مجله مدیریت اطلاعات سلامت
سال بیستم/ شماره دوم/ تابستان ۱۴۰۲

Published by *Isfahan University of Medical Sciences*
DOI: 10.48305/him.2023.41593.1100

مدل‌سازی شباهت بیمار با استفاده از بازنمایی هوشمند خلاصه پرونده برای پیش‌بینی تشخیص نهایی

هدی معمارزاده[۱]، ناصر قدیری[۲]، مریم لطفی شهرضا[۳]

مقاله پژوهشی

چکیده

مقدمه: داده‌های متنی ثبت شده در پرونده الکترونیک سلامت Electronic health record (EHR) در بردارنده اطلاعات مهمی از شرح حال بیمار و مسیر درمانی اوست ولی به دلیل آن‌که بدون ساختار ذخیره می‌شود نمی‌تواند به‌صورت مستقیم در الگوریتم‌های تحلیل داده مورد استفاده قرار گیرد. یکی از راه‌های ساختارمند کردن داده‌های متنی تولید بردار بازنمایی از آن‌هاست. این مطالعه چهارچوبی به‌منظور تولید بردار بازنمایی از متن‌های خلاصه پرونده ارائه داده است.

روش بررسی: در این مطالعه پیمایشی از بازنمایی متن خلاصه پرونده بیماران برای تولید بردار متناظر با هر متن استفاده شده است. برای بازنمایی از مدل‌های زبانی که از آخرین روش‌های پردازش متن هستند استفاده شده است. مجموعه داده شامل متن خلاصه پرونده بیش از ۲۶۰۰۰ بیمار از پایگاه داده Medical Information Mart for Intensive Care III (MIMIC-III) است. برای تحلیل کیفیت بردارهای بازنمایی از مسئله پیش‌بینی تشخیص استفاده شده و معیارهای ارزیابی برای هر مدل زبانی گزارش شده است.

یافته‌ها: از بین مدل‌های زبانی استفاده شده در طراحی بهترین مدل بازنمایی برای متن خلاصه پرونده مدل BIO-BERT و سپس مدل SciBERT است که به ترتیب نتایج ۰/۷۱۵ و ۰/۷۱۳ را برای معیار ارزیابی ROC_AUC تولید کرده‌اند. این معیار ارزیابی برای بررسی کیفیت مدل‌های پیش‌بینی استفاده می‌شود. استفاده از پیش‌پردازش متن بالینی و نگاشت موجودیت‌های بالینی به اسامی استاندارد آن‌ها در پایگاه دانش The Unified Medical Language System - UMLS معیارهای ارزیابی برای مدل‌های زبانی خاص حوزه زیست پزشکی بهبود یافته است و بیشترین بهبود مربوط به مدل UMLSBERT است که روی اسامی استاندارد پایگاه دانش آموزش دیده است.

نتیجه‌گیری: بر اساس یافته‌های این مطالعه مدل‌های زبانی BIO-BERT و SciBERT که روی داده‌های مقالات بالینی آموزش دیده‌اند به‌عنوان بهترین گزینه برای بازنمایی اطلاعات نهفته متن خلاصه پرونده به بردارها پیشنهاد می‌شوند. با این وجود به دلیل آنکه متن خلاصه پرونده از نظر ساختار و محتوا با متن مقالات علمی متفاوت است، پیش‌پردازش متن‌های بالینی به‌منظور شناسایی موجودیت‌ها و نگاشت آن‌ها به منابع دانش برای استفاده از اسامی استاندارد مفاهیم بالینی باعث بهبود نتایج به‌دست آمده در مدل‌های زبانی می‌گردد.

واژه‌های کلیدی: پردازش متن بالینی، انفورماتیک پزشکی، مدل‌های زبانی

پیام کلیدی: پردازش داده‌های بدون ساختار ثبت شده در پرونده بیماران با بهره‌گیری از روش‌های هوشمند پردازش متن‌های بالینی می‌تواند در طراحی سامانه‌های شباهت بیماران و کمک به تعیین تشخیص نهایی مؤثر باشد.

دریافت مقاله:۱۴۰۱/۱۲/۱۱ پذیرش مقاله:۱۴۰۲/۳/۱۶ تاریخ انتشار: ۱۴۰۲/۴/۱۵

ارجاع: معمارزاده هدی، قدیری ناصر، لطفی شهرضا مریم. **مدل‌سازی شباهت بیمار با استفاده از بازنمایی هوشمند خلاصه پرونده برای پیش‌بینی تشخیص نهایی.** مدیریت اطلاعات سلامت۱۴۰۱؛۲۰ (۲):۷۱-۶۵.

مقدمه

همزمان با رشد روزافزون فناوری‌های مرتبط باEHR ، امکان استفاده از اطلاعات ذخیره شده برای کاربردهای ثانویه‌ای که منجر به بهبود کیفیت ارائه خدمات و تصمیم‌گیری‌های کلان می‌شود افزایش یافته است (۱). یافتن بیماران شبیه به یکدیگر در میان جمعیت‌های بیماران یکی از ارکان بنیادی این نوع از کاربردهای ثانویه است. محاسبه شباهت بیماران نیازمند تبدیل انواع اطلاعات ثبت شده در EHR و سایر منابع داده به فرمی مقایسه پذیر است (۲،۳).

EHR شامل داده‌های عددی و متنی است. داده‌های متنی برای مقایسه پرونده‌ها چالش برانگیز هستند (۴). از جمله روش‌های مقایسه متن‌ها به‌صورت خودکار بازنمایی متن به صورت بردار عددی است. یک بردار بازنمایی خوب از متن، اطلاعات کافی برای حل مسئله را حفظ می‌کند (۵). بازنمایی متن، امکان استفاده از این داده‌ها را مدل‌های یادگیری ماشین که با هدف پیش‌بینی شرایط بیمار (پیش‌بینی تشخیص، طول مدت بستری، احتمال فوت بیمار) فراهم می‌آورد .

شرکت گوگل، نسل جدیدی از روش‌های بازنمایی متن را با عنوان مدل زبانی توسعه داده است (۶).

۱- دانشجوی دکتری ، نرم افزار، گروه نرم افزار، دانشکده برق و کامپیوتر، دانشگاه صنعتی اصفهان، ایران
۲- دانشیار، نرم افزار، گروه نرم افزار، دانشکده برق و کامپیوتر، دانشگاه صنعتی اصفهان، ایران
۳- استادیار، نرم فزار، گروه مهندس کامپیوتر، پردیس شهرضا، دانشگاه اصفهان، ایران
نویسنده طرف مکاتبه: ناصر قدیری؛ دانشیار، نرم افزار، گروه نرم افزار، دانشکده برق و کامپیوتر، دانشگاه صنعتی اصفهان، ایران
Email: nghadiri@iut.ac.ir

مدل زبانی می‌تواند روابط معنایی بین کلمات متن را در تولید بردار عددی حفظ کنند. «حفظ روابط معنایی» به این معناست که فاصله بردارهای حاصل از بازنمایی متن‌هایی که از نظر معنایی شباهت بیشتری دارند، کمتر از فاصله آن‌ها با بردارهای حاصل از بازنمایی سایر متن‌ها شده و در فضای برداری نزدیک‌تر به هم قرار گیرند. این ویژگی امکان پیاده‌سازی الگوریتم‌های یادگیری ماشین مانند دسته‌بندی و خوشه‌بندی را فراهم می‌کند. مدل‌های زبانی، مبنای جدیدترین ابزارهای هوش مصنوعی مانند ChatGPT نیز هستند (۷).

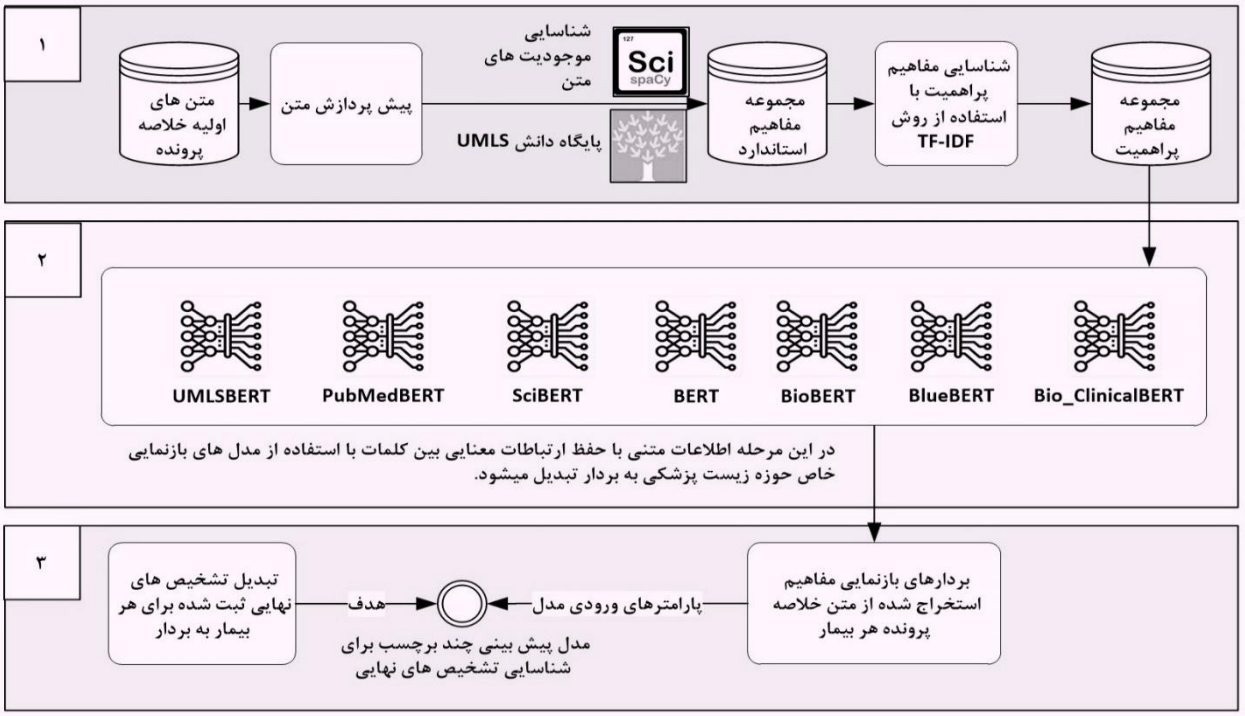
استفاده از مدل زبانی در زمره یادگیری انتقالی Transfer learning (۸) محسوب می‌شود. در یادگیری انتقالی وزن‌های یک مدل با آموزش روی دامنه بزرگی از داده‌ها محاسبه و سپس این وزن‌ها بر روی یک مجموعه داده کوچک و برای یک هدف خاص که می‌تواند توسعه یک مدل یادگیری ماشین باشد، تنظیم می‌شوند (۹).

برای متن بالینی نیاز است تا عبارات متنوعی که برای ثبت اصطلاحات تخصصی توسط ارائه‌دهندگان خدمات به‌کار رفته یکسان‌سازی شود. لذا شناسایی موجودیت‌ها و انتساب اسامی استاندارد به آن‌ها در افزایش کیفیت متن بالینی اهمیت دارد (۱۰). در مطالعه Kang و همکاران از نگاشت موجودیت‌های متن با منبع دانش استفاده شده و تأثیر آن در افزایش کیفیت مدل‌های یادگیری عمیق مورد بررسی قرار گرفته است (۱۱). در مطالعه دارابی و همکاران از یک مدل زبانی برای بازنمایی متن بالینی استفاده گردید و بردارهای تولید شده در مسئله‌های پیش‌بینی احتمال فوت درون بیمارستانی و طول مدت بستری بیمار مورد استفاده قرار گرفته است (۱۲). مدل‌های زبانی متعددی در حوزه زیست پزشکی توسعه داده شده است که روی منابع داده‌ای متفاوت آموزش دیده‌اند از آن جمله می‌توان به مدل‌های BioBERT (۱۳)، BlueBERT (۹)، Bio-clinical BERT (۱۴)، SciBERT (۱۵)، PubMedBERT (۱۶) و UMLSBERT (۱۷) اشاره کرد.

مقایسه عملکرد این مدل‌ها در پردازش متن‌های بالینی به‌عنوان یک خلأ مطالعاتی نیازمند یک پژوهش مجزا است. مدل ارائه شده در مطالعه جاری به منظور مقایسه توأمان به‌کارگیری مدل زبانی و نگاشت موجودیت‌های بالینی به منبع دانش و یافتن ترکیب مطلوب در تولید بردارهای بازنمایی طراحی شده است. به منظور مقایسه عملکرد مدل‌های زبانی مختلف، بردارهای حاصل از آن‌ها در پیش‌بینی تشخیص نهایی مورد استفاده قرار گرفته‌اند. نتایج حاصل از این مطالعه می‌تواند در توسعه سامانه‌های پردازش متن‌های بالینی مورد استناد قرار گیرد.

روش بررسی

مطالعه حاضر یک مطالعه مقطعی است. مجموعه داده‌های این پژوهش برگرفته از پایگاه داده MIMIC-III است که شامل متن‌های بالینی و آزمایش‌ها و داروها و تشخیص‌های بیش از ۴۰ هزار بیمار مراجعه کننده به یک مرکز درمانی در فاصله سال‌های ۲۰۰۱ تا ۲۰۱۲ است (۱۸). انتشار مجموعه داده‌های مذکور پس از حذف همه اسامی و آدرس‌ها انجام شده است. پژوهشگر به‌منظور ارائه مدل از ۲۶۰۰۰ فرم خلاصه پرونده مربوط به بیماران بین ۱۸ تا ۹۹ سال این پایگاه داده استفاده کرده است. مدل پیشنهادی در این پژوهش در سه بخش طراحی شده است. گام‌های اجراشده در شکل ۱ نشان داده شده‌اند:



شکل ۱ مراحل انجام فرایند پیش‌بینی تشخیص با استفاده از بازنمایی متن خلاصه پرونده

ابتدا متن‌های خلاصه پرونده بیماران از پایگاه داده استخراج و سپس به منظور شناسایی موجودیت‌های متن پیش‌پردازش شده‌اند. اهمیت شناسایی موجودیت‌ها به وجود اطلاعات تخصصی و مخففهای متعدد در متن برمی‌گردد (۱۹). به این منظور از پایگاه دانش UMLS که برای استانداردسازی مفاهیم تعریف شده در استانداردهای کدگذاری مختلف طراحی شده و تا کنون بالغ بر ۲۲۰ استاندارد کدگذاری را شامل می‌شود، با کمک کتابخانه پایتون (۲۰) Scispacy استفاده گردید. در مرحله شناسایی موجودیت‌ها امکان تولید مفاهیم تکراری وجود دارد. به منظور مشخص کردن مفاهیم پراهمیت‌تر و حذف موارد تکراری از الگوریتم TFIDF (۲۱) با حد آستانه ۰/۲ استفاده شده است. روش محاسبه امتیاز TFIDF بر اساس رابطه ۱ است (۲۱).

$$TF - IDF_t = TF \times \log\left(\frac{|D|}{|d \in D : t \in d|}\right) \quad (1)$$

در بخش دوم از مدل‌های زبانی که برای حوزه‌های زیست پزشکی توسعه داده شده‌اند به منظور بازنمایی متن پیش‌پردازش شده خلاصه پرونده‌ها استفاده گردید.

این مدل‌ها روی مجموعه داده‌های زیست پزشکی متفاوتی آموزش دیده‌اند. ابتدا پارامترهای مدل‌ها روی داده‌های آموزشی بهینه‌سازی Fine-tune شد و سپس مدل با وزن‌های تنظیم‌شده روی داده‌های اعتبارسنجی و آزمودن بکار گرفته شد. در این مرحله ۸۰ درصد نمونه‌ها برای آموزش مدل و از ۱۰ درصد برای اعتبارسنجی پارامترها و از ۱۰ درصد باقیمانده برای آزمودن مدل استفاده شده است.

تفاوت عملکرد مدل‌ها به منابع دانشی که روی آن‌ها آموزش دیده‌اند و نیز نحوه تقسیم کلمات به زیر کلمه‌ها که اصطلاحاً توکن‌بندی گفته می‌شود مربوط است. منابع دانش استفاده شده برای هر مدل در جدول ۱ نشان داده شده است. همچنین نمونه‌ای از نحوه عملکرد متفاوت مدل‌ها در توکن‌بندی در جدول ۱ برای عبارت ورودی «coronary arteriosclerosis»" نشان داده شده است. همان‌طور که مشاهده می‌شود لزوماً همه زیرکلمه‌های تولید شده دارای معنا نیستند و این مسئله می‌تواند فرایند یادگیری را در کاربردهای دست پایین مختل کند.

جدول ۱: مدل‌های بازنمایی حوزه زیست پزشکی

مدل	مرجع	دامنه کلمات	حوزه اطلاعاتی خاص	توکن های تولیدشده برای “coronary arteriosclerosis”
BioBERT	(۱۳)	دامنه کلمات مدل پایه‌ای BERT	خلاصه مقالات PubMed وPMC	[‘corona’, ‘##ry’, ‘arte’, ‘##rio’, ‘##sc’, ‘##ler’, ‘##osis’]
BlueBERT	(۹)	دامنه کلمات مدل پایه‌ای BERT	خلاصه مقالات PubMed و متن‌های بالینی دادگان MIMIC	[‘corona’, ‘##ry’, ‘arte’, ‘##rio’, ‘##sc’, ‘##ler', '##osis']
Bio-clinical BERT	(۱۴)	BioBERT دامنه کلمات مدل پایه‌ای BERT	متن‌های بالینی دادگان MIMIC	[‘co’, ‘##rona’, ‘##ry’, ‘art’, ‘##eri’, ‘##os’, ‘##cle’, ‘##rosis’]
SciBERT	(۱۵)	SciVocab	مقالات حوزه بالینی و کامپیوتر	[‘coronary’, ‘arterios’, ‘##cle’, ‘##rosis’]
PubMedBERT	(۱۶)	دامنه کلمات مدل پایه‌ای BERT	خلاصه و متن کامل مقالات PubMed	['coronary', 'arteri', '##osclerosis']
UMLSBERT	(۱۷)	Bio-clinical BERT	متن‌های بالینی و تشخیصی دادگان MIMIC	['co', '##rona', '##ry', 'art', '##eri', '##os', '##cle', '##rosis']

در پایان این مرحله، به ازای خلاصه پرونده هر بیمار یک بردار بازنمایی با طول یکسان به ازای هر مدل تولید شده است که امکان به‌کارگیری الگوریتم‌های یادگیری ماشین را فراهم می‌آورد.

در بخش سوم از بردار بازنمایی تولید شده به عنوان ورودی یک مدل دسته‌بندی برای پیش‌بینی تشخیص نهایی بیماران استفاده شده است. مدل دسته‌بندی طراحی شده به صورت چند برچسب است. در مدل چند برچسب هر نمونه می‌تواند به رده‌های مختلف تعلق داشته باشد در مقابل در مدل چند کلاس هر نمونه تنها به یک کلاس تعلق خواهد داشت. دلیل طراحی به‌صورت دسته‌بندی چند برچسب وجود بیش از یک تشخیص برای هر بیمار است. برای ارزیابی نتایج این دسته‌بندی از معیارهای F1-Score و ROC-AUC استفاده شده است

دسترسی به مجموعه داده MIMIC-III از طریق سایت https://physionet.org/ امکان‌پذیر است. این داده‌ها به‌صورت بی‌نام شده و پس از کسب مجوز توسط پژوهشگر در اختیار وی قرار گرفته است.

یافته‌ها

نتایج حاصل از اجرای چهارچوب طراحی شده در جدول شماره ۲ ارائه شده است.

جدول ۲: نتایج به‌دست‌آمده از پیش‌بینی تشخیص بیمار با استفاده از بازنمایی متن خلاصه پرونده توسط مدل‌های زبانی معرفی‌شده در جدول ۱

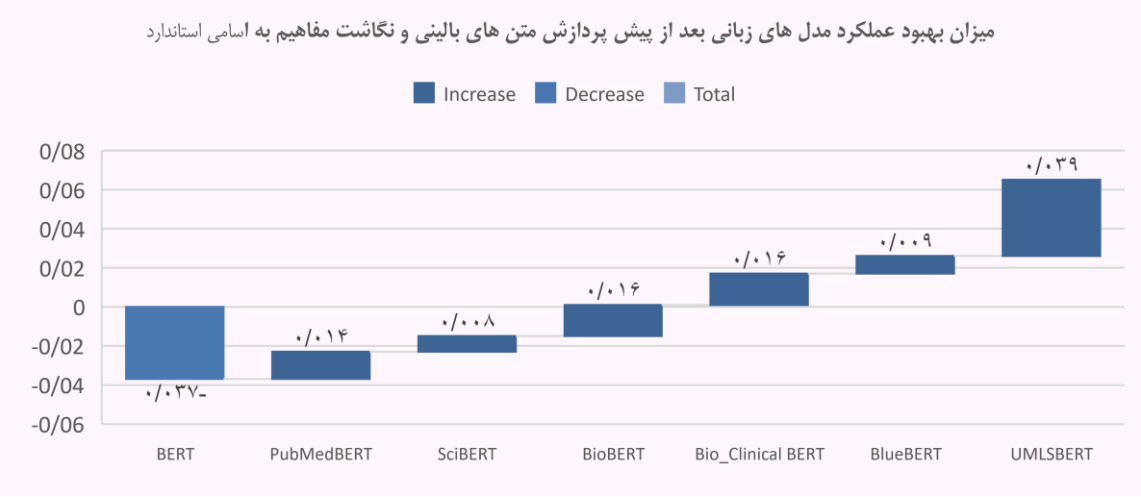
نام مدل زبانی استفاده شده برای بازنمایی	BERT	PubMedBERT	SciBERT	BioBERT	Bio-clinical BERT	BlueBERT	UMLSBERT
ROC-AUC							
مجموعه همه مفاهیم استخراج‌شده	۰/۶۹۲	۰/۷۰۷	۰/۷۱۳	۰/۷۱۵	۰/۶۹۹	۰/۶۹۳	۰/۷۰۲
مجموعه مفاهیم پراهمیت‌تر شناسایی‌شده با استفاده از الگوریتم TF-IDF	۰/۶۵۵	۰/۷۲۱	۰/۷۲۱	۰/۷۳۱	۰/۷۱۵	۰/۷۰۲	۰/۷۲۹
F1*							
مجموعه همه مفاهیم استخراج‌شده	۰/۵۱۹	۰/۵۴۵	۰/۵۵۵	۰/۵۶۱	۰/۵۵۲	۰/۵۲۶	۰/۵۳۶
مجموعه مفاهیم پراهمیت‌تر شناسایی‌شده با استفاده از الگوریتم TF-IDF	۰/۴۴۵	۰/۵۶۸	۰/۵۶۸	۰/۵۶۹	۰/۵۵۳	۰/۵۳۴	۰/۵۵۹

روش محاسبه F1-Score به‌صورت رابطه ۲ است (۲۲).

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

در این رابطه P نشان دهنده دقت (Precision) و R نشان دهنده یادآور (Recall) است . AUC - ROC یک معیار اندازه‌گیری عملکرد برای مسئله‌های کلاس‌بندی است و میزان تمایز مدل بین کلاس‌های مختلف را نشان می‌دهد. در مسئله‌های چند برچسب این معیار تعمیم پیدا کرده و میانگین مقدار به‌دست آمده برای برچسب‌های مختلف گزارش می‌شود (۲۳). بررسی نتایج حاصل از تولید بازنمایی از مفاهیم استخراج شده نشان می‌دهد که بهترین عملکرد در بین مدل‌های زبانی در شناسایی تشخیص نهایی بیمار توسط مدل BIO-BERT با ۰/۷۱۵ و سپس مدل SciBERT با ۰/۷۱۳ به‌دست آمده است. همچنین بررسی عملکرد مدل‌ها با دریافت ورودی‌هایی که امتیاز TFIDF آن‌ها حداقل ۰/۲ بوده است از حالتی که همه مفاهیم به عنوان ورودی دریافت شده است در همه مدل‌های خاص منظوره بهتر بوده است.

شکل ۲ نشان می‌دهد که استفاده از مفاهیم منحصر به فردی که امتیاز TF-IDF آن‌ها از ۰/۲ بیشتر بوده است، می‌تواند منجر به بهبود عملکرد مدل‌های زبانی در مسئله پیش‌بینی تشخیص نهایی شود



شکل ۲ مراحل انجام فرایند پیش‌بینی تشخیص با استفاده از بازنمایی متن خلاصه پرونده

این بهبود ناشی از حذف موارد تکراری است که می‌توانند دقت مدل را کاهش دهند. یافته دیگر عملکرد بهتر مدل در زمان استفاده از مجموعه مفاهیم استخراج شده نسبت به استفاده از متن‌های پیش‌پردازش نشده است. دلیل این امر وجود عبارت‌های غیراستاندارد و عمومی در متن‌های اصلی است که پس از شناسایی موجودیت‌ها حذف می‌شوند. همان‌طور که انتظار می‌رود بیشترین بهبود مربوط به مدل UMLSBERT است که روی پایگاه دانش UMLS پیش آموزش شده است.در خصوص تفاوت عملکرد مدل‌های زبانی

بهترین نتیجه از طریق مدل BIO-BERT و مدل SciBERT به دست آمده است. این نتیجه ناشی از پیش آموزش مدل‌ها روی مقالات علمی است. مدلی که در جایگاه سوم قرار می‌گیرد مدل UMLSBERT است.

بحث

اولین یافته این مطالعه نشان می‌دهد با استفاده از بازنمایی اطلاعات می‌توان داده‌های متنی را به فرم بردار تبدیل کرد تا امکان مقایسه آن‌ها فراهم شود. مطالعه Dligachو همکاران (۲۴) و نیز مطالعه دارابی و همکاران (۱۲) این یافته را تائید می‌کند. از دیگر یافته‌های این مطالعه این است که شناسایی موجودیت‌های متن‌های بالینی می‌تواند منجر به بهبود معیارهای ارزیابی مدل گردد. مقایسه معیار AUC-ROC در مسئله شناسایی تشخیص در این مطالعه (مقدار ۰/۷۱۵) با مطالعه دارابی و همکاران (۱۲) (مقدار ۰/۶۷۱) این یافته را تائید می‌کند .

یافته دیگر این مطالعه این است که حذف مفاهیم شناسایی‌شده کم ارزش و تکراری می‌تواند منجر به بهبود معیار AUC-ROC از ۰/۷۱۵ به ۰/۷۳۱ شده است. مطالعه Duque و همکاران (۲۵) این یافته را تائید می‌کند. این مطالعه مسئله شناسایی مفاهیم کم‌ارزش و تکراری را به عنوان مهم‌ترین چالش فرایند شناسایی موجودیت‌ها و اتصال آن‌ها به پایگاه دانش عنوان کرده است.

یافته دیگر مطالعه این است که کاربرد الگوریتم‌های بازنمایی روی داده‌های متنی می‌تواند مستقل از نوع بیماری بوده و برای همه مجموعه داده‌ها مورد استفاده قرار گیرد. با این وجود انتظار می‌رود جداسازی مجموعه داده‌ها به گروه‌های اصلی بیماری دقت معیارهای ارزیابی را افزایش دهد. مطالعه Shen و همکاران (۲۶) این یافته را تائید می‌کند در این مطالعه از مدل‌های بازنمایی برای پیش‌بینی دو سبک زندگی بیماران آلزایمر استفاده شده است و به مقدار ۰/۹۰ برای معیار F1 رسیده است.

یافته دیگر این مطالعه تفاوت عملکرد مدل‌های زبانی است. بهترین نتیجه از طریق مدل BIO-BERT و مدل SciBERT به‌دست آمده است. این نتیجه توسط مطالعه (۱۵) نیز تائید شده است. این برتری ناشی از پیش آموزش مدل‌ها روی مقالات علمی است. مدلی که در جایگاه سوم قرار می‌گیرد مدل UMLSBERT است. ترتیب برتری نتایج نشان می‌دهد که آموزش مدل روی مقالات علمی نسبت به روابط پایگاه دانش UMLS منجر به تولید بردارهایی باکیفیت بالاتر شده است.

نتیجه‌گیری

پردازش زبان طبیعی یکی از مورد توجه‌ترین حوزه‌های هوش مصنوعی است. از آنجا که بخش مهمی از اطلاعات بیماران در EHR در متن‌های بالینی ذخیره می‌شود پردازش آن‌ها در ارتقا کیفیت سامانه‌های تصمیم‌یار تأثیرگذار است. این مطالعه رویکرد نوینی برای پردازش متن‌های خلاصه پرونده خلاصه پرونده پیشنهاد می‌دهد که در طراحی آن از مدل‌های زبانی استفاده شده است. این مدل‌ها بر اساس دانش نهفته در پایگاه‌های دانش متعددی آموزش‌دیده و می‌توانند برای شناسایی ارتباطات بین عبارت‌های متن‌های علمی به‌صورت مؤثری مورد استفاده قرار بگیرند. در این مطالعه از این مدل‌ها برای پردازش متن خلاصه پرونده که ازنظر محتوایی و ساختاری با پیچیدگی‌های بسیار مواجه است با هدف پیش‌بینی تشخیص نهایی استفاده شده است همچنین به‌منظور یکسان‌سازی عبارات علمی ثبت شده در متن‌ها از روش‌های نوین استخراج مفاهیم بالینی در مرحله پیش‌پردازش استفاده گردید. مدل BIO-BERT که روی متن مقالات علمی آموزش دیده است نسبت به سایر مدل‌ها نتایج مطلوبی در پیش‌بینی تشخیص نهایی با مقدار ۰/۷۱۵ برای معیار ROC-AUC ارائه کرده است. این نتیجه با حذف مفاهیم تکراری استخراج‌شده از خلاصه پرونده‌ها به ۰/۷۳۱ ارتقا پیدا کرده است که نشان می‌دهد تلاش برای انتخاب عبارت‌های کلیدی از متن‌ها در غنی‌تر کردن مدل‌ها تأثیرگذار خواهد بود.

ازجمله محدودیت‌های این پژوهش عدم دسترسی به متن‌های فارسی خلاصه پرونده است زیرا مجموعه داده فقط شامل متن‌های انگلیسی است .

پیشنهادها

به‌منظور دستیابی به یک بردار بازنمایی مطلوب از داده‌های EHR لازم است همه مفاهیم بالینی به منبع دانش UMLS نگاشت شوند و سپس از آنجا که حجم زیادی از کلمات و اشارات تخصصی در خلاصه پرونده وجود دارد و اهمیت آن‌ها از نظر بالینی یکسان نیست، مفاهیم بالینی پراهمیت‌تر با روش TF-IDF انتخاب شود. همچنین پیشنهاد مدل زبانی مناسب برای بازنمایی SciBERT است.

تشکر و قدردانی

نویسندگان این مقاله، از کلیه کارشناسان که ما را در انجام این پژوهش یاری نمودند، تشکر و قدردانی می‌نمایند.

تضاد منافع

در انجام پژوهش حاضر، نویسندگان هیچ‌گونه تضاد منافعی نداشتند.

References

1. Shah SM, Khan RA. Secondary use of electronic health record: Opportunities and challenges. IEEE Access [Internet]. 2020;8:136947–65. Available from: http://dx.doi.org/10.1109/ACCESS.2020.3011099
2. Pokharel S, Zuccon G, Li X, Utomo CP, Li Y. Temporal tree representation for similarity computation between medical patients. Artif Intell Med. 2020 Jun 11;108:101900.
3. Memarzadeh H, Ghadiri N, Samwald M, Lotfi Shahreza M. A study into patient similarity through representation learning from medical records. Knowl Inf Syst. 2022;64(12):3293–324.
4. Hosseini Pozveh Z, Monadjemi A, Ahmadi A. FNLP-ONT: A feasible ontology for improving NLP tasks in Persian. Expert Syst. 2018 Aug;35(4):e12282.
5. Koroleva A, Kamath S, Paroubek P. Measuring semantic similarity of clinical trial outcomes using deep pre-trained language representations. J Biomed Inform [Internet]. 2019;100:100058. Available from: https://www.sciencedirect.com/science/article/pii/S2590177X19300575
6. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv Prepr arXiv181004805. 2018; Oct 11.
7. OpenAI TB. Chatgpt: Optimizing language models for dialogue. OpenAI. 2022.
8. Pan W, Zhong E, Yang Q. Transfer learning for text mining. Min text data. 2012;223–57.

9. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. BioNLP 2019 - SIGBioMed Work Biomed Nat Lang Process Proc 18th BioNLP Work Shar Task. 2019;58–65.
10. Hu Y, Nie T, Shen D, Kou Y, Yu G. An integrated pipeline model for biomedical entity alignment. Front Comput Sci [Internet]. 2021;15(3):153321. Available from: https://doi.org/10.1007/s11704-020-8426-4
11. Kang T, Perotte A, Tang Y, Ta C, Weng C. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Informatics Assoc. 2021 Apr;28(4):812–23.
12. Darabi S, Kachuee M, Fazeli S, Sarrafzadeh M. TAPER: Time-aware patient EHR representation. IEEE J Biomed Heal Informatics [Internet]. 2020 Dec [cited 2019 Dec 24];24(11):3268–75. Available from: http://arxiv.org/abs/1908.03971
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40.
14. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, et al. Publicly available clinical BERT embeddings. arXiv Prepr arXiv190403323. 2019; Apr 6.
15. Beltagy I, Lo K, Cohan A. SCIBERT: A pretrained language model for scientific text. EMNLP-IJCNLP 2019 - 2019 Conf Empir Methods Nat Lang Process 9th Int Jt Conf Nat Lang Process Proc Conf. 2019;3615–20.
16. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc. 2021;3(1):1–23.
17. Michalopoulos G, Wang Y, Kaka H, Chen H, Wong A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. arXiv Prepr arXiv201010391. 2021;1744–53.
18. Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3.
19. Allvin H, Carlsson E, Dalianis H, Danielsson-Ojala R, Daudaravičius V, Hassel M, et al. Characteristics of Finnish and Swedish intensive care nursing narratives: a comparative analysis to support the development of clinical language technologies. In: Journal of Biomedical Semantics. Springer; 2011;(2): 1–11.
20. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural language processing. BioNLP 2019 - SIGBioMed Work Biomed Nat Lang Process Proc 18th BioNLP Work Shar Task. 2019;319–27.
21. Sammut C, Webb GI, editors. TF–IDF BT - Encyclopedia of Machine Learning. In Boston, MA: Springer US; 2010. p. 986–7. Available from: https://doi.org/10.1007/978-0-387-30164-8_832
22. Goutte C, Gaussier E. A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation BT - Advances in Information Retrieval. In: Losada DE, Fernández-Luna JM, editors. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005: 345–59.
23. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit. 1997;30(7):1145–59.
24. Dligach D, Miller T. Learning Patient Representations from Text. NAACL HLT 2018 - Lex Comput Semant SEM 2018, Proc 7th Conf [Internet]. 2018;119–23. Available from: https://aclanthology.org/S18-2014
25. Duque A, Fabregat H, Araujo L, Martinez-Romo J. A keyphrase-based approach for interpretable ICD-10 code classification of Spanish medical reports. Artif Intell Med [Internet]. 2021;121:102177. Available from: https://www.sciencedirect.com/science/article/pii/S0933365721001706
26. Shen Z, Schutte D, Yi Y, Bompelli A, Yu F, Wang Y, et al. Classifying the lifestyle status for Alzheimer's disease from clinical notes using deep learning with weak supervision. BMC Med Inform Decis Mak. 2022;22(1):1–11.

Health Information Management

Vol 20/ No 2/ Summer 2023

Published by *Isfahan University of Medical Sciences*

DOI: 10.48305/him.2023.41593.1100

Patient Similarity Model Using Discharge Sheet Representation and Final Diagnosis Prediction

Hoda Memarzadeh[1], Nasser Ghadiri [2] , Maryam Lotfi Shahreza[3]

Original Article

Abstract

Introduction: The clinical trials recorded in the electronic health record (EHR) contains important information about the patient's history and the treatments performed. Since clinical notes are stored unstructured, they cannot be applied directly in machine learning algorithms. One way to structure textual data is to represent them as vectors.

Methods: In this research, the discharge sheets are used to generate the vector corresponding for each patient. Language models are used to represent the latest text processing methods. The dataset contains the discharge sheets of more than 26,000 patient records from the Medical Information Mart for Intensive Care III (MIMIC-III) database. To analyze the quality of representation framework, the diagnosis prediction downstream task is used and the evaluation criteria are reported for each language model.

Results: Among the LLMs used in the framework, the best one for the discharge sheets is the BIO-BERT model and then the SciBERT model, which produced the ROC_AUC 0.715 and 0.713 respectively. This evaluation criterion is used to check the quality of forecasting models. The use of clinical text preprocessing and mapping of clinical entities to their standard names in the UMLS knowledge base has improved the evaluation criteria for specific language models in the biomedical field, and the greatest improvement is related to the UMLSBERT model, which is trained on the standard names of the knowledge base.

Conclusion: BIO-BERT and SciBERT language models that trained on the data of clinical papers are suggested as the best option for representing the discharge sheet to vectors. However, since the for discharge sheets are different from the scientific paper in terms of structure and content, the preprocessing of clinical trials in order to identify entities and map them to knowledge sources to fetch the standard names of clinical concepts improves the results obtained in clinical LLMs.

Keywords: Natural language processing; Health informatics; Large Language Model

Received: 5 March; 2023 Accepted: 6 June; 2023 Published: 5 July; 2023

Citation: Memarzadeh H, Ghadiri N, Lotfi Shahreza M. **Patient Similarity Model Using Discharge Sheet Representation and Final Diagnosis Prediction.** Health Inf Manage 2023; 20(2):65-71.

Article resulted from an independent research without financial support.

1.PhD Student, Engineering, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan
2 Associate Professor, Engineering, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan
3-Assisstant Professor, Engineering Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan
Corresponding author: Nasser Ghadiri; Associate Professor, Engineering, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan. Email: nghadiri@iut.ac.ir