مقاله مروری

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره پنجم شماره دوم. ص ۳۲٤-۳۱٤، ۱۳۹۷

مروری بر رویکردها و تکنیک‌های متن‌کاوی در مستندات بالینی

فرزانه فیض‌منش[۱]، علی‌اصغر صفائی[۲]*

• **دریافت مقاله:** ۹۵/۱۲/۱۸ • **پذیرش مقاله:** ۹۶/۹/۵

مقدمه: با به کارگیری گسترده سیستم‌های الکترونیکی مدارک پزشکی، حجم بسیار زیادی از داده‌های متنی پزشکی در بیمارستان و سایر محیط‌های درمانی به صورت روزانه تولید می‌شوند که سازمان‌دهی این اطلاعات متنی امری مهم و ضروری است و نیاز به بازیابی خودکار دانش مفید از این داده‌ها برای کمک به متخصصان بالینی کاملاً احساس می‌شود. به‌منظور استخراج ارزش‌های نهفته در مستندات متنی پزشکی، می‌توان از تکنیک‌های متن‌کاوی در حوزه سلامت بهره برد.

روش: در این پژوهش مروری پایگاه‌های اطلاعاتی Science Direct، IEEE، PubMed central، Google Scholar، SID و Magiran با استفاده از کلید واژه‌های "Text Mining" AND "Medicine"، "Clinical Text Mining" AND "Predict"، "knowledge discovery in medical text"، "Text Mining for Medical and Healthcare" در پایگاه‌های اطلاعاتی انگلیسی و از ترکیب "متن کاوی و کشف دانش در پزشکی" در پایگاه‌های اطلاعاتی فارسی، مورد جستجو قرار گرفتند. سپس، همه مقالاتی که به نوعی به کشف دانش پزشکی و کاربردهای متن‌کاوی در حوزه سلامت اشاره داشتند، انتخاب شدند.

نتایج: متن‌کاوی از تکنیک‌های مهم و قدرتمند برای استخراج اطلاعات از سیستم‌های اطلاعات بهداشتی و درمانی می‌باشد. متن‌کاوی داده‌های کلینیکی، توان بالقوه‌ای برای اکتشافات جدید و همچنین بهبود کارایی و ارتباطات در سیستم‌های بیمارستان برای پزشکان و مدیران بیمارستان فراهم می‌کند.

نتیجه‌گیری: متن‌کاوی در مستندات بالینی از جمله تکنولوژی‌های توسعه یافته کشف دانش پزشکی در عصر حاضر است که استفاده از آن در پایگاه داده‌های پزشکی به‌منظور دستیابی سریع به منابع مهم سلامت، امری انکارناپذیر است و به کارگیری آن موجب بهبود مراقب بیمار و کاهش هزینه‌های درمانی می‌شود.

کلید واژه‌ها: متن کاوی در مستندات بالینی، کاربردهای متن‌کاوی، کشف دانش پزشکی، استخراج اطلاعات

• **ارجاع:** فیض‌منش فرزانه، صفائی علی‌اصغر. مروری بر رویکردها و تکنیک‌های متن‌کاوی در مستندات بالینی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۷؛ ۵(۲): ۳۱۴-۳۲۳.

۱. کارشناس ارشد انفورماتیک پزشکی، گروه انفورماتیک پزشکی، دانشکده علوم پزشکی، تهران، دانشگاه تربیت مدرس، تهران، ایران

۲- دکتری کامپیوتر- نرم افزار، استادیار، گروه انفورماتیک پزشکی، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، تهران، ایران

* **نویسنده مسئول:** تهران، دانشکده علوم پزشکی، دانشگاه تربیت مدرس، گروه انفورماتیک پزشکی

• **شماره تماس:** ۸۲۸۸۴۵۸۱ - ۰۲۱

• **Email:** aa.safaei@modares.ac.ir

مقدمه

اطلاعات و دانش، دارایی‌های با ارزش در هر سازمان محسوب می‌شوند. امروزه بیمارستان‌های مدرن با کامپیوترها و دیگر وسایل جمع‌آوری داده تجهیز شده‌اند و دسترسی سریع و صحیح به منابع مهم و مورد علاقه، یکی از دغدغه‌های استفاده از این منبع اطلاعاتی بسیار بزرگ است. مقدار وسیع داده‌های جمع شده در پایگاه داده‌های پزشکی نیاز به ابزارهای تخصصی جهت ارزیابی، نگهداری و تجزیه و تحلیل دارند. به ویژه افزایش مقادیر داده، باعث به وجود آمدن مشکلات زیادی در استخراج اطلاعات مفید جهت پشتیبانی از تصمیمات می‌باشد و تنها تجزیه و تحلیل داده‌ها به صورت سنتی کافی نیست و نیازمند تجزیه و تحلیل مبتنی بر کامپیوتر می‌باشد. به عنوان مثال، گزارش‌های رادیولوژی حاوی اطلاعات غنی از توصیف مشاهدات رادیولوژیست از شرایط پزشکی بیمار در ارتباط با تصاویر پزشکی می‌باشد؛ اگرچه همانند بسیاری از گزارش‌ها در شکل‌های متنی رایگان هستند، ولی یک مانع بزرگ بین گزارش رادیولوژی و متخصصان پزشکی (رادیولوژی، پزشک و پژوهشگران)، محسوب می‌شوند. که برای آن‌ها بازیابی و استفاده از اطلاعات مفید و دانش از گزارش‌های رادیولوژی را مشکل ساخته است؛ بنابراین به‌منظور ارائه اطلاعات موردنیاز برای متخصصان پزشکی و امکان استفاده از اطلاعات، متن کاوی در گزارش رادیولوژی یک راه حل برای این مشکل فراهم می‌کند [۱].

در تشخیص بسیاری از بیماری‌ها، نظیر اسکیزوفرنی که علائم منفی بالینی، به عنوان بزرگترین عامل وقوع آن به حساب می‌آید و این علائم به عنوان یک هدف مهم برای اهداف درمانی به خصوص پس از پیش‌بینی وقوع عواقب بالینی طولانی مدت محسوب می‌شوند، توسعه روش‌های عملی برای ارزیابی آن‌ها مشکل است، ولی با استفاده از یک رویکرد متن‌کاوی به راحتی می‌توان حضور علائم منفی بالینی در پرونده الکترونیک سلامت بیماران را تشخیص داد و از این یافته‌ها به منظور توسعه روش‌های درمانی جدید که به تسکین علائم منفی بیماری کمک می‌کند، استفاده کرد. علاوه بر این با افزایش بهره‌گیری از پرونده الکترونیک سلامت و استفاده از رویکرد متن‌کاوی، می‌توان به منظور به دست آوردن اطلاعات برای تحقیق و توسعه و پشتیبانی تصمیم‌گیری بالینی در سایر مناطق و مراکز درمانی بهره برد [۲]. هچنین در طبقه‌بندی بسیاری از بیماری‌ها منطبق بر طبقه‌بندی بین‌المللی بیماری‌های سازمان بهداشت جهانی مانند صرع که روند دسته‌بندی بیماری به صورت دستی، وقت‌گیر و نیازمند تحقق آزمایش‌های مکمل است، به منظور نادیده گرفتن این فرآیند دشوار و طاقت‌فرسا، می‌توان از یک فرآیند خودکار طبقه‌بندی تشخیص بیماری بر اساس کدهای (International Classification Disease) ICD استفاده کرد. بمنظور برآورده کردن این نیاز اطلاعات پزشکی، فناوری‌های توسعه یافته در زمینه‌های جدید را به کار می‌برند و در انفورماتیک پزشکی از فناوری ایجاد شده جدید به نام کشف دانش از پایگاه داده استفاده می‌کنند[۳].

اطلاعات پزشکی در قالب الکترونیکی، به طور فزاینده‌ای در حال گسترش هستند، دسته‌ای از این اطلاعات نه تنها شامل اطلاعات دموگرافیکی بیماران، همچنین دربرگیرنده اطلاعات تشخیص و شدت بیماری، نتایج آزمایش‌های بالینی، تست‌های عملکردی و داروهای تجویز شده برای بیمار و جزئیات بیشتری در ارتباط با اطلاعات تماس بیمار با سیستم مراقبت بهداشتی و درمان می‌باشد و دسته دیگر شامل مقالات علمی و همچنین بررسی مدیریت بالینی و حتی مدارک پزشکی موجود در مؤسسات بهداشتی و مراکز درمانی که با داده‌های بیمار سر و کار دارند، می‌باشد. سه مزیت مهم در ثبت این مقدار عظیم از داده‌ها در قالب دیجیتال وجود دارد: ۱- بهبود کیفیت نگهداری داده‌ها، ۲- کاهش چشمگیر در زمان بلااستفاده و غیر مفید کارکنان بهداشت و درمان در انجام وظایف خود، ۳- داده‌ها می‌توانند در سیستم‌های خودکار مانند سیستم‌های متن‌کاوی و داده‌کاوی استفاده شوند؛ با این حال ابزارهای دستی، استفاده کمی برای استخراج اطلاعات مناسب چه در زمینه بالینی و چه در زمینه تحقیقاتی دارند. متن‌کاوی می‌تواند این حجم عظیم از اطلاعات را به درستی مدیریت کند. با استفاده از سیستم‌های پردازشی، آن‌ها را از منابع مختلف استخراج کرده و در نهایت آن‌ها را یکپارچه کند و منجر به ایجاد دانش جدید شود. ارائه ابزارهایی که با بررسی متون بتواند تحلیلی روی آن‌ها انجام دهند منجر به شکل‌گیری این زمینه در هوش مصنوعی شده که به متن کاوی معروف است [۴].

پلتفرم (Medical Language Extraction and Encoding)MEDLEE در سطح سازمانی، نمونه بارز سیستم متن‌کاوی خودکار است که این امکان را فراهم می‌سازد اطلاعات مرتبط و مفید را از گزارش‌های بالینی استخراج کند [۴].

نمونه دیگر از سیستم‌های خودکار استخراج دانش، سیستم clinical Text Analysis and Knowledge Extraction System) cTAKES)، با قابلیت زبان‌شناختی و توسعه روایی معنایی قوی یافته‌های بالینی در پرونده‌های الکترونیکی پزشکی است [۵].

متن‌کاوی در مطالعات پزشکی با اهداف مختلفی، شامل طبقه‌بندی خودکار بیماری‌ها از خلاصه پرونده بیماران، پیش‌بینی و تعیین عوامل خطرزای بیماری به منظور پیشگیری و تشخیص به‌موقع، پیش‌بینی پذیرش بیماران اورژانس در بخش‌های بیمارستان و تجزیه و تحلیل اسناد پزشکی و بالینی برای شناسایی وابستگی‌های دارویی به‌کار گرفته شده است [۶-۸].

با توجه به اهمیت علم پزشکی و بهداشت و درمان در زندگی انسان و نیاز به تحلیل و آنالیز خودکار، در این مقاله، درابتدا به متن‌کاوی و مفاهیم مرتبط با آن پرداخته می‌شود. و رویکردهای اصلی آن مانند بازیابی اطلاعات (Information Recovery)، استخراج اطلاعات (Information extraction) و تفسیر نتایج شرح داده می‌شود. هدف از این مقاله، استفاده از تکنیک‌های متن‌کاوی در حوزه پزشکی است. در نهایت، در یک مرور کلی، کاربردهای اصلی و روش‌های متن‌کاوی در حوزه پزشکی بیان می‌شود و به بررسی نتایج و یافته‌های حاصل از آن می‌پردازیم.

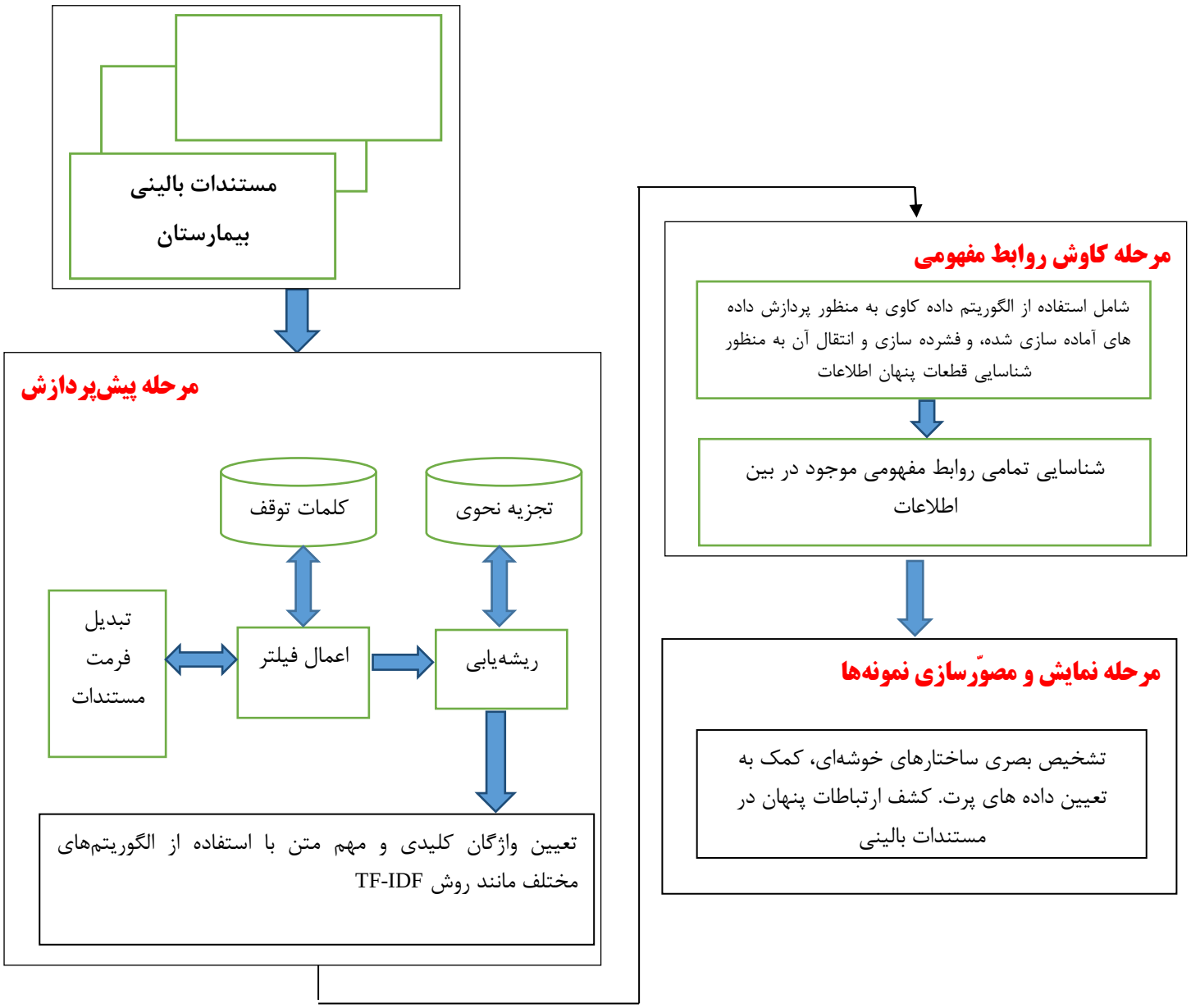
روش

مطالعه حاضر به روش مروری انجام شد. برای تحلیل و ارزیابی پژوهش‌های انجام شده در مورد سیستم‌های متن کاوی پزشکی، رویکردها و تکنیک‌های استفاده شده، پایگاه‌های اطلاعاتی بین‌المللی ScienceDirect، IEEE، PubMedCentral، Google Scholar، SID و Magiran تا سال ۲۰۱۷ به‌طور نظام‌مند مورد جستجو قرار گرفتند. جستجو با استفاده از کلید واژه‌های "Text Mining" AND "Medicine"، "Clinical Text Mining"، "Predict" AND "Knowledge Discovery in Medical Text"، "Text Mining for Medical and Healthcare" در پایگاه‌های اطلاعاتی انگلیسی و ترکیب "متن‌کاوی" و " کشف دانش در پزشکی" برای جستجو در پایگاه‌های فارسی استفاده شد. نوع مقاله در جستجوی پیشرفته، مقاله اصیل پژوهشی انتخاب شده بود و بازه زمانی جستجو نیز را برای مقالات انگلیسی از ۲۰۰۰ تا ۲۰۱۷ و در پایگاه‌های فارسی از ۱۳۸۵ تا ۱۳۹۵ محدود شد. همچنین، یکی از شرایط انتخاب مقالات، دسترس‌پذیر بودن مقالات تمام متن بود که ۹۰ مقاله جمع‌آوری شد. سپس با حذف مقالات تکراری و با استفاده از یک ارزیابی منتقدانه، بر اساس بیشترین ارتباط با موضوع مورد پژوهش، در نهایت ۳۰ مقاله مرتبط با بالاترین میزان تناسب با موضوع مورد پژوهش، برای مطالعه انتخاب شدند و مورد بررسی قرار گرفتند. در ابتدا به بررسی چکیده مقالات پرداخته شد، با بررسی چکیده مقالات انتخابی، همه مقاله‌هایی که به نوعی به بررسی و تحلیل، مرور یا یک نوآوری در مورد سیستم‌های متن کاوی پزشکی پرداخته بودند، در مطالعه شرکت داده شدند. همچنین مقالاتی که به بازخورد مرتبط بودن، ارائه راه‌حل و رویکردی جدید در زمینه متن کاوی متون و مستندات پزشکی پرداخته بودند نیز در این مرحله وارد مطالعه شدند. سپس با استفاده از یک چک لیست بررسی، ارزیابی نهایی روی مقالات انتخابی صورت پذیرفت و به دنبال آن تمام مقالات غیر مرتبط از مطالعه خارج شدند. معیارهای خروج مقالات از مطالعه، این بود که مقالات مورد بررسی قرار گرفته به متن کاوی مستندات غیر بالینی و غیر مرتبط با حوزه پزشکی پرداخته باشند یا به موضوع متن کاوی در آن‌ها پرداخته نشده بود. همچنین مطالعاتی که به تکرار مطالعه دیگر پرداخته بودند و به عبارتی سلیس‌تر، تکراری بودند و آن دسته از مقالاتی که دسترسی به تمام متن مقاله در دسترس نبود، در مطالعه شرکت داده نشدند.

اصول متن کاوی

متن کاوی که به «کشف دانش در متن»، «داده کاوی متنی»، «تحلیل هوشمند متن» مشهور است، به طور کلی به فرآیند استخراج دانش و اطلاعات مورد علاقه و مهم از مجموعه متنی غیرساخت یافته اشاره دارد؛ به عبارت دیگر متن کاوی فرآیند تحلیل طبیعی متن به منظور کشف و ثبت اطلاعات معنایی برای درونداد و ذخیره‌سازی در یک ساختار سازمان‌دهی شده دانش است [۹].

با توسعه مفهوم متن‌کاوی مفاهیم دیگری نیز پا به پای آن رشد کردند بازیابی اطلاعات از مجموعه متون «بازیابی اطلاعات از یک متن»، «کشف دانش از بانک‌های اطلاعاتی» مدیریت دانش در سازمان‌ها، «نمایش (مصوّرسازی) داده‌ها و اطلاعات» این مفاهیم توسط Kostoff و Demarco در سال ۲۰۰۰ در چند مقاله منتشر و سعی شد تا میان این مفاهیم تمایز قایل شوند [۱۰]. (شکل۱)



شکل ۱: معماری متن‌کاوی [۱۱]

بر این اساس چنین برداشت می‌شود که یک سیستم متن کاوی بایستی از چهار مشخصه زیر برخوردار باشد تا بتوان آن را در زمره سیستم‌های متن کاوی دانست در غیر این صورت آن سیستم در مجموعه بازیابی اطلاعات و یا سایر موارد قرار می‌گیرد. این چهار مشخصه عبارت‌اند از:

- باید قادر باشد بر روی مجموعه‌های حجیم و متون مبتنی بر زبان طبیعی کار کند.
- باید از الگوریتم‌های اصولی بیشتر از الگوریتم‌های ابتکاری و نظرهای کاربر استفاده کند.
- باید اطلاعات رخدادها، موجودیت‌ها و .... را از متن استخراج کند.
- باید دانش جدیدی استخراج کند.

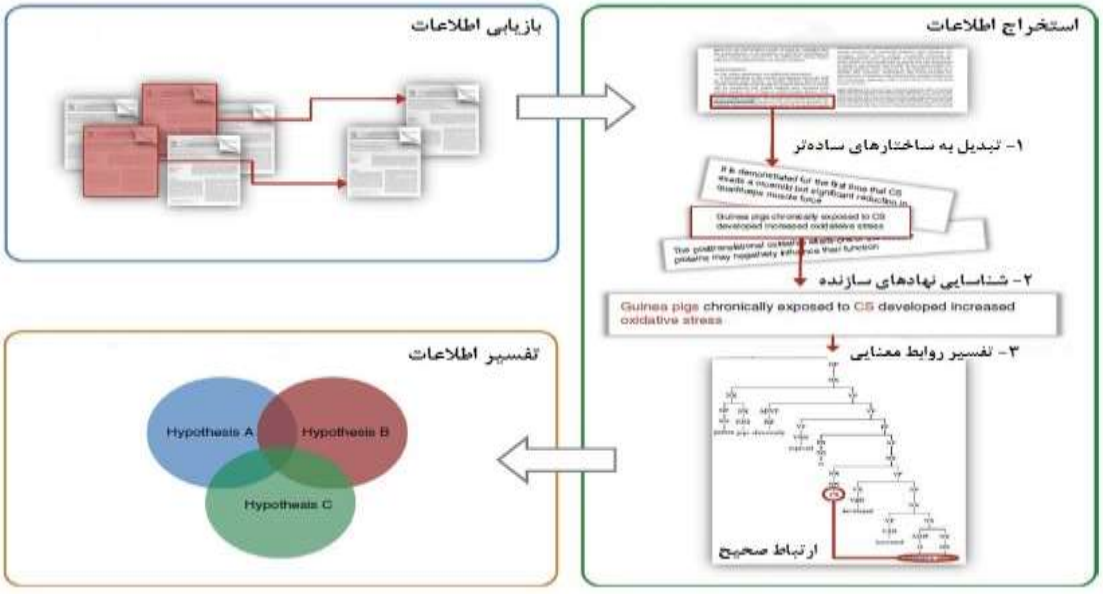
این حوزه تمام فعالیت‌هایی که به نوعی به دنبال کسب دانش از متن هستند را شامل می‌گردد [۱۲]. این تکنیک‌ها در ابتدا در مورد داده‌های ساخت‌یافته به کار گرفته شدند و علمی به نام

داده‌کاوی را به وجود آوردند. داده‌های ساخت‌یافته به داده‌هایی گفته می‌شود که به طور کاملاً مستقل از همدیگر، ولی از لحاظ ساختاری به صورت یکسان در یک محل گردآوری شده‌اند. انواع بانک‌های اطلاعاتی را می‌توان نمونه‌هایی از این دسته اطلاعات نام برد؛ اما در مورد متون که عمدتاً غیرساخت‌یافته یا نیمه‌ساخت‌یافته هستند، ابتدا باید توسط روش‌هایی آن‌ها را ساختارمند نمود و سپس از این روش‌ها برای استخراج اطلاعات و دانش ضمنی نهفته در آن‌ها استفاده کرد. ازجمله مشکلاتی که در زمینه داده‌کاوی وجود دارد کشف دانش مفید از متون نیمه ساخت یافته یا غیر ساخت‌یافته است که توجه زیادی را به خود جلب کرده است. روش‌های داده‌کاوی سنتی فرض بر این دارند که اطلاعات به فرم متنی پایگاه داده‌های رابطه‌ای هستند؛ به همین دلیل برای بسیاری از کاربردها مانند اطلاعات الکترونیکی قابل دسترس به فرم متنی نیمه ساخت‌یافته یا غیر ساخت‌یافته مفید نیستند. متن کاوی، شبیه داده‌کاوی است، به جزء این که ابزارهای داده‌کاوی برای مدیریت داده‌های ساخت‌یافته پایگاه داده‌ها طراحی می‌شود؛ اما متن‌کاوی می‌تواند با داده نیمه ساخت‌یافته یا ساخت نیافته مانند اسناد متون سروکار داشته باشد. در نتیجه متن‌کاوی یک راه‌حل بسیار بهتر برای سازمان‌ها است؛ با این وجود بیشتر تحقیقات و تلاش‌های گسترده روی کوشش‌های داده‌کاوی که داده ساخت‌یافته استفاده می‌کند، متمرکز است [۱۳].

متن کاوی فرآیندی است که به کمک آن، اطلاعات ارزشمند و مفید نهفته در داده‌های غیر ساخت‌یافته، استخراج می‌شود. در متن کاوی تحلیل محتوای داده‌های غیرساخت‌یافته بر اساس آنالیز کمی متن صورت می‌پذیرد [۹].

رویکردهای متن‌کاوی

بر طبق تعریف، هدف از متن‌کاوی، بازیابی، استخراج و تفسیر اطلاعات ذخیره شده در قالب الکترونیکی با استفاده از ابزارهای خودکار و نیمه خودکار است که در آرشیو اسناد و مدارک با حجم اطلاعاتی بالا و در پایگاه داده‌ها ذخیره می‌شوند (شکل ۲) [۴].



شکل ۲: شماتیک کلی روش‌های مورد استفاده در متن‌کاوی و داده‌کاوی: (۱) بازیابی اطلاعات، (۲) استخراج اطلاعات، (۳) تفسیر اطلاعات (ادغام فرضیه‌های مختلف برای تولید فرضیه ترکیبی جدید) [۴]

الف) بازیابی اطلاعات

معمولاً در بازیابی اطلاعات با توجه به نیاز مطرح شده از سوی کاربر، مرتبط‌ترین متون و مستندات و یا در واقع «کیسه کلمات» (Bag of Words) از میان دیگر مستندات یک مجموعه بیرون کشیده می‌شود. بازیابی اطلاعات یافتن دانش نیست، بلکه تنها آن مستنداتی را که مرتبط‌تر به نیاز اطلاعاتی جستجوگر تشخیص داده به او تحویل می‌دهد. این روش در

واقع هیچ دانش و حتی هیچ اطلاعاتی را بـه ارمغـان نمـی‌آورد [۱۴].

ب) استخراج اطلاعات

استخراج اطلاعات عبارت است از نگاشت متن‌های زبان طبیعی به یک نمایش ساخت‌یافته و از پیش تعریف شده که منتخبی از اطلاعات کلیدی از متن اصلی را نشان می‌دهنـد. تکنیک‌هـای استخراج اطلاعات، داده‌هـای سـاختارمند را از داده‌هـای بـدون ساختار استخراج می‌کند. در واقع هدف از این مرحله، شناسـایی بخش‌های مرتبط از اطلاعات بازیابی شده در مرحله قبل است. که این کار در سه مرحله انجام می‌شود:

۱- پردازش: با تبدیل متون پیچیـده و غیـر سـاخت‌یافته بـه ساختارهای ساده (به عنوان مثال کلمات یا جملات کوتـاه) کـه می‌تواند توسط سیستم‌های کامپیوتری تفسیر شود.

۲- شناسایی نهادهـای سـازنده: ایـن مرحلـه بـه منظـور شناسایی موجودیت‌های بالینی یا فرآیندهای بیولـوژیکی اشـاره شده در سند به کار می‌رود.

۳- تفسیر روابط معنایی: در این مرحله تفسیر روابط معنـایی بـین سـاختارهای متنـوع یـا موجودیت‌هـای مختلـف صـورت می‌پذیرد. بدین منظـور بـرای اسـتخراج اطلاعـات، سیسـتم‌ها می‌بایست مبتنی بر الگوریتم‌های هوش مصنوعی باشند [۱۵].

ج) تفسیر نتایج

گام نهایی در فرآیند متن‌کاوی، تفسیر نتایج است. این گـام بـا هدف یکپارچه‌سازی اطلاعات به دست آمده از گام‌های قبل، در نهایـت بـرای بـه دسـت آوردن ارتباطـات معنـادار بـین موجودیت‌های بیماری، پدیده‌های بیولوژیکی یـا بـالینی کـه در ابتدا با استفاده از روش‌های سنتی غیر قابل شناسایی بودند، بـه کار می‌رود [۱۶].

نتایج

بررسـی‌هـای انجـام شـده نشـان داد کـه متن‌کـاوی بـه طـور گسترده‌ای در تحقیقات سرطان مورد استفاده قرار گرفته اسـت؛ با این حال برای بهره‌گیری کامل از متن‌کـاوی، توسـعه روش-های جدید برای کاوش کامل متون و اسناد بـالینی و همچنـین سیستم عامل‌هایی برای ادغـام پایگاه‌هـای اطلاعـات پزشـکی لازم است. به‌رغم دارا بودن پتانسیل بالقوه متن‌کاوی در زیست پزشکی، هنوز هم نیاز بـه توسـعه بیشـتر اسـت و سیسـتم‌های متن‌کـاوی زیسـت پزشـکی، در حـال حاضـر بـه عنـوان ابـزار استاندارد طلایـی محققـان زیسـت پزشـکی بـه‌منظور بازیـابی اطلاعات شناخته نمی‌شوند. از مهم‌ترین بحث‌هـای مـرتبط بـا متن‌کاوی، همکاری و هماهنگی با چنـدین موضـوع اسـت؛ بـه عبارت دیگر متن‌کاوی متـون زیست‌پزشـکی، همـراه بـا سـایر روش‌ها و ابزارها باید نتایج سـازگار، قابـل انـدازه‌گیری و قابـل آزمون با هم داشته باشـند [۱۷]. جـدول ۱ برخـی از تحقیقـات صورت گرفته در زمینه متن‌کـاوی پزشـکی را نشـان می‌دهـد.

جدول ۱: یافته‌های کلیدی از پژوهش‌های انجام شده در متن‌کاوی مستندات بالینی

نویسندگان	موضوع پژوهش	روش‌شناسی پژوهش	یافته‌های کلیدی
Pereira و همکاران [۳]	طبقه‌بندی تشخیص صرع به یک کد استاندارد ICD-9 مبتنی بر تکنیک‌های متن کاوی	از هستان‌شناسی به منظور استخراج و کشف دانش ضمنی از سوابق پرونده سلامت کودکان که منجر به طبقه‌بندی این بیماری و انواع مختلف آن می‌شود، بهره بردند.	ارائه یک سیستم خودکار طبقه‌بندی تشخیص صرع مبتنی بر ICD-9 بر اساس سوابق پرونده سلامت کودکان زیر ۱۶ سال پرتغال پرداختند.
Jonnagaddala و همکاران [۱۸]	شناسایی و پیشرفت عوامل خطر بیماری قلبی در بیماران دیابتی از پرونده الکترونیک سلامت	رویکرد ترکیبی؛ یادگیری ماشین و تکنیک‌های متن کاوی بالینی به منظور استخراج اطلاعات مرتبط در مورد عوامل خطر بیماری‌های قلبی از پرونده الکترونیک سلامت مربوط به ۱۳۰۴ بیمار دیابتی	در این پژوهش، بیماران در یکی از سه گروه: بیمارانی که مبتلا به بیماری عروق کرونری بودند، بیمارانی که به تازگی به این بیماری مبتلا شده بودند و بیمارانی که در بیش از یک دوره زمانی بیماری عروق کرونری نگرفته بودند، طبقه بندی شدند.
Gong و همکاران [۱]	متن کاوی در گزارش‌های رادیولوژی	یک سیستم متن کاوی به منظور استخراج خودکار و استفاده از اطلاعات در گزارش‌های رادیولوژی پیشنهاد کردند.	متن کاوی در گزارش رادیولوژی، امکان بازیابی و استفاده از اطلاعات مفید و دانش از گزارش‌های رادیولوژی، به منظور ارائه اطلاعات مورد نیاز برای متخصصان پزشکی و استفاده از اطلاعات
Jonnagaddala و همکاران[۱۹]	ارزیابی خطر ابتلا به بیماری‌های عروق کرونر از پرونده الکترونیک سلامت، مبتنی بر متن کاوی	پیشنهاد سیستمی مبتنی بر قاعده برای استخراج عوامل خطر فرامینگهم نظیر سن، جنس، میزان کلسترول، فشار خون، دیابت و ..... به منظور ارزیابی ریسک خطر ابتلا به بیماری‌های عروق کرونری در گروهی از بیماران مبتلا به دیابت	خروجی سیستم مذکور، قابل اعتماد بود و همچنین تجزیه و تحلیل حاصله از ارزیابی ریسک عوامل خطر فرامینگهم نشان داد که اکثر بیماران مبتلا به دیابت در معرض خطر ابتلا به بیماری‌های عروق کرونر هستند.
Sumathi و همکاران [۱۱]	پیش‌بینی وقوع بیماری‌های قلبی با استفاده از متن کاوی	پیشنهاد یک روش مؤثر برای استخراج داده‌ها از تعداد زیادی اسناد بالینی با رویکرد متن کاوی	با بهره‌گیری از تکنیک‌های متن‌کاوی و الگوریتم‌های استقرایی توان به نتایج دقیق و کارآمد به منظور پیش‌بینی وقوع بیماری‌ها از اسناد بالینی دست یافت.
Wagland و همکاران [۲۰]	توسعه و آزمایش یک رویکرد مبتنی بر متن کاوی به منظور تجزیه و تحلیل نظرات بیماران از تجارب خود در مراقبت از سرطان روده بزرگ	یک رویکرد متن کاوی مبتنی بر یادگیری ماشین، به منظور تسهیل در فرآیند تجزیه و تحلیل تجارب بیماران و توسعه یک مدل توضیحی	تکنیک‌های متن کاوی مبتنی بر یادگیری، ابزارهای مفید و عملی برای شناسایی نظرات فرمت آزاد خاص، در یک مجموعه بزرگ هستند. همچنین کیفیت مراقبت از سرطان روده بزرگ، تأثیر مستقیمی بر کیفیت سلامت زندگی این بیماران دارد.
Holzinger و همکاران [۲۱]	اطلاعات معنایی در سیستم‌های پزشکی: بهره‌گیری از تکنیک‌های متن کاوی به منظور تجزیه و تحلیل تشخیص‌های پزشکی	به طراحی و توسعه یک سیستم توصیفی- کاربردی به منظور تجزیه و تحلیل نظرات کارشناسان خبره در حوزه تصاویر MRI پرداخته شد	با به کارگیری ابزارهای متن کاوی، همبستگی منطقه ای تصاویر MRI سنجیده شود که به تبع آن، برآورد مهمی از احتمال وقوع همزمان بیماری در نواحی تعریف شده بدن انسان، به منظور شناسایی خطرات احتمالی برای سلامتی دست یافته شود.
Lee و همکاران [۲۲]	متن کاوی از پرونده‌های بالینی به منظور تشخیص سرطان	پیشنهاد خودکار ساختن فرآیند استخراج روابط بین بیماری سرطان و فاکتورهای بالقوه از پرونده‌های بالینی با ادغام آنتولوژی سرطان، تکنیک‌های متن کاوی توسعه یافته، الگوریتم (Self-Organizing Maps) SOM و روش (Support Vector Machines)SVM	ادغام هستان‌شناسی پزشکی و تکنیک‌های متن کاوی، قادر به استخراج الگوهای بالقوه و دسته‌بندی مجدد پرونده‌های بالینی است.

با توجه به افزایش تخصص‌ها و حجم ادبیات تخصصی حوزه پزشکی، محققان در حال تلاش به منظور استخراج حقایق کلی به منظور تولید فرضیه‌هایی قابل آزمون و پذیرفتنی هستند. بر اساس یافته‌های حاصل از پژوهش، مشخص شد، هیچ تکنیک متن‌کاوی منفردی وجود ندارد که برای تمامی انواع داده‌های سیستم‌های اطلاعات بهداشت و درمان نتایج ثابتی ارائه کند. عملکرد تکنیک‌های متن‌کاوی به نوع مجموعه داده‌ای بستگی دارد که برای انجام پژوهش به کار برده شد؛ بنابراین می‌توان از تکنیک متن‌کاوی ترکیبی و حتی ترکیب توأم تکنیک‌های متن‌کاوی و داده‌کاوی برای دستیابی به عملکردی بهتر استفاده شود.

بحث و نتیجه‌گیری

با توجه به حجم بسیار بالای ادبیات علمی که هر ساله تولید می‌شوند، متن‌کاوی برای تحقیقات علمی ضروری است [۲۳]. امروزه بیمارستان‌های مدرن با کامپیوترها و دیگر وسایل جمع‌آوری داده تجهیز شده‌اند که وسایل نسبتاً ارزانی برای جمع‌آوری و ذخیره داده‌ها در سیستم‌های اطلاعاتی درون و

بیرون بیمارستان می‌باشند. مقدار وسیع داده‌های جمع شده در پایگاه داده‌های پزشکی نیاز به ابزارهای تخصصی جهت ارزیابی، نگهداری و تجزیه و تحلیل دارند. به ویژه افزایش مقادیر داده، باعث به وجود آمدن مشکلات زیادی در خارج کردن اطلاعات مفید جهت حمایت از تصمیمات می‌باشد و تنها تجزیه‌وتحلیل داده‌ها به صورت سنتی کافی نیست و نیازمند تجزیه و تحلیل مبتنی بر کامپیوتر می‌باشد. متن‌کاوی در کاربردهای زیست‌پزشکی، بسیاری از تکنیک‌های محاسباتی مانند یادگیری ماشین، پردازش زبان طبیعی، آمارزیستی، فناوری اطلاعات و شناخت الگو را به‌منظور یافتن خروجی‌های تلویحی در متون زیست‌پزشکی غیرساختار یافته به کار می‌گیرد. کاربردهای زیادی از متن‌کاوی متون مربوط به سرطان مانند شناسایی تومورهای بدخیم مرتبط با علوم زیست‌پزشکی (ژن‌ها، پروتئین‌ها و غیره)، یافتن ارتباط بین نهادهای بیولوژیکی (پروتئین- پروتئین، ژن- بیماری و غیره)، استخراج دانش از متون و تولید فرضیه‌های معنی‌دار وجود دارد [۱۷]. اثبات شده است که متن‌کاوی و داده‌کاوی از تکنیک‌های مهم و قدرتمند برای استخراج اطلاعات و بینش از سیستم‌های اطلاعات بهداشتی و درمانی هستند [۲۴-۲۷].

با افزایش استفاده از پرونده‌های پزشکی الکترونیکی در سایه تلاش‌های گسترده داده‌کاوی برای کشف روند در داده‌های سلامت، بخش قابل توجهی از اطلاعات که در پرونده‌های الکترونیک باقی می‌ماند به صورت متن ذخیره می‌شود و غیر قابل استفاده با روش‌های داده‌کاوی معمولی است. این داده‌ها نیمه ساختار یافته یا بدون ساختار شامل یادداشت‌های بالینی، دسته خاصی از نتایج آزمایش‌هایی مانند اکوکاردیوگرام و گزارش رادیولوژی و دیگر اسناد و مدارک مهم است. برای استفاده کامل از EMRs، نیاز به استفاده از هر دو تکنیک‌های داده‌کاوی و متن کاوی برای کشف نتایج از پرونده بیماران است.

طبق یافته‌ها، اکثر برنامه‌های کاربردی متن‌کاوی بالینی در گذشته از یک منبع داده متنی منفرد، مانند گزارش‌های رادیولوژی برای شناسایی یا کاوش اطلاعات مرتبط به یک وضعیت واحد استفاده کرده‌اند. با این حال افزایش ارتباط داده‌ها (به عنوان مثال منابع داده‌ای متعدد از بیمار با یک شناسه منحصر به فرد) در سیستم‌های اطلاعات بیمارستانی، ایجاد فرصت‌هایی برای تکنیک‌های قدرتمند و دقیق متن‌کاوی است که از اطلاعات حاصل از منابع مختلف داده استفاده می‌کنند [۱۱].

متن کاوی در اسناد پزشکی بسیار حائز اهمیت است با این حال در این حوزه با مشکلات متعددی مواجه است. از مهم‌ترین چالش‌های متن کاوی می‌توان موارد ذیل را برشمرد:

- چالش مهم و اصلی ارائه یک چارچوب استخراج دانش پزشکی منحصر به فرد، فقدان مجموعه استانداردهای مناسب برای ارسال و انتقال اطلاعات در سیستم‌های اطلاعات پزشکی است. علیرغم این که استانداردهایی نظیر ICD10 برای یکپارچه کردن اطلاعات بیماری وجود دارد.
- یک چالش مکرر که متن‌کاوی بالینی با آن مواجه است، در مدیریت ماهیت ناهمگون داده‌های بدون ساختار پرونده‌های پزشکی هستند. تشخیص دقیق وضعیت بیماری از متن بالینی، نیازمند درک دقیقی از الگوها و عبارات کلیدی در سوابق پزشکی است که به‌طور گسترده‌ای متفاوت می‌باشند.
- با توجه به این که داده‌های پزشکی شامل اطلاعات شخصی می‌باشند؛ در معرض سوء استفاده‌اند، مانع اصلی و عمده برای متن‌کاوی پزشکی محرمانه بودن اسناد و استفاده اخلاقی از اطلاعات بیماران است که محدودیت‌های قانونی و اخلاقی باید در مورد آن‌ها رعایت شود. بدین منظور قبل از شروع فرآیند متن‌کاوی، مؤسسات بهداشت درمانی باید سیاست مشخصی را مربوط به محرمانگی و امنیت رکوردهای اطلاعاتی بیماران تنظیم کنند.

با این که متن‌کاوی از مستندات بالینی، به کشف دانش تخصصی نهفته پزشکی منجر می‌شود، با این‌حال، ماهیت غیرساختاری پرونده‌های الکترونیک سلامت، به خودی خود متن‌کاوی بالینی را با چالش‌های بسیاری روبه‌رو می‌کند.

References

1. Gong T, Tan CL, Leong TY, Lee CK, Pang BC. Text mining in radiology reports. Proceedings: Eighth IEEE International Conference on Data Mining; 2008 Dec 15-19; Los Alamitos, CA: IEEE Computer Society; 2008. p. 815-20.

2. Patel R, Jayatilleke N, Jackson R, Stewart R, Mcguire P. Investigation of negative symptoms in schizophrenia with a machine learning text-mining approach. The Lancet 2014; 383: S16.

3. Pereira L, Rijo R, Silva C, Agostinho M. ICD9-based text mining approach to children epilepsy classification. Procedia Technology, 2013;9:1351-60.

4. Piedra, D., & Ferrer, A. Text Mining and Medicine : Usefulness in Respiratory Diseases. Archivos *de Bronconeumología*, 2014 ;*50*(3), 113–119.

5. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17(5):507-13.

6. Lucini FR, Fogliatto FS, da Silveira GJ, Neyeloff JL, Anzanello MJ, Kuchenbecker RD, et al. Text mining approach to predict hospital admissions using early medical records from the emergency department. International Journal of Medical Informatics 2017;100:1-8.

7. Allahyari M, Pouriyeh S, Assefi M, Safaei S, Trippe ED, Gutierrez JB, et al. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. Halifax, Canada: KDD Bigdas; 2017.

8. Casillas A, Gojenola K, Perez A, Oronoz M. Clinical text mining for efficient extraction of drug-allergy reactions. IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2016 Dec 15-18; Shenzhen, China: IEEE; 2016. p. 946–52.

9. Pons-Porrata A, Berlanga-Llavori R, Ruiz-Shulcloper J. Topic discovery based on text mining techniques. Information Processing & Management 2007; 43(3): 752-68.

10. Ramezani H, Alipour Hafezi M, Momeni E. Scientific maps: methods and techniques. Journal of the Popularization of Science 2014; 5(6): 53-84. Persian

11. Sumathi S, Mohanapriya S, Nagasandhiyalakshmi B, Shanmugapriya N. Prediction of outbreak of heart disease using text mining. Discovery 2016; 52(245): 1070-7.

12. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Research Synthesis Methods 2011; 2(1):1-14.

13. Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence 2009; 1(1): 60-76.

14. Ananiadou S, McNaught J. Text Mining for Biology and Biomedicine. Boston, London: Artech House; 2006.

15. Hatzivassiloglou V, Duboue PA, Rzhetsky A. Disambiguating proteins, genes, and RNA in text: a machine learning approach. Bioinformatics 2001;17 Suppl 1:S97-106.

16. Swanson DR. Migraine and magnesium: eleven neglected connections. Perspect Biol Med 1988;31(4):526-57.

17. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46(2):200-11.

18. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Dai HJ, Hsu CY. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. Biomed Res Int 2015;2015:636371.

19. Jonnagaddala J, Liaw ST, Ray P, Kumar M, Chang NW, Dai HJ. Coronary artery disease risk assessment from unstructured electronic health records using text mining. J Biomed Inform 2015;58 Suppl:S203-10.

20. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Qual Saf 2016;25(8):604-14.

21. Holzinger A, Geierhofer R, Mödritscher F. Semantic information in medical information systems: utilization of text mining techniques to analyze medical diagnoses. Journal of Universal Computer Science 2008; 14(22):3781-95.

22. Lee CH, Wu CH, Yang HC. Text mining of clinical records for cancer diagnosis. Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007); 2007 Sep 5-7; Kumamoto, Japan: IEEE; 2007.

23. Gutierrez JB, Galinski MR, Cantrell S, Voit EO. From within host dynamics to the epidemiology of infectious disease: Scientific overview and challenges. Math Biosci 2015;270(Pt B):143-55.

24. Kocbek S, Cavedon L, Martinez D, Bain C, Manus CM, Haffari G, et al. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. J Biomed Inform 2016;64:158-67.

25. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010;17(4):440-5.

26. Nguyen A, Moore J, Zuccon G, Lawley M, Colquist S. Classification of pathology reports for cancer registry notifications. Stud Health Technol Inform 2012;178:150-6.

27. Coden A, Savova G, Sominsky I, Tanenblatt M, Masanz J, Schuler K, et al. Automatically extracting cancer disease characteristics from pathology reports into a Disease Knowledge Representation Model. J Biomed Inform 2009;42(5):937-49.

Review Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 5(2): 314-323

Review of the Approaches and Techniques of Text Mining in Clinical Documentation

Feizmanesh Farzaneh[1], Safaei Ali Asghar[2]*

• **Received:** 8 Mar, 2017 • **Accepted:** 26 Nov, 2017

Introduction: With the extensive use of electronic medical records systems, a large amount of medical text data is produced daily in the hospitals and other medical environments that organizing this text information is important and necessary. Also, a need to automatically retrieve useful knowledge from this data to help clinicians is felt. In order to extract the hidden values in the medical text documents, text mining can be used in the field of health.

Methods: In this review study, SID, Magiran, Pubmed, ScienceDirect, IEEE, and Google Scholar databases were searched with the keywords including "Text Mining" AND "Medicine" ،"Clinical Text Mining" AND "Predict"، "knowledge discovery in medical text" and "Text Mining for Medical and Healthcare" in the English databases and keywords such as "text mining" AND " Discovering Knowledge in Medicine" in the Persian databases. Then, all articles that somehow refer to Medical knowledge discovery and text mining applications in the field of health were selected.

Results: Text mining is one of the important and powerful techniques for extracting information from health information systems. Text mining in clinical data provides potential for new discoveries and it also improves efficiency and communication in hospital systems for doctors and hospital administrators.

Conclusion: Nowadays, text mining in clinical documentation, is one of the developed technologies for the discovery of medical knowledge that its use in medical databases is essential to achieving immediate access to important health resources, and its application can improve patient care and reduce medical costs.

Keywords: Text mining in clinical documentation, Text mining applications, Medical knowledge Discovery, Information extraction

• **Citation:** Feizmanesh F, Safaei AA. Review of the Approaches and Techniques of Text Mining in Clinical Documentation. Journal of Health and Biomedical Informatics 2018; 5(2): 314-323.

1. MSc in Medical Informatics, Medical Informatics Dept., Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran
2. Assistant Professor, Computer engineering-software, Medical Informatics Dept., Faculty of Medical Sciences, Tarbiat Modares University Tehran, Tehran, Iran

***Correspondence**: Medical Informatics Dept., Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran.
• **Tel:** 021- 82884581 • **Email:** aa.safaei@modares.ac.ir