژوهشگران قرار گرفته، اما در زبان فارسی کمتر به آن پرداخته شده است. پس از معرفی الگوریتم BERT، مدل‌های مختلفی ارائه گردیدند که با بهره‌گیری از این الگوریتم به درون‌سازی واژگان و متون پرداختند. مدل پیش‌آموزش‌یافته چندزبانه BERT[۱۱]، مدلی است که ۱۰۲ زبان زنده دنیا را که فارسی هم جزو آنهاست پوشش می‌دهد. این مدل، شامل ۱۲ لایه مخفی می‌باشد و توسط تیم توسعه‌دهنده BERT ارائه گردیده است [۱۱]. همچنین مدل پیش‌آموزش‌یافته پارس‌برت[۱۲] بر مبنای معماری مشابه و آموزش بر روی مجموعه دادگان فارسی به دست آمده است. بخش اعظم داده مورد آموزش در این مدل، مربوط به ویکی‌پدیای فارسی، اخبار فارسی و کتاب‌های الکترونیکی فارسی است. این مدل از پیکربندی استاندارد BERT استفاده کرده و شامل ۱۲ لایه پشته‌شده روی یکدیگر است [۲۰].

تمام کارهای انجام‌گرفته تا این لحظه، مبتنی بر درون‌سازی واژگان با استفاده از مدل‌های اصلی و استفاده از مجموعه دادگان پراکنده است. در این پژوهش از مدل BERT جهت آموزش بر روی اسناد وب فارسی استفاده گردیده است. در این راستا برخی از پارامترهای اصلی دچار تغییر شده و مدل بر مبنای چندین فرایند متوالی و متفاوت، تنظیم می‌گردد. هدف از این کار، آموزش مدل بر روی اسناد وب فارسی و بررسی تأثیر آن در بازیابی اسناد وب و رتبه‌بندی دقیق‌تر آنها بر اساس پرس‌وجوی کاربر است.

۳- چهارچوب پیشنهادی مبتنی بر درون‌سازی BERT

الگوریتم BERT شامل یک معماری چندلایه است که لایه‌ها روی یکدیگر پشته شده‌اند و با دریافت دنباله ورودی و عبور از لایه‌های مختلف به خروجی مناسب دست می‌یابد. در معماری BERT استاندارد از ۱۲ لایه استفاده شده که این مدل بر مبنای ترنسفورمرها[۱۳] می‌باشد که مبتنی بر شبکه‌های کاملاً متصل طراحی گردیده است. جمله ورودی با استفاده از یک سری نشانه، رمزگذاری می‌گردد و یادگیری بر پایه تکنیک ماسک مدل زبانی[۱۴] و تکنیک پیش‌بینی جمله بعدی[۱۵] با شیوه خودتوجهی[۱۶] است. خودتوجهی که گاهی درون‌توجه[۱۷] نیز نامیده می‌شود، مکانیزمی است که با درنظرگرفتن موقعیت‌های مختلف هر واژه در دنباله ورودی به بررسی ارتباط واژگان با یکدیگر می‌پردازد. در واقع این مدل شامل مکانیزم توجه و رمزگذاری مکانی در بدنه ترنسفورمر است. پیاده‌سازی BERT مستلزم دو مرحله پیش‌آموزش[۱۸] و تنظیم دقیق[۱۹] مدل بر مبنای وظیفه مورد نظر است [۱۱].

هدف از این پژوهش، ارائه مدلی غنی‌تر برای درک بهتر مفهوم جملات فارسی است. در این راستا بهره‌گیری از الگوریتم BERT مورد

11. Multilingual BERT
12. Pars BERT
13. Transformer
14. Mask Language Model
15. Next Sentence Prediction
16. Self Attention
17. Intra-Attention
18. Pre-Training
19. Fine-Tuning

جدول ۱: جزئیات مجموعه‌های دادگان.

ردیف	مجموعه دادگان	تعداد	برچسب	کاربرد
۱	مجموعه دادگان صفحات وب	۶۸۱ میلیون عنوان صفحه وب	-	پیش‌آموزش
۲	مجموعه دادگان پرس‌وجو و سند	۲۰۰ پرس‌وجو، برای هر پرس‌وجو متوسط ۱۵ سند برچسب‌دار	مرتبط (۵۵ درصد) غیرمرتبط (۴۵ درصد)	تنظیم دقیق اول
۳	مجموعه دادگان پرس‌وجو و سند	۳۰۰۰ پرس‌وجو، برای هر پرس‌وجو سه سند برچسب‌دار	متناقض (۳۳ درصد) خنثی (۳۳ درصد) مستلزم (۳۴ درصد)	تنظیم دقیق دوم
۴	مجموعه دادگان پرس‌وجو و سند	۱۰۰ پرس‌وجو، برای هر پرس‌وجو متوسط ۱۰ سند برچسب‌دار	کاملاً مرتبط (۱۴ درصد) مرتبط (۱۳ درصد) نسبتاً مرتبط (۱۳ درصد) کمی مرتبط (۲۰ درصد) خیلی کم مرتبط (۲۰ درصد) نامرتبط (۲۰ درصد)	ارزیابی رتبه‌بندی

بررسی قرار گرفت. به صورت کلی دو راهکار برای بهره‌گیری از BERT مطرح می‌گردد. روش اول استفاده از مدل‌های موجود است که از قبل بر روی حجم وسیعی از مجموعه دادگان مورد آموزش قرار گرفته است. روش دوم، پیاده‌سازی و آموزش مدل از پایه است که کار پرهزینه‌ای محسوب می‌شود. همان گونه که انتظار می‌رفت، مدل پیش‌آموزش‌یافته مناسبی برای زبان فارسی ارائه نگردیده و مدل‌های ارائه‌شده بر پایه این زبان، اکثراً با محدودیت‌هایی مواجه هستند. از مدل‌های ارائه‌شده می‌توان به مدل پیش‌آموزش‌یافته چندزبانه BERT و مدل پیش‌آموزش‌یافته پارس‌برت اشاره کرد. مدل اول، یک مدل چندزبانه است که زبان‌های مختلفی را پوشش می‌دهد که فارسی هم جزو آنها محسوب می‌شود. این مدل توسط تیم توسعه‌دهنده BERT در سال ۲۰۱۸ ارائه گردیده است [۱۱]. دومین مدل، پارس‌برت است که بر مبنای آموزش بر روی مجموعه دادگان فارسی در سال ۲۰۲۱ منتشر شده است. هر دو مدل شامل ۱۲ لایه مخفی هستند و بر پایه معماری مشابه ارائه گردیدند [۲۰].

بنا بر توصیه ارائه‌دهندگان الگوریتم BERT، بهترین راهکار با صرف هزینه کمتر، انتخاب یک مدل مناسب از مجموعه مدل‌های موجود و تنظیم دقیق آن مدل جهت استفاده برای یک منظور خاص است [۲۱]. بنابراین تنظیم دقیق مدل با ساختار پیشنهادی بر مبنای دو مدل موجود در دستور کار قرار گرفت، اما به‌منظور درک بهتر فرایند آموزش و همچنین تلاش در جهت ارائه مدلی غنی‌تر بر مبنای زبان فارسی، پیاده‌سازی و آموزش مدل از پایه نیز مورد بررسی، پیاده‌سازی و ارزیابی قرار گرفت.

بنابراین در این پژوهش، دو راهکار مورد بررسی قرار گرفته است. در ابتدا مدل بر اساس مجموعه دادگان وب فارسی بر روی ۱۰ میلیون سند مورد آموزش از پایه قرار گرفته و یک مدل برت سفارشی ارائه شده است. همچنین در کنار این مدل از دو مدل آموزش‌یافته مبتنی بر زبان فارسی به عنوان مدل‌های پایه در دسترس استفاده گردیده است. سپس راهکار تنظیم دقیق مدل با ساختار پیشنهادی در راستای بهبود مدل‌های پایه و ارائه بردارهای بامعناتر از واژگان و عبارات ارائه گردیده است. این فرایند بر روی مدل پایه برت سفارشی و همچنین دو مدل پایه پیش‌آموزش‌یافته چندزبانه و پارس‌برت مورد آموزش و ارزیابی قرار گرفت.

۳-۱ انواع مجموعه‌های دادگان

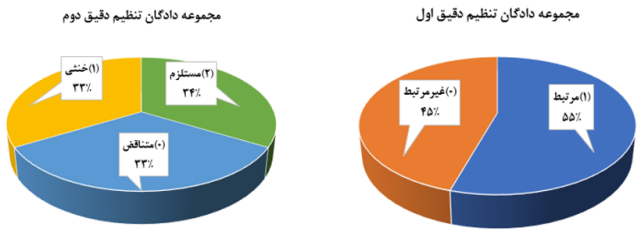
در این مقاله از چهار مجموعه دادگان استفاده شده و جزئیات مربوط به هر مجموعه در جدول ۱ آمده است. مجموعه دادگان اول توسط موتور جستجوی پارسی‌جو بر روی وب فارسی، خزش شده است. این مجموعه با ۶۸۱ میلیون عنوان سند و بدون برچسب است و جهت پیش‌آموزش مدل مورد استفاده قرار می‌گیرد. سه مجموعه دیگر در فرمت‌های متفاوت و با برچسب‌های مجزا تهیه گردیده‌اند. این مجموعه‌ها شامل جفت پرس‌وجو و اسناد هستند که از سایت‌های مختلفی همچون دیجی‌کالا و آپارات جمع‌آوری و توسط تیم متخصص، برچسب‌دهی شده‌اند.

مجموعه دادگان دوم شامل ۲۰۰ پرس‌وجو و برای هر پرس‌وجو، متوسط ۱۵ سند با برچسب مرتبط و غیرمرتبط است. این مجموعه دادگان جهت فرایند تنظیم دقیق اول مورد استفاده قرار می‌گیرد. همچنین مجموعه دادگان دوم شامل ۳۰۰۰ پرس‌وجو و برای هر پرس‌وجو سه سند برچسب‌دار است. این مجموعه دادگان در فرایند تنظیم دقیق دوم مورد استفاده قرار می‌گیرد. نهایتاً مجموعه دادگان چهارم شامل ۱۰۰ پرس‌وجو و سند با ۶ برچسب است. این برچسب‌ها بیانگر میزان مرتبط‌بودن هر سند به پرس‌وجوست. این مجموعه جهت استفاده در بخش ارزیابی مدل‌های پیشنهادی به‌کار می‌رود. جزئیات دو مجموعه دوم و سوم در شکل ۱ و جزئیات مربوط به مجموعه دادگان چهارم در شکل ۶ آمده است. مجموعه‌های دادگان برچسب‌دار در هر فرایند به صورت تصادفی به سه مجموعه آموزش، ارزیابی و آزمون تقسیم می‌گردند.

۳-۲ نوآوری

پیاده‌سازی مدل به فرم پیش‌آموزش‌یافته از پایه، کار دشوار و هزینه‌بری است. به‌عنوان راهکار برای پیاده‌سازی مدل بر روی مجموعه دادگان غنی وب فارسی، با اعمال برخی تغییرات در تنظیمات و پیکربندی مدل مانند کاهش تعداد لایه‌ها و همچنین طول دنباله ورودی، آموزش مدل از پایه مورد بررسی و پیاده‌سازی قرار گرفت. این مسیر با هدف درک بهتر فرایند آموزش بر مبنای الگوریتم BERT و تلاش در بهبود درون‌سازی جملات با استفاده از داده غنی‌تر و ارائه نتایج بهتر اعمال گردید. بنابراین به‌عنوان اولین نوآوری در این پیاده‌سازی، یک سری از پارامترها به‌صورت نوآورانه تغییر یافته است. در واقع با کاهش تعداد لایه‌ها و طول دنباله ورودی و واژه‌نامه، مدل جدیدی نسبت به نسخه استاندارد ارائه می‌گردد. این مدل با آموزش مناسب بر مبنای مجموعه دادگان غنی در سطوح مختلف مبتنی بر ساختار پیشنهادی، توانسته دقت خوبی نسبت به مدل‌های مشابه با ساختار استاندارد ارائه دهد که جالب توجه است.

وظیفه مورد استفاده در مرحله تنظیم دقیق این پژوهش، دسته‌بندی است. دسته‌بندی در این وظیفه می‌تواند به‌صورت جمله(ها) انجام گیرد، اما ارتباط بین جملات در نظر گرفته نمی‌شود و فقط به دسته‌بندی هر جمله



شکل ۱: نمایی از جزئیات مجموعه‌های دادگان برچسب‌دار جهت تنظیم دقیق.

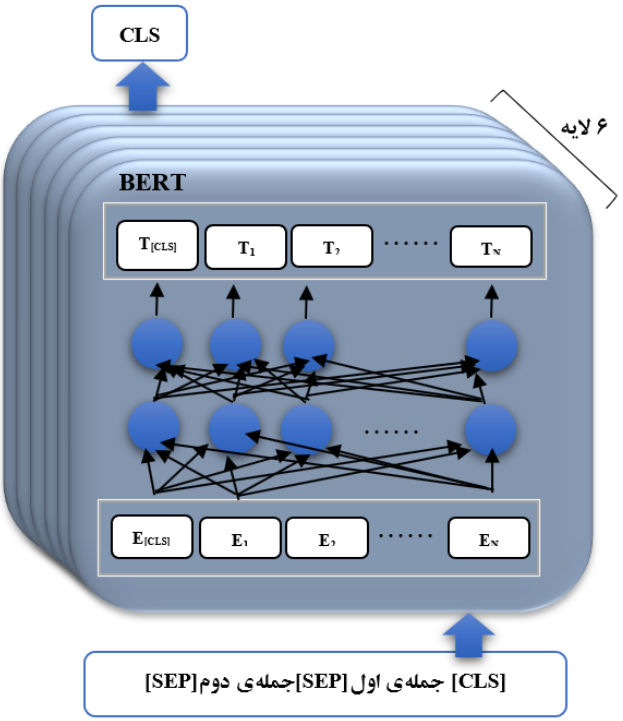
یا جملات خواهد پرداخت. بنابراین می‌توان گفت که این وظیفه به ازای جمله ورودی و دسته‌بندی آن بر مبنای برچسب‌های کلاسی ارائه می‌گردد. به‌عنوان نوآوری دوم در این پژوهش، به‌جای یک جمله از دو جمله پرس‌وجو و سند استفاده گردیده و دسته‌بندی در هر مرحله به ازای سند اول و ارتباط آن با سند دوم لحاظ شده است. در واقع در این روش، تعداد دسته‌ها مشخص است اما دسته‌بندی جملات دوم وابسته به جملات اول و مبتنی بر آن در نظر گرفته شده که با توجه به ساختار متفاوت با وظیفه موجود، به‌صورت سفارشی پیاده‌سازی گردید. در واقع به ازای هر پرس‌وجو، مجموعه‌ای از اسناد با برچسب‌های متفاوت وجود دارند که مرتبط با آن پرس‌وجو هستند. در ابتدا اسناد مربوط به هر پرس‌وجو دسته‌بندی می‌گردند. سپس در مرحله تزریق جملات به مدل، با تکرار پرس‌وجو و انتخاب سند جدید از دسته همان پرس‌وجو، فرایند ادامه می‌یابد و تنظیم دقیق مدل بر مبنای ارتباط دو جمله ورودی، صورت می‌پذیرد.

در فرایند تنظیم دقیق، لازم است که مدل در ابتدا با پارامترهای پیش‌آموزش‌یافته مقداردهی گردد و سپس این پارامترها بر مبنای وظیفه مورد نظر، تنظیم دقیق شوند. در فرایند تنظیم دقیق، اول از مدل پیش‌آموزش‌یافته استفاده گردیده، اما ابتکار دیگر مقاله در استفاده از مدل استخراج‌شده از فرایند تنظیم دقیق اول به‌عنوان مدل پایه در فرایند تنظیم دقیق دوم است. در واقع، فرایند تنظیم دقیق مدل بر مبنای دو وظیفه با معماری‌های متفاوت صورت می‌پذیرد که از خروجی مدل اول به‌عنوان ورودی مدل دوم استفاده شده است. بنابراین برخلاف نسخه‌های رایج، فرایندهای تنظیم متوالی با استفاده از خروجی نهایی فرایند قبلی صورت پذیرفته است. همچنین رتبه‌بندی اسناد وب جهت ارزیابی بر اساس پرس‌وجو و اسناد درون‌سازی‌شده در فضای n بعدی اعمال می‌گردد. در این مرحله از بردارهای درون‌یابی به صورت مستقیم در رتبه‌بندی استفاده گردیده است. به عبارت دیگر در مرحله رتبه‌بندی از هیچ پارامتر و ساختار دیگری به‌جز بردارهای معنایی استفاده نشده، اما دقت به صورت قابل قبول ارائه گردیده است.

٤- پیاده‌سازی و آموزش مدل از پایه

جهت پیاده‌سازی BERT از پایه، از کتابخانه ترنسفورمر استفاده شده و کدنویسی با استفاده از زبان پایتون و از طریق محیط یکپارچه توسعه نرم‌افزار پای‌چارم[1] و آزمایشگاه مشترک گوگل[2] صورت پذیرفته است. اولین و مهم‌ترین قسمت، ایجاد نشانه‌سازی می‌باشد که مسئول ساخت درون‌سازی نشانه‌ها به ازای دنباله ورودی است. همچنین پیاده‌سازی مدل BERT شامل دو قسمت است؛ قسمت اول، درون‌سازی است که برای ساخت نمایش مناسب از ورودی مدل شامل نشانه، قسمت و درون‌سازی مکانی به کار می‌رود. قسمت دوم رمزگذار است که پشته اصلی جهت

1. PyCharm
2. Google Colaboratory



شکل ۲: نمایی از معماری BERT.

رمزگذاری در ترنسفورمرهاست.

٤-١ معماری پیاده‌سازی‌شده

شکل ۲ به نمایش معماری مدل BERT می‌پردازد. این معماری شامل لایه‌های پشته‌شده روی یکدیگر و دریافت دنباله ورودی و تولید خروجی است. این مدل بر مبنای ترنسفورمرها می‌باشد و یادگیری بر پایه تکنیک ماسک مدل زبانی و تکنیک پیش‌بینی جمله بعدی صورت می‌پذیرد. لایه رمزگذار در ترنسفورمر، دنباله‌ای از واژگان ورودی را به‌صورت یک‌جا می‌خواند. این خصوصیت منجر به یادگیری بافت[3] هر واژه بر اساس واژه‌های نزدیک به آن در سمت چپ و راست می‌گردد.

معماری پیاده‌سازی‌شده، شامل ۶ لایه است که روی یکدیگر پشته شده‌اند و نسبت به طراحی استاندارد با ۱۲ لایه، کاهش یافته است. همچنین استاندارد طول دنباله ورودی برابر با ۵۱۲ است که در اینجا برابر با ۱۲۸ در نظر گرفته شده است. جملات ورودی طبق فرمت استاندارد ایجاد گردیده و با نشانه‌های اختصاصی به مدل تزریق می‌گردند.

٤-٢ ایجاد نشانه‌ساز

فرایند ایجاد ورودی مناسب جهت درون‌سازی، طی دو مرحله ایجاد مجموعه دادگان و نشانه‌ساز صورت می‌پذیرد. استفاده از داده غنی در فرایند آموزش مدل BERT بسیار حائز اهمیت است، زیرا در طی فرایند نشانه‌گذاری، واژگان یکتا از مجموعه دادگان استخراج و در حین آموزش بر مبنای جملات موجود، وزن‌دهی و درون‌سازی می‌شوند. در این پژوهش از عنوان صفحات وب به عنوان مجموعه دادگان استفاده گردیده است. این مجموعه توسط موتور جستجوی پارسی‌جو بر روی وب، مورد خزش قرار گرفته، ۶۸۱ میلیون عنوان سند وب، جمع‌آوری و استخراج گردیده و شامل متون به زبان‌های فارسی، انگلیسی و سایر زبان‌هاست. از آنجایی که مدل‌های پیش‌آموزش‌یافته BERT موجود، اغلب روی زبان‌های غیر

3. Context

شکل ۳: فرمت دنباله ورودی.

جدول ۲: اطلاعات آماری فرایند پیش‌آموزش مدل.

عنوان	مقدار
بیشینه طول دنباله ورودی	۱۲۸
تعداد لایه‌ها	۶
تعداد واژگان واژه‌نامه	۳۰۵۲۲
تعداد دوره	۳
تعداد سندهای مورد آموزش	۱۰ میلیون
زمان آموزش	۷۰ ساعت

از فارسی است، این مجموعه دادگان می‌تواند به‌عنوان یک منبع غنی از متون فارسی مورد استفاده قرار گیرد. این مجموعه طی چند مرحله، مورد پیش‌پردازش قرار گرفته است. در فرایند پیش‌پردازش، کاراکترهای اضافی از متن حذف گردیده و سپس فرایند نرمال‌سازی واژگان و متون، اعمال شده است. نهایتاً جملات استخراج گردیده و در هر سطر به صورت جداگانه نوشته شده است. فرمت دنباله ورودی در شکل ۳ آمده است. ورودی پایین‌ترین لایه BERT، دنباله‌ای از نشانه‌هاست که لزوماً واژه نیست. این الگوریتم از مدل نشانه‌گذاری قطعه واژه[1] استفاده می‌کند تا با ایجاد نشانه‌های جدید به بهبود عملکرد آموزش کمک کند. در واقع نشانه‌های قطعه واژه به‌عنوان واحدهای زیرواژه که شامل کاراکترهای مشخصی هستند، تبدیل می‌شوند. همچنین واژگان نادر که در مجموعه لغات مدل نیستند به زیرواژگان پرتکرار تبدیل می‌شوند. در زمان ایجاد نشانه‌ساز[2]، بعد از دریافت مجموعه دادگان، تعداد واژگان در نشانه‌ساز و نشانه‌های خاص تعیین می‌گردند [۲۱].

طول واژگان بعد از اعمال تکنیک نشانه‌گذاری قطعه‌کردن واژگان برابر با ۳۰۵۲۲ در نظر گرفته شده و ابعاد برابر با ۷۶۸ است. وزن‌های این ماتریس در طول فرایند آموزش، یاد گرفته می‌شوند.

۴-۳ آموزش مدل

با مقداردهی اولیه نشانه‌ساز با فایل‌های ایجادشده، مرحله بعدی آغاز می‌گردد. آموزش مدل از طریق رمزگذاری داده‌ها با کمک نشانه‌ساز و مدل ماسک زبانی انجام می‌پذیرد. در آموزش بر پایه وظیفه ماسک زبانی، تابع هزینه بر مبنای واژگان ماسک‌شده محاسبه می‌گردد. بنابراین مدل یاد می‌گیرد تا به پیش‌بینی واژه‌های ماسک‌شده بر مبنای واژه‌های اطراف آن واژه در آن لایه و لایه‌های قبلی بپردازد.

واضح است که مدل‌های BERT به موفقیت‌های بزرگی در پردازش زبان طبیعی دست یافتند اما یکی از چالش‌هایی که اغلب در مواجه با این مدل‌ها مطرح می‌شود، آن است که استفاده از آنها در سیستم‌های با منابع محدود به دلیل مشکلات حافظه و پردازش، دشوار است. دلیل آن می‌تواند درون‌سازی دنباله ورودی با ابعاد بزرگ و عظیم واژگان و پردازش‌های متوالی در لایه‌های مختلف مدل باشد. بنابراین برای میسرشدن مسیر آموزش بر روی یک سیستم کاملاً معمولی، ابعاد واژگان و تعداد لایه‌های مخفی کاهش یافت. لذا تعداد واژگان در این آموزش به دلیل هزینه زیاد، برابر با مقدار پیش‌فرض ۳۰۵۲۲ تنظیم گردید. این در حالی است که سایر مدل‌ها با افزایش ابعاد واژگان به آموزش بهتر مدل بر مبنای واژگان غنی‌تر می‌پردازند. در واقع ۳۰۵۲۲ نشانه[3] در نظر گرفته شده که بردار نمایش هر نشانه شامل ۷۶۸ مؤلفه است. تعداد لایه‌ها برابر با ۶ و بیشینه طول دنباله ورودی برابر با ۱۲۸ تعیین گردیده است. عدد ۱۲۸ بیانگر طول جدول درون‌سازی مکانی است و ورودی با طول بیشتر از این عدد، برش داده می‌شود تا بقیه آن نادیده گرفته شود. دلیل انتخاب این عدد به‌جای مقدار پیش‌فرض ۵۱۲، کمبود حافظه در زمان پردازش و درون‌سازی دنباله ورودی است. همچنین با وجود مجموعه دادگان غنی با ۶۸۱ میلیون عنوان سند از وب فارسی به دلیل هزینه فراوان، آموزش مدل تنها بر روی ۱۰ میلیون سند صورت پذیرفت. در فرایند آموزش بر روی سیستم معمولی، هر دوره آموزش، ۲۳ ساعت زمان به طول انجامید و در مجموع برای سه دوره، حدود ۷۰ ساعت زمان سپری شد که در جدول ۲ ذکر گردیده است.

در هر فرایند آموزش، امکان تزریق ۱۰ میلیون سند جدید به مدل فراهم گردید. فرایند آموزش با این ساختار، سه روز زمان لازم دارد تا مدل را مقداردهی کند و به ساخت وزن‌های معنایی متناسب با هر جمله و نشانه دست یابد. طولانی‌بودن زمان آموزش در هر دسته سند به دلیل پردازش‌های فراوان در لایه‌های مختلف ترنسفورمر است که منجر به صرف هزینه بالاتر برای آموزش می‌گردد. در ساختارهای موجود، مدل یک بار مورد آموزش قرار می‌گیرد و سپس روی داده برچسب‌دار برای وظیفه مشخص، تنظیم دقیق می‌شود. اما در اینجا با ادامه فرایند آموزش به‌صورت تکنیک ماسک زبانی و پیش‌بینی جمله بعدی، فرایند آموزش مدل ادامه می‌یابد. بنابراین می‌توان مدل آموزش‌یافته در هر مرحله را ذخیره کرد و فرایند آموزش را روی همان مدل و با مجموعه اسناد جدیدتر در بازه زمانی دیگری ادامه داد. به دلیل هزینه فراوان از ادامه

1. Word Piece
2. Tokenizer
3. Token

جدول ۳: مقایسه مدل پیش‌آموزش‌یافته با مدل‌های موجود.

مدل	تعداد واژگان واژه‌نامه	ویژگی‌ها (تعداد لایه‌های مخفی)	بیشینه طول دنباله ورودی	تعداد لایه‌ها
مدل BERT چندزبانه	۱۰۵۸۷۹	۷۶۸	۵۱۲	۱۲
مدل پارس برت	۱۰۰۰۰۰	۷۶۸	۵۱۲	۱۲
مدل BERT سفارشی	۳۰۵۲۲	۷۶۸	۱۲۸	۶

آموزش مدل به این صورت صرف نظر شد و ادامه آموزش منوط به تنظیم دقیق مدل بر اساس مجموعه دادگان هدفمند، طرح‌ریزی گردید. نهایتاً مدل آموزش‌یافته بر روی وب فارسی، تحت عنوان مدل پایه سفارشی برت، تولید و جهت استفاده در مراحل بعدی ذخیره گردید. با توجه به اینکه مدل‌های مختلفی از الگوریتم BERT ارائه گردیده است، می‌توان مقایسه را بر اساس انواع مدل‌های آن مورد ارزیابی قرار داد. دو مدل برت چندزبانه[1] و پارس‌برت[2]، جزو معدود مدل‌های BERT می‌باشند که بر روی واژگان فارسی آموزش دیده‌اند. این دو مدل می‌توانند به عنوان گزینه مناسب جهت مقایسه با مدل پیشنهادی در مراحل پیش‌آموزش و تنظیم دقیق مورد ارزیابی قرار گیرند. جزئیاتی از مدل برت سفارشی و دو مدل پایه موجود در جدول ۳ قابل مشاهده است.

مرحله بعد، تنظیم دقیق مدل بر مبنای وظیفه مورد نیاز است تا فرایند آموزش با نرخ بسیار پایین‌تر اما به صورت هدفمند بر روی مجموعه دادگان برچسب‌گذاری‌شده ادامه یابد.

۵- تنظیم دقیق مدل

تنظیم دقیق مدل به هزینه بسیار کمتری نسبت به آموزش مدل از پایه نیاز دارد. می‌توان با انتخاب یک مدل مناسب پیش‌آموزش‌یافته پایه به آموزش هدفمند مدل در راستای یک وظیفه مشخص پرداخت. این فرایند نیازمند مجموعه دادگان ساختاریافته و مناسب به‌منظور آموزش جهت‌دار مدل در مسیر هدف است. در این مقاله از داده برچسب‌دار جهت تنظیم دقیق مدل طی دو فرایند متمایز و متوالی استفاده گردیده است. این مجموعه شامل اسنادی است که به پرس‌وجوی مورد نظر کاربر مرتبط هستند. به‌منظور اعمال بررسی بیشتر از اسناد غیرمرتبط به‌عنوان نویز در مجموعه دادگان استفاده شده و این مجموعه با اسناد مرتبط و غیرمرتبط مورد ارزیابی قرار گرفته است.

۵-۱ تنظیم دقیق اول

وظیفه انتخاب‌شده در این مرحله، دسته‌بندی است [۲۲]. با تنظیم دقیق مدل در راستای دسته‌بندی اسناد بر مبنای پرس‌وجوی کاربر، می‌توان به آموزش مدل بر اساس تعیین شباهت دو عبارت در قالب پرس‌وجو و سند پرداخت. برچسب‌های کلاسی مجموعه دادگان مورد استفاده در این مرحله به‌صورت دودویی و بیانگر مرتبط یا غیرمرتبطبودن سند به پرس‌وجوی مربوط است. بنابراین دسته‌بندی اسناد هر پرس‌وجو به ازای دو دسته مرتبط و غیرمرتبط مورد بررسی و آموزش قرار خواهد گرفت که این وظیفه با افزودن چندلایه در بالای مدل پایه مورد استفاده قرار می‌گیرد.

برای تنظیم دقیق مدل، مجموعه‌ای از پرس‌وجو و سندهای واقعی، جمع‌آوری و سپس بر مبنای مرتبطبودن سند به پرس‌وجو برچسب‌گذاری گردید. این مجموعه شامل ۲۰۰ پرس‌وجو است و برای هر پرس‌وجو، متوسط ۱۵ سند مرتبط و غیرمرتبط در نظر گرفته شده است. برچسب صفر بیانگر غیرمرتبطبودن سند به پرس‌وجو و برچسب یک بیانگر ارتباط آن سند به پرس‌وجوی مربوط است. این مجموعه دادگان در ابتدا مورد پیش‌پردازش و نرمال‌سازی قرار گرفته و سپس از کتابخانه سایکیت‌لرن[3] برای تقسیم داده‌ها به مجموعه‌های آموزش، ارزیابی و آزمون استفاده شده است. در این فرایند، ۷۰ درصد از دادگان برای فرایند آموزش و ۳۰ درصد از آنها شامل دو مجموعه ۱۵ درصدی جهت ارزیابی و آزمون به صورت تصادفی استفاده می‌شوند.

۵-۱-۱ ایجاد نشانه‌ساز

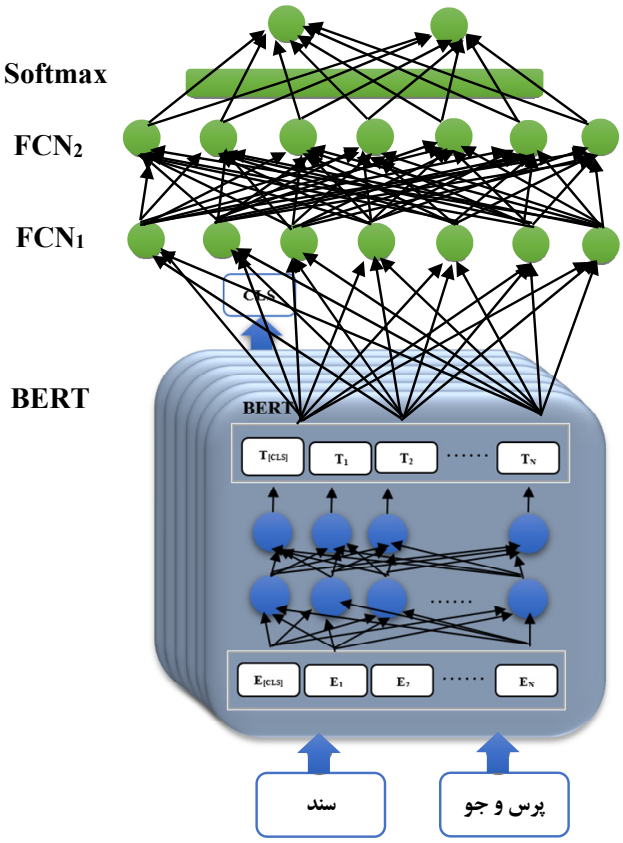
برای پردازش متون جدید و تعیین شیوه لایه‌گذاری[4] و برش جملات در جهت مدیریت طول متغیر دنباله ورودی به یک نشانه‌ساز نیاز است. در ابتدا نشانه‌ساز مدل پیش‌آموزش‌یافته BERT بارگذاری می‌شود و سپس عمل نشانه‌سازی برای سه مجموعه آموزش، ارزیابی و آزمون انجام می‌گیرد. روال نشانه‌سازی در این مرحله به این صورت است که با دریافت جمله ورودی با فرمت مناسب مطابق مرحله قبل، به قطعه‌سازی واژگان و سپس جستجو در مجموعه نشانه‌ساز می‌پردازد. افزودن واژگان جدید به نشانه‌ساز امکان‌پذیر است، اما از آنجایی که این واژگان در مرحله آموزش در فاز قبل، وزن‌دهی نشده‌اند می‌تواند منجر به کاهش دقت در تأثیر این واژگان با وزن درون‌سازی نامناسب بر روی سایرین گردد و بنابراین از آن پرهیز می‌شود.

۵-۱-۲ ایجاد مدل

با توجه به اینکه برچسب‌ها بر مبنای تعیین شباهت اسناد به پرس‌وجو در نظر گرفته شده است، شیوه دسترسی و پردازش مجموعه دادگان به صورت سفارشی ایجاد می‌گردد. سپس مجموعه دادگان به دسته‌های کوچک‌تر تقسیم می‌شوند و بعد از تزریق به شبکه، درهم‌سازی انجام می‌پذیرد. در هر دوره آموزش، وزن‌ها با استفاده از پارامترهای ثابت‌شده و لایه‌های اضافه‌شده به بالای مدل، به‌روز می‌گردند و میانگین هزینه در حین فرایند آموزش، محاسبه می‌شود. در معماری این مدل، طبق شکل ۴ از دو لایه شبکه کاملاً متصل برای وزن‌دهی لایه مخفی و از تابع بیشینه فعال‌ساز[5] برای خروجی لایه آخر استفاده شده است. این مدل دو خروجی برمی‌گرداند که اولی مربوط به مدل ماسک زبانی و دومی مربوط به پیش‌بینی جمله بعدی است [۲۲].

خروجی پیش‌بینی جمله بعدی در لایه اول مدل به شبکه کاملاً متصل داده می‌شود. سپس با عبور از تابع هزینه، یک درصد از وزن‌های لایه مخفی به‌صورت تصادفی به فراموشی سپرده می‌شود تا برای منظم‌سازی مناسب‌تر باشد و خطای مرحله آزمون را کاهش دهد. این خروجی در لایه دوم به شبکه کاملاً متصل داده شده و سپس از تابع بیشینه فعال‌ساز عبور داده می‌شود. خروجی نهایی مدل، بیانگر مرتبطبودن جمله دوم به اول یا عدم ارتباط آن است.

1. https://huggingface.co/bert-base-multilingual-uncased
2. https://huggingface.co/HooshvareLab/bert-base-parsbert-uncased
3. Scikit Learn
4. Padding
5. Softmax



شکل ۴: معماری تنظیم دقیق اول.

۵-۱-۳ آموزش مدل

آموزش مدل با اختصاص تعداد برچسب‌های کلاسی مورد انتظار، آغاز می‌گردد. در طول این فرایند، برخی از وزن‌های پیش‌آموزش‌یافته بدون استفاده باقی می‌مانند و برخی دیگر به‌صورت تصادفی، مقداردهی می‌شوند. لایه سر[1] پیش‌آموزش‌یافته مدل از بین می‌رود و با یک سر دسته‌بندی که به‌صورت تصادفی مقداردهی اولیه گردیده است، جایگزین می‌شود. در حین فرایند آموزش، این سر جدید روی دنباله وظایف دسته‌بندی، تنظیم دقیق می‌گردد و دانش مدل پیش‌آموزش‌یافته به آن منتقل می‌شود. فرایند آموزش طی سه دوره ادامه می‌یابد و نهایتاً مدل به‌دست‌آمده با وزن‌های جدید برای استفاده‌های بعدی ذخیره می‌گردد.

۵-۲ تنظیم دقیق دوم

به‌منظور افزایش دقت، ساختار دیگری با مجموعه دادگان متفاوت و فرمت مجزا برای تنظیم دقیق مدل در نظر گرفته شد. این روش در ادامه روش قبلی است و به همین دلیل از مدل تنظیم‌شده مرحله قبل به‌عنوان ورودی این مرحله استفاده می‌شود. در واقع، خروجی تنظیم دقیق شده از مرحله اول به‌عنوان ورودی تنظیم دقیق دوم به کار می‌رود تا وزن‌ها مطابق با هر دو ساختار بهبود یابد که در شکل ۵ قابل مشاهده است. مجموعه دادگان استفاده‌شده در این روش شامل ۳۰۰۰ پرس‌وجو می‌باشد که به زبان انگلیسی و فارسی تهیه گردیده است. این مجموعه توسط تیم متخصص، بررسی و برچسب‌گذاری شده است. به دلیل افزایش دقت مدل در زمان آموزش از سه نوع برچسب بر مبنای میزان مرتبط‌بودن پرس‌وجو و سند تحت عناوین مستلزم، متناقض و خنثی استفاده گردیده است. برچسب مستلزم به سندی تخصیص داده شده که با پرس‌وجوی مورد نظر در یک راستا و از نظر مفهومی کاملاً مرتبط باشد. در نقطه مقابل، برچسب متناقض به سندی اختصاص داده شده که از نظر ظاهری به پرس‌وجوی مورد نظر، نزدیک اما از نظر معنایی کاملاً متفاوت است. به عنوان مثال با درنظرگرفتن پرس‌وجوی "آخرین مدل ماشین سراتو"، سندی با محتوای "آخرین مدل ماشین بوگاتی" به‌عنوان متناقض در نظر گرفته می‌شود. همچنین برچسب خنثی به سندی اطلاق می‌شود که با محتوای کلی و بدون جهت‌گیری ارائه گردیده است. به عنوان مثال، سندی با محتوای "اخبار ماشین‌های روز دنیا" نمونه‌ای از این برچسب محسوب می‌شود.

۵-۲-۱ ایجاد نشانه‌ساز

فرایند نشانه‌گذاری در این روش، مشابه روش اول تنظیم دقیق است؛ با این تفاوت که در این مجموعه از جملات انگلیسی نیز در کنار جملات فارسی استفاده شده است. یکی از مراحل اصلی در نشانه‌سازی، برش متن و ثابت‌کردن طول دنباله ورودی است. طول سندهای انگلیسی بلندتر از فارسی است اما با توجه به اینکه هدف از این آموزش، مانور بیشتر روی متون فارسی است، بیشینه طول سند، متناسب با متون فارسی برابر با ۱۰ در نظر گرفته شد تا میزان صفرهای لایه‌گذاری کاهش یابد.

۵-۲-۲ ایجاد مدل

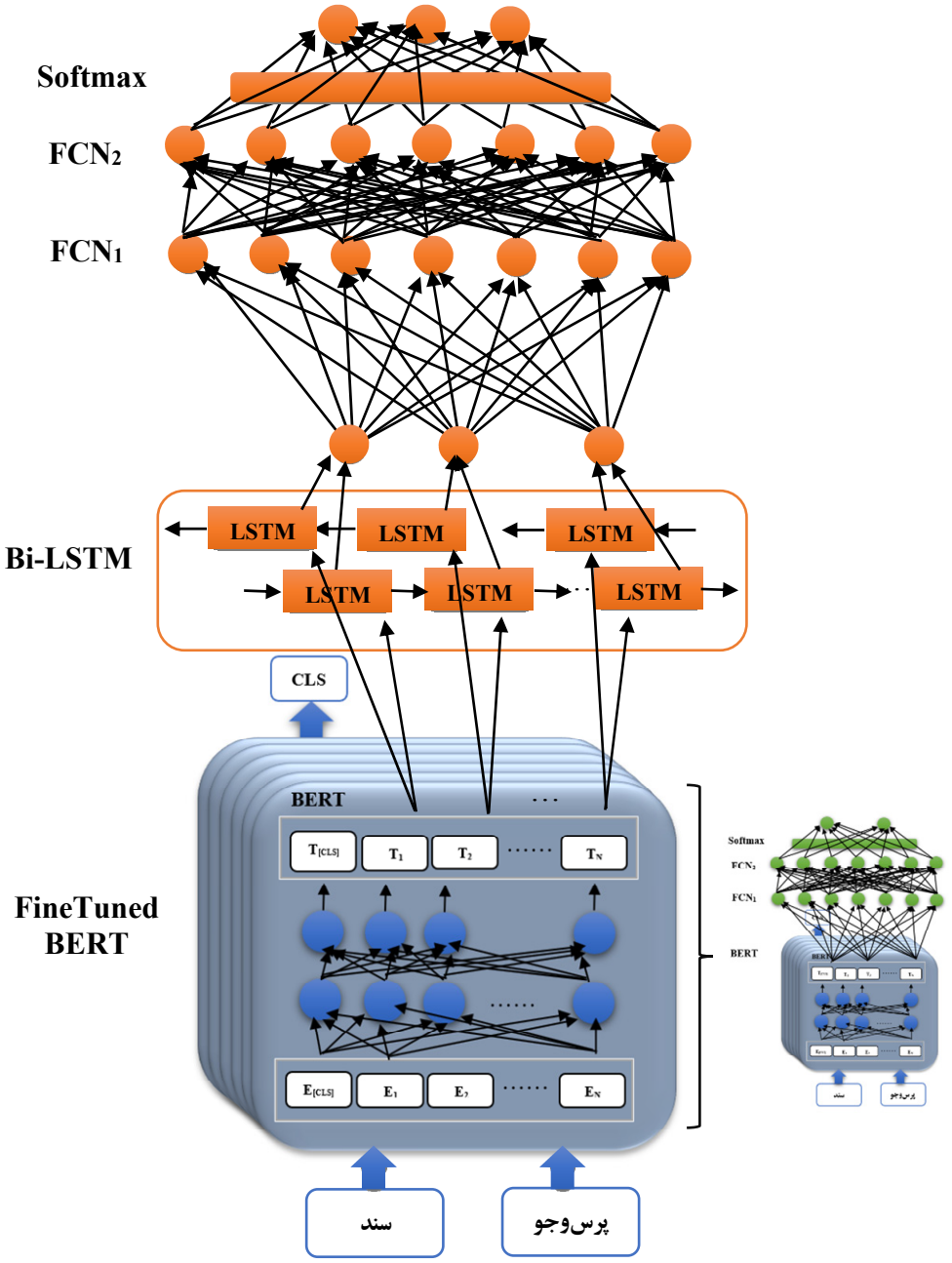
در این مرحله با بازنویسی شیوه تولید داده معنایی به‌صورت سفارشی، پرس‌وجوها به‌صورت دسته‌ای انتخاب می‌شوند و شناسه نشانه‌ها و برچسب‌ها تولید می‌گردند. سپس با بارگذاری مدل پیش‌آموزش‌یافته، فرایند رمزگذاری متون صورت می‌پذیرد. در مرحله بعد به ساخت مدل و رمزگذاری نشانه‌ها پرداخته می‌شود و مکانیزم خودتوجهی تعیین خواهد کرد که در هر جمله، کدام نشانه‌ها مورد توجه قرار گیرند. سپس مدل پیش‌آموزش‌یافته به‌منظور استفاده مجدد از ویژگی‌هایی که قبلاً آموزش دیده‌اند و بدون تغییر آنها به‌صورت ثابت درمی‌آید. در نهایت نوبت به افزودن لایه‌های قابل آموزش در بالای لایه‌های ثابت‌شده می‌رسد تا با ویژگی‌های قبلاً آموزش‌یافته تطابق یابند. شکل ۵ معماری این مدل را نشان می‌دهد. در این مدل از یک لایه LSTM دوطرفه در بالای مدل استفاده گردیده است [۲۳] و [۲۴].

این لایه که گسترش‌یافته LSTM سنتی است برای آموزش دوطرفه روی دنباله ورودی به کار می‌رود. LSTM دوم به صورت آینه‌وار از اولی عمل می‌کند و فرایند آموزش بر اساس گذشته و آینده ویژگی‌های ورودی در هر گام زمانی اعمال می‌گردد. دنباله ورودی به‌صورت هم‌زمان از دو جهت از LSTM عبور داده می‌شود. بعد از الحاق خروجی‌ها، سه درصد از وزن‌های لایه مخفی به‌صورت تصادفی به فراموشی سپرده می‌شوند و سپس با گذر از لایه‌های شبکه کاملاً متصل، از تابع بیشینه فعال‌ساز عبور داده می‌شوند.

۵-۲-۳ آموزش مدل

در این مرحله، آموزش فقط برای لایه‌های بالایی اعمال می‌شود تا ویژگی‌ها استخراج گردند، اما امکان استفاده از مدل پیش‌آموزش‌یافته نیز فراهم باشد. بعد از استخراج ویژگی‌های مدل پیش‌آموزش‌یافته به تنظیم دقیق مدل بر مبنای وظیفه مورد نظر پرداخته تا داده‌های جدید نیز پوشش داده شود. بنابراین ابتدا مدل از حالت ثابت خارج می‌شود و به وضعیت قابل آموزش درمی‌آید. این فرایند به دلیل تداوم آموزش با یادگیری کمتر اعمال می‌گردد تا داده‌های جدید به‌صورت تدریجی با ویژگی‌های از پیش‌آموزش‌یافته تطابق پیدا کنند. این فرایند می‌تواند به بهبود مدل منجر شود.

1. Head Layer



شکل ۵: معماری تنظیم دقیق دوم.

جدول ۴: دقت حاصل از آزمون مدل‌ها بر مبنای تنظیم دقیق.

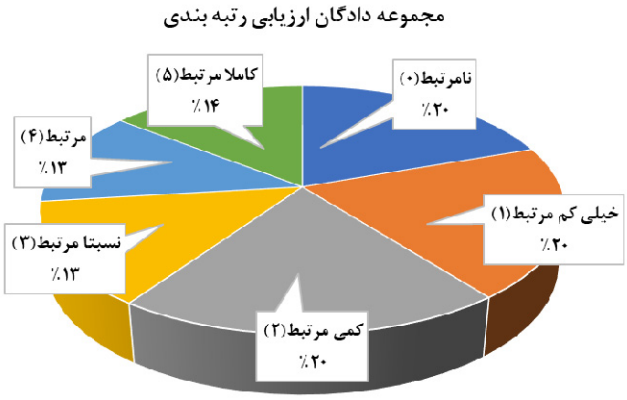
مدل	تنظیم دقیق اول (مجموعه دادگان دوکلاسه)	تنظیم دقیق دوم (مجموعه دادگان سه‌کلاسه)
	دقت	دقت
برت چندزبانه	۰/۷۹	۰/۸۰
برت سفارشی	۰/۸۰	۰/۸۱
پارس برت	۰/۸۱	۰/۸۳

۶- ارزیابی مدل‌های پیشنهادی

به‌منظور ارزیابی مدل‌های به‌دست‌آمده حاصل از تنظیم دقیق، معیار دقت در نظر گرفته شده است. برای این کار از مجموعه دادگان آزمون مربوط به هر مجموعه دادگان و به صورت مجزا استفاده گردیده است. این ارزیابی در زمان آموزش مدل و بر مبنای تقسیم مجموعه دادگان مورد استفاده به سه مجموعه آموزش، ارزیابی و آزمون محاسبه شده است.

جدول ۴ جزئیاتی از ارزیابی مدل بر مبنای مجموعه دادگان آزمون را نمایش می‌دهد. در ابتدا مجموعه دادگان آزمون با دو برچسب مرتبط و غیرمرتبط به منظور بررسی مدل به‌دست‌آمده از تنظیم دقیق اول مورد ارزیابی قرار می‌گیرد. دقت حاصل از مدل سفارشی آموزش‌یافته از پایه تحت عنوان برت سفارشی برابر با ۰/۸۰ محاسبه شده است. همین ارزیابی برای مدل برت چندزبانه و پارس‌برت اعمال می‌گردد که به ترتیب به دقت‌های ۰/۷۹ و ۰/۸۱ دست می‌یابند. در این ارزیابی، پارس‌برت بهترین دقت را کسب کرده و مدل برت سفارشی در جایگاه دوم و با دقت بالاتری از برت چندزبانه به دست آمده است. بنابراین دقت حاصل از مدل برت سفارشی با یک درصد افزایش نسبت به برت چندزبانه محاسبه گردیده که قابل توجه است.

در فرایند تنظیم دقیق دوم به‌جای انتخاب مدل پیش‌آموزش‌یافته پایه، از مدل به‌دست‌آمده از تنظیم دقیق اول به‌عنوان مدل پایه استفاده شده تا با تنظیم دقیق دوم، منجر به بهبود وزن‌ها در راستای هدف دسته‌بندی گردد. برای این ارزیابی از مجموعه دادگان آزمون با سه برچسب استفاده



شکل ۶: نمایی از جزئیات مجموعه دادگان جهت رتبه‌بندی اسناد.

گردیده است. دقت حاصل از این ارزیابی به ازای سه مدل برت سفارشی، چندزبانه و پارس‌برت به ترتیب برابر با ۰٫۸۱، ۰٫۸۰ و ۰٫۸۳ محاسبه گردیده است. بهبود دقت در مدل برت سفارشی نسبت به برت چندزبانه در این فرایند نیز به وضوح قابل مشاهده است که به دلیل مدل پایه غنی از واژگان مناسب فارسی می‌باشد. مدل پارس‌برت به عنوان مدل پیش‌آموزش‌یافته مناسب به دلیل آموزش مدل بر روی حجم عظیمی از مجموعه دادگان در زمان پیش‌آموزش و دیدن واژگان بیشتر در مدل خود، توانسته که در هر دو مرحله تنظیم دقیق، بهترین دقت را کسب نماید. اما مدل سفارشی برت نیز توانسته با تعداد لایه‌های کمتر و آموزش مناسب بر روی مجموعه دادگان وب فارسی به دقت مناسبی نسبت به پارس‌برت دست یابد. ضمن اینکه این مدل توانسته دقت حاصل از مدل چندزبانه برت را بهبود بخشد.

با توجه به اینکه مجموعه دادگان مورد ارزیابی در فرایند تنظیم دقیق اول و دوم متفاوت است، لذا نمی‌توان نتایج حاصل از دو فرایند تنظیم دقیق را با یکدیگر مقایسه کرد. به همین دلیل، ارزیابی دوم بر مبنای سنجش معیار nDCG [۲۵] در رتبه‌بندی اسناد بر مبنای پرس‌وجوی کاربر محاسبه می‌گردد. از این رو مطابق شکل ۶ از یک مجموعه دادگان مجزا با شش برچسب که بیانگر میزان مرتبط‌بودن سند به پرس‌وجوست استفاده گردیده است.

فرمول رتبه‌بندی مورد استفاده در این ارزیابی بر مبنای محاسبه کسینوس بردار درون‌سازی سند و پرس‌وجو است. این بردار درون‌سازی بر اساس جمله ورودی T که می‌تواند مطابق (۱) شامل سند d یا پرس‌وجوی q باشد، از روی مدل M_{BERT} مورد ارزیابی، مطابق (۲) استخراج می‌گردد

$$q,d \in T \qquad (۱)$$

$$\vec{e} = M_{BERT}(T) \qquad (۲)$$

سپس شباهت دو بردار معنایی طبق (۳) از طریق محاسبه کسینوس زاویه بین بردارها محاسبه می‌شود. در واقع این شباهت به ازای هر پرس‌وجو و سند محاسبه می‌گردد و سپس اسناد بر مبنای شباهت به‌دست‌آمده، مرتب و رتبه‌بندی می‌شوند

$$Similarity = \cos\theta = \frac{\vec{e}_q . \vec{e}_d}{|\vec{e}_q||\vec{e}_d|} \qquad (۳)$$

برای ارزیابی و مقایسه کیفیت رتبه‌بندی در بازیابی اطلاعات از معیار nDCG طبق (۴) استفاده گردیده است. در این رابطه که برای n نتیجه اول استفاده شده، r_j بیانگر درجه ارتباط سند j با پرس‌وجوی مربوط است [۲۵]

جدول ۵: ارزیابی مدل‌ها بر مبنای رتبه‌بندی.

مدل	(مجموعه دادگان پنج‌کلاسه)		
	پیش‌آموزش‌یافته (پایه)	تنظیم دقیق اول	تنظیم دقیق دوم
	nDCG	nDCG	nDCG
برت چندزبانه	۰٫۸۰	۰٫۸۲	۰٫۸۳
برت سفارشی	۰٫۷۹	۰٫۸۳	۰٫۸۴
پارس برت	۰٫۷۹	۰٫۸۳	۰٫۸۵

$$nDCG@n = \sum_{j=1}^{n} \frac{2^{r_j} - 1}{\log(1+j)} \qquad (۴)$$

مطابق نتایج حاصل از این ارزیابی که در جدول ۵ به آن اشاره شده است، دقت مدل‌ها با هر بار تنظیم دقیق، بهبود یافته است. بهترین دقت در هر مدل، مربوط به تنظیم دقیق دوم است که به خوبی، رشد مدل و بهبود وزن‌های مدل را بعد از هر بار فرایند آموزش نشان می‌دهد. برت سفارشی در ابتدا با کسب دقت ۰٫۷۹ محاسبه گردیده، اما بعد از هر مرحله تنظیم دقیق، این دقت بهبود یافته و در نهایت به دقت ۰٫۸۴ رسیده که نسبت به دقت اولیه مدل، ۵ درصد بهبود داشته است.

همین روند در خصوص مدل چندزبانه برت و پارس‌برت رخ داده و دقت این دو مدل به ترتیب از ۰٫۸۰ و ۰٫۷۹ به ۰٫۸۳ و ۰٫۸۵ افزایش یافته است. بنابراین به‌وضوح می‌توان به عملکرد مثبت فرایند تنظیم دقیق اول و دوم با معماری پیاده‌سازی‌شده و ساختار پیشنهادی در جهت بهبود دقت رتبه‌بندی بر مبنای معیار nDCG دست یافت که ارزشمند است. همان گونه که از نتایج برمی‌آید، دقت رتبه‌بندی بر مبنای مدل برت سفارشی در زمان تنظیم دقیق اول و دوم، بهتر از مدل چندزبانه برت محاسبه گردیده که نشان‌دهنده مدل آموزش‌یافته سفارشی غنی‌تر بر پایه زبان فارسی است. همچنین دقت رتبه‌بندی مدل برت سفارشی بر مبنای مدل پایه و تنظیم دقیق اول با مدل پارس‌برت یکسان به دست آمده، اما بر پایه مدل حاصل از تنظیم دقیق دوم، مدل پارس‌برت توانسته با یک درصد افزایش نسبت به مدل برت سفارشی عمل کند. در واقع می‌توان از نتایج دریافت که فرایند تنظیم دقیق در هر مرحله به‌خوبی عمل کرده و دقت را بهبود بخشیده است. همچنین دقت حاصل از مدل برت سفارشی با تعداد لایه‌ها و پیچیدگی کمتر اما مجموعه دادگان فارسی غنی‌تر نسبت به مدل چندزبانه برت، بهتر عمل کرده و توانسته که در هر مرحله از تنظیم دقیق، دقت را نسبت به آن بهبود بخشد. ضمن آنکه منجر به افزایش ۵ درصدی دقت در رتبه‌بندی اسناد وب فارسی از برت چندزبانه با دقت ۰٫۸۰ به پارس‌برت با دقت ۰٫۸۵ بر پایه فرایندهای تنظیم دقیق پیشنهادی شده است.

پیچیدگی زمانی در BERT به دو مؤلفه پیچیدگی زمانی هر لایه و تعداد عملیات متوالی وابسته است. با توجه به استفاده از مکانیزم خودتوجهی در الگوریتم BERT، پیچیدگی زمانی هر لایه برابر با $O(n^2 d)$ محاسبه می‌گردد. n برابر با طول دنباله یعنی تعداد واژگان و d بیانگر ابعاد درون‌سازی است. $O(n)$ عملیات برای هر واژه در نظر گرفته می‌شود که به دلیل توجه هر واژه به سایر واژگان در دنباله است. بنابراین $O(n^2)$ عملیات برای تمام واژگان لحاظ می‌شود. همچنین پیچیدگی تعداد گام‌های متوالی برابر با $O(1)$ است، زیرا تمامی n عملیات در یک گام زمانی اتفاق می‌افتد. بیشینه طول دنباله در مدل برت سفارشی برابر با ۱۲۸ در نظر گرفته شده که نسبت به مدل برت چندزبانه و پارس‌برت با طول ۵۱۲ کاهش یافته است. بنابراین پیچیدگی زمانی مدل برت سفارشی برابر با $O(n^2.d/3)$ محاسبه می‌گردد.

۷- نتیجه‌گیری

هدف از این پژوهش، ارائه راهکاری در درک بهتر منظور کاربر از پرس‌وجوی واردشده است. در این راستا به درون‌سازی پویای BERT به‌منظور آموزش واژگان و متون و استخراج بردار معنایی آنها پرداخته شده است. بهره‌گیری از الگوریتم BERT از دو طریق امکان‌پذیر است. در روش اول، مدل از پایه مورد آموزش قرار می‌گیرد و سپس بر مبنای یک وظیفه مشخص، تنظیم دقیق می‌گردد. روش دوم بر مبنای استفاده از مدل‌های پیش‌آموزش‌یافته موجود و تنظیم دقیق آنهاست. به‌منظور بررسی بیشتر بر روی متون فارسی از دو مدل پیش‌آموزش‌یافته چندزبانه برت و پارس‌برت استفاده شد و همچنین آموزش مدل از پایه بر روی مجموعه‌ای از دادگان اعمال گردید. در نهایت سه مدل موجود بر مبنای دو معماری متفاوت جهت تنظیم دقیق مدل مورد آموزش قرار گرفت. معماری‌های پیاده‌سازی‌شده برای فرایندهای تنظیم دقیق متوالی، مبتنی بر دسته‌بندی و بر اساس میزان ارتباط دو عبارت پرس‌وجو با سند است. در اولین مرحله تنظیم دقیق از مدل پیش‌آموزش‌یافته به عنوان مدل پایه استفاده شده و سپس با افزودن لایه‌های مختلف در بالای مدل به آموزش لایه‌ها بر مبنای مجموعه دادگان متشکل از پرس‌وجو و اسناد با دو برچسب مرتبط و غیرمرتبط می‌پردازد. این فرایند به بهبود مدل و وزن‌دهی هدفمند آن منجر می‌شود. در این معماری از دو لایه شبکه کاملاً متصل استفاده گردیده است. همچنین با استفاده از تابع بیشینه فعال‌ساز به پیش‌بینی و تنظیم دقیق مدل پرداخته می‌شود. این فرایند برای هر یک از سه مدل پیش‌آموزش‌یافته مورد نظر به‌صورت جداگانه اعمال گردیده است. نتایج حاصل از آزمون مدل بر مبنای مجموعه دادگان آزمون، بیانگر بهبود دقت در برت سفارشی نسبت به برت چندزبانه و نزدیک به مدل پارس‌برت است. فرایند تنظیم دقیق دوم از خروجی مدل آموزش‌یافته از تنظیم دقیق اول استفاده می‌کند و با درنظرگرفتن آن به عنوان مدل پایه، فرایند تنظیم دقیق دوم آغاز می‌شود. در معماری تنظیم دقیق دوم که مبتنی بر دسته‌بندی است از یک لایه LSTM دوطرفه استفاده گردیده است. این مدل با افزودن دو لایه شبکه کاملاً متصل به آموزش و بهبود وزن‌ها می‌پردازد. این فرایند از مجموعه دادگان متفاوت با سه برچسب استفاده می‌کند. نتایج حاصل از این فرایند بیانگر بهبود دقت برت سفارشی نسبت به برت چندزبانه تا حداقل یک درصد است.

به‌منظور ارزیابی بیشتر و مقایسه بین دو فرایند تنظیم دقیق از یک مجموعه دادگان با شش برچسب بر مبنای میزان مرتبط‌بودن اسناد به پرس‌وجوی کاربر استفاده گردید. در این بررسی، رتبه‌بندی اسناد بر مبنای محاسبه کسینوس زاویه بین دو بردار درون‌سازی حاصل از عبارت پرس‌وجو و سند محاسبه شد. نتایج بیانگر بهبود دقت بر مبنای معیار nDCG در هر مرحله از تنظیم دقیق نسبت به مرحله قبلی به ازای تمام مدل‌های مورد ارزیابی است. همچنین مدل برت سفارشی توانسته با بهبود دقت نسبت به مدل برت چندزبانه به ازای فرایندهای تنظیم دقیق اول و دوم ظاهر گردد. در واقع نتایج رتبه‌بندی بر مبنای مدل‌های نهایی، بیانگر بهبود دقت رتبه‌بندی وب فارسی نسبت به مدل‌های پایه مورد ارزیابی با افزایش حدود ۵ درصدی دقت در بهترین حالت است. بنابراین هرچه مدل آموزش‌یافته از پایه غنی‌تر باشد و فرایند تنظیم دقیق با ساختار مناسب به درستی بر روی آن اعمال گردد، می‌تواند اثر قابل توجهی در رتبه‌بندی بهتر و مرتب‌سازی دقیق‌تر اسناد داشته باشد.

هرچه واژه‌نامه و طول دنباله ورودی بزرگ‌تر باشد، واژگان وسیع‌تری را در بر می‌گیرد و می‌تواند به بهبود دقت کمک کند. اما ذکر این نکته ضروری است که با افزایش طول دنباله ورودی و واژه‌نامه، هزینه پردازش نیز افزایش خواهد یافت [۱۱]. در این مقاله علی‌رغم تغییر پارامترها و کاهش آنها با استفاده از مجموعه دادگان متمرکز در حوزه وب و استفاده از آن در همان حوزه و نوآوری در فرایندهای تنظیم دقیق متوالی، دقت بهبود یافته و منجر به کاهش هزینه پردازش نیز گردیده که ارزشمند است. بنابراین در صورتی که واژه‌نامه و طول دنباله ورودی، مشابه سایر روش‌های رایج افزایش یابد، می‌تواند به بهبود نسبی دقت، نسبت به نتایج فعلی نیز منجر گردد، اما به دلیل هزینه پردازشی از آن صرف‌نظر گردیده است. ضمن اینکه دقت در نتایج به‌دست‌آمده، نسبت به مدل‌های موجود به‌صورت قابل قبول ارائه شده است.

مجموعه دادگان وب مورد استفاده جهت فرایند پیش‌آموزش، شامل ۶۸۱ میلیون عنوان سند است. در آموزش مدل BERT به فرمت مناسبی از دادگان نیاز است و بنابراین مجموعه دادگان طبق استاندارد مورد نظر، فرمت‌دهی شده است. فرایند آموزش به صورت دسته‌ای اعمال می‌گردد. هر دسته شامل ۱۰ میلیون سند بوده و مدل بر اساس جملات با فرمت مناسب مورد آموزش قرار گرفته است. نتایج به‌دست‌آمده در این مقاله بر اساس پیش‌آموزش مدل روی دسته اول از مجموعه دادگان ارائه گردیده است. امکان ادامه آموزش بر مبنای بقیه مجموعه دادگان در دسته‌های بعدی وجود دارد و به عنوان کارهای آینده در حال انجام است.

از محدودیت‌های مدل BERT می‌توان به بزرگ‌بودن مدل به دلیل فرایند آموزش وسیع آن اشاره نمود. به‌دلیل وجود وزن‌های فراوانی که باید در فرایند آموزش به‌روزرسانی گردند، فرایند آموزش زمان‌بر می‌باشد. همچنین به دلیل محاسبات فراوان در لایه‌های مختلف، هزینه بالایی در بر دارد. به همین دلیل در این مقاله به‌صورت نوآورانه بر روی مجموعه دادگان در حوزه وب تمرکز گردید. در این راستا یک سری از پارامترها در فرایند آموزش تغییر داده شد تا به بهبود آموزش و سرعت اجرا کمک کند. سپس فرایند آموزش بر روی مجموعه دادگان وب فارسی مورد آموزش قرار گرفت و کارایی مدل بر اساس همان حوزه بررسی گردید که نتایج جالبی دربرداشت. برای استفاده از مدل BERT در حوزه‌های دیگر و سایر کارهای پردازش زبانی، لازم است که از مجموعه دادگان مناسب آن حوزه استفاده گردد که در آینده به آن پرداخته خواهد شد.

الگوریتم BERT به‌عنوان یک مدل زبانی پویا معرفی گردیده است. در مدل‌های پویا برای هر واژه، متناسب با جمله‌ای که آن واژه در آن ظاهر می‌شود، بردار معنایی متفاوتی تولید می‌گردد. این الگوریتم به دلیل پردازش‌های متعدد در زمان دریافت هر جمله به صورت آنلاین، نسبت به الگوریتم‌های ایستایی که تمام پردازش‌ها را به فاز برون‌خط انتقال می‌دهند، هزینه بیشتری را متحمل می‌شود. به عنوان یک راهکار می‌توان از الگوریتم‌های درون‌سازی ایستا همچون الگوریتم موفق Word۲vec در کنار الگوریتم BERT استفاده کرد و برای واژگان تک‌معنی و واژگان کم‌اهمیت‌تر از نمونه ایستای آن بهره‌مند گردید. این روش می‌تواند منجر به کاهش هزینه پردازش‌های آنلاین شود و در صورتی که به کاهش دقت منجر نشود، ارزشمند خواهد بود. این مسئله به عنوان کارهای آینده مورد بررسی قرار خواهد گرفت.

مراجع

[1] A. Bidoki, *Effective Web Ranking and Crawling*, Ph.D. Thesis, University of Tehran, 2009.

[2] W. Qader, M. Ameen, and B. Ahmed, "An overview of bag of words; importance, implementation, applications, and challenges," in *Proc. IEEE Int. Engineering Conf., IEC'19,* pp. 200-204, Erbil, Iraq, 23-25 Jun. 2019.

[3] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management,* vol. 24, no. 5, pp. 513-523, 1988.
[4] Y. Benjio and R. Ducharme, "A neural probabilistic language model," *The J. of Machine Learning Research*, vol. 3, pp. 1137-1155, 2003.
[5] T. Mikolov, K. Chen, G. Corrado, and J. Dea, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. on Learning Representations, ICLR'13*, pp. 1137-1155, Scottsdale, AZ, USA, 2-4 May 2013.
[6] T. Mikolov, I. Sutskever, K. Chen, and G. Corr, "Distributed representations of words and phrases and their compositionality," In C. J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (ed.), *Annual Conf. on Neural Information Processing Systems, NIPS'13*, vol. 2, pp. 3111-3119, Lake Tahoe, NV, USA, 5-10 Dec. 2013.
[7] J. Pennington, R. Socher, C. Ma, and C. Manning, "GloVe: global vectors for word representation," in *Proc. Conf. on Empirical Methods in Natural Language Processing, EMNLP'14*, pp. 1532-1543, Doha, Qatar, Oct. 2014.
[8] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. of the Association for Computational Linguistics (TACL)*, vol. 5, pp. 135-146, 2017.
[9] S. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010.
[10] M. Peters, *et al*., "Deep contextualized word representations," in *Proc. Conf. of the North American Chapter of the Association of Computational Linguistics, HLT-NAACL'18*, vol. 1, pp. 2227-2237, New Orleans, LA, USA, Jun. 2018.
[11] J. Devlin, M. Chang, and K. Kristina, "BERT: pre-training of deep bidirectional transformers for language understanding," in Proc. *Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, HLT-NAACL'19*, pp. 4171-4186, Minneapolis, MN, USA, 2-7 Jun. 2019.
[12] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, *Improving Language Understanding by Generative Pre-Training*, Technical Report, OpenAI, 11 Jun. 2018.
[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, Nov. 1997.
[14] T. Mikolov, S. Kombrink, L. Burget, and J. Cernocky, "Extensions of recurrent neural network language model," in *Proc. IEEE Int. Speech and Signal Processing, ICASSP'11,* pp. 5528-5531, Prague, Czech Republic, 22-27 May 2011.
[15] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, Nov. 1997.
[16] A. Vaswani, *et al*., "Attention is all you need," In *Proc. 31st Annual Conf. on Neural Information Processing Systems, NIPS'17*, 11 pp., Long Beach, CA, USA, 4-9 Dec. 2017.
[17] Z. Lan, et al., A Lite BERT for Self-Supervised Learning of Language Representations, arXiv preprint arXiv:1909.11942, 2019.
[18] Y. Liu, *et al*., *A Robustly Optimized BERT Pretraining Approach*, arXiv preprint arXiv:1907.11692, 2019.
[19] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *DistilBERT, A Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*, arXiv preprint arXiv:1910.01108, 2019.
[20] M. Farahani, M. Gharachorloo, M. Farahani, and M. Manthouri, " ParsBERT: transformer-based model for persian language understanding," *Neural Processing Letters*, vol. 53, pp. 3831-3847, 2021.
[21] BERT, "huggingface," 2018. Available: https://huggingface.co/docs/transformers/.
[22] C. Sun, X. Qiu, Y. Xu, and X. Huang, "How to fine-tune BERT for text classification?" in *Proc. China National Conf. on Chinese Computational Linguistics*, *CCL'19,* pp. 194-206, Kunming, China, 18-20 Oct. 2019.
[23] D. Viji and S. Revathy, "A hybrid approach of weighted fine-tuned BERT extraction with deep siamese bi-LSTM model for semantic text similarity identification," *Multimedia Tools and Applications*, vol. 81, pp. 6131-6157, 2022.
[24] A. Agarwal and P. Meel, "Stacked bi-LSTM with attention and contextual BERT embeddings for fake news analysis," in *Proc. 7th Int. Conf. on Advanced Computing and Communication Systems, ICACCS'21*, pp. 233-237, Coimbatore, India, 19-20 Mar. 2021.
[25] K. Jarvelin and J. Kekalainen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. on Information Systems*, vol. 20, no. 4, pp. 422-446, Oct. 2002.

شکوفه بستان دانشجوی دکتری در رشته مهندسی کامپیوتر با گرایش نرم‌افزار در دانشگاه یزد است. او در حال حاضر به عنوان مدرس در دانشکده مهندسی کامپیوتر دانشگاه یزد و همچنین به عنوان توسعه‌دهنده نرم‌افزار در یک شرکت برجسته جستجوی ابری فعالیت دارد. زمینه های تحقیقاتی مورد علاقه ایشان شامل یادگیری عمیق، بازیابی معنایی اطلاعات و تحلیل معنایی شبکه‌های اجتماعی است.

علی‌محمد زارع بیدکی تحصیلات خود را در مقطع کارشناسی در سال ۱۳۷۸ از دانشگاه صنعتی اصفهان و مقاطع کارشناسی ارشد و دکتری کامپیوتر به‌ترتیب در سال‌های ۱۳۸۱ و ۱۳۸۸ از دانشکده فنی دانشگاه تهران به پایان رسانده است و هم‌اکنون عضو هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه یزد می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان شامل بازیابی اطلاعات، موتورهای جستجو، رتبه بندی و پردازش زبان‌های طبیعی است.

محمدرضا پژوهان، استادیار گروه مهندسی کامپیوتر دانشگاه یزد است. او دکترای خود را در بخش علوم کامپیوتر از دانشگاه ساینس مالزی (USM) و دانشگاه ملی سنگاپور (NUS) اخذ کرده است. ایشان فارغ‌التحصیل کارشناسی و کارشناسی ارشد مهندسی کامپیوتر از دانشگاه صنعتی شریف است. علایق تحقیقاتی ایشان شامل پایگاه داده، داده کاوی، علوم داده و حفظ حریم خصوصی در انتشار داده‌هاست.