ساخت پیکره برچسب خورده گزارش‌های آسیب‌شناسی

مسلم سمیعی پاقلعه[1]، مهرنوش شمس‌فرد[2]
[1]دانشگاه شهید بهشتی، m.samiee@sbu.ac.ir
[2] دانشگاه شهید بهشتی، m-shams@sbu.ac.ir

چکیده - تجزیه و تحلیل سریع و قابل اعتماد متن و استخراج اطلاعات از متن‌های فاقد ساختار از مباحث ضروری تحلیل داده‌های بزرگ است. جهت انجام این موضوع مجموعه‌ای مناسب و به اندازه کافی بزرگ از متن‌ها مورد نیاز است. ساخت چنین مجموعه‌ای از متن‌ها با دو چالش مهم رو به رو است. چالش اول تولید بسیار مشکل و پرهزینه چنین مجموعه‌ای به ویژه در حوزه پزشکی است. چالش دوم برچسب‌زنی دستی و غیرخودکار متن‌های مذکور است که یک عمل زمانبر و خسته‌کننده است. هدف از این نوشتار گزارش ساخت پیکره گزارش‌های آسیب‌شناسی و برچسب‌زنی خودکار آن‌ها است که بتوان از آن جهت آموزش مدل‌ها به منظور استخراج اطلاعات استفاده کرد. مجموعه اولیه گزارش‌ها که به زبان انگلیسی می‌باشد از چهار مرکز درمانی جمع‌آوری شده است. پس از جمع‌آوری گزارش‌ها پیش‌پردازش-های اولیه انجام شده و تمامی آن‌ها پاک‌سازی می‌شوند. سپس براساس کدهای استاندارد سازمان بهداشت جهانی منطبق بر دو کتاب رده-بندی بین‌المللی بیماری‌ها و رده‌بندی بین‌المللی سرطان‌ها چهار برچسب مکان اولیه، زیرمکان، رفتار سرطان و نوع بافت‌شناسی به گزارش‌ها زده شده و پیکره برچسب خورده ساخته می‌شود. به دلیل اینکه برچسب‌زنی گزارش‌های پیکره به صورت نیمه‌خودکار صورت گرفته است، ارزیابی عملیات برچسب‌زنی توسط خبره انسانی و با تایید دقت ۹۷% انجام شده است.

کلید واژه - گزارش آسیب‌شناسی، برچسب‌زنی، کد مکان‌شناسی، کد ریخت‌شناسی، ساخت پیکره

۱- مقدمه

گزارش آسیب‌شناسی یک سند پزشکی است که توسط خبره انسانی که به طور خاص می‌تواند متخصص آسیب‌شناسی باشد تولید می‌شود. این گزارش شامل توصیف و تشریح وضعیت و شرایط سلول‌ها و بافت‌های نمونه‌ای از بدن است که آسیب‌شناس با استفاده از میکروسکوپ مشاهده و ثبت کرده است. برخی از خصوصیات نمونه که توسط آسیب‌شناس ثبت می‌شود عبارتند از اندازه، شکل، قطر، رنگ، ظاهر نمونه و غیره[۱]. آسیب‌شناس خبره انسانی است که متخصص تفسیر گزارش‌های آزمایشگاهی و ارزیابی سلول‌ها، بافت‌ها و اندام‌های بدن بیمار به منظور تشخیص بیماری است. گزارش‌های آسیب‌شناسی نقش بسیار مهمی در تشخیص و توصیف بیماری‌ها به ویژه سرطان دارند. این تشخیص و توصیف بر تعیین گزینه‌های درمانی بیماری کمک می‌کند. گزارش‌های آسیب‌شناسی دارای چندین بخش به نام‌های نوع بافت، ماکروسکوپی، میکروسکوپی، تشخیص نهایی، پیشینه پزشکی و توضیحات هستند.

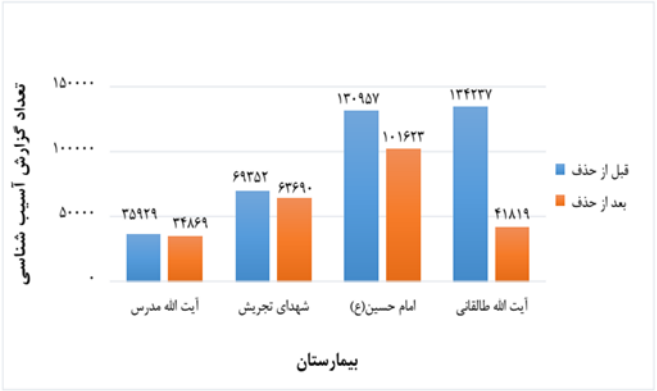
براساس طبقه‌بندی سازمان بهداشت جهانی در گزارش‌های آسیب‌شناسی بیماری‌ها با استفاده از کدهایی به نام کدهای رده‌بندی بین‌المللی بیماری‌ها ثبت می‌شود[۲]. کدهای مکان‌شناسی و ریخت-شناسی دو نمونه از این کدها هستند که هر بخش از این کدها شامل اطلاعاتی مهمی راجع به بیماری و بافتی از بدن که درگیری بیماری شده است، می‌باشد. کد مکان‌شناسی مکان دقیق بیماری در بدن را مشخص می‌کند در حالی که کد ریخت‌شناسی حاوی اطلاعاتی در رابطه با سرطان نظیر نوع بافت و نوع سلول، رفتار و همچنین درجه آن است. گزارش‌های آسیب‌شناسی به زبان طبیعی نوشته شده و بدون ساختار هستند[۳]. لذا پردازش چنین گزارش‌هایی و استخراج اطلاعات ساخت‌یافته از آن‌ها با استفاده از تکنیک پردازش زبان طبیعی مورد پژوهش محققان بسیار بوده است. برای بهبود کارایی رویکردهای پردازش این گزارش‌ها نیاز است که پردازش چنین گزارش‌هایی از رویکردهای مبتنی بر قاعده به سمت رویکردهای مبتنی بر یادگیری ماشین حرکت کند. رویکردهای مبتنی بر یادگیری ماشین به ویژه رویکردهای مبتنی بر یادگیری عمیق به پیکره برچسب خورده بزرگی نیاز دارند.

یکی از اولین پیکره‌های موجود در زمینه آسیب‌شناسی پیکره‌ای است که شامل ۵۶۴۲۶ گزارش آسیب‌شناسی سرطان بوده که تنها دارای برچسب مکان اولیه می‌باشد. گزارش‌های موجود در این پیکره از ۳۵ آزمایشگاه متفاوت جمع‌آوری شده است. بعلاوه گزارش‌های موجود در این پیکره در بخش تشخیص نهایی خود تنها دارای یک مکان اولیه هستند. در این پیکره ۷۰ کلاس متمایز مکان اولیه وجود دارد که ۱۳ کلاس آخر آن کمتر از ۵۰ گزارش آسیب‌شناسی را در خود جای داده‌اند. این در حالی است که ۱۰ کلاس پرتکرار مکان اولیه در این پیکره نزدیک به ۷۰ درصد از گزارش‌های آسیب‌شناسی را شامل می‌شود. [۴]. پیکره دیگری در این زمینه دارای ۳۷۴۸۹۹ گزارش آسیب‌شناسی سرطان است که در آن هر گزارش دارای حداقل ۱ و حداکثر ۶ برچسب از برچسب‌های مکان اولیه با ۷۰ کلاس، زیرمکان با ۳۰۶ کلاس متمایز، نوع بافت شناسی با ۵۱۶ کلاس متمایز، رفتار با ۴ کلاس متمایز و درجه سرطان با ۹ کلاس متمایز می‌باشد. گزارش‌های موجود در این پیکره تنها از یک مرکز ثبت سرطان جمع‌آوری شده است[۵]. پیکره بعدی در این زمینه دارای ۵۷۴۰۲۷ گزارش آسیب‌شناسی است که از دو مرکز مستقل ثبت سرطان جمع‌آوری شده و دارای دو برچسب مکان اولیه و زیرمکان می‌باشد. گزارش‌های جمع‌آوری شده از مرکز اول مربوطه به سال‌های ۲۰۱۸ - ۲۰۰۹ و دارای ۳۰۶ کلاس متمایز زیرمکان و گزارش‌های جمع‌آوری از مرکز دوم مربوط به سال‌های ۲۰۱۸ – ۲۰۰۴ و دارای ۲۹۹ برچسب متمایز می‌باشد[۶]. پیکره‌ای نیز دارای ۳۶۰۲۰۲ گزارش آسیب‌شناسی می‌باشد که مربوط به سال‌های ۲۰۱۷ - ۲۰۰۴ بوده و تنها از یک مرکز ثبت سرطان جمع‌آوری شده است. تعداد گزارش‌های باقی مانده در این پیکره پس از انجام عملیات پیش-پردازش‌های اولیه ۹۵۲۳۱ می‌باشد هر گزارش دارای حداقل ۱ و حداکثر ۶ برچسب از برچسب‌های مذکور است[۷].

دو چالش بر سر راه استفاده از پیکره‌های مذکور جهت ساخت مدل به منظور پردازش گزارش‌های آسیب‌شناسی وجود دارد. چالش اول این است که به دلیل محرمانه و خصوصی بودن اطلاعات سلامت بیماران و حفظ حریم خصوصی آن‌ها پیکره‌های مذکور عمومی و رایگان نبوده و تنها توسط گروه جمع‌آوری‌کننده قابل استفاده می-باشند. چالش دوم یکسان نبودن شیوه نگارش گزارش‌های آسیب-شناسی در ایران با شیوه نگارش گزارش‌های مذکور است. فقدان وجود پیکره‌ای از گزارش‌های آسیب‌شناسی که بتوانیم از آن به صورت عمومی و آزاد جهت پردازش و استخراج اطلاعات استفاده کنیم و تفاوت شیوه نگارش گزارش‌ها در ایران با موارد مشابه در سایر کشورها ما را بر آن داشت تا به ساخت یک پیکره برچسب خورده از گزارش-های آسیب‌شناسی به صورت نیمه خودکار بپردازیم. این در حالی است که تا کنون پیکره برچسب خورده‌ای از گزارش‌های آسیب-شناسی سرطان به صورت خودکار یا نیمه‌خودکار ساخته نشده است. استفاده عمومی از این پیکره منوط به اخذ مجوز از مراکز درمانی تامین‌کننده گزارش‌ها است.

۲- پیکره گزارش‌های آسیب‌شناسی

مجموعه داده مورد استفاده در این نوشتار پیکره بزرگی از گزارش‌های آسیب‌شناسی به زبان انگلیسی می‌باشد که از چهار مرکز بهداشتی، درمانی و آموزشی آیت‌الله مدرس، شهدای تجریش، امام حسین(ع) و آیت‌الله طالقانی که زیرمجموعه دانشگاه علوم پزشکی و خدمات درمانی شهید بهشتی هستند گردآوری شده و با شناسه ملی اخلاق در پژوهش‌های زیست پزشکی IR.SBU.REC.1398.032 ثبت شده است. به این ترتیب پیکره ساخته شده از رده پیکره‌های چند منبعی می‌باشد.گزارش‌های موجود در پیکره مربوط به سال‌های ۱۳۹۷ - ۱۳۸۶ است که توسط متخصصین و خبره‌های انسانی حوزه آسیب‌شناسی با سلیقه‌های نگارشی متفاوت نوشته و تولید شده است. تعداد اولیه گزارش‌ها پیکره ۳۷۰۴۷۵ و تعداد نهایی آن‌ها ۲۴۲۰۰۱ است. شکل ۱ اطلاعات مربوط به این پیکره را قبل و بعد از حذف گزارش‌های خالی و تکراری به تفکیک هر مرکز درمانی نشان می‌دهد.



شکل ۱ : اطلاعات پیکره گزارش های آسیب شناسی

۳- پیش‌پردازش‌ها

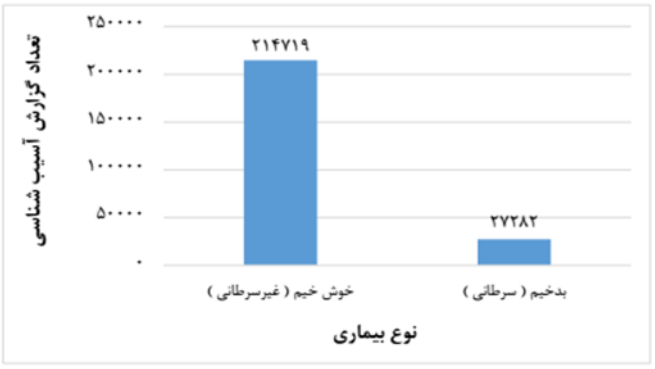
جهت انجام پیش پردازش‌ها ابتدا دو بخش میکروسکوپی و ماکروسکوپی[۱] هر گزارش با هم ادغام شده و به عنوان بخش جدیدی از گزارش به نام بخش ماکروسکوپی- میکروسکوپی ذخیره می‌شود. سپس بر روی متن موجود در این بخش پیش‌پردازش‌هایی

نظیر حذف تمامی کاراکترهای غیرالفبایی نظیر اعداد، علائم نگارشی، تک‌واژه‌سازی متن، حذف ایست‌واژه‌ها، حذف کلمه‌های تکراری به منظور تمیزسازی گزارش‌ها و آماده‌سازی آن‌ها جهت استخراج برچسب‌های مورد نظر انجام می‌شود. گرچه کاراکترهای غیرالفبایی می‌توانند از نظر معنایی دارای ارزش بوده و مهم باشند، اما در مجموعه داده موجود این نمادها اغلب مبهم هستند. یک نمونه از این کاراکترها نقطه است که در جایی با آن یک جمله پایان می‌پذیرد و در جایی دیگر در یک عدد اعشاری ظاهر می‌شود.

۴- تعیین برچسب گزارش‌های آسیب‌شناسی

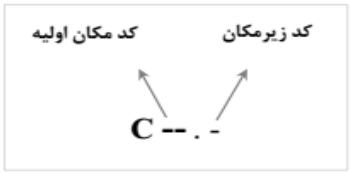
۴-۱- کدهای مکان‌شناسی و ریخت‌شناسی

گزارش‌های موجود در پیکره گردآوری شده به دو دسته گزارش به نام گزارش‌های مربوط به بیماری‌های خوش‌خیم و گزارش‌های مربوط به بیماری‌های بدخیم(سرطان) تقسیم می‌شوند (شکل۲). همانطوری که بیان شد گزارش‌های آسیب‌شناسی می‌توانند دارای دو کد مهم به نام‌های کد مکان‌شناسی و کد ریخت‌شناسی باشند که براساس این کدها می‌توان به یک گزارش برچسب‌های مختلفی زد. گزارش‌های مربوط به بیماری‌های خوش‌خیم می‌توانند تنها دارای کد مکان‌شناسی باشند. در حالی که گزارش‌های مربوط به بیماری‌های بدخیم می‌توانند دارای هر دو کد مکان‌شناسی و ریخت‌شناسی باشند. لذا در این نوشتار ابتدا به همه گزارش‌های موجود در پیکره (گزارش-های مربوط به بیماری‌های خوش‌خیم و بدخیم) برچسب مکان‌شناسی زده شده و سپس به گزارش‌های مربوط به بیماری‌های بدخیم برچسب دیگری به نام ریخت‌شناسی نیز زده می‌شود.



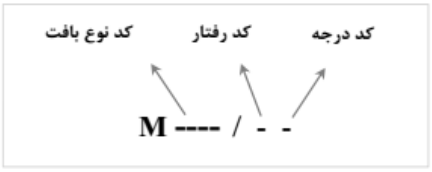
شکل ۲ : تعداد گزارش ها به تفکیک نوع بیماری

می‌توان براساس کد مکان‌شناسی دو برچسب دیگر به نام‌های مکان اولیه و زیرمکان را به گزارش آسیب‌شناسی نسبت داد. دو برچسب مکان اولیه و زیرمکان به ترتیب عضو اصلی بدن و بخش جزئی‌تر و دقیق‌تری از عضو اصلی را که درگیر بیماری شده است مشخص می‌کنند. به عنوان مثال، ریه یکی از اعضای اصلی بدن انسان است که در طبقه‌بندی سازمان جهانی بهداشت به سه بخش شاخه اصلی، ریه راست و ریه چپ تقسیم می‌شود. ریه راست خود به سه زیرمکان به نام‌های لُپ بالایی، لُپ وسطی و لُپ پایینی و ریه چپ به دو زیرمکان به نام‌های لُپ بالایی و لُپ پایینی تقسیم می‌شود. لذا، ریه می‌تواند مکان اولیه و یک سوم بالایی ریه یا همان لپ بالایی می‌تواند به عنوان زیرمکان در نظر گرفته شود[۲]. ساختار که کد مکان‌شناسی به صورت شکل ۳ است.



شکل ۳ : ساختار کد مکان شناسی

همچنین می‌توان براساس کد ریخت‌شناسی نیز دو برچسب دیگر به نام‌های نوع بافت‌شناسی و رفتار را به گزارش آسیب‌شناسی نسبت داد. ساختار کد ریخت‌شناسی به صورت شکل ۴ است. چهار رقم اول در کد ریخت‌شناسی نوع بافت و نوع سلولی از بدن را که درگیر سرطان شده مشخص می‌کند. رقم پنجم در کد ریخت‌شناسی تعیین‌کننده رفتار سرطان است. مقادیر ممکن برای این رقم عبارتند از : ۰ ، ۱ ، ۲ ، ۳ ، ۶ که هر عدد نماینده نوع خاصی از رفتار سرطان است. رقم ششم در کد ریخت‌شناسی نیز نشان‌دهنده درجه سرطان است.



شکل ۴ : ساختار کد ریخت شناسی

۴-۲- تولید نیمه خودکار برچسب‌های مکان اولیه و زیرمکان

در تعداد زیادی از گزارش‌های آسیب‌شناسی عضوی از بدن که درگیر بیماری شده به صورت صریح توسط خبره انسانی تعیین نشده است. در واقع این گزارش‌ها فاقد هرگونه حاشیه‌نویسی لازم جهت برچسب مکان‌شناسی و به تبع آن دو برچسب مکان اولیه و زیرمکان مربوط به بیماری است. این حاشیه‌نویسی عموماً توسط خبره انسانی یا متخصص آسیب‌شناسی انجام می‌شود. این در حالی است که چنین حاشیه‌نویسی بسیار اهمیت داشته و وجود آن ضروری است. لذا به دلیل عدم انجام حاشیه‌نویسی توسط خبره انسانی و به منظور انجام

خودکار این عمل و زدن سه برچسب مذکور به هر گزارش آسیب-شناسی به صورت زیر عمل می‌شود.

- جمع‌آوری مجموعه‌ای از واژگان تخصصی آسیب‌شناسی و بافت‌شناسی بدن از کتاب‌های تخصصی [۸] موجود در این حوزه و دو کتاب طبقه‌بندی بین‌المللی بیماری‌ها [۹] و طبقه‌بندی بین‌المللی سرطان‌ها [۲] که توسط سازمان بهداشت جهانی نگارش و چاپ شده‌اند. این مجموعه دارای ۳۰۰۰ عبارت کلیدی و تخصصی تک‌کلمه‌ای، دو کلمه‌ای، سه کلمه‌ای و چهار کلمه‌ای علم آسیب‌شناسی و بافت‌شناسی است که به صورت دستی از دو کتاب مذکور استخراج شده است. در مجموعه داده‌گان مورد نظر مقابل هر یک از این عبارت‌های تخصصی بیماری متناظر با آن ثبت شده است. در کنار این مجموعه از مجموعه‌ای دیگر از کدهای مکان‌شناسی که توسط سازمان بهداشت جهانی و جهت شناسایی برخی از بیماری‌ها تولید شده است، استفاده شده است.

- استخراج و ثبت کد مکان‌شناسی از بخش ماکروسکوپی- میکروسکوپی هر گزارش با استفاده از مجموعه واژگان جمع‌آوری شده و پیمانه مربوطه. در این مرحله ابتدا متن موجود در دو بخش مذکور از گزارش آسیب‌شناسی با هم ادغام شده و با استفاده از پیمانه مورد نظر تک‌کلمه‌ای، دو کلمه‌ای، سه کلمه‌ای و چهار کلمه‌ای آن تولید شده و سپس با تطبیق آن‌ها با عبارت-های تخصصی آسیب‌شناسی و بافت‌شناسی کد استخراج می‌شود.

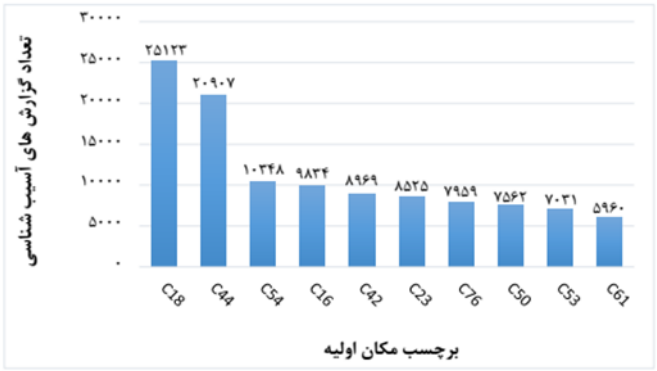
- بررسی و اصلاح برچسب استخراج شده توسط متخصص آسیب شناسی. در بررسی‌های انجام شده توسط خبره انسانی دقت ۹۷% برای برچسب مکان‌شناسی تایید شده است. دقت بدست آمده با انتخاب تصادفی مجموعه‌های ۱۰۰ تایی از گزارش آسیب‌شناسی صورت گرفته است.

- تولید دو برچسب مکان اولیه و زیرمکان برای هر گزارش آسیب‌شناسی از روی برچسب مکان‌شناسی استخراج شده. بدیهی است که براساس دقت برچسب مکان‌شناسی دقت برچسب‌های مکان اولیه و زیرمکان نیز همان ۹۷٪ می‌باشد.

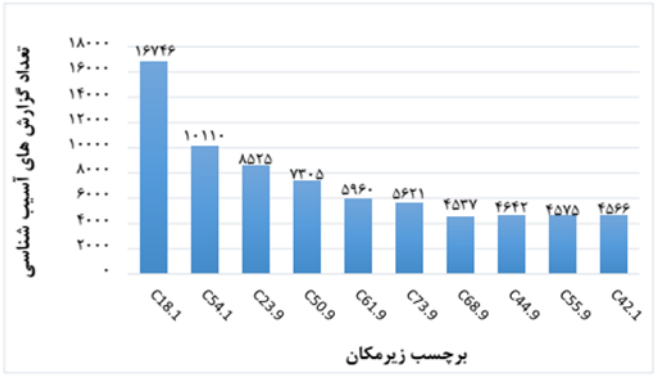
پس از مراحل فوق پیکره‌ای از گزارش‌های آسیب‌شناسی در اختیار داریم که دارای دو برچسب مکان اولیه و زیرمکان است. با آگاهی از این موضوع که متن موجود در یک گزارش آسیب‌شناسی لزوماً به تشریح و تفسیر یک نمونه بافت نمی‌پردازد و ممکن است نمونه‌گیری جهت انجام آزمایش آسیب‌شناسی از چند عضو بدن صورت گرفته باشد، به روشنی درک می‌شود که یک گزارش آسیب‌شناسی ممکن است دارای چند برچسب مکان‌شناسی و در ضمن آن چند برچسب مکان اولیه و زیرمکان باشد. شکل ۵ و شکل ۶ به ترتیب ۱۰ کلاس پرتکرار برچسب‌های مکان اولیه و زیرمکان را نشان می‌دهد.

۳-۴- تولید نیمه خودکار برچسب‌های رفتار و نوع بافت-شناسی

پیش‌تر بیان شد کد ریخت‌شناسی در گزارش آسیب‌شناسی تنها برای بیماری‌های بدخیم نظیر سرطان تعریف می‌شود که توسط متخصص آسیب شناسی و در بخش تشخیص نهایی گزارش نوشته می‌شود. به همین دلیل نیاز است گزارش‌هایی که دارای چنین کدی هستند از سایر گزارش‌ها تفکیک شده و کد ریخت شناسی آن‌ها از بخش تشخیص نهایی استخراج شود. در واقع همانند کد مکان‌شناسی باید به گزارش آسیب‌شناسی یک برچسب دیگر به نام برچسب ریخت‌شناسی زده شود. همانطوری که بیان شد کد ریخت‌شناسی در یک گزارش آسیب‌شناسی دارای بخش‌های مختلفی است که هر یک از این بخش‌ها بیان‌کننده یکی از خصوصیات سرطان است. با این کد می‌توان نوع بافت و نوع سلول درگیر با سرطان شده در ۶۱ گروه بافت‌شناسی مختلف، رفتار سرطان را در میان ۵ گروه مختلف و همچنین درجه سرطان را در میان ۵ گروه مختلف تعیین نمود[۲].



شکل ۵ : ده کلاس پرتکرار برچسب مکان اولیه



شکل ۶ : ده کلاس پرتکرار برچسب زیرمکان

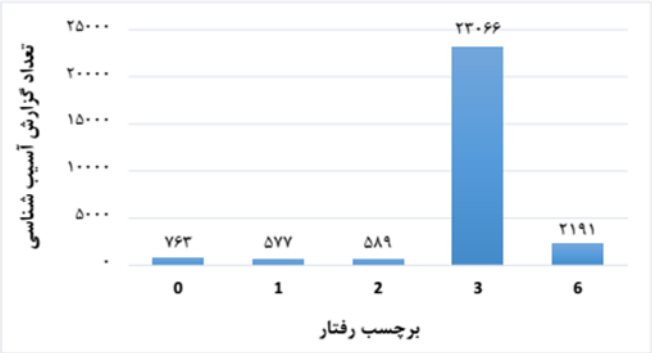
شایان ذکر است که گزارش‌های موجود در پیکره مذکور دارای کد درجه سرطان نیستند. زدن برچسب‌های رفتار و نوع بافت‌شناسی به گزارش‌ها در دو مرحله زیر انجام می‌شود.

- در مرحله اول پس از پردازش بخش تشخیص نهایی گزارش‌های موجود در پیکره، گزارش‌های مربوط به بیماری‌های خوش‌خیم و بدخیم از هم تفکیک می‌شوند(شکل ۲). همانطوری که بیان شد گزارش های مربوط به بیماری های خوش‌خیم فاقد کد ریخت‌شناسی هستند. در حالی که گزارش‌های مربوط به بیماری‌های بدخیم دارای این کد هستند. لذا، بعد از پردازش کد ریخت‌شناسی هر گزارش استخراج شده و در بخش جدیدی به نام برچسب ریخت شناسی برای آن گزارش ثبت می‌شود.

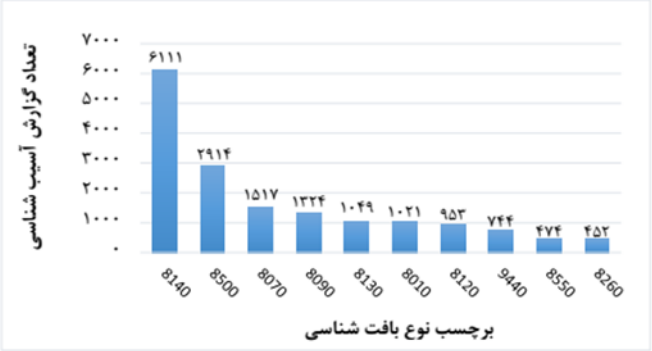
- در این مرحله کد ریخت شناسی گزارش‌هایی که دارای چنین کدی هستند پردازش شده و دو بخش اصلی آن از هم تفکیک می‌شوند. بخش اول قسمتی از کد ریخت‌شناسی است که بیان‌کننده نوع بافت یا همان نوع سلول عضو از بدن است که به سرطان مبتلا شده است. بخش دوم نیز آن قسمتی از کد ریخت‌شناسی است که بیان‌کننده رفتار سرطان است. هر یک از این دو قسمت در بخش جدیدی به همان نام برای هر گزارش ثبت می‌شود. در واقع پس از این مرحله به هر یک از ۲۷۲۸۲ گزارش آسیب شناسی که دارای کد ریخت شناسی هستند، دو برچسب جدید به نام‌های نوع بافت و رفتار زده می‌شود. برای انجام این کار کد ریخت‌شناسی هر یک از گزارش های آسیب‌شناسی به عنوان ورودی به پیمانه تجزیه‌کننده داده می‌شود. برای تولید برچسب نوع بافت یک گزارش آسیب‌شناسی چهار رقم اول بعد از حرف M از کد ریخت شناسی جدا شده و استخراج می‌شود. برای تولید برچسب رفتار سرطان نیز رقم پنجم در کد ریخت‌شناسی یعنی اولین رقم بعد نشانه « / » جدا شده و استخراج می‌شود. شکل ۷ تمامی کلاس‌های مختلف برچسب رفتار و شکل ۸ کلاس‌های پُرتکرار برچسب نوع بافت‌شناسی را به همراه تعداد آن‌ها نشان می‌دهد.

۵- نتیجه‌گیری و کارهای آینده

گزارش‌های آسیب‌شناسی به عنوان اولین و اصلی‌ترین سندهای پزشکی هستند که بسیاری از بیماری‌های خوش‌خیم و بدخیم به ویژه سرطان در آن توصیف و ثبت می‌شود[۱۰] [۱۱]. این گزارش‌ها فاقد ساختار بوده و استخراج دستی اطلاعات از آن‌ها بسیار زمان‌بر، پرهزینه و خسته‌کننده است[۷]. لذا پردازش خودکار این گزارش‌ها و استخراج اطلاعات مفید و مرتبط با بیماری‌ها از آن‌ها می‌تواند کمک شایان توجهی جهت کدگذاری و ساختاردهی به آن‌ها کند. پردازش خودکار گزارش‌های آسیب شناسی جهت استخراج اطلاعات نیازمند ساخت مدل‌های هوشمند و ساخت این مدل‌ها نیز نیازمند پیکره برچسب خورده بزرگی جهت آموزش آن‌ها است.



شکل ۷ : کلاس های مختلف برچسب رفتار



شکل ۸ : ده کلاس پُرتکرار برچسب نوع بافت شناسی

لذا بر همین اساس در این نوشتار گزارشی از ساخت پیکره برچسب خورده‌ای از گزارش‌های آسیب‌شناسی ارائه شد. پیکره حاصل به چهار پیکره کوچکتر به نام‌های پیکره مکان اولیه، پیکره زیرمکان، پیکره نوع بافت‌شناسی سرطان و پیکره رفتار سرطان تقسیم می‌شود. آمار بدست آمده از این پیکره‌ها به این صورت است که پیکره مکان اولیه دارای ۱۸۸۷۱۱ گزارش آسیب‌شناسی با ۶۹ کلاس متمایز، پیکره زیرمکان دارای ۱۷۵۳۳۲ گزارش آسیب‌شناسی با ۲۲۲ کلاس متمایز ، پیکره نوع بافت‌شناسی سرطان دارای ۲۷۱۹۸ گزارش آسیب‌شناسی با ۵۴۸ کلاس متمایز و در نهایت پیکره رفتار سرطان دارای ۲۷۱۸۸ گزارش آسیب‌شناسی با ۵ کلاس متمایز است. کاربرد چنین پیکره‌ای را می‌توان ساخت مدل یادگیری ماشین و به طور خاص مدل یادگیری عمیق جهت پردازش گزارش‌های آسیب‌شناسی

با هدف استخراج خودکار اطلاعات ساخت‌یافته از متن آن‌ها است. استخراج چنین اطلاعاتی از این گزارش‌ها می‌تواند جهت رمزگذاری و ساختاردهی به این سندهای پزشکی که فاقد ساختار هستند بسیار مفید و کمک‌کننده باشد.

همانطوری که بیان شد مجموعه بزرگی از گزارش های آسیب‌شناسی از چهار مرکز درمانی و آموزشی جمع‌آوری و برچسب زده شد تا بتوان از آن‌ها جهت ساخت مدل به منظور پردازش و استخراج اطلاعات ساخت‌یافته استفاده کرد. در مسیر جمع‌آوری و ساخت پیکره برچسب خورده گزارش‌های آسیب‌شناسی چالش‌های بسیار وجود دارد که در ادامه تعدادی از آن‌ها تشریح می‌شود. ابتدایی‌ترین چالش عدم امکان انتشار مجموعه گزارش های جمع‌آوری شده به صورت آزاد، عمومی و رایگان است. چالش بسیار مهم دیگری که بر سر راه ساخت پیکره گزارش‌های آسیب شناسی وجود دارد، فقدان استانداردی یکسان و مشخص جهت نگارش متن گزارش است. متغیر و متنوع بودن گزارش‌های آسیب‌شناسی به دلیل متفاوت بودن نویسندگان و سلیقه‌های نگارشی آنها، جمع‌آوری گزارش‌ها از آزمایشگاه‌ها و مراکز متفاوت و متنوع ثبت سرطان، خطای علمی و عملی خبره انسانی در حاشیه‌نویسی و نگارش متن گزارش‌ها، شیوه‌های نگارشی متفاوت خبره‌های انسانی نظیر خلاصه‌نویسی و کوته‌نویسی‌های غیراستاندارد نمونه‌هایی از مسائلی هستند که منجر شده‌اند پردازش گزارش‌های آسیب‌شناسی و برچسب‌زنی آن‌ها به سختی انجام شود. برخلاف بسیاری از حوزه‌های عمومی پردازش متن که شاید هر شخصی بتواند حاشیه‌نویسی لازم را برای مجموعه داده به منظور تولید استاندارد طلایی انجام دهد، حاشیه‌نویسی گزارش‌ها در حوزه آسیب‌شناسی به منظور تولید استاندارد طلایی نیاز به دانش پزشکی و آسیب‌شناسی دارد.

افزایش مراکز درمانی و آموزشی جهت جمع‌آوری گزارش‌های آسیب‌شناسی با هدف بزرگ‌تر، غنی‌تر و تکمیل کردن پیکره مذکور می‌تواند یکی از کارهای آینده مطرح شود. افزایش اندازه پیکره می‌تواند با متنوع‌تر کردن گزارش‌ها در نتیجه آن افزودن بیماری‌های بیشتر در لیست بیماری‌های تحت پوشش پیکره نیز همراه باشد. افزایش اندازه مجموعه عبارت‌های تخصصی و کلیدی آسیب‌شناسی و بافت‌شناسی که جهت زدن برچسب مکان‌شناسی به گزارش‌های آسیب‌شناسی استفاده می‌شود نیز می‌تواند به عنوان یکی دیگر از کارهای آینده پیش روی این پژوهش در نظر گرفته شود. این موضوع می‌توان سبب افزایش دقت در زدن برچسب مکان‌شناسی شود.

۶- سپاس‌گزاری

در پایان این نوشتار از دانشگاه شهید بهشتی، دانشگاه علوم پزشکی شهید بهشتی، بیمارستان امام حسین(ع)، بیمارستان شهدای تجریش، بیمارستان آیت‌الله طالقانی و بیمارستان آیت‌الله مدرس جهت همکاری و مساعدت در گردآوری مجموعه گزارش‌های آسیب-شناسی مورد نیاز سپاس‌گزاری می‌نمائیم.

مراجع

[1] "Pathology Reports - National Cancer Institute," Dec. 06, 2007. https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet.

[2] A. Fritz *et al.*, *International classification of diseases for oncology*. World Health Organization, 2000.

[3] J. X. Qiu, H.-J. Yoon, P. A. Fearn, and G. D. Tourassi, "Deep Learning for Automated Extraction of Primary Sites From Cancer Pathology Reports," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 1, pp. 244–251, Jan. 2018.

[4] Ramakanth Kavuluru and Isaac Hands, "Automatic Extraction of ICD-O-3 Primary Sites from Cancer Pathology Reports," *AMIA Jt Summits Transl Sci Proc.Published online 2013 Mar 18*, pp. 112–116.

[5] J. X. Qiu *et al.*, "Semi-Supervised Information Extraction for Cancer Pathology Reports," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, May 2019, pp. 1–4.

[6] M. Alawad *et al.*, "Deep Transfer Learning Across Cancer Registries for Information Extraction from Pathology Reports," in *2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, May 2019, pp. 1–4.

[7] M. Alawad *et al.*, "Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks," *J Am Med Inform Assoc*, vol. 27, no. 1, pp. 89–98, Jan. 2020.

[8] A. Mescher, *Junqueira's Basic Histology: Text and Atlas, Thirteenth Edition*, 13th Edition. New York, NY: McGraw-Hill Education / Medical, 2013.

[9] "WHO | ICD-10 online versions," *WHO*. http://www.who.int/classifications/icd/icdonlineversions/en/

[10] M. Alawad, H.-J. Yoon, and G. D. Tourassi, "Coarse-to-fine multi-task training of convolutional neural networks for automated information extraction from cancer pathology reports," in *2018 IEEE EMBS International Conference on Biomedical Health Informatics (BHI)*, Mar. 2018, pp. 218–221.

[11] H. Yoon, A. Ramanathan, and G. Tourassi, "Multi-task Deep Neural Networks for Automated Extraction of Primary Site and Laterality Information from Cancer Pathology Reports," 2016.