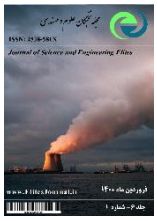
مجله نخبگان علوم و مهندسی

Journal of Science and Engineering Elites

جلد ۶- شماره ۱- سال ۱۴۰۰

www.ElitesJournal.ir



طراحی سیستم پاسخ به پرسش‌های کوید ۱۹ با استفاده از مدل برت–اسکواد

محسن یزدی‌نژاد[۱*]، حسن کمالو[۲]، احسان مجاهدی[۳]، عباس کلوانی[۴]

۱–کاندیدای دکتری هوش مصنوعی، دانشگاه اصفهان

۲–کارشناسی ارشد مدیریت اجرایی، دانشگاه اصفهان

۳–کارشناسی ارشد مهندسی دانش و علوم تصمیم، دانشگاه تهران

۴– کارشناسی ارشد معماری کامپیوتر، دانشگاه آزاد واحد علوم و تحقیقات

*mohsen.yazdinejad@gmail.com

ارسال: اسفند ۹۹ پذیرش: فروردین ۱۴۰۰

چکیده

مقدمه: از زمان شیوع ویروس کرونا، بیماری کرونا تهدیدی برای سلامتی و زندگی مردم جهان شده است، چرا که این سندرم حاد تنفسی به یک بیماری همه‌گیر جهانی تبدیل شده است. گسترش جهانی کوید۱۹ فشار فزاینده‌ای بر سازمان بهداشت جهانی، وارد کرده است. برای بررسی روند درمان و سایر اطلاعات پزشکی در مورد این بیماری نیاز به جستجو در حجم انبوه مقالات و دستاوردهای علمی تولید شده و بروز است. محتویات فنی و طولانی بودن مقالات زیست پزشکی کار جستجو را پیچیده‌تر می‌کند. سیستم پاسخ به پرسش در زیست پزشکی چالشی است که در سال‌های اخیر مورد توجه قرار گرفته است. به دلیل پیچیدگی زبان طبیعی، پیشرفت قابل توجهی در حوزه زیست پزشکی صورت نگرفته است.

روش: در این تحقیق یک سیستم پاسخ به پرسش‌های کوید ۱۹ پیشنهاد می‌شود. روش پیشنهادی برای پاسخ به پرسش‌ها از جدیدترین یافته‌های علمی در پایگاه داده تحقیقات باز کوید ۱۹ و خزش در سایت‌های معتبر علمی و پزشکی که به صورت روزانه بروز رسانی می‌شوند، استفاده می‌کند. برای یافتن پاسخ‌ها از مدل برت–اسکواد بهره می‌گیرد و در نهایت، نتایج رتبه‌بندی می‌شود و پاسخ نهایی به صورت خلاصه صادر می‌شود.

نتایج: ارزیابی روش پیشنهادی نشان می‌دهد، سیستم پاسخ به پرسش‌های کوید ۱۹ با استفاده از مدل برت–اسکواد با دقت ۹۲/۳ درصد، کارایی مناسبی دارد.

کلید واژه ها: سیستم پاسخ به پرسش، سیستم بازیابی سند، کوید۱۹، مدل برت–اسکواد.

۱–مقدمه

از زمان شیوع ویروس کرونا[1] در ووهان چین، در دسامبر ۲۰۱۹، بیماری کرونا تهدیدی برای سلامتی و زندگی مردم جهان شده است، چرا که این سندرم حاد تنفسی به یک بیماری همه‌گیر جهانی تبدیل شده است. شیوع ویروس جدید کرونا (SARS- CoV2) بنام کوید ۱۹، تلاش‌های بین‌المللی بهداشت عمومی و مراقبت‌های پزشکی را به چالش کشانده است[۱].

در ۱۱ مارس ۲۰۲۰، سازمان بهداشت جهانی[2]، کوید ۱۹ را یک بیماری همه‌گیر اعلام کرد و از آن زمان روند بی‌سابقه‌ای را طی کرده است[۲]. تا ۲۱ ژانویه ۲۰۲۱، بیش از ۱۰۵ میلیون مورد ابتلا و بیش از دو میلیون مورد مرگ مرتبط با کوید ۱۹در سراسر جهان گزارش شده است [۳].

گسترش جهانی کوید۱۹، فشار فزاینده‌ای بر سازمان بهداشت جهانی وارد کرده است. در این زمینه، چندین استراتژی برای مقابله با بیماری همه‌گیر از جمله افزایش بودجه سازمان بهداشت جهانی اتخاذ شده است. با این حال، افزایش بودجه به خودی‌خود باعث تسهیل بار کار متخصصان بهداشت نمی‌شود، زیرا علاوه بر تعداد زیاد بیماران، تعداد زیادی از کارمندان آلوده نیز وجود دارند که انجام کار را با مشکل روبرو می‌کند [۴]. تعداد بیماران اخیر بی‌سابقه است و بسیاری از خدمات بهداشتی را تحت تأثیر قرار می‌دهد[۶].

برای بررسی روند درمان و سایر اطلاعات پزشکی در مورد این بیماری نیاز به جستجو در حجم انبوه مقالات و دستاوردهای علمی تولید شده و بروز است. محتویات فنی و طولانی بودن مقالات زیست پزشکی کار جستجو را پیچیده‌تر می‌کند [۵]. مطالعات اخیر در مورد استراتژی‌های جستجوی تخصصی در میان متخصصان اطلاعات مراقبت‌های بهداشتی نشان می‌دهد که برای یک جستجوی خاص، متخصصان به طور متوسط ۶۰ دقیقه در هر مجموعه یا پایگاه داده، ۳ دقیقه برای بررسی ارتباط هر سند و در کل حدود ۴ ساعت زمان صرف می‌کنند [۶]. با فراوانی منابع اطلاعاتی در حوزه پزشکی و ویروس کوید۱۹، متخصصان با چالش روبرو می‌شوند، چالشی که نیاز به راه حل‌های اختصاصی دارد که می‌تواند با ناهمگنی و مشخصات اطلاعات مربوط به سلامت سازگار شود.

سیستم‌های پاسخ به پرسش[3] تخصصی یکی از راه‌حل‌های مناسب برای حل این مشکل است؛ زیرا برای درک پرسش‌ها زبان طبیعی بدون اتکا به اطلاعات خارجی در مورد کاربران طراحی شده است [۷]. پاسخ دادن به پرسش، شناسایی پاسخ‌های کوتاه به پرسش‌های کاربر که با عبارات زبان طبیعی نوشته شده‌اند، مسئله‌ای دیرینه است که طی دهه‌های گذشته مورد بررسی قرار گرفته [۸].

باوجود در دسترس بودن منابع و مجموعه داده‌های متعدد برای بازیابی پاسخ، پاسخ به پرسش به دلیل دشواری درک پرسش و استخراج پاسخ، همچنان یک مشکل چالش برانگیز است [۹] و سیستم پاسخ به پرسش هنوز به عنوان یک چالش واقعی در حوزه زیست پزشکی باقی مانده است؛ زیرا اکثر سیستم‌های موجود از تعداد محدودی از انواع پرسش و پاسخ پشتیبانی می‌کنند و برای بهبود عملکرد آنها از نظر دقت در پرسش‌ها پشتیبانی شده، به تلاش بیشتری نیاز دارند.

انواع پرسش‌ها و پاسخ آنها را می‌توان در چهار گروه دسته بندی کرد[۸]:

- پرسش‌های بله/ خیر: پرسش‌هایی که به پاسخ «بله» یا «نه» نیاز دارند.
- پرسش‌ها فاکتوئیدی[4]: پرسش‌هایی که به عنوان موجودی خاص (به عنوان مثال، از یک بیماری، دارو یا ژن)، یک عدد یا یک عبارت کوتاه مشابه به عنوان پاسخ نیاز دارند.
- پرسش‌های لیستی: پرسش هایی که لیستی از نام موجودیت‌ها (به عنوان مثال، لیستی از نام ژن‌ها، لیستی از نام داروها)، اعداد یا عبارات کوتاه مشابه را به عنوان پاسخ انتظار دارند.

[1] COVID-19
[2] World Health Organization (WHO)
[3] Question answering system
[4] Factoid

• پرسش‌های خلاصه: پرسش‌هایی که پاسخ خلاصه و کوتاه دارند.

در این تحقیق یک سیستم پاسخ به پرسش برای بیماری کوید ۱۹ پیشنهاد می‌شود. سیستم پیشنهادی[۵] می‌تواند بطور موثر در جدیدترین مقالات و پایگاه‌های علمی بیماری کوید ۱۹ جستجو کند و پس از بازیابی اسناد مرتبط، با استفاده از مدل برت محدوده‌های پاسخ را شناسایی و رتبه بندی کند و مرتبط‌ترین پاسخ‌ها را به صورت پاسخ خلاصه به کاربر تحویل دهد.

سازمان‌بندی این تحقیق به شرح زیر است: در بخش ۲، مروری بر کارهای پیشین در زمینه سیستم‌های پاسخ به پرسش زیست پزشکی انجام می‌شود. در بخش ۳، روش پیشنهادی بطور کامل شرح داده می‌شود. در بخش ۴، روش پیشنهادی مورد ارزیابی قرار می‌گیرد و در بخش ۵ بحث و نتیجه گیری انجام می‌شود و پیشنهاداتی برای تحقیق‌های آتی ارائه می‌شود.

۲- مرور کارهای پیشین

سیستم پاسخ به پرسش در زیست پزشکی چالشی است که در سال‌های اخیر مورد توجه قرار گرفته است. به دلیل پیچیدگی زبان طبیعی، پیشرفت قابل توجهی در حوزه زیست پزشکی صورت نگرفته است. با توجه به اینکه از زمان پیدایش بیماری کوید ۱۹، زمان کمی سپری شده است، طبق تحقیق ما سیستم پاسخ به پرسش مربوط به این بیماری هنوز طراحی نشده است. بنابراین در این بخش مروری بر تعدادی از تحقیق‌ها در مورد سیستم‌های پرسش و پاسخ زیست پزشکی انجام می‌شود.

لی و همکاران در [۱۰]، یک سیستم پاسخ به پرسش پزشکی به نام MedQA ایجاد کردند که از پنج مولفه تشکیل شده است که شامل (۱) طبقه‌بندی پرسش‌ها، (۲) ایجاد پرس‌وجو، (۳) بازیابی سند، (۴) استخراج پاسخ و (۵) خلاصه‌سازی متن است. در MedQA، در ابتدا، مولفه طبقه‌بندی پرسش‌ها به طور خودکار پرسش‌های پزشکی را در دسته‌بندی ایجاد شده بر اساس رویکردهای یادگیری ماشین تحت نظارت طبقه‌بندی می‌کند. در مرحله بعد، مولفه بازیابی اسناد از اصطلاحات پرس‌وجو برای بازیابی اسناد مربوطه یا از اسناد وب با استفاده از گوگل[۵] یا شرکت‌های MEDLINE با نمایه محلی از لوسن[۶] استفاده می‌کند. سپس، مولفه استخراج پاسخ، از اسناد بازیابی شده جملات مربوطه را شناسایی می‌کند که براساس الگوهای نحوی/واژه‌ای به پرسش‌ها پاسخ می‌دهند. سرانجام، مولفه جمع‌بندی متن، جملات زائد را حذف می‌کند و جملات را به یک خلاصه منسجم تبدیل می‌کند که به عنوان پاسخ در نظر گرفته می‌شود. اگرچه سیستم MedQA خلاصه‌های کوتاهی را ارائه می‌دهد که به طور بالقوه می‌توانند به پرسش‌های پزشکی پاسخ دهند، اما ظرفیت فعلی MedQA محدود است و فقط پاسخ پرسش‌های تعریف شده را ارائه می‌دهد.

نویسندگان در [۱۱] یک سیستم پاسخ به پرسش پزشکی معنایی به نام مینز[۷] ارائه دادند. مینز از سه فاز اصلی تشکیل شده است: (۱) حاشیه نویسی، (۲) تجزیه و تحلیل و طبقه‌بندی پرسش، و (۳) جستجوی پاسخ. نویسندگان از روش‌های پردازش زبان طبیعی برای شناسایی نهاد و استخراج رابطه نامگذاری استفاده کردند تا حاشیه نویسی‌های RDF[۸] از اسناد منبع و پرس‌وجوهای اسپارکل[۸] را که یک نوع زبان پرسمان معنایی برای پایگاه‌های داده است را نشان دهند که پرسش‌های کاربران را نشان می‌دهد. آنها بیشتر از هستی‌شناسی MESA برای نمایش مفاهیم و روابط بین آنها به منظور ساخت ترجمه های اسپارکل، از پرسش‌های زبان طبیعی استفاده کردند. برای استخراج پاسخ، پرس و جوهای اسپارکل به منظور بازیابی از سه گانه RDF ساخته شده در مرحله حاشیه‌نویسی اجرا شد. نویسندگان، بر روی پرسش‌های فکتوئید و بله/ خیر متمرکز شدند، درحالیکه فرایندهای خاص‌تری برای رسیدگی به پرسش‌های پیچیده مورد نیاز است (به عنوان مثال چرا، چه زمانی).

5 Google
6 Lucene
7 MEANS
8 SPARQL

کراوس و همکاران [۱۲]، سیستم اولئو[9] را برای کاوش بصری در ادبیات زیست پزشکی توسعه دادند. سیستم اولئو از سه ماژول اصلی تشکیل شده است: (۱) پردازش پرسش، (۲) بازیابی سند/گذرگاه و (۳) پردازش پاسخ. در اولئو، ماژول پردازش پرسش بر اساس سیستمی است که در [۱۳] شرح داده شده است. در ماژول دوم، اولئو ابتدا از نشانه‌ها و اصطلاحات همسان برای پرس و جو در پایگاه داده استفاده می‌کند تا خلاصه‌ای از اسناد بازیابی شده را بیابد و سپس خلاصه‌های به دست آمده را با توجه به وقوع و اهمیت نشانه‌های جستجو شده رتبه‌بندی می‌کند. سرانجام، بسته به نوع پرسش توسط ماژول سوم، پاسخی به کاربر صادر می‌شود. اگرچه اولئو ثابت کرده است که در پاسخ دادن به پرسش‌های زیست پزشکی کاملاً موفق است، اما در حال حاضر، اولئو فقط از سه نوع پرسش از جمله پرسش‌ها فکتوئید، لیستی و خلاصه پشتیبانی می‌کند. در واقع، پرسش‌های بله/ خیر را پشتیبانی نمی‌کند که یکی از پیچیده‌ترین پرسش‌ها برای پاسخ دادن است، زیرا آنها به دنبال پاسخ روشن «بله» یا «خیر» هستند. علاوه بر این برای بهبود عملکرد خود از نظر دقت به انواع پرسش‌ها پشتیبانی شده، هنوز به تلاش‌های بیشتری نیاز دارد.

بر اساس رویکردهای مبتنی بر یادگیری، یانگ و همکاران [۱۴]، توسعه یک سیستم با کیفیت پزشکی زیست پزشکی را توصیف کرده است که فقط پاسخ‌های دقیق پرسش‌ها فاکتوئید و لیست را برمی‌گرداند. آنها دو الگوریتم تحت نظارت طراحی کردند که در هر دو نوع پرسش مشترک هستند: (۱) پیش بینی نوع پاسخ و (۲) امتیازدهی پاسخ کاندیدا. آنها از چندین ویژگی (به عنوان مثال، لما، نوع معنایی هر مفهوم و غیره) و رگرسیون لجستیک به عنوان یک طبقه‌بندی کننده چندطبقه‌ای برای شناسایی نوع پاسخ معنایی یک پرسش استفاده کردند. از طرف دیگر، نویسندگان مفاهیم را به عنوان پاسخ‌های کاندیدا در نظر گرفته و آنها را بر اساس روش ضرب امتیازدهی کردند تا از این طریق نمره اجبار نوع و تعداد وقوع و همچنین رگرسیون لجستیک را برای یادگیری عملکرد امتیازبندی ترکیب کنند.

در [۱۵]، معماری سیستم و ارزیابی سیستم پاسخ به پرسش پزشکی با نام SemBioNLQA، که شامل ماژول‌های طبقه‌بندی پرسش، بازیابی سند، بازیابی گذرگاه و استخراج پاسخ است، توصیف می‌شود. پرسش‌ها مربوط به زبان طبیعی را به عنوان ورودی در نظر می‌گیرد، و هم جواب های کوتاه و هم خلاصه‌ها را به عنوان نتیجه ارائه می‌دهد. سیستم SemBioNLQA که به چهار نوع پرسش می‌تواند پاسخ دهد، مبتنی بر چهار گام است: (۱) الگوهای ترکیبی دست ساز و الگوریتم یادگیری ماشین برای طبقه بندی پرسش‌ها، (۲) موتور جستجوی PubMed و شباهت UMLS برای بازیابی سند، (۳) مدل BM25 و کلمات بنیادی و مفاهیم UMLS برای بازیابی گذرگاه، و (۴) متاثوروس[10] UMLS ، مترادف BioPortal، تجزیه و تحلیل احساسات و معیار فرکانس اصطلاح برای استخراج پاسخ. این سیستم قادر به پاسخ به چهار نوع پرسش فکتوئیدی، لیستی، خلاصه و بله و خیر است.

در [۱۶] یک رویکرد جدید مبتنی بر شناخت رسمی پرسش (RQE)[11] پیشنهاد می‌شود و سیستم و منابع پاسخ به پرسش که از پرسش‌های پزشکی واقعی طراحی و ارزیابی شده توصیف می‌شود. ابتدا، از روش‌های مختلف رگرسیون لجستیک و یادگیری عمیق برای RQE با استفاده از انواع مختلف مجموعه داده شامل استنتاج متن، شباهت پرسش، و نتیجه در هر دو حوزه باز و بالینی استفاده می‌شود. در مرحله دوم نویسندگان، مدل‌های بازیابی اطلاعات را با بهترین روش RQE ترکیب می‌کنند تا پرسش‌ها الزامی را انتخاب کنند و پاسخ‌های بازیابی شده را رتبه‌بندی کنند.

از آنجا که مدل‌های یادگیری عمیق به مقدار زیادی از داده های آموزشی نیاز دارند، یک مدل یادگیری انتقالی مبتنی بر سلسله مراتب در [۱۷]، برای پاسخگویی به پرسش‌ها پزشکی ارائه شده است. از مدل برت[12] (ترانسفورماتور نمایندگی رمزگذار دو طرفه) که توانایی یادگیری از داده های بدون نظارت در مقیاس بزرگ دارد، برای غنی سازی نمایشی معنایی در این مدل استفاده شده است. از روش انطباق

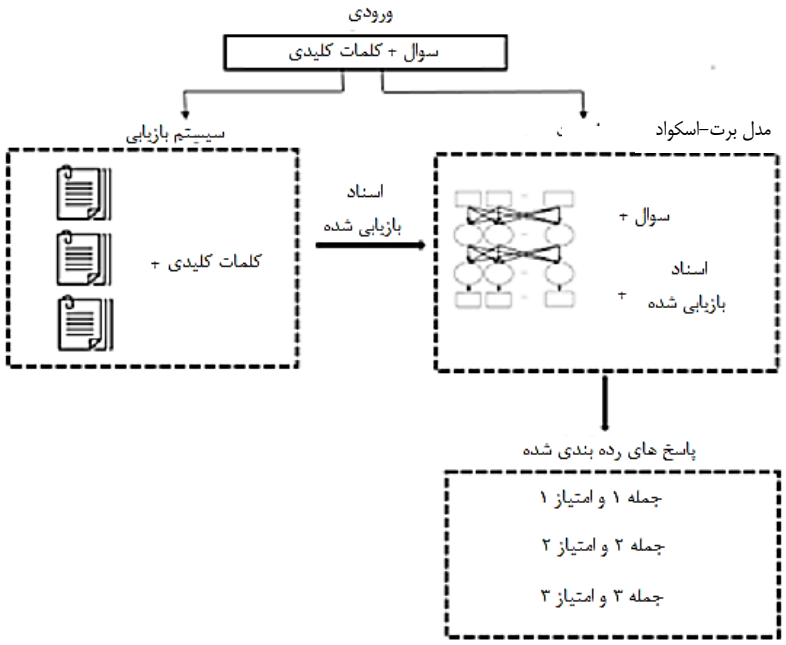
[9] Olelo
[10] metathesaurus
[11] Recognizing Question Entailment (RQE)
[12] BERT

دامنه برای تقویت عملکرد در داده های کوچک BioASQ استفاده شده است. این مدل با مجموعه داده دامنه باز SQuAD در مقیاس بزرگ آموزش دیده و تنظیم دقیق آن روی داده کوچک هدف BioASQ انجام شده است. نتایج تجربی نشان می‌دهد که این مدل عملکرد مناسبی را در مورد پاسخ به سولات فکتوئیدی زیست پزشکی بدست می‌آورد. اما در پاسخ به پرسش‌ها خلاصه عملکرد مناسبی ندارد.

با توجه به آنچه مرور شد، تا کنون سیستم پاسخ به پرسش مناسبی خاص پاسخ به پرسش‌های کوید ۱۹ ارائه نشده است، در ادامه روش پیشنهادی که سیستم پرسش و پاسخ تخصصی کوید ۱۹ می‌باشد، شرح داده می‌شود.

۳-روش پیشنهادی

در این تحقیق روشی برای پاسخ به پرسش‌های گوید ۱۹ ارائه می‌شود. روش پیشنهادی مانند بسیاری مقالات دیگر، از دو مرحله اصلی تشکیل یافته است؛ یک سیستم بازیابی سند و یک مدل پیش آموزشی برت-اسکوا[۱۳]د برای تشخیص نهایی پاسخ‌ها. سیستم بازیابی سند، برای محدودکردن تعداد مقالات مرتبط با پرسش مطرح شده توسط کاربر استفاده می‌شود. خروجی این سیستم، ورودی مدل برت-اسکواد است و پاسخ نهایی از بین این مجموعه سندها استخراج می‌شود. در شکل ۱ ساختار کلی سیستم پاسخ به پرسش تخصصی کوید۱۹ پیشنهادی را نشان می‌دهد. ابتدا پرسش به زبان طبیعی به همراه کلمات کلیدی توسط کاربر به عنوان ورودی به سیستم ارسال می‌شود، از سیستم بازیابی اسناد برای بازیابی اسنادی که حاوی پاسخ هستند، استفاده می‌شود. اسناد حاوی پاسخ به مدل برت-اسکواد ارسال می‌شوند و در نهایت پاسخ‌های دریافتی از مدل برت- اسکواد رتبه‌بندی شده و به صورت خلاصه به کاربر به عنوان پاسخ صادر می‌شوند. در ادامه سیستم بازیابی سند و مدل برت-اسکواد و رتبه بندی پاسخ‌ها شرح داده می‌شود.



شکل۱- معماری مدل پیشنهادی

[۱۳] Stanford Question Answering Dataset (SQuAD 2.0)

۳-۱- سیستم بازیابی سند

یکی از مراحل مهم سیستم پیشنهادی، مرحله بازیابی سند است. این مرحله کیفیت پاسخ را تضمین می‌کند. پاسخ‌های صحیح را فقط وقتی می‌توان یافت که مجموعه اسناد بازیابی شده به درستی تعیین شده باشد[۸]. سیستم بازیابی اسناد، شامل مراحلی است که درآن بازیابی مجموعه‌ای از اسناد مرتبط که احتمالاً حاوی پاسخ کاربر است، انجام می‌شود.

در این پژوهش به جای خزیدن در تمامی مقالات، از پایگاه داده تحقیقات باز کوید ۱۹[۱۸] ، که توسط موسسه آلن برای هوش مصنوعی[۱۴] جمع‌آوری شده، استفاده شده است. این مجموعه داده بیش از ۵۲ هزار، مقاله علمی در زمینه کوید ۱۹ و ویروس کرونا می‌باشد که به صورت مرتب و هفتگی از منابع زیر بروز می‌شود:

- PubMed's PMC open access corpus
- World Health Organization (WHO) corpus
- bioRxiv and medRxiv pre-prints

یکی از نقاط قوت روش پیشنهادی این است که علاوه بر استفاده از پایگاه داده تحقیقات باز کوید ۱۹، با خزیدن در سایت PubMed [۱۹] به روزترین مقالاتی که ممکن است در دیتاست کوید ۱۹ اضافه نشده باشند، نیز جمع‌آوری می شود. این پایگاه داده، که توسط کتابخانه پزشکی ملی[۱۵] ایالات متحده نگهداری می‌شود، به عنوان منبع معتبر شواهد پزشکی برای متخصصان پزشکی، محققان زیست پزشکی و بسیاری از کاربران دیگر در نظر گرفته می‌شود [۲۰].

در بازیابی سند، هر پاراگراف به عنوان یک بخش[۱۶] در نظر گرفته شده و تمامی بخش‌ها با توجه به کلمه های کلیدی مشخص شده مورد بررسی قرار می‌گیرند. سپس به کمک مدل رتبه بندی مورد استفاده در[۲۱] انسرینی[۱۷]، K بخشِ مرتبط‌تر استخراج می‌شوند. انسرینی مجموعه ابزار بازیابی اطلاعاتی است که بر روی لوسن[۱۸] ساخته شده است. لوسن دارای یک پایگاه توسعه دهنده فعال، تعداد زیادی از کاربران و قابلیت‌های متنوع برای کار با مجموعه های ناهمگن در مقیاس گسترده فراهم کرده است. با این حال، فاقد پشتیبانی سیستماتیک برای آزمایش موقت با استفاده از مجموعه آزمون استاندارد است. انسرینی به محققان اجازه می‌دهد تا با مدل های مدرن رتبه‌بندی کلمات کلیدی در مجموعه های مختلف نتایج رتبه‌بندی به‌راحتی تولید کنند.

۳-۲- مدل برت- اسکواد

برت یکی از الگوریتم‌های از پیش آموزش داده شده کلمات تعبیه شده[۱۹] است که امروزه بخش بسیار مهمی از پردازش زبان طبیعی را تشکیل می‌دهند. این مدل زبانی در واقع یک مدل شبکه عصبی چند لایه دو جهتی تبدیلی کدگذاری شده[۲۰] است که در زمینه‌های مختلف پردازش زبان طبیعی مورد استفاده قرار می‌گیرد[۲۲].

در این پژوهش برای استخراج پاسخ پرسش‌ها از اسناد انتخاب شده، از مدل برت-اسکواد[۲۳]، استفاده شده است. مدل برت-اسکواد، یک مدل پاسخ به پرسش بر پایه مدل برت است که بر روی مجموعه داده‌ای که توسط دانشگاه استنفورد جمع‌آوری، آموزش و ارزیابی شده است. مجموعه داده اسکواد شامل پرسش، پاسخ و پاراگراف مفهوم می‌باشد که تمامی پاراگراف‌ها از ویکی پدیا جمع آوری شده‌اند. در این مجموعه داده حدود ۱۵۰هزار پرسش وجود دارد که تقریبا برای نیمی از پرسش‌ها، پاسخی در متن‌ها وجود ندارد. این

[14] Allen Institute for AI
[15] National Library of Medicine (NLM)
[16] segment
[17] Anserini
[18] Lucene
[19] pre-trained word embedding
[20] Multi-Layer Bidirectional Encoder Transformer

مسئله باعث می‌شود مدل برت-اسکواد به خوبی یاد بگیرد که اگر پاسخ مناسبی برای پرسش وجود نداشت، پاسخ نامربوطی را انتخاب نکند.

شایان ذکر است، در بسیاری از پژوهش‌ها از روش تنظیم دقیق مدل[21] برای بالابردن دقت مدل پیش آموزش داده شده در زمینه خاص مورد نظر استفاده می‌شود، اما بعلت اینکه داده‌ی برچسب‌دار در زمینه پرسش و پاسخ کوید۱۹ موجود نبود، در این پژوهش این مرحله حذف شده است.

۳-۳- رتبه‌بندی پاسخ‌ها

پس از بررسی پرسش‌های مطرح شده بر روی تمام اسناد بازیابی شده با استفاده از مدل برت-اسکواد و تشخیص پاسخ پرسش‌ها در مقالات، با استفاده از رمزگذار عمومی جملات گوگل[22]، تشابه معنایی بین پاسخ‌های مشخص شده و پرسش مربوط به آن محاسبه شده پاسخ‌هایی با بیشترین امتیاز به عنوان مجموعه پاسخ نهایی در نظر گرفته می‌شوند. وظیفه تطبیق مبتنی بر محتوا چالش برانگیز است، دلیل اصلی آن مسئله تعیین شباهت معنایی متون است. امروزه، مدل‌های بسیاری برای تعبیه سازی جملات وجود دارد که با آموزش در شرکت‌های عظیم، با هدف ارائه یادگیری انتقالی به طیف گسترده ای از وظایف پردازش زبان طبیعی، نمایش معنایی عمیق را یاد می گیرند. در [۲۴] ، یک روش برای ارزیابی چنین مدل هایی در بازپرداخت توصیه های مبتنی بر BM25 پیشنهاد شده است.

در مرحله رتبه بندی مجدد، شباهت کسینوس را بین جمله تعبیه شده عنوان مقاله ورودی و تعبیه شده کلیه عناوین مقاله منتخب محاسبه می‌شود. این معیار تشابه، بیانگر تشابه معنایی است. یک ترکیب خطی بین نمرات اولیه BM25 پس از نرمال سازی و نمرات شباهت معنایی تعبیه شده در جمله را با جمع بندی نمرات (با وزن های یکنواخت تنظیم شده روی ۰.۵) برای تولید نتایج رتبه بندی نهایی انجام می‌شود.

۴- ارزیابی روش پیشنهادی

۴-۱- مجموعه داده

بمنظور ارزیابی روش پیشنهادی، از [۲۵]، کمک گرفته شده است. در [۲۵] مجموعه ای از ۱۶۹۰ پرسش و پاسخ راجع به کوید۱۹ از ۱۳ منبع ارائه شده است، که آنها از لحاظ موضوعی در ۱۵ دسته پرسش و ۲۰۷ خوشه پرسش قرار دارند. جدول ۱ نمونه‌ای از پرسش‌ها و نتایج را نشان می‌دهد. ستون اول پرسش‌های موجود در مجموعه داده را نشان می‌دهد. ستون دوم پاسخ مربوطه در مجموعه داده می‌باشد و ستون سوم پاسخی است که سیستم به عنوان پاسخ نهایی صادر می کند. ستون‌های تطابق و وضعیت، وضعیت انطباق یا عدم انطباق پاسخ سیستم با پاسخ مورد انتظار را نشان می‌دهد.

جدول ۱- نمونه ای از پرسش‌ها و نتایج سیستم پیشنهادی

پرسش	پاسخ اصلی	پاسخ مدل	تطابق	وضعیت
When was the Middle East Respiratory Syndrome Coronavirus isolated first?	was first isolated in 2012, in a 60-year-old man who died in Jeddah	in june 2012	بله	تطابق دقیق
Which medical comorbidities most profoundly influenced MERS-CoV outcomes?	if they were obese, immunocompromised, diabetic or had cardiac disease	diabetes, hypertension, cardiac and chronic kidney disease	بله	تطابق نسبی

[21] fine-tuning
[22] Google's Universal Sentence Encoder

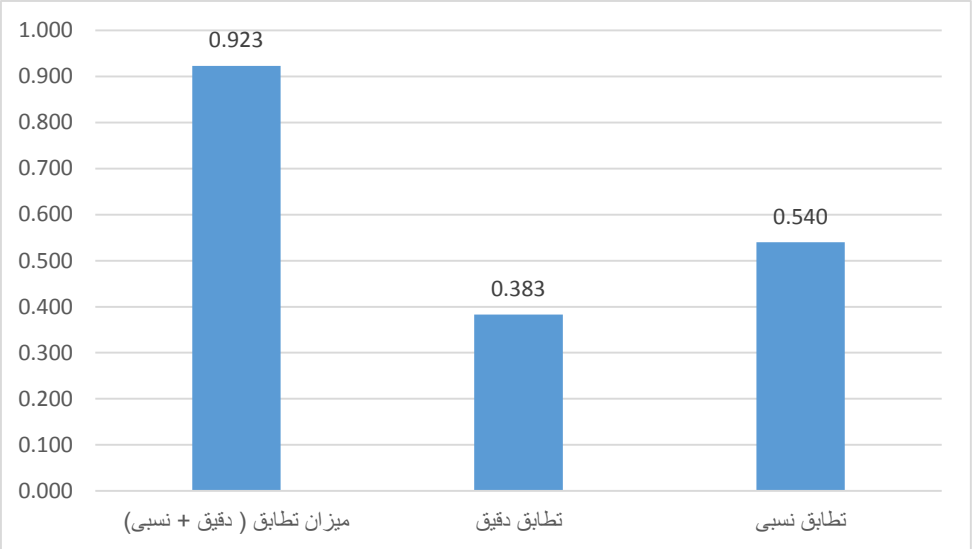
What is the estimated R0 of COVID-19?	2.2	be between 2 and 3	بله	تطابق نسبی
What is the relationship between SARS-CoV and acute lung injury (ALI)?	Many infected patients have acute lung injury (ALI)	cause fatal acute lung injury (ali)	بله	تطابق نسبی
What is the genus of the SARS coronavirus?	Betacoronavirus	betacoronavirus	بله	تطابق دقیق
What is important for understanding the origin and evolution of this novel lineage B betacoronavirus?	Learning from the roles of civet in SARS and camel in MERS, hunting for the animal source of 2019-nCoV and its more ancestral virus	hunting for the animal source	بله	تطابق دقیق
What is the difference between the Tokachi09 and Irish coronavirus genomic sequences?	148-nucleotide insertion	excluding the 3 ’- poly (a) tails	خیر	غیر مرتبط
Where does the NLRP3 inflammasome activate after a SARS-CoV infection?	in lipopolysaccharide-primed macrophages	in lipopolysaccharide -primed macrophages	بله	تطابق دقیق
What was the fatality rate of the SARS coronavirus outbreak between November 2002 and July 2003?	9. 6 %	9. 6 %	بله	تطابق دقیق
What animals can carry coronavirus?	uding birds, farm animals, pets, camels, an	mainly bats, birds, genic sheep, cattle, cat and primates	بله	تطابق نسبی
How long is the incubation time for COVID19?	between 2 and 14 days	in average 5. 08 days	بله	تطابق نسبی
What viruses are similar to the COVID-19 coronavirus?	MERS-CoV and SARS-CoV.	sars-cov-2	بله	تطابق نسبی
What is the relation between the new2019-nCOV and the conserved orf8?	the new 2019-nCoV orf8 is distant from the conserved orf8	8 likely encodes a secreted protein with an alpha-helix, following with a beta-sheet (s) containing six strands	بله	تطابق نسبی

۴-۲- معیارهای ارزیابی

با توجه به اینکه سیستم پیشنهادی ویژه پاسخ خلاصه می‌باشد، برای ارزیابی اثربخشی، کارایی و قابلیت استفاده از معیارهای دقت و سرعت اجرا برای ارزیابی استفاده شده است.

ارزیابی به دو صورت سختگیرانه و غیرسختگیرانه انجام شد. در فاز غیرسختگیرانه بررسی میکنیم که آیا پاسخ ارائه شده توسط سیستم ارتباط معنایی و مفهومی با پاسخ اصلی دارد یا خیر. اما در فاز سختگیرانه بررسی میکنیم که آیا دقیقا همان پاسخ (دقیقا همان پاراگراف)

موجود در مجموعه داده را ارائه داده است یا خیر. برای مثال در اولین سطر جدول ۱ مشاهده می‌کنید که پاسخ اصلی "۲۰۱۲" بوده و سیستم هم "۲۰۱۲" را ارائه داده است، اما در سطر دوم سیستم دقیقا همان پاسخ اصلی را ارائه نکرده است بلکه پاسخ سیستم جواب قابل قبولی بوده که ارتباط معنایی و مفهومی زیادی با پاسخ اصلی دارد. سیستم در ارائه پاسخ مرتبط ۹۲/۳ درصد موفق بوده که در ۳۸/۳ درصد اوقات پاسخ دقیق[۲۳] ارائه می‌دهد. شکل ۲ نمودار مربوط به کارایی روش پیشنهادی را نشان می‌دهد.

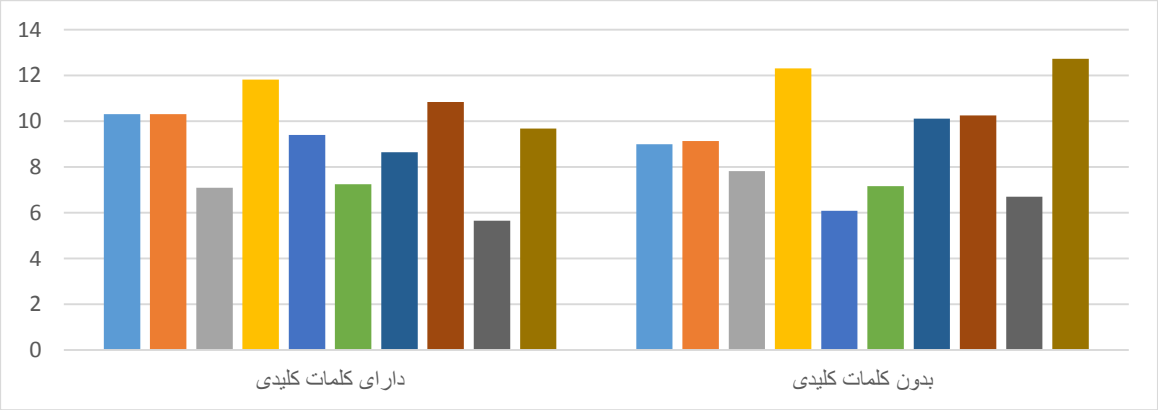


شکل۲- میزان کارایی مدل پیشنهادی

۴-۳- زمان اجرا

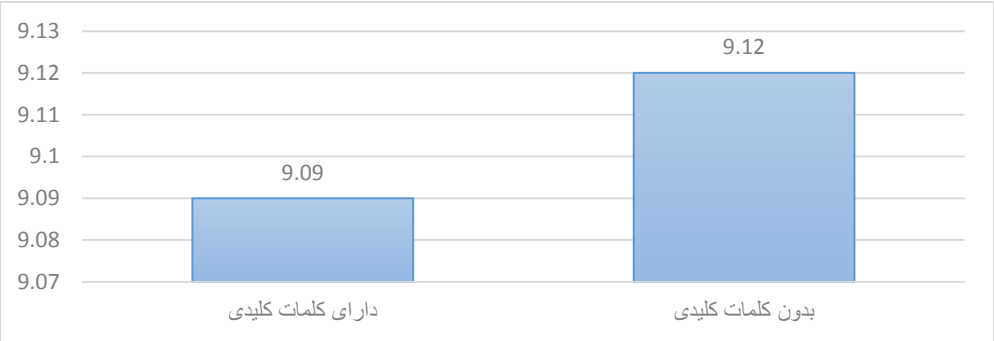
زمان اجرای هر پرسش و پاسخ را به ازای دو حالت «استفاده از کلمات کلیدی» و «بدون استفاده از کلمات کلیدی» در سیستمی با Gpu 2080 ti، پردازنده core i7، و حافظه اصلی (RAM) ۳۲ گیگابایت، ارزیابی شد. شکل ۳ زمان اجرای الگوریتم را به ازای ۱۰ پرسش مختلف با کلمه کلیدی و بدون کلمه کلیدی برحسب ثانیه نشان می‌دهد. میانگین زمان اجراها در شکل ۴ نشان داده شده است.

همانطور که در شکل ۳ دیده می‌شود استفاده یا عدم استفاده از کلمات کلیدی تاثیر چندانی در زمان اجرا ندارد بلکه زمان اجرا، بستگی به پرسش و تعداد مقالات دارد.



شکل ۳ - زمان اجرای روش پیشنهادی به ازای ۱۰ پرسش مختلف با کلمه کلیدی و بدون کلمه کلیدی بر حسب ثانیه

[۲۳] Exact match



شکل ۴- میانگین زمان اجرا با کلمه کلیدی و بدون کلمه کلیدی برحسب ثانیه

۵- بحث و نتیجه گیری

گسترش جهانی کوید۱۹ فشار فزاینده ای بر سازمان بهداشت جهانی وارد کرده است. در این زمینه، چندین استراتژی برای مقابله با بیماری همه گیر از جمله افزایش بودجه سازمان بهداشت جهانی اتخاذ شده است. تعداد بیماران آلوده این زمان‌ها بی‌سابقه است و بسیاری از خدمات بهداشتی را تحت تأثیر قرار می‌دهد. برای بررسی روند درمان و سایر اطلاعات پزشکی در مورد این بیماری نیاز به جستجو در حجم انبوه مقالات و دستاوردهای علمی تولید شده و بروز است. محتویات فنی و طولانی بودن مقالات زیست پزشکی کار جستجو را پیچیده تر می کند. در صورتی که متخصصی بخواهد پاسخ یک پرسش را از میان چندین پایگاه علمی بیابد نیاز به صرف ساعتها زمان و هزینه دارد، در این راستا برای کمک به متخصصان دانشمندان علم داده با طراحی سیستم پاسخ به پرسش های خاص این بیماری می‌توانند کمک در خور توجه به پزشکان کنند. در این راستا در این تحقیق یک سیستم پاسخ به پرسش ویژه بیماری کوید ۱۹، پیشنهاد شده است. سیستم پیشنهادی می‌تواند بطور موثر در جدیدترین مقالات و پایگاه های علمی بیماری کوید ۱۹ جستجو کند و پس از بازیابی اسناد مرتبط، محدوده پاسخ های مرتبط را با استفاده از مدل برت-اسکواد رتبه بندی کند و مرتبط ترین پاسخ‌ها را به صورت پاسخ خلاصه به کاربر تحویل دهد. ارزیابی روش پیشنهادی نشان می‌دهد، سیستم پاسخ به پرسش برای کوید ۱۹ با استفاده از برت-اسکواد با دقت ۹۲/۳ درصد، کارایی مناسبی دارد.

برای کارهای آتی پیشنهاد می‌شود تا با کمک روش‌های فراابتکاری دقت روش پیشنهادی را بهبود داد. از طرف دیگر با توجه به اینکه روش پیشنهادی ویژه پرسش‌های خلاصه طراحی شده است، می‌توان با افزودن مراحلی به روش پیشنهادی، آنرا جهت پاسخ به پرسش‌های لیستی و بله/ خیر، ارتقا داد.

۶- منابع

1. Popescu, S., Roadblocks to infection prevention efforts in health care: SARSCoV- 2/COVID-19 response, Disaster Med. Public Health Prep., 2020; 1-3.

2. Perondi B, Miethke-Morais A, C. Montal A, Harima L, C. Segurado A, Setting up hospital care provision to patients withCOVID-19: lessons learnt at a 2400-bed academictertiary center in São Paulo, Brazil, The Brazilian Journal of INFECTIOUS DISEASES, 2020.

3. "Coronavirus (Covid-19)," news.google, 21 Jan 2021. [Online]. Available: https://news.google.com/covid19/map?hl=en-US&gl=US&ceid=US%3Aen. [Accessed 2021].

4. Jeyabaladevan P., COVID-19: an FY1 on the frontline., Med. Educ.2020; 25(1): 1-5.

5. Bashir S., Saeed Khattak A., Alshara M., Automatically transforming full length biomedical articles into search queries for retrieving related articles, Egyptian Informatics Journal, 2020; 10-18.

6. Russell-Rose T, Chamberlain J., Expert Search Strategies: The Information Retrieval Practices of Healthcare Information Professionals., JMIR Med Inform.2017;5(4): 254-260.

7. Ben Abach A., Demner-Fushman D., A question-entailment approach to question answering, BMC Bioinformatics, 2019; 1-23.

8. Sarrouti M., Ouatik El Alaoui S.,SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions, Artificial Intelligence In Medicine, 2019; 14-20.

9. Neves M., Leser U., Question answering for Biology, Methods, 2015; 74: 36–46.

10. Lee M., Cimino J., Zhu J., Sable C., Shanker V., Ely J., Yu H., Beyond information retrieval-medical question answering, in: AMIA Annual Symposium Proceedings, 2006; 469–473.

11. Abacha A. B., Zweigenbaum P., MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies, Information Processing & Management, 2015; 51(5):570–594.

12. Kraus M., Niedermeier J., Jankrift M., Tietbohl S., Stachewicz T., Folkerts H., Uflacker M., Neves M., Olelo: a web application for intuitive exploration of biomedical literature, Nucleic Acids Research, 2017; 45(w1): 478–483.

13. Schulze F., Schuler R., Draeger T., Dummer D., Ernst A., Flemming P., Perscheid C., Neves M., HPI question answering system in BioASQ 2016, in: Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics,2016.

14. Schulze F., Schuler R., Draeger T., Dummer D., Ernst A., Flemming P., Perscheid C., Neves M., HPI question answering system in BioASQ 2016, in: Proceedings of the Fourth BioASQ workshop at the Conference of the Association for Computational Linguistics,2016.

15. Yang Z., Gupta N., Sun X., Xu D., Zhang C., Nyberg E., Learning to answer biomedical factoid & list questions: OAQA at BioASQ 3B, 2015; CLEF (Working Notes).

16. Sarrouti M., Ouatik El Alaoui S.,SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions, Preprint submitted to Artificial Intelligence in Medicine, 2019; 102: 48-60.

17. Abacha B, Fushman D., A question-entailment approach to question answering, BMC Bioinformatics, 2019; 20(511):1-23.

18. Du Y., Pei B., Zhao X., Ji J.,Deep scaled dot-product attention based domain adaptation model for biomedical question answering, Methods, 2020; 15(173):69-74 .

19. COVID-19 Open Research Dataset, semantic scholar, 2020. [Online]. Available: https://pages.semanticscholar.org/coronavirus-research.

20. PUBMED, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/.

21. Demner-Fushman D., Lin L., Answering clinical questions with knowledge-based and statistical techniques, Computational Linguistics, 2007; 33(1):63-103.

22. Yang P., Fang H., Lin J., Anserini: Reproducible Ranking Baselines Using Lucene, ACM Journal of Data and Information Quality,2018; 10(4):1-20.

23. Devlin J., Chang M., Lee K., Toutanova K., Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805, 2018.

24. Yuwen Zhang, Zhaozhuo Xu, "BERT for Question Answering on SQuAD 2.0," Engineering, pp. 1-9, 2019.

25. Hebatallah A. Mohamed Hassan , Giuseppe Sansonetti ,Fabio Gasparetti, Alessandro Micarelli, Joeran Beel,, "BERT, ELMo, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation?," in ACM RecSys 2019 Late-breaking Results, Copenhagen, Denmark, 2019.

26. Jerry Wei, Chengyu Huang, Soroush Vosoughi, Jason Wei, "What Are People Asking About COVID-19? A Question Classification Dataset," in Published in Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, 2020.