فصلنامه رویکردهای نوین در تحقیقات علوم پایه، فنی و مهندسی
سال پنجم، شماره ۱۸، پاییز ۱۴۰۱
ISSN: ۲۶۴۵-۵۴۳۹
www.jonarbset.ir

سازمان دهی معنایی اسناد پزشکی با استفاده از رویکرد مبتنی بر هستان شناسی

سینا دامی[۱]، کامران تاجیک[۲]

[۱] استادیار، گروه مهندسی کامپیوتر، واحد تهران غرب، دانشگاه آزاد اسلامی، تهران، ایران(نویسنده مسئول)
[۲] دانش‌آموخته کارشناسی ارشد، گروه مهندسی کامپیوتر، واحد تهران غرب، دانشگاه آزاد اسلامی، تهران، ایران

چکیده

در زمینه سازمان‌دهی متون پزشکی، فعالیت‌های پژوهشی محدودی انجام شده که هرکدام با رویکردهای متفاوتی سعی در بهبود سازماندهی دسترسی به این متون را داشته‌اند. با این وجود یک شکاف مهم درپژوهش‌های انجام شده در این حوزه ملاحظه شده و آن این است که بهره‌گیری چندانی از تحلیل‌های معنایی برای سازماندهی اسناد پزشکی صورت نگرفته است. در این پژوهش به منظور رفع این شکاف تحقیقاتی به ارایه روشی برای سازماندهی معنایی اسناد پزشکی با استفاده از رویکرد معنایی مبتنی بر هستان‌شناسی پرداخته شد که توانست سازماندهی اسناد پیچیده پزشکی را با تکنیک‌های معنایی بطور موثرتر و کارآمدتر به انجام برساند. برای بخش داده‌های متنی مورد نیاز این پژوهش از مجموعه داده متنی پزشکی MeDAL استفاده شده است. مجموعه داده متنی پزشکی MeDAL، یک مجموعه داده آموزشی بزرگ برای استفاده در پردازش زبان طبیعی است که برای درک زبان طبیعی پیش از آموزش در حوزه پزشکی طراحی شده است. برای سنجش نتایج حاصله برحسب متریکهای ارزیابی از مقایسه روش پیشنهادی با سه روش دیگر مبتنی بر خوشه‌بندی FCM بدون آنتولوژی و الگوریتم خوشه‌بندی EM با/ بدون آنتولوژی استفاده شده است. نتایج تجربی نشان داد که روش پیشنهادی در بیشتر حالت‌های آزمون با متریک‌های صحت، دقت و معیار-F از سایر روش‌های مورد مقایسه بهتر عمل کرده و تنها در بخش کوچکی از آزمایشات، همپای سایر روش‌ها یا اندکی ضعیف‌تر عمل کرده است؛ اما بطور میانگین مشاهده شد که روش پیشنهادی دارای برتری قابل ملاحظه‌ای نسبت به روش‌های مورد مقایسه بوه است.

واژه‌های کلیدی: پردازش زبان طبیعی، زبان شناسی رایانشی، خوشه‌بندی اسناد پزشکی، تحلیل معنایی متن، هستان‌شناسی

۱. مقدمه

امروزه با وجود حجم فراوان از بیماری‌های مختلف که در سراسر جهان وجود دارد نیاز است کادر پزشکی و متخصصان حوزه سلامت بتوانند به اسناد پزشکی موردنظرشان بطور کارآمدی دسترسی داشته باشند و این نیازمند سازماندهی موثری از اسناد پزشکی است .(Mohammed et al., 2021) حوزه‌های دانشی مرتبط با رشته‌های پزشکی از آنجا که با جان مردم سروکار دارند، جزء مهم‌ترین حوزه‌های تخصصی بوده و به تبع آن اسناد پزشکی نیز دارای جایگاه ویژه‌ای بین اسناد تخصصی رشته‌های مختلف می‌باشد. استفاده کارآمد از این اسناد برای کادر پزشکی دارای اهمیت حیاتی است (Kolling et al., 2021) و هرگونه بهبود در سازماندهی بهتر این اسناد، باعث خواهد شد متخصصان حوزه‌های مختلف پزشکی بتوانند بطور موثرتری به محتوای این اسناد که عموما محتوای دانشی است، دسترسی پیدا کرده و با آنها کار کنند. این امر باعث افزایش کارایی کادر پزشکی و نهایتاً بهبود خدمات‌رسانی به بیماران است.(Mallick et al., 2021)

با توجه به پیچیدگی‌هایی که اسناد پزشکی دارند نیاز است تا مفاهیم معنایی به نحو بکار گرفته شوند تا سازماندهی اسناد پزشکی به نحوه موثرتری انجام شده و به تبع آن دسترسی به آنها نیز ساده‌تر و در عین حال موثرتر گردد. خوشه‌بندی معنایی یک تکنیک برای توسعه‌ی واژگان کلیدی است. این تکنیک با تقسیم‌بندی نقاط یک مجموعه داده به گروه‌های مجزا (خوشه‌ها) به گونه‌ای عمل می‌کند که دو نقطه از یک خوشه به لحاظ معنایی مشابه یکدیگر بوده، اما دو نقطه از خوشه‌های مجزا با یکدیگر متفاوت باشند. این تکنیک می‌تواند برای سازماندهی اسناد پزشکی در قالب خوشه‌های معنایی، بسیار مفید باشد .(Tang et al., 2021)

در زمینه سازماندهی متون پزشکی، فعالیت‌های پژوهشی محدودی انجام شده که هرکدام با رویکردهای متفاوتی سعی در بهبود سازماندهی دسترسی به این متون را داشته‌اند ولی یک شکاف مهم درپژوهش‌های انجام شده در این حوزه ملاحظه شده و آن این است که بهره‌گیری چندانی از تحلیل‌های معنایی برای سازماندهی اسناد پزشکی انجام نگرفته است. در این مقاله به منظور رفع این شکاف تحقیقاتی به ارایه روشی برای سازماندهی معنایی اسناد پزشکی با استفاده از رویکرد مبتنی بر هستان‌شناسی پرداخته شده که می‌تواند سازماندهی اسناد پیچیده پزشکی را با تکنیک‌های معنایی بطور موثرتر و کارآمدتر به انجام برساند.

۲. کارهای مرتبط

Hemanjaliو همکاران (Hemanjali et al., 2021) پژوهشی برای سازماندهی اسناد پزشکی مرتبط با کووید-۱۹ با بهره‌گیری از تکنینک خوشه‌بندی انجام دادند. آنها از خوشه‌بندی k-menas استفاده کردند و با استفاده از برچسب‌گذاری خوشه‌ها و بصری‌سازی نتایج، سعی کردند تا نتایج را ملموس تر سازند Pandey .و همکاران در ۲۰۲۱ پژوهشی برای طبقه بندی سوابق پزشکی بیماران ارایه کردند. آنها از الگوریتم کلونی زنبور عسل فازی استفاه کردند. نتایج آنها از الگوریتم LDA نیز بهتر شد. برای سنجش نتایج هم از K-Fold به ازای K=5 استفاده کردند .(Pandeyet al., 2021)

یکی از حوزه‌هایی که در آن داده کاوی متون به شدت مورد استفاده قرار می‌گیرد، علوم زیست پزشکی است. کوهن و هانتر (Bretonnel et al., 2008)ادبیات پزشکی را در حال رشد می‌دانند، نشان می‌دهد که رشد در نشریات /MEDLINE PubMedپدیده‌ای است که باعث می‌شود دانشمندان علوم پزشکی برای جذب نشریات جدید و انتشار نشریات مربوطه در حوزه تحقیقاتی خود اقدام کنند.

سیستم UMLS (Liyang et al., 2011) جامع ترین دانشی است که بیش از ۱۰۰ واژه نامه، اصطلاحات و هستی‌شناسی را در متازاروس متحد می‌کند و توسط کتابخانه ملی پزشکی طراحی و نگهداری می‌شود. این مکانیسم برای یکپارچه‌سازی کلیه واژگان کلیدی زیست پزشکی مانند مش، طبق شرایط بالینی پزشکی نامگذاری سیستماتیک اصطلاحات بالینی پزشکی (SNOMED CT)، هستی‌شناسی ژن (GO) و غیره فراهم شده است.

جدا از هستی‌شناسی و منابع دانش که در بالا ذکر شد، هستی‌شناسی‌های گوناگونی به‌طور خاص در حوزه‌های زیست‌پزشکی مورد توجه قرار گرفته‌اند. به عنوان مثال، پایگاه دانش فارماکوگنومیسک شامل اطلاعات بالینی حاوی دستورالعمل‌های دوز و برچسب‌های دارویی، به طور بالقوه مرتبط با داروهای ژن دارویی قابل درمان و روابط ژنتیکی و فنوتیپ است (Olivier et al., 2004). هستی‌شناسی‌ها و پایگاه‌های اطلاعاتی که قبلا توصیف شده‌اند به طور وسیع توسط تکنیک‌های متفاوتی از متن استخراج شده از جمله استخراج اطلاعات و خوشه‌بندی در حوزه‌پزشکی استفاده می‌شود.(Mc Cray et al., 1993)

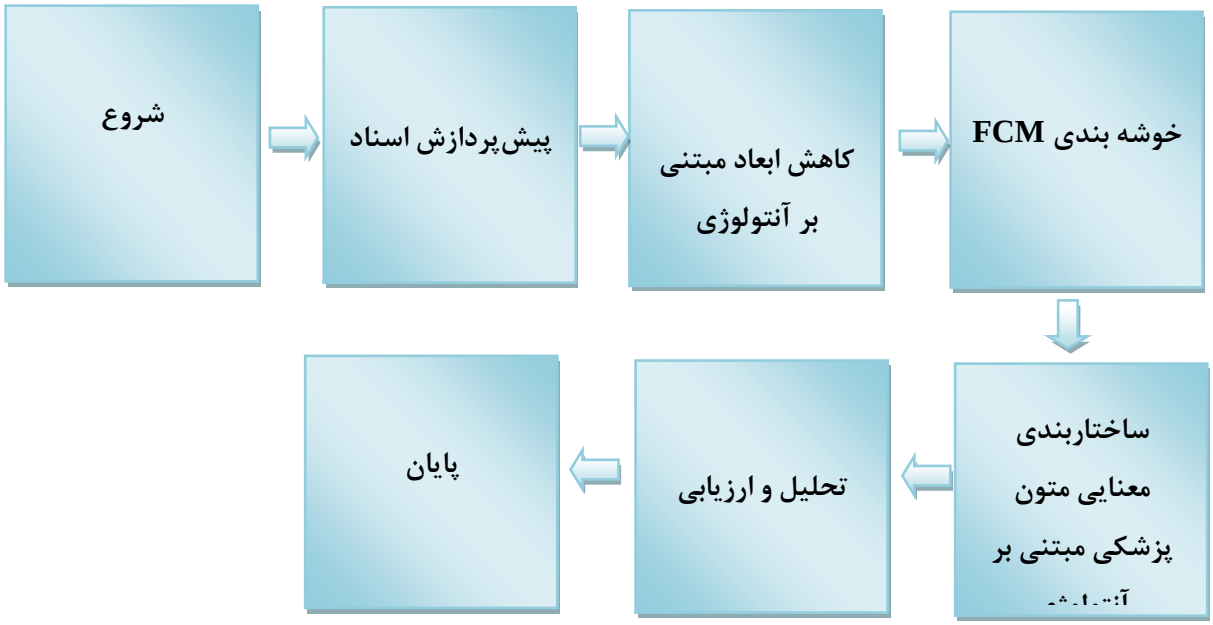
یکی از مأموریت‌های متداول استخراج متون پزشکی زیست‌پزشکی که عمدتا از وظایف استخراج اطلاعات بهره می‌برد، خلاصه‌سازی است. خلاصه سازی، وظیفه شناسایی جنبه‌های مهم یک یا چند اسناد است و به طور خودکار آنها را بصورت یکپارچه نشان می‌دهد. به تازگی توجه زیادی را به دلیل رشد عظیم اطلاعات غیرواقعی در زمینه زیست پزشکی از قبیل مقالات علمی و اطلاعات بالینی به دست آورده است.(Illhoi et al., 2008)

خلاصه‌سازی پزشکی اغلب کاربردی است و ممکن است برای اهداف متفاوت استفاده شود. بر اساس اهداف خود، خلاصه‌ای از خلاصه‌های اسناد مختلف را می‌توان ایجاد کرد، مانند خلاصه‌سازی تک‌سندی که محتوای محتویات اسناد فردی و خلاصه‌های چند سند را در بر می‌گیرد که درآن محتویات اطلاعات چندین اسناد در نظر گرفته می‌شود.

ارزیابی روش‌های خلاصه‌سازی در حوزه پزشکی بسیار دشوار است. از آنجا که تصمیم‌گیری درباره اینکه آیا یک خلاصه "خوب" است یا خیر، اغلب ذهنی است و همچنین ارزیابی دستی از خلاصه‌ها بسیار دشوار است؛ تکنیک ارزیابی خودکار برای خلاصه‌سازی وجود دارد که به اختصار ROUGE نامیده می‌شود. ROUGE کیفیت یک خلاصه‌ی به صورت خودکار تولید شده را در مقایسه با خلاصه‌ی ایده‌آل که توسط انسان‌ها ایجاد می‌شود، اندازه‌گیری می‌کند. این اندازه‌گیری با شمارش لغات هم‌پوشانی بین خلاصه‌ی تولیدشده توسط کامپیوتر و خلاصه‌ی ایده‌آل انسانی تولید شده محاسبه می‌شود. برای یک مرور جامع از تکنیک‌های مختلف خلاصه‌سازی زیست پزشکی به (Aggarwal et al., 2012) مراجعه کنید.

۳. روش پیشنهادی

نمودار رسم شده در شکل ۱، گام‌های‌های انجام پژوهش را نشان می‌دهد. در بخش پیش‌پردازش، ابتدا داده‌های ورودی به سیستم، مورد پردازش اولیه قرار می‌گیرند و پس از نرمال‌سازی مقادیر ویژگی‌ها، برای اینکه دقت محاسبات افزایش و زمان محاسبات نیز کاهش یابد، اقدام به کاهش تعداد ابعاد داده‌های ورودی با یک شبکه باور عمیق (DBN) خواهیم نمود تا ویژگی‌هایی که اضافی هستند و در محاسبات موردنیاز نیستند محذوف شوند و تنها ویژگی‌هایی از داده‌های ورودی باقی بمانند که در انجام محاسبات موثر هستند. در ادامه با استفاده از الگوریتم FCM اقدام به خوشه‌بندی اسناد می‌شود. سپس با رویکرد مبتنی به آنتولوژی اقدام به تعیین برچسب معنایی خوشه‌ها خواهیم نمود که هدف نهایی این سیستم است. در پایان با توجه به خروجی‌های سیستم، ارزیابی می‌کنیم که سازماندهی اسناد پزشکی به چه میزان بطور صحیح انجام شده است و دقت را بدست خواهیم آورد.



شکل ۱ – فرآیند انجام پژوهش

بعد از بدست آمدن خوشه‌ها، ابتدا مطابق با رابطه (۱) برای هر خوشه، مرکز خوشه را محاسبه می‌کنیم:

$$CC_k = (\sum_{i=1}^{n_k} X_{i,1} \ , \sum_{i=1}^{n_k} X_{i,2} \ , \dots \ , \sum_{i=1}^{n_k} X_{i,m}) \quad (1)$$

رابطه (۱)، نشان دهنده مرکز محاسبه شده برای هر خوشه است. نمادهای استفاده در رابطه فوق به شرح مندرج در جدول ۱ هستند.

جدول ۱- شرح نمادهای رابطه

نماد	شرح
CC_k	مرکز خوشه ام K
n_k	تعداد اعضای خوشه ام K
X	ماتریس TF-IDF اسناد متنی
m	تعداد ستونهای ماتریس

حال با محاسبه فاصله بین تک تک واژگان در گراف هستان‌شناسی به ماتریس زیر می‌رسیم:

OntologyMinLengthMatrix =

$$\begin{pmatrix} o_{1,1}/(\sum_{k=1}^{m} o_{1,k}) & \cdots & \cdots & \cdots & o_{1,m}/(\sum_{k=1}^{m} o_{1,k}) \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & o_{i,j}/(\sum_{k=1}^{m} o_{i,k}) & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ o_{n,1}/(\sum_{k=1}^{m} o_{n,k}) & \cdots & \cdots & \cdots & o_{n,m}/(\sum_{k=1}^{m} o_{n,k}) \end{pmatrix}$$

شکل ۲– ماتریس فاصله واژگان در آنتولوژی

در ماتریس فوق مقدار $o_{i,j}$ برابر با میزان فاصله نرمال شده بین تک تک واژگان در گراف هستان‌شناسی است.

ادامه کار به شرح گام‌های زیر است:

۱- وزن هر واژه در داخل گراف هستان‌شناسی به نسبت تعداد کل واژگانی که به فاصله NLen (حداکثر فاصله نودهای همجوار به نود مفروض) با توجه به ماتریس شکل قبل محاسبه می‌شود. پارامتر NLen می‌تواند با استفاده آزمایشات عملی ویا فرضیات هیوریستیک تعیین شود.

۲- برای هرخوشه تعدادی از نزدیکترین اسناد به مرکز خوشه به نسبت تعداد اسناد داخل خوشه به تعداد کل اسناد منتخب می‌شود.

۳- سپس در اسناد منتخب برای هرخوشه پر وزن‌ترین واژگان (با با توجه به خروجی گام ۱) که پرتکرار هم باشند منتخب می شوند.

۴- اجتماع واژگان در اسناد منتخب گام ۴ محاسبه می‌شود به اینصورت برای هر واژه، مقدار حاصل‌ضرب فرکانس حضور آن واژه در سند در وزن آنتولوژیک آن واژه بعنوان وزن نهایی آن واژه محاسبه میشود.

۵- برای مجموعه حاصله از گام ۴ که برای هر خوشه بطور جداگانه محاسبه شده، تعدادی از پر وزن‌ترین واژگان، منتخب شده بعنوان توصیف‌گر معنایی هر خوشه لحاظ می‌گردد.

۴. نتایج تجربی

د برای بخش داده‌های متنی موردنیاز این پژوهش از مجموعه داده متنی پزشکی MeDAL استفاده شده است. مجموعه داده متنی پزشکی MeDAL، یک مجموعه داده آموزشی بزرگ برای استفاده در پردازش زبان طبیعی است که برای درک زبان طبیعی پیش از آموزش در حوزه پزشکی طراحی شده است. این مجموعه داده‌ها در کارگاه ClinicalNLP در EMNLP منتشر شد. این داده‌ها به صورت عمومی قابل دسترسی هستند [۱].

در این پژوهش، از یک هستان‌شناسی برای علم پزشکی عمومی (OGMS) نیز بهره گرفته شده است. این هستان‌شناسی برای مطالعات بالینی، درمان بیماری‌ها، تشخیص سرطان و سایر واحدهای آسیب‌شناسی توسعه داده شد. OGMS سعی می‌کند به برخی از مسائل مطرح شده در کارگاه هستی‌شناسی بیماری‌ها (دالاس، تگزاس) و کارگاه علائم، نشانه‌ها و یافته‌ها

[۱] https://www.kaggle.com/general/99783

[۲] Ontology for General Medical Science

(میلان، ایتالیا) رسیدگی کند. هستان‌شناسی OGMS سابقا هستان‌شناسی فنوتیپ بالینی نامیده می‌شد. این هستان‌شناسی نیز به صورت عمومی قابل دسترسی می‌باشد [۳].

از داده‌های MeDAL به تعداد ۵ زیرمجموعه داده مطابق جدول ۲ برای این پژوهش منتخب شدند.

جدول ۲- مجموعه داده‌های مورد استفاده

کد اختصاری	نام مجموعه داده
A	Allergies
B	Colds and Flu.
C	Conjunctivitis
D	Diarrhea
E	Headaches

برای شبیه‌سازی روش پیشنهادی از نرم افزار متلب استفاده شد. همچنین از متریک‌های صحت، دقت و معیار F- برای ارزیابی و تحلیل مقایسه‌ای عملکرد روش پیشنهادی در مقایسه با روش‌های پایه استفاده کردیم.

برای سنجش نتایج حاصله برحسب متری‌کهای ارزیابی از مقایسه روش پیشنهادی با سه روش دیگر مبتنی بر خوشه‌بندی FCM بدون آنتولوژی، خوشه‌بندی EM[۴] با/ بدون آنتولوژی استفاده شده است. نتایج ارزیابی برای مقایسه دقت روش پیشنهادی در حالتهای با/ بدون هستان‌شناسی مطابق جدول ۳ است.

جدول ۳- مقایسه دقت روش پیشنهادی با حالت بدون آنتولوژی

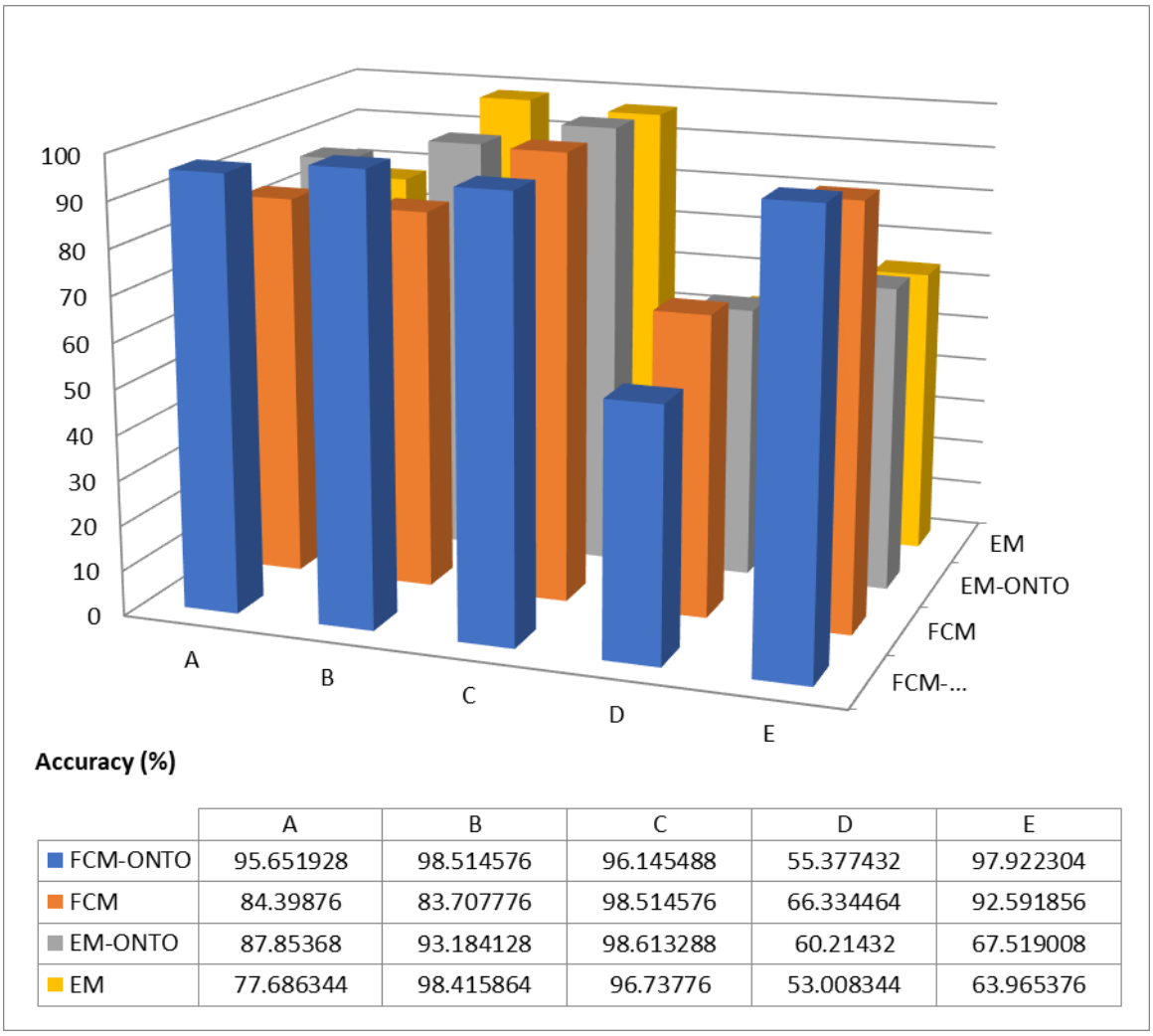
مجموعه داده	FCM + Ontology	FCM
A	۹۷.۲۳۳۳	۸۹.۲۲۶۳
B	۹۴.۰۹۸۹	۹۱.۹۸۹۲
C	۸۳.۲۳۲۶	۸۹.۹۹۲۳
D	۹۴.۹۲۷۵	۸۲.۸۲۸۵
E	۹۸.۸۹۶۵	۷۸.۸۹۹۲
میانگین	***۹۳.۶۷۷۷۶***	***۸۶.۵۸۷۱***

[۳] https://old.datahub.io/dataset/bioportal-ogms

[۴] Naïve bayzian

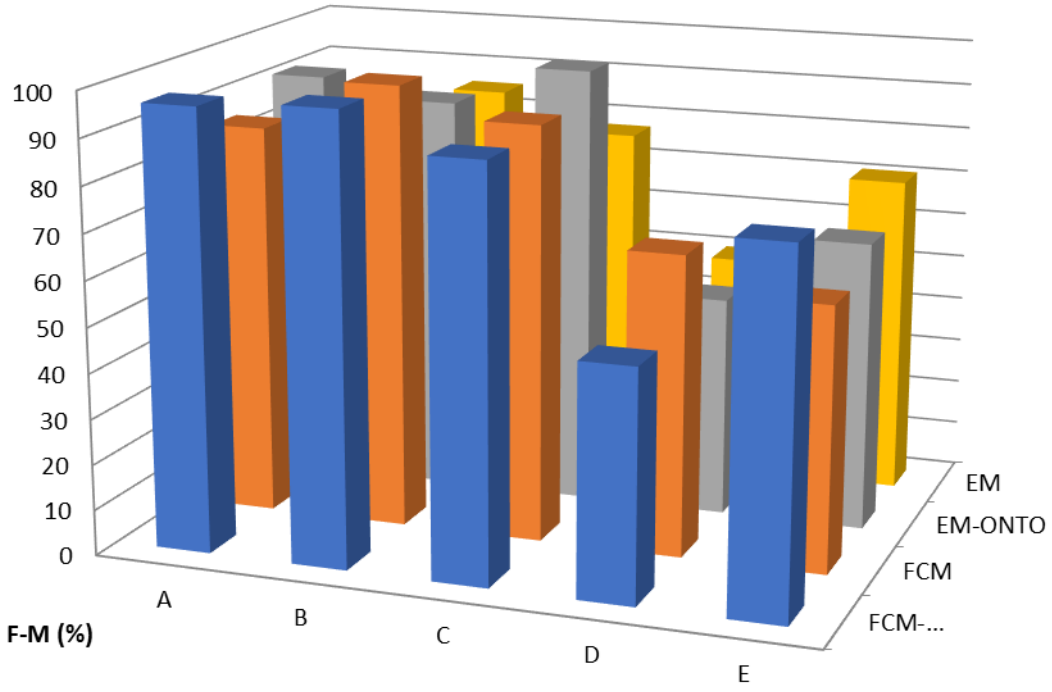
همان‌گونه که در جدول فوق قابل مشاهده است در بیشتر موارد، روش پیشنهادی در حالت با استفاده از آنتولوژی بهتر از بدون آنتولوژی عمل کرده است. این امر به ویژه در مورد مجموعه داده‌های E و A به وضوح مشهود است. تنها موردی که روش پیشنهادی از روش بدون آنتولوژی ضعیف‌تر عمل کرده در مورد مجموعه داده‌های C است. در کل همان‌طور که میانگین دقت هم نشان می‌دهد روش پیشنهادی با آنتولوژی به میزان قابل‌توجهی از روش بدون آنتولوژی بهتر عمل کرده است. شکل ۳، مقایسه صحت روش پیشنهادی را با سایر روش‌های مورد مقایسه نشان می‌دهد.

همچنین، نمودار مقایسه‌ای روش پیشنهادی با دو روش مورد مقایسه از نظر معیار F- نیز در شکل ۴ نشان داده شده است. آنچه مشهود است این است که روش پیشنهادی به‌طور میانگینم دارای عملکرد بهتری نسبت به سایر روش‌ها می‌باشد.



	A	B	C	D	E
FCM-ONTO	95.651928	98.514576	96.145488	55.377432	97.922304
FCM	84.39876	83.707776	98.514576	66.334464	92.591856
EM-ONTO	87.85368	93.184128	98.613288	60.21432	67.519008
EM	77.686344	98.415864	96.73776	53.008344	63.965376

شکل ۳ – مقایسه صحت برای روش‌های مورد مقایسه



	A	B	C	D	E
FCM-ONTO	96.446509	97.927264	89.733753	50.444387	78.085147
FCM	86.179941	97.433679	91.017074	65.903405	58.045596
EM-ONTO	92.201678	88.227338	97.236245	48.568764	63.672465
EM	81.720923	85.390205	77.236245	50.049519	70.780089

شکل ۴ – مقایسه معیار F برای روش‌های مورد مقایسه

۵. نتیجه‌گیری و پیشنهادات

خوشه‌بندی معنایی یک تکنیک برای توسعه‌ی واژگان کلیدی است. خوشه‌بندی معنایی داده‌ها با تقسیم‌بندی نقاط یک مجموعه داده به خوشه‌ها به گونه‌ای عمل می‌کند که دو نقطه از یک خوشه به لحاظ معنایی مشابه یکدیگر بوده، اما دو نقطه از خوشه‌های مجزا با یکدیگر متفاوت باشند. این تکنیک می‌تواند برای سازماندهی اسناد پزشکی در قابل خوشه‌های معنایی، بسیار مفید باشد. در زمینه سازمان‌دهی متون پزشکی، فعالیت‌های پژوهشی محدودی انجام شده که هرکدام با رویکردهای متفاوتی سعی در بهبود سازماندهی دسترسی به این متون را داشته‌اند ولی یک شکاف مهم درپژوهش‌های انجام شده در این حوزه ملاحظه شده این است که بهره‌گیری چندانی از تحلیل‌های معنایی برای سازماندهی اسناد پزشکی انجام نگرفته است. در این پژوهش به منظور رفع این شکاف تحقیقاتی به ارایه روشی برای سازماندهی معنایی اسناد پزشکی با استفاده از رویکرد مبتنی بر هستان‌شناسی پرداخته شد که توانست سازماندهی اسناد پیچیده پزشکی را با تکنیک‌های معنایی بطور موثرتر و کارآمدتر به انجام برساند.

برای ارزیابی روش پیشنهادی از مجموعه داده متنی پزشکی MeDAL استفاده شده است. این مجموعه داده، یک مجموعه‌ی آموزشی بزرگ برای استفاده در کاربردهای پردازش زبان طبیعی است که برای درک زبان طبیعی پیش از آموزش در حوزه

پزشکی طراحی شده است. این مجموعه داده در کارگاه ClinicalNLP در EMNLP منتشر شد. ۵ زیرمجموعه از این مجموعه داده‌ها برای انجام این پژوهش انتخاب شد.

برای سنجش نتایج حاصله برحسب معیارهای ارزیابی از مقایسه روش پیشنهادی با سه روش دیگر مبتنی بر خوشه‌بندی FCM بدون آنتولوژی و الگوریتم خوشه‌بندی EM با/ بدون آنتولوژی استفاده شده است. روش پیشنهادی با معیار صحت برای مجموعه داده‌های A و E به میزان قابل توجهی از سایر روش‌های مورد مقایسه بهتر عمل کرده است. در مجموعه داده‌های B و C نیز تقریباً همپای سایر روش‌ها بوده و تنها در مجموعه داده D است به میزان جزیی ضعیف‌تر عمل کرده است. در مجموع به طور میانگین مشاهده شد که روش پیشنهادی دارای صحت و دقت بهتری نسبت به سایر روش‌ها است.

روش پیشنهادی با معیار F در مجموعه داده‌های A، D و E دارای عملکرد بهتری نسبت به سایر روش‌های مورد مقایسه است. در مورد مجموعه داده B، روش پیشنهادی بطور مشترک با روش FCM از بقیه بهتر عمل کرده‌اند. تنها در مورد مجموعه داده C است که تا حدی نسبت به سایر روش‌های مورد مقایسه ضعیف‌تر عمل کرده است. در مجموع، به طور میانگین مشاهده شد که روش پیشنهادی دارای سنجش F بهتری نسبت به سایر روش‌های پایه است.

از جمله پژوهش‌های آتی که در راستای این پژوهش باشند می‌توان به بهره‌گیری از ترکیب روش‌های یادگیری عمیق ویژگی‌ها با مکانیزم امتیازدهی برای کاهش ابعاد اشاره کرد. از جمله مواد دیگر می‌توان به توسعه روش پیشنهادی جهت کاربرد در سیستم‌های موازی و توزیع شده اشاره نمود.

مراجع

AT Mc Cray. 1993. A. The Unifed Medical Language System. Meth Inf Med 34 (1993), 281–291.

Charu C Aggarwal and ChengXiang Zhai. 2012. Mining text data. Springer.

Cohen KB, Hunter L. Getting started in text mining. PLoS computational biology. 2008 Jan;4(1):e20.

Hemanjali A, Revathy S, Anu VM, MaryGladence L, Jeyanthi P. Document Clustering on COVID literature using Machine Learning. In2021 5th International Conference on Computing Methodologies and Communication (ICCMC) 2021 Apr 8 (pp. 1530-1535). IEEE.

Illhoi Yoo and Min Song. 2008. Biomedical Ontologies and Text Mining for Biomedicine and Healthcare: A Survey. JCSE 2, 2 (2008), 109–136.

Kolling ML, Furstenau LB, Sott MK, Rabaioli B, Ulmi PH, Bragazzi NL, Tedesco LP. Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development. International Journal of Environmental Research and Public Health. 2021 Jan;18(6):3099.

Liyang Yu. 2011. A developer’s guide to the semantic Web. Springer.

Mallick C, Das AK, Ding W, Nayak J. Ensemble summarization of bio-medical articles integrating clustering and multi-objective evolutionary algorithms. Applied Soft Computing. 2021 Jul 1;106:107347.

Mohammed SM, Jacksi K, Zeebaree SR. A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms. Indonesian Journal of Electrical Engineering and Computer Science. 2021 Apr;22(1):552-62.

Olivier Bodenreider. 2004. The unifed medical language system (UMLS): integrating biomedical terminology. Nucleic acids research 32, suppl 1 (2004), D267–D270.

Pandey SC, Kumar UK. A Noval Fuzzy-Based Artificial Bee Colony Algorithm for Medical Records Classification. InProceedings of International Conference on Big Data, Machine Learning and their Applications 2021 (pp. 73-83). Springer, Singapore.

Reyes-Peña C, Tovar Vidal M, Lavalle Martínez JD. Document Clustering by Relevant Terms: An Approach. InProceedings of the Future Technologies Conference 2019 Oct 24 (pp. 610-617). Springer, Cham.

Tang C, Plasek JM, Xiong Y, Zhang Z, Bates DW, Zhou L. A clustering algorithm based on document embedding to identify clinical note templates. Annals of Data Science. 2021 Sep;8(3):497-515.