Scientometrics Research Journal

Scientific Bi-Quarterly of Shahed University
Vol. 8, No. 1, Spring & Summer 2022 (Serial 15)

Research Article

DOI: 10.22070/RSCI.2020.5813.1432

Topic Modeling of Endocrinology and Metabolism Articles by Iranian Researchers in the Web of Science

Omolbanin Asadi Qadiklaei [1]

Nadjla Hariri [2*]

Maryam Khademi [3]

Fahimeh Babalhavaeji [4]

1. *PhD student in Information and Knowledge Science, Department of Communication and Knowledge Sciences, Science and Research Branch, Islamic Azad University, Tehran, Iran. Email: oasady@gmail.com*
2. *Professor, Department of Communication Science and Science, Islamic Azad University, Tehran, Iran. (Corresponding Author)*
3. *Assistant Professor of Applied Mathematics, Islamic Azad University, South Tehran Branch. Email: khademi@azad.ac.ir*
4. *Associate Professor, Islamic Azad University ,Science & Research Branch. Email: fbabalhavaeji@gmail.com*

Email: nadjlahariri@gmail.com

Date of Reception: 18/09/2020

Date of Acceptation: 06/02/2021

Abstract

Purpose: Probabilistic topic modeling methods consist of a set of algorithms whose main purpose is to discover the hidden subject structure in a large volume of documents. The purpose of this study is to thematically model the articles of Iranian researchers in the field of endocrinology and metabolism in the citation database of Web of Science.

Methodology: The present research is of applied type and has been done by text mining and content analysis method. In this study, all required data were retrieved from the Web of Science Citation Database using the keywords registered in the medical subject heading without a time limit until November 6, 2018. Then, using a hidden allocation algorithm, the whole set of documents in MATLAB was analyzed.

Findings: Subject categories were extracted as groups of 20 words in 10 subject categories. Then, by endocrinologists, the subject categories were named based on their relationship to various topics in the field of endocrinology and metabolism, and each category was assigned a subject title.

Conclusion: The results indicate that the implementation of the latent Dirichlet allocation model has an acceptable performance in presenting the categories of endocrinology and metabolism. The extracted subject categories have good homogeneity and thematic relevance with each other.

Keywords: Endocrinology and metabolism, Topic modeling, LDA, Text mining, Iran

پژوهش‌نامه علم‌سنجی
دوفصلنامه علمی دانشگاه شاهد/ دوره ۸/ شماره ۱/ بهار و تابستان ۱۴۰۱ (پیاپی ۱۵)

مقاله پژوهشی

DOI: 10.22070/RSCI.2020.5813.1432

مدل‌سازی موضوعی مقالات پژوهشگران ایرانی در حوزه غدد درون‌ریز و متابولیسم در پایگاه استنادی وب علوم

ام‌البنین اسدی قادیکلایی [۱]

نجلا حریری [*۲]

مریم خادمی [۳]

فهیمه باب‌الحوائجی [۴]

۱. دانشجوی دکتری علم اطلاعات و دانش‌شناسی، گروه علوم ارتباطات و دانش‌شناسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. Email: oasady@gmail.com

۲. استاد گروه علوم ارتباطات و دانش‌شناسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. (نویسنده مسئول)

۳. دانشیار گروه ریاضی کاربردی، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران. Email: khademi@azad.ac.ir

۴. دانشیار گروه علوم ارتباطات و دانش‌شناسی، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران. Email: fbabalhavaeji@gmail.com

Email: nadjlahariri@gmail.com

صفحه ۶۸-۴۹

دریافت: ۱۳۹۹/۰۶/۲۸

پذیرش: ۱۳۹۹/۱۱/۱۸

چکیده

هدف: روش‌های مدل‌سازی موضوعات احتمالاتی متشکل از مجموعه‌ای از الگوریتم‌هایی است که هدف اصلی آنها کشف ساختار پنهان موضوعی در حجم وسیعی از اسناد است. هدف از انجام این پژوهش مدل‌سازی موضوعی مقالات پژوهشگران ایرانی در حوزه غدد درون‌ریز و متابولیسم در پایگاه استنادی وب علوم است.

روش‌شناسی: پژوهش حاضر از نوع کاربردی است که با روش متن‌کاوی و تحلیل محتوا به انجام رسیده است. در این پژوهش کلیه داده‌های مورد نیاز، از پایگاه استنادی وب علوم با استفاده از کلیدواژه‌های ثبت‌شده در سرعنوان موضوعی پزشکی بدون محدودیت زمانی تا ۱۵ آبان ۹۷ بازیابی شدند. سپس با استفاده از الگوریتم تخصیص پنهان دریکله مجموعه اسناد در محیط متلب تجزیه و تحلیل شدند.

یافته‌ها: دسته‌های موضوعی به‌صورت دسته‌هایی از ۲۰ واژه و در ۱۰ دسته موضوعی استخراج شدند. سپس توسط فوق‌تخصصان غدد دسته‌های موضوعی بر اساس ارتباط آنها به موضوعات مختلف حوزه غدد درون‌ریز و متابولیسم نام‌گذاری شدند و به هر دسته عنوان موضوعی اختصاص یافت.

نتیجه‌گیری: نتایج بیانگر این است که اجرای مدل تخصیص پنهان دریکله عملکرد قابل قبولی در ارائه دسته‌های موضوعات حوزه غدد داشته است. دسته‌های موضوعی استخراج‌شده دارای تجانس و ارتباط موضوعی خوبی با یکدیگر هستند.

واژگان کلیدی: غدد درون‌ریز و متابولیسم، مدل‌سازی موضوعی، تخصیص پنهان دریکله، متن‌کاوی، ایران.

مقدمه و بیان مسئله

بیماری‌های غدد درون‌ریز و متابولیسم از جمله بیماری‌های مهم در جهان هستند. یکی از مهم‌ترین و شایع‌ترین بیماری‌ها در این حوزه بیماری دیابت است که سالانه افراد زیادی به آن مبتلا شده و بر اساس پیش‌بینی‌های مطالعات و مقالات متعدد تا سال ۲۰۲۹ شیوع آن به بیش از ۳۰۰ میلیون نفر در سراسر جهان خواهد رسید. با پیشرفت بیماری، آسیب‌های عروقی و بافتی منجر به عوارض شدیدی مثل رتینوپاتی، نوروپاتی، نفروپاتی، عوارض قلب-عروقی، و زخم پای دیابتی خواهد شد (حشمتی، ۱۳۹۲).

حدود ۴۲۲ میلیون نفر در سراسر جهان مبتلا به دیابت هستند که اکثریت آنها در کشورهای با درآمد کم و متوسط زندگی می‌کنند و سالانه ۱.۶ میلیون مرگ به‌طور مستقیم به دیابت نسبت داده می‌شود. طی چند دهه گذشته هم تعداد موارد و هم شیوع دیابت به‌طور پیوسته در حال افزایش است (سازمان جهانی بهداشت، ۲۰۲۰).

در ایران نیز، طی ۳ دهه گذشته شیوع دیابت دوبرابر شده است. در سال ۲۰۱۴، ۳۸۰۷۹ نفر به علت دیابت در ایران جان باختند که بیشترین میزان مرگ در پی عوارض قلبی و عروقی ناشی از این بیماری بوده است (چارچوب ملی ارائه خدمت در بیماری دیابت، ۱۳۹۵).

همچنین با توجه به افزایش شیوع چاقی این حوزه از اهمیت زیادی نیز برخوردار است. ازآنجایی‌که شیوع چاقی در دنیا رو به افزایش است در این حوزه از اهمیت زیادی نیز برخوردار است. طبق بررسی‌های مرکز ملی آمار سلامت (آمریکا) طی ۳۰ سال درصد افزایش وزن در مردان از ۲۲.۸ درصد به ۳۱.۷ درصد و در زنان از ۲۵.۷ درصد به ۳۴.۹ درصد بوده است و بیشتر افزایش، با تخمین ۳۰ درصد در دهه ۸۰ بوده است. شیوع استئوپروز نیز در جهان رو به افزایش است. بر اساس تخمین بنیاد بین‌المللی استئوپروز، در حال حاضر ۲۰۰ میلیون نفر از زنان سراسر دنیا به این بیماری مبتلا هستند و بیشترین رشد این بیماری در گروه سنی مسن مشاهده می‌شود (گلدن[1]، ۲۰۰۹).

با رشد چشمگیر حجم اطلاعاتی که در دنیای وب قرار می‌گیرد، دیگر نیروی انسانی قابلیت مطالعه و دسته‌بندی اسناد به‌صورت دستی را ندارد. مدل‌سازی موضوعی شامل روش‌هایی است که به کمک ماشین، ما را به سازمان‌دهی، فهمیدن و جستجوی بهینه اسناد متنی قادر می‌سازد.

بسیاری از این الگوریتم‌ها روش‌هایی آماری هستند که با تحلیل متن‌ها تلاش دارند تا زمینه‌ها و موضوعاتی که در متن ها نهفته است را کشف کرده و به بررسی چگونگی ارتباط این موضوعات با هم و یا تغییرات آنها در طول زمان بپردازند (سند هی کومار[2]، ۲۰۱۳).

پژوهشگران در رشته های هوش مصنوعی و یادگیری ماشین جهت رفع مشکلات درخصوص جستجو در حجم وسیعی از داده‌ها مجموعه‌ای از الگوریتم‌ها با عنوان مدل‌سازی موضوعی آماری را توسعه داده‌اند. این الگوریتم‌ها بر اساس روش های آماری‌ای هستند که کلمات موجود در متن را تحلیل کرده، بدون نیاز به فرض اولیه از آنها و دخالت انسانی موضوعات را استخراج می‌کنند و سازمان‌دهی آرشیوهای الکترونیکی را در ابعاد بسیار وسیع که تخصیص موضوع به‌صورت دستی در آنها امکان‌پذیر نیست ممکن می‌کنند (شکرچیان، ۱۳۹۵).

مجموعه‌هایی که با این روش مورد بررسی قرار می‌گیرند معمولاً ساختارنیافته هستند و این روش‌ها کمک می‌کنند تا اسناد متنی از لحاظ موضوعی سازمان‌دهی شوند. این الگوریتم‌ها نیازی به تفسیر و برچسب اولیه روی اسناد ندارند

1 . Golden
2 . Sendhilkumar

و با تکیه بر محتوای اسناد، موضوعات آنها استخراج می‌شوند. مدل‌سازی موضوعات در زمینه‌های مختلفی استفاده دارند. از جمله این زمینه‌ها می‌توان به متن‌کاوی، تکنولوژی‌های جستجو، تکنولوژی‌های نرم‌افزاری، بینایی ماشین، بیوانفورماتیک و اقتصاد اشاره کرد (کاندولا[1]، ۲۰۱۱).

با کمک کشف الگوهای پنهان و برقراری ارتباط معنایی میان مجموعه‌ها، اسناد و کلمات، می‌توانیم به سازمان‌دهی و ساختاربخشی به این مجموعه‌ها بپردازیم. درنتیجه می‌توان اسناد را گروه‌بندی موضوعی کرده و با سهولت بیشتر به بررسی اطلاعات و بهبود نتایج جستجو پرداخت. الگوریتم های موضوعی با روش‌های مختلفی موضوعات موجود در مجموعه‌های اسناد را به نمایش می‌گذارند. اختصاص موضوعات به هر موضوع و دسته‌بندی آنها را الگوریتم‌های مدل‌سازی موضوعی تعیین می‌کنند (لیو[2]، ۲۰۱۶).

تولیدات علمی یکی از رایج‌ترین و مهم‌ترین معیارهای سنجش علمی در جوامع هستند. پایگاه استنادی علوم یکی از معتبرترین مراجع رتبه‌بندی علمی پژوهشگران است. تحلیل مقالات یکی از روش‌های ارزیابی تولیدات علمی در یک کشور است. این تحلیل‌ها با استفاده از ابزارهای مختلفی صورت می‌پذیرد و با استفاده از این نتایج می‌توان توان علمی کشور را افزایش داد که نتیجه آن در حوزه پزشکی افزایش سطح بهداشت و بهبود سیاست‌های کلان پزشکی کشور است (صابری و اسفندیاری مقدم،۱۳۹۰).

برای هرگونه برنامه‌ریزی و سیاست‌گذاری، در اختیار داشتن اطلاعات در مورد تولیدات علمی مورد نیاز است. تسهیل بازیابی اطلاعات در این حوزه می‌تواند بسیار مهم باشد. بررسی تولیدات علمی و مدل‌سازی موضوعی حوزه غدد درون‌ریز و متابولیسم[3] به پژوهشگران خواهد گفت که چگونه برای رقابت در عرصه جهانی توانمند شوند. بدیهی است بهبود وضعیت علمی در حوزه غدد درون ریز در طول زمان به پیشرفت در زمینه پیشگیری، درمان و کاهش مرگ‌ومیر منجر خواهد گردید. انتظار می‌رود با انجام مدل‌سازی موضوعی بتوان وضعیت مطالعات و پژوهش‌ها را در این رشته مجسم‌تر نمود و مسیر آینده برای پژوهشگران روشن‌تر گردد همچنین می‌توان در بهبود نظام سلامت و درمان بیماران مفید واقع شد.

یکی از بهترین روش‌های بررسی تولیدات علمی یک کشور تحلیل و بررسی مقالات علمی است و با درنظرگرفتن هدف اصلی مدل‌سازی موضوعی که نمایش الگوهای پنهان موجود در اسناد متنی بدون دخالت انسان است در این پژوهش بر آنیم تا به این سؤال پاسخ دهیم "آیا توزیع موضوعات حوزه غدد درون‌ریز و متابولیسم در مقالات پژوهشگران ایرانی متناسب با جایگاه این رشته در کشور است؟"

سؤال‌های پژوهش

۱. توزیع موضوعات حوزه غدد درون‌ریز و متابولیسم چگونه است؟
۲. واژگان با بیشترین وابستگی موضوعی در حوزه غدد درون‌ریز و متابولیسم کدام است؟
۳. واژگان و دسته‌های موضوعی خارج از حوزه غدد درون‌ریز و متابولیسم کدام است؟
۴. آیا روش مورد استفاده و نرم‌افزار متلب[4] برای مدل‌سازی موضوعی قابل اعتماد هستند؟

1 . Kandula
2 . Liu
3 . Endocrinology and metabolism
4 . MATLAB

چارچوب نظری

مدل‌سازی موضوعی

مدل‌سازی موضوعی برای اولین‌بار در پژوهش پاپادیمیتریو و همکاران[1] معرفی شد. پس از آن هافمن در پژوهش دیگری روش PLSA[2] را در سال ۱۹۹۹ ارائه نمود. این روش اساس پیدایش مدل‌سازی موضوعی قرار گرفت که توسط بلی[3] و همکاران[4] (۲۰۰۳) توسعه و گسترش پیدا کرد. در ابتدا به موضوعات به‌صورت بستهٔ[5] لغات نگاه می‌شد. گسترش‌هایی بر این نگاه صورت گرفت. به‌دست‌آوردن موضوعات به‌صورت سلسله‌مراتبی به جای حالت تک‌لایه، به‌کارگیری موجودیت‌ها در کنار متن، استفاده از عبارات به جای لغات و استفاده از دانش داخلی و خارجی در مدل‌سازی از جمله این گسترش‌ها می‌باشند (هافمن[6]، ۲۰۲۰).

مدل‌سازی موضوعی یک روش احتمالاتی زایاست[7] که به‌صورت وسیعی در رشته کامپیوتر کاربرد دارد و در سال‌های اخیر بر داده‌کاوی و بازیابی اطلاعات متمرکز شده است. مدل احتمالاتی زایا برای اسناد بر اساس یک سری قانون نمونه‌گیری احتمالاتی عمل می‌کند. این قانون‌ها مشخص می‌کنند کلمات اسناد چگونه ممکن است بر پایه متغیرهای نهان تولید شوند. به جز داده‌کاوی این رشته همچنین کاربردهای موفقی در رشته‌های کامپیوتر، ژنتیک و شبکه‌های اجتماعی داشته است (دیروست[8]، ۱۹۹۰). پیدایش مدل‌سازی موضوعی با نمایه‌سازی معنایی پنهان[9] بود که یکی از پایه‌های توسعه مدل‌سازی موضوعی است (هافمن[10]، ۲۰۰۱). این مدل یک مدل احتمالاتی نیست بنابراین به عنوان مدل سازی موضوعی اعتبار ندارد. بر پایه این الگوریتم، الگوریتم احتمالی آنالیز معنایی پنهان[11] پیشنهاد شد (هاوز[12]، ۲۰۰۱). این مدل پایه اصلی مدل‌سازی موضوعی قرار گرفت. پس از آن مدل تخصیص پنهان دیریکله[13] توسط بلی و دیگران در سال ۲۰۰۳ ارائه شد. این مدل کامل‌تر، احتمالاتی و زایا و بر پایه PLSA بود. امروزه تعداد مدل‌های احتمالاتی که بر پایه تخصیص پنهان دریکله به وجود آمده‌اند بسیار افزایش پیدا کرده‌اند (بلی[14]، ۲۰۱۷).

مدل تخصیص پنهان دیریکله

روش‌های مدل‌سازی موضوعات که مبتنی بر احتمالات هستند، متشکل از مجموعه‌ای از الگوریتم‌هاست که هدف اصلی آنها کشف ساختار پنهان موضوعی در حجم بسیار زیادی از اسناد است. یکی از پرکاربردترین و اساسی‌ترین روش‌های مدل‌سازی موضوعی احتمالاتی مدل تخصیص پنهان دریکله است (بلی[15]، ۲۰۰۳). هدف از اجرای این روش یافتن بهترین مجموعه از شاخصه‌های پنهان در اسناد است که مشاهدات را توصیف می‌کنند (کلمات موجود در

1 . Papadimitriou, Raghavan, Tamaki and Vempala
2 . probabilistic latent semantic analysis (PLSA)
3 . Blei
4 . David Blei, Andrew Ng, and Michael I. Jordan
5 . Bag of words
6 . Hofmann
7 . generative model
8 . Deerwester
9 . latent semantic indexing (LSI)
10 . Hofmann
11 . Probabilistic hidden semantic analysis algorithm
12 . Howes
13 . latent Dirichlet allocation (LDA)
14 . Blei
15 . Blei

اسناد)، با فرض تولید داده‌ها توسط مدل به دست آمده. یکی از اهداف اصلی مدل‌های احتمالاتی زایا این است که اسناد توزیعی از موضوعات هستند. تفاوت این مدل‌ها بیشتر در فرضیات آماری است رامیج[1]، (۲۰۰۹). اگر فرض کنیم $P(Z)$ برای یک سند نشان‌دهندهٔ توزیع روی تمام موضوعات z است در این صورت $P(W|Z)$ توزیع کلمات را روی موضوع z نشان می‌دهد و توزیع کلمه موضوع[2] ارائه می‌شود. در این روش تولید هر کلمه به ازای هر سند در دو مرحله صورت می‌پذیرد. برای تولید هر کلمه W_i در یک سند، ابتدا یک نمونه‌گیری روی توزیع موضوعات صورت می‌گیرد و یک موضوع Z موضوع Z می‌شود. سپس یک کلمه از توزیع کلمه-موضوع $P(W|Z)$ انتخاب می‌شود. از $P(Z_i = j)$ برای نشان‌دادن احتمال انتخاب موضوع j برای کلمه i ام در نمونه‌گیری و از $P(W_i|Z_i)$ برای احتمال کلمه W_i در موضوع j ام استفاده می‌شود. برای یک سند، احتمال تولید کلمات آن مطابق با فرمول ۲-۱ است. که T در اینجا تعداد؛ که موضوعات است. فرمول ۱ به‌طور خلاصه بیان می‌کند که احتمال تولید یک کلمه برای یک سند، برابر با احتمال تولید کلمه توسط موضوعات است (شکرچیان، ۱۳۹۵).

فرمول ۱.
$$P(W_i) = \sum_{j-1}^{T} P(W_i|Z_i = j)P(Z_i = j)$$

پیشینه پژوهش

پیشینه پژوهش در داخل

مسعودی و راحتی (۱۳۹۴) در پژوهشی مدلی برای رفع ابهام از واژگان مبهم فارسی بر اساس استخراج ویژگی‌های جدید پیشنهاد داده‌اند. در این مقاله از روش بدون نظارت تخصیص پنهان دریکله استفاده شده است. نتایج آزمایشات برای چهار واژه مبهم پرتکرار در زبان فارسی که از پیکره پژوهش‌های پردازش هوشمند علائم استخراج شد، دقّت حدود ۶۶.۵۱ درصد را نشان می‌دهد که بیانگر مؤثربودن این روش در یافتن معنی مناسب واژگان مبهم است.

قاضی میرسعید و صنیعی (۱۳۹۴) در پژوهشی جایگاه علمی مراکز تحقیقاتی غدد درون ریز، دیابت و متابولیسم دانشگاه‌های علوم پزشکی کشور را با روش Exergy ارزیابی کردند. نتایج این پژوهش بیانگر اهمیت تأثیر ارتقای جایگاه علمی مراکز تحقیقاتی در کنار کمیت مقالات در این حوزه است.

شکرچیان چالشتری (۱۳۹۵) در پژوهشی برای اطمینان از انسجام موضوعات خوشه‌ها، از کلمات خود اسناد برای بازیابی اطلاعات استفاده کرد. موضوعات خوشه ها برای هموارسازی مدل زبانی اسناد مورد استفاده قرار گرفتند. نتایج نشان داد که موضوعات به‌دست‌آمده باعث بهترشدن نتایج بازیابی شده‌اند. این روش نسبت به روش شباهت پرسش و همچنین روشی که از موضوعات کل مجموعه برای هموارسازی استفاده می‌کند بهتر عمل کرد. تقسیم مجموعه به چند خوشه موضوعی و استفاده از موضوعات خوشه‌ها در بازیابی اطلاعات به‌عنوان روش مقیاس‌پذیر جدیدی در بازیابی اطلاعات معرفی شد. همچنین از عبارت‌های پرتکرار اسناد به‌عنوان منبع دانشی مورد استفاده قرار گرفت. در این تحقیق روش پایه تنها از متن اسناد برای مدل‌سازی استفاده می‌کرد و فرض بر آن بود که جز متن اسناد اطلاعات دیگری در دسترس نبوده و در صورت وجود اطلاعات بیشتری از مجموعه سندی از این اطلاعات در کنار دانش

1 . Ramage
2 . Topic-word distribution

داخلی مجموعه برای بالابردن انسجام موضوعات استفاده شد.

پیشینه در خارج

ازموتلو و کاودر[1] (۲۰۰۵) در پژوهشی به ارائه یک شبکه عصبی مصنوعی برای شناسایی تغییرات موضوع به‌صورت خودکار با استفاده از ویژگی‌های آماری پرس‌وجو، مانند فواصل زمانی و الگوهای اصلاح پرس‌وجو پرداختند. داده های ورودی از موتور جستجوی نروژی FAST برای آموزش شبکه عصبی انتخاب شده و سپس شبکه عصبی برای شناسایی تغییرات موضوع در ورود اطلاعات مورد استفاده قرار گرفت. یافته ها نشان داد درمجموع ۵۴.۴ درصد از تغییرات موضوع و ۴۶.۶ درصد از تمهیدات موضوع به‌درستی برآورد شده است.

مهلر و واتینگر[2] (۲۰۰۹) پژوهشی را با هدف ارائه یک مدل طبقه‌بندی موضوعی با استفاده از طبقه‌بندی ده‌دهی دیوی به انجام رساندند. این کار با کاوش ابرداده‌ای که توسط ابتکار آرشیو باز (OAI) ارائه شده است برای استخراج سند انجام شد. علاوه‌براین، این مقاله با هدف انتخاب و گسترش ویژگی‌ها با استفاده از هستی‌شناسی‌های اجتماعی و منابع واژگان وابسته به وب انجام شد. یافته‌ها نشان داد که طبقه‌بندی‌های مبتنی بر SVM با کاوش انتخاب‌های خاصی از ابرداده سند OAI بهتر عمل می‌کنند.

وانگ[3] و همکاران (۲۰۱۷) تکنیک خاصی را جهت استخراج داده و تشخیص شرایط اضطراری شهری در خطرات طبیعی، بلایای طبیعی و سایر موارد اضطراری پیشنهاد داده‌اند. این روش هم ابعاد معنایی و هم جغرافیایی پدیده‌ها را با استفاده از ماژول‌های تشخیص جغرافیایی و ارزیابی سطح بحران بر اساس شدت نگرش منفی ماژول رتبه‌بندی بیان می‌کند. این مدل برای توییت‌های جغرافیایی از اپلیکیشن توییتر طراحی شده است. برای ارزیابی این تکنیک آزمایشی با توییت‌های مربوط به شهرهای مختلف در دوره‌های ۴ تا ۶ ساعته صورت گرفت و مدل مورد نظر شرایط اضطراری انواع مختلف جغرافیایی را شناسایی کرد.

رابینسون[4] (۲۰۱۹) در مقاله‌ای با استفاده از روش تخصیص پنهان دریکله داده‌های ۱۴ ساله گزارش‌های حمل و نقل هوایی را جهت استخراج داده‌های تجاری خطوط هوایی فیلتر کرده است. موضوعات به همراه وزن کلمات و استفاده‌های معنایی هر موضوع به‌صورت موضوعی به سه متخصص موضوعی ارائه شدند. متخصصان در خصوص مضامین موضوعات انتخابی به میزان بسیار زیادی توافق داشته اند. کارشناسان موضوعی قادر به تشخیص صلاحیت خبرنگاران، مقررات زیست‌محیطی، نظارتی و صنعتی مطابق با روندهای زمانی در زمینه استفاده از موضوع است بوده‌اند. نتایج بیانگر افزایش نگرانی‌های ایمنی خدمه پرواز در صورت آگاهی از سیستم مشاوره‌ای در باند بین کابین خلبان و فرودگاه بوده است. به‌علاوه نتایج سبب تسریع در تغییرات راه‌های ایمن‌ترکردن صنعتی است. این پژوهش سبب ارزیابی عملی استفاده از پردازش زبان طبیعی در شناسایی و تعیین دقیق روند لازم برای اولویت بندی فعالیت‌های ایمنی شد.

باستانی[5] و دیگران (۲۰۱۹) در پژوهشی یک روش هوشمند مبتنی بر تخصیص پنهان دیریکله برای تجزیه و تحلیل شکایات مصرف‌کنندگان پیشنهاد دادند. رویکرد پیشنهادی با هدف استخراج موضوعات نهفته در شکایات و بررسی

1 . Özmutluan and Çavdur
2 . Mehler and Waltinger
3 . Wang
4 . Robinson
5 . Bastani

روندهای مرتبط با آنها در طول زمان انجام شد. سپس از روند زمان برای ارزیابی اثربخشی مقررات و انتظارات از مؤسسات مالی در ایجاد فرهنگ مصرف‌گرا استفاده کردند. نتایج خوشه‌های منسجم و معنادار از شکایات مصرف‌کننده ایجاد کرد. این موضوعات نه تنها خلاصه‌ای از شکایات را در تجزیه متون قابل تفسیر توسط انسان خلاصه کردند، بلکه به پزشکان نیز کمک کردند تا مطالب جدیدی را که ممکن است توسط سازمان حمایت از مصرف‌کننده نادیده گرفته شود کشف کنند.

هدایت‌الله[1] و همکاران (۲۰۱۹) مطالعه‌ای با هدف استفاده از روش مدل‌سازی موضوع با استفاده از تخصیص پنهان دریکله برای مجموعه داده‌های توییتر به اشتراک گذاشته‌شده توسط حساب رسمی توییتر در جزیره جاوا به انجام رساندند. نتیجه این پژوهش وضعیت آب و هوا و فاجعه‌ای را که در جزیره جاوا رخ داده است را نشان دهد. بر اساس نتیجه مدل‌سازی موضوعی، پنج موضوع قابل توجه از حساب‌های توییتر یافته شد که در مورد موضوعات مورد بحث، حساب‌های توییتر اطلاعات مربوط به اطلاعات آب و هوا و پیش‌بینی آب و هوا، آخرین اطلاعات آب و هوا در منطقه یوگیاکارتا، پیش‌بینی و هشدار هوا در جاوه مرکزی و جاوه غربی بود. در این مطالعه، همچنین با تجزیه و تحلیل متداول‌ترین کلمات از هر منطقه در جاوا، روند آب و هوا و فاجعه نشان داده شد.

جمع‌بندی از مرور پیشینه

مرور پیشینه پژوهش نشان می‌دهد برای ارزیابی تولیدات علمی، تحلیل مقالات یکی از رایج‌ترین روش‌ها و مورد توجه بسیاری از پژوهشگران است. با توجه بررسی‌هایی که توسط پژوهشگران این مقاله انجام شد پژوهشی تاکنون به بررسی مقالات حوزه غدد درون‌ریز و متابولیسم نپرداخته است و از نرم‌افزار متلب با توجه به دقت بالا و همچنین وجود بسته ابزاری مخصوص مدل‌سازی موضوعی استفاده نشده است. نوآوری مقاله حاضر از لحاظ موضوع و از لحاظ ابزار مورد استفاده در تحلیل مقالات است.

روش‌شناسی پژوهش

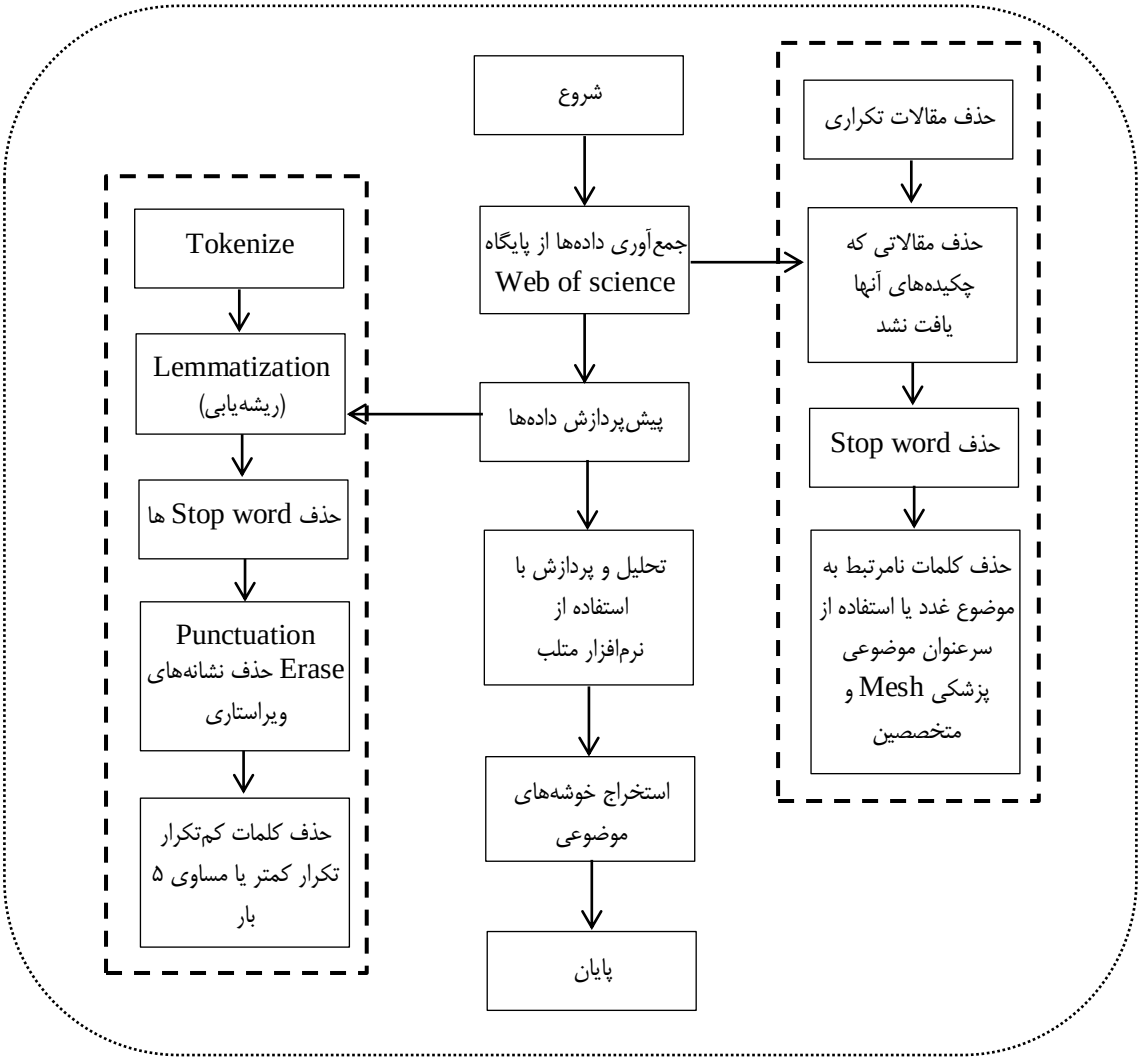
پژوهش حاضر از نوع کاربردی است که با استفاده از روش متن‌کاوی و تحلیل محتوا صورت گرفته است.

انتخاب نمایه استنادی علوم و وب علوم به دلیل اهمیت این نمایه‌نامه که دربرگیرنده برون‌دادهای علمی (کتاب، مقاله مجلات و مقالات کنفرانس) در تعداد زیادی از منابع و نشریات معتبر در زمینه‌های مختلف نظیر پزشکی و سایر رشته‌هاست می‌باشد.

جامعه پژوهش شامل کلیه مقالات حوزه غدد درون‌ریز و متابولیسم بود که بر اساس اصطلاحات انتخاب شده در سرعنوان‌های موضوعی پزشکی یا مش کلیدواژه‌های منتخب و مترادف بازیابی شدند.

برای دستیابی به اعضای جامعه پژوهش یعنی مقالات در این پژوهش، نیاز به توصیفگرهایی بود که با استفاده از آنها به بازیابی مقالات پرداخت. بدین منظور کلیه واژه‌های منتخب و مترادف تحت مدخل غدد درون‌ریز و متابولیسم که شامل ۱۲۱۵ کلیدواژه هستند در بخش جستجوی پیشرفته پایگاه استنادی علوم با محدودشدن در فیلد CU به کشور ایران و محدودکردن به‌عنوان مقالات بدون محدودیت تاریخی مورد جستجو قرار گرفتند. درنتیجه کلیه برون‌دادهای علمی با جستجوی تمامی توصیف‌گرها و مدخل‌ها بازیابی شدند. سپس داده‌ها در قالب فایل متن ساده ذخیره شدند تا در مراحل بعدی مورد تجزیه و تحلیل قرار گیرد.

1 . Hidayatullah



شکل ۱. روند کلی مدل‌سازی موضوعی مقالات حوزه غدد پژوهشگران ایرانی در پایگاه استنادی وب علوم

روند کلی

شکل شماره ۱ روند کلی مدل‌سازی موضوعی مقالات حوزه غدد را نشان می‌دهد. در گام اول چکیده کلیه مقالات با استفاده از کلیدواژه‌های ثبت‌شده در سرعنوان موضوعی پزشکی از پایگاه استنادی وب علوم استخراج شد. مقالات بازیابی شده به نرم‌افزار مدیریت منابع اِندنوت ۷[1] وارد شدند. درمجموع ۷۸۹۰ مقاله از تاریخ ۱۹۷۷ تا تاریخ ۲۰۱۸ بازیابی شد که از این تعداد ۲۳۳۸ مقاله دارای چکیده تکراری و یا فاقد چکیده حذف شدند. درنهایت تعداد ۵۵۵۲ مقاله باقی ماند. درنهایت خروجی داده‌ها به‌صورت متن ساده در گام دوم عملیات مورد پیش‌پردازش قرار گرفتند که شامل مراحل زیر است:

۱. پیش‌پردازش متن

۲. تبدیل متن به حروف کوچک؛

۳. تبدیل متن به توکن[2]؛

1 . Endnote
2 . Tokenize

۴. حذف نشانه‌های ویراستاری[1]؛

۵. حذف کلمات کمتر از ۲ و یا بیشتر از ۱۵ حرف؛

۶. ریشه‌یابی[2] کلمات با استفاده از نرمالیزه‌کردن؛

حذف کلمات ایست[3]: واژه ها و لغاتی که با وجود تکرار مکرر در متن مقالات، از نظر معنایی دارای اهمیت کمی هستند مانند "اما"، "ولی"، "که"، "با" و غیره. بسیاری از افعال، اسامی، قیود، صفات و کلمات ربط و تعریف نیز ایست‌واژه شناخته شده‌اند. حذف این کلمات، موجب بهبود نتایج خواهد شد و همچنین سبب کاهش بار محاسبات و افزایش سرعت پردازش خواهد گردید. بر این اساس این کلمات در مرحله پیش پردازش حذف می‌شوند (تقوا[4]، ۲۰۰۳).

عملیات پیش‌پردازش با استفاده از نرم‌افزار متلب انجام شد.

پس از پیش پردازش متن، واژه‌های کلی که به حوزه موضوعی غدد مرتبط نبودند با استفاده از سرعنوان موضوعی پزشکی و همچنین تأیید فوق‌تخصصان غدد درون‌ریز و متابولیسم حذف شدند. در گام سوم داده‌ها به‌صورت بانک داده به‌عنوان ورودی به نرم‌افزار متلب 11 داده شد، سپس با استفاده از الگوریتم تخصیص پنهان دریکله که یکی از مهم‌ترین الگوریتم‌ها در تکنیک متن‌کاوی است مجموعه اسناد مورد تجزیه و تحلیل قرار گرفتند. همان طور که در پیوست شماره ۱ به نمایش درآمده است، در گام چهارم خوشه‌های موضوعی به‌صورت دسته‌های ۲۰ تایی و در ۱۰ دسته موضوعی مورد استخراج قرار گرفتند.

با کمک فوق تخصصان رشته غدد درون‌ریز و متابولیسم با توجه به ارتباط موضوعی واژه‌ها به حوزه غدد عناوین موضوعی برای هر دسته از موضوعات انتخاب شد.

در کل تعداد ۲۹۵۸۰ واژه یکتا و تعداد ۱۶۸۵۵۸۰ توکن مورد تحلیل قرار گرفتند. پس از اجرای مدل‌سازی موضوعی با استفاده از الگوریتم تخصیص پنهان دریکله تعداد ۱۰ دسته موضوعی و در هر دسته موضوعی ۲۰ واژه استخراج شد.

یافته‌های پژوهش

پاسخ به سؤال اول پژوهش. توزیع موضوعات حوزه غدد درون‌ریز و متابولیسم چگونه است؟

پیوست شماره ۱ شمای کلی دسته‌بندی موضوعات را نمایش می‌دهد.

دسته موضوعی علائم و درمان (Disease and treatment)

شامل واژه‌های patient, treatment, disease, infection, therapy ,test, diagnosis, disorder, symptom, diagnose, leukemia, regimen, HSCT Treat, chronic, GVHD , mortality, marrow, sexual, thalassemia.

دسته موضوعی بیماری‌های قلبی و عروقی (Cardiovascular disease)

شامل واژه‌های patient, diabetes, diabetic, disease, age, coronary, mellitus, artery, heart,

1 . Punctuations
2 . Stemming
3 . stop words
4 . Taghva

.death, eye, gender, old, man ,mortality, hypertension, Retinopathy, cardiovascular, stroke, male
واژه‌های این دسته موضوعی بیشترین ارتباط را با بیماری‌های قلبی و عروقی دارند.

دسته موضوعی عوارض دیابت (Complications of Diabetes)

شامل واژه‌های treatment, test, medicine, chronic, therapeutic, global, therapy, endocrine, gastric, periodontal secretion, adult, rural, pathogenesis, melatonin, relate, diseases, periodontitis, inflammatory, basic. در این دسته موضوعی واژه‌ها به عوارض ناشی از بیماری دیابت مرتبط هستند.

دسته موضوعی سندرم متابولیک (Metabolic Syndrome)

شامل واژه‌های metabolic, mets, age, bmi, obesity, man, Cardiovascular, cvd, hypertension, adult, weight, obese, diabetes, predict, systolic, sex, predictor, anthropometric, diastolic, criterion. در این دسته موضوعی واژه‌هایی حضور دارند که به سندرم متابولیک ارتباط دارند.

دسته موضوعی دیابت نوع ۲ (Type2 Diabetes)

شامل واژه‌های glucose, serum, diabetes, fast, patient, lipid, resistance, cholesterol, plasma, t2dm, hba1c, lipoprotein, diabetic, triglyceride, adiponectin, mellitus, glycemic, sugar, hdl, leptin. واژه‌های این دسته موضوعی به بیماری دیابت نوع ۲ مرتبط هستند.

دسته موضوعی بیماری‌های متابولیک استخوان (Metabolic bone disease)

شامل واژه‌های bone, tumor, disease, diagnosis, lesion, tnfalpha, adrenal, lung, injury, brain, tissue, diagnostic, biopsy, inflammatory, diagnose, mri, drug, tooth, symptom, pathological. در این دسته موضوعی واژه‌ها به بیماری‌های متابولیک استخوان مرتبط هستند.

دسته موضوعی سرطان‌ها در حوزه غدد (Endocrine cancers)

شامل واژه‌های Cancer, disease, mscs, immune, endothelial, cytokine, breast, growth, chronic, oxide, nitric, hepatitis, tumor, organ, prostate, lymphocyte, antigen, vegf, death, malignant. واژه‌های مرتبط با سرطان‌ها در این دسته موضوعی حضور دارند.

دسته موضوعی دیابت (Diabetes)

شامل واژه‌های Diabetic, diabetes, glucose, treatment, animal, treat, mellitus, male, injection, hyperglycemia, streptozotocin, pancreatic, stz, oil, weight, enzyme, liver, islets, test, acid. موضوع دیابت در این دسته موضوعی بیشترین ارتباط را با واژه‌ها دارد.

دسته موضوعی سبک زندگی و رژیم غذایی (Lifestyle and diet)

شامل child, dietary, diet, age, food, adolescent, t1dm, fat, girl, healthy, nutritional, weight, lifestyle, disease, adult, selfcare, nutrition, milk, gender, vegetable. واژه‌های مرتبط با سبک و رژیم غذایی در این دسته موضوعی قرار گرفته‌اند.

دسته موضوعی بیماری‌های مزمن در حوزه غدد (Endocrine Chronic disease)

شـامل واژه‌هـای disease, disorder, cardiac, ckd, skin, diagnosis, pulmonary, congenital, respiratory, limb, symptom, chronic, organ, male, cah, tuberculosis, curcumin, pon1, hyperplasia, bmmscs. واژه‌های مرتبط با بیماری‌های مزمن در حوزه غدد در این دسته موضوعی قرار گرفته‌اند.

پاسخ به سؤال دوم پژوهش. واژه هایی که دارای بیشترین احتمال وابستگی به موضـوعات هسـتند چـه واژه‌هایی هستند و دلیل این امر چیست؟

جدول ۱. واژه‌هایی که بیشترین احتمال وابستگی را در هر دسته موضوعی داشته‌اند

Number	Word	Frequency	Probability
1	patient	12282	0.213422
2	patient	12282	0.134784
3	treatment	3026	0.059828
4	diabetes	7145	0.01818
5	patient	12282	0.046996
6	disease	4492	0.028427
7	disease	4492	0.040359
8	diabetes	7145	0.072167
9	disease	4492	0.011062
10	disease	4492	0.061458

همان‌طور که جدول شماره ۱ نشان می‌دهند واژه‌هـای بیمـاری، درمـان و دیابـت دارای بیشـترین میـزان احتمـال وابستگی به موضوعات در دسته‌های موضوعی را دارند. با توجه به شیوع و اهمیت بیماری دیابت نتایج جدول شماره ۱ مورد تأیید است و نشان‌دهنده این است که پژوهشگران ایرانی در زمینه درمان، پیشگیری و علائم تشخیصی بیماری دیابت بیشتر پرداخته‌اند.

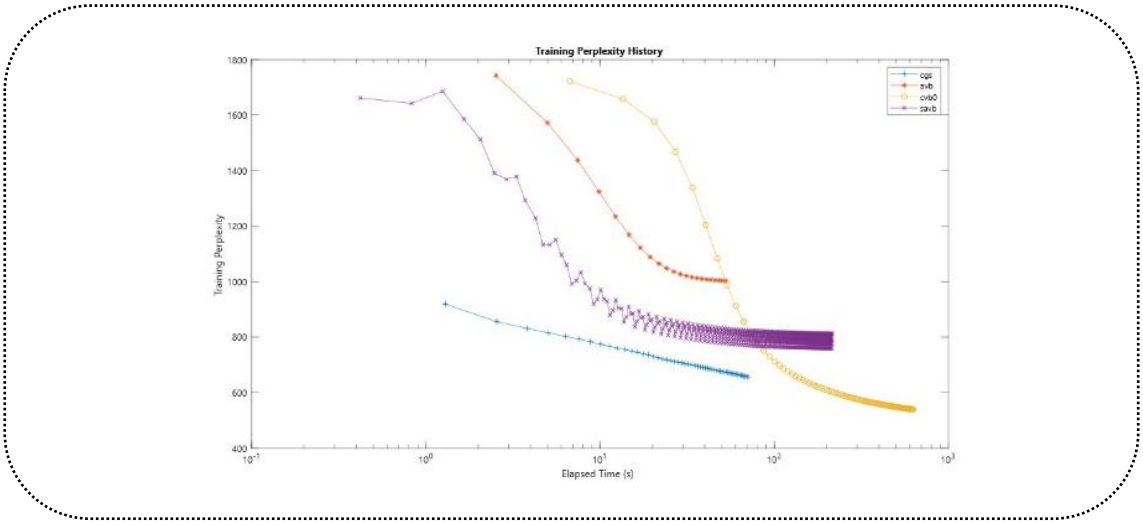
پاسخ به سؤال سوم. واژگان و دسته‌های موضوعی خارج از حـوزه غـدد درون‌ریـز و متابولیسـم کـدام است؟

واژه‌های مربوط به اختلالات جنسی، بیمـاری سـل-غـدد، بیمـاری‌هـای آدرنـال و بیمـاری‌هـای خـود ایمنـی در بیماری‌های غدد درون‌ریز و متابولیسم دسته موضوعی مستقلی به آنها اختصاص نیافته که این امر بیانگر این است کـه این دو حوزه موضوعی با توجه به بحث شیوع و اهمیت دیابت کمتر مورد توجـه پژوهشـگران در حـوزه غـدد قـرار گرفته‌اند.

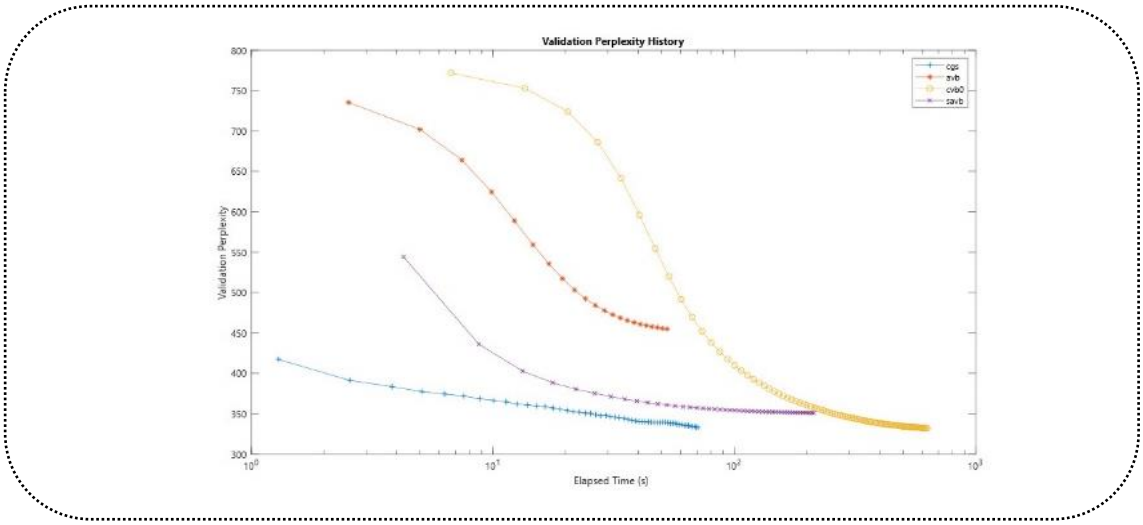
مباحث علائم و درمان، بیماری‌های قلبی و عروقی و سبک زندگی و رژیم غذایی در تحلیل مقـالات پژوهشـگران ایرانی دسته های موضوعی مستقلی اختصاص یافته که در سرعنوان موضوعی پزشکی به‌عنوان موضـوعات مسـتقل در حوزه غدد نمایش داده نشده است. این امر بیانگر توجه بیشتر پژوهشگران ایرانی به این حوزه‌های موضوعی و اهمیت این موضوعات در جامعه ایران با توجه به شیوع و درگیری جامعه پزشکی کشور و محققان است.

پاسخ به سؤال چهارم پژوهش. آیا روش مورد استفاده و نرم‌افزار متلب برای مدل‌سازی موضوعی قابل اعتماد هستند؟

رایج‌ترین راه جهت ارزیابی مدل‌های احتمالی محاسبه لگاریتم احتمال وقوع برای یک مجموعه داده‌ای آزمون که قبلاً به سیستم ارائه نشده است. در این روش داده‌ها به دو مجموعه آموزش و ارزیابی تقسیم می‌شوند. مجموعه دادهٔ ارزیابی مجموعه‌ای است که توسط مدل‌ساز مشاهده نشده و مجموعه داده‌ای آموزش، که لگاریتم احتمال وقوع آنها باید بر اساس مجموعه ارزیابی محاسبه شود. رایج‌ترین معیار محاسبه لگاریتم احتمال وقوع در مدل‌های موضوعی معیار سرگشتگی[1] است (ژاو[2]، ۲۰۱۵).

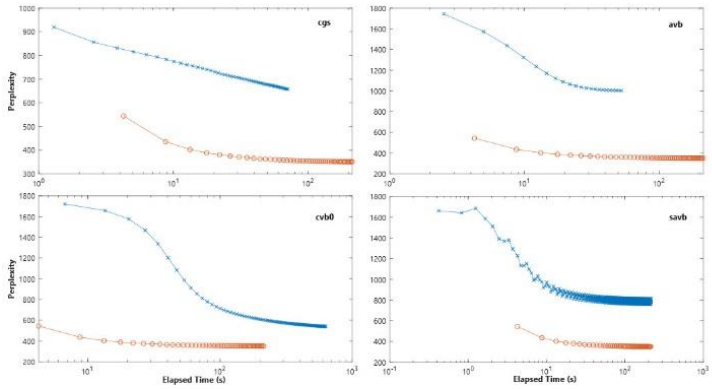


نمودار ۱. تاریخچه معیار سرگشتگی در مرحله آموزش برای ۴ روش



نمودار ۲. تاریخچه سرگشتگی در مرحله ارزیابی (تست) برای ۴ روش

1 . Perplexity
2 . zhao



نمودار ۳. مقایسه مقدار سرگشتگی برای ۴ روش به‌صورت جدا در فازهای آموزش و ارزیابی

جدول ۲. میزان معیار سرگشتگی بر اساس مدت‌زمان محاسبه در مرحله آموزش و ارزیابی (تست)

Method	Train	Validation	Time (s)
Cgs	657.3	333.3	71
avb	1002.8	454.9	53
cvb0	539.9	331.9	629
savb	790.7	351	215

همان‌گونه که در جدول (۲) و نمودار شماره ۱ و ۲ و ۳ نشان داده شده است مقادیر به‌دست‌آمده سرگشتگی در روش‌های مختلف در مرحله آموزش به‌ترتیب ۶۵۷.۳، ۱۹۰۲.۸، ۵۳۹.۹، ۷۹۰.۷ و در مرحله ارزیابی میزان سرگشتگی به‌ترتیب۳۳۳.۳، ۴۵۴.۹، ۳۳۱.۹، ۳۵۱ است که در نمودار ۱-۳ نشان داده شده است.

با توجه به نتایج به دست آمده به نظر می‌رسد روش cgs[1] بهترین عملکرد را نسبت به بقیه روش‌ها داشته است (نمودار شماره ۳) و با توجه به پایین‌تربودن میزان سرگشتگی در مرحله ارزیابی، عملکرد مدل تأیید می‌شود.

بحث و نتیجه‌گیری

این پژوهش با هدف متن‌کاوی و مدل‌سازی موضوعی مطالعات متخصصان ایرانی حوزه غدد درون‌ریز و متابولیسم به انجام رسید. در این پژوهش به بررسی و تحلیل مقالات حوزه غدد که توسط متخصصان ایرانی نوشته و در پایگاه استنادی وب علوم نمایه شده است پرداخته‌ایم. مراحل تحلیل شامل استخراج اطلاعات، پیش‌پردازش، پردازش و سپس تحلیل نتایج به انجام رسید. نتایج بیانگر این است که الگوریتم تخصیص پنهان دریکله و روش cgs با توجه به ارائه کمترین میزان سرگشتگی، کارایی قابل قبولی در انجام مدل‌سازی موضوعی دارند. عملکرد الگوریتم پنهان دریکله در پژوهش پارک[2] و همکاران (۲۰۱۴) تأیید شده است. در این پژوهش که یک مدل احتمالاتی موضوع بیماری پزشکی برای کشف دانش مرتبط با بیماری‌ها و داروها پیشنهاد شده است، از خصیصه پنهان دیریکله، به‌عنوان مدل

1 . collapsed Gibbs samples
2 . Park

پایه استفاده شده است. سپس، کیفیت موضوعات کمی و کیفی مورد مقایسه قرار گرفت. نتایج مقایسه نشان داد که موضوعات مشتق‌شده به شناسایی واضح‌تر الگوهای خاص در بیماری‌ها و داروها منجر شده و یک الگوی خاص در بیماری‌ها و داروها پدید آمد.

همچنین در پژوهشی که توسط وانگ[1] و همکاران (۲۰۱۷) به انجام رسید نتایج پژوهش حاضر تأیید می‌شود و مدل‌سازی موضوعی از طریق الگوریتم پنهان دریکله سبب بهبود عملکرد جهت شناسایی خودکار فایل‌های منابع مرتبط برای گزارش‌های مشکلات ارائه‌شده گردید. مدل پیشنهادی راه‌های نظارت را بهبود بخشید و به اشکال گزارش شده و فایل‌های منبع، هم شباهت‌های متنی و هم شباهت‌های معنایی را آموزش داده، علاوه بر آن انواع مختلف اشکالات گزارش‌شده در نظر گرفته شد. نتایج بر سه دسته داده واقعی نشان داد مدل پیشنهادی می تواند به بیش از ۲۳.۶ درصد در پیش‌بینی دقت عملکرد و مقیاس خطی حجم داده‌ها بهبود حاصل کند.

سانگ[2] و همکاران (۲۰۱۷) پژوهشی را با هدف توسعه داده‌کاوی در داده‌های با حجم عظیم پزشکی به کار گرفتند. مدل پیشنهادی با استفاده از مدل‌سازی موضوعی و الگوریتم تخصیص پنهان دریکله به خوشه‌بندی موضوعات پرداخت. سازمان‌دهی داده‌های پزشکی در این مدل به‌گونه‌ای انجام‌شده در تشخیص‌های پزشکی، پشتیبانی تصمیم‌گیری بالینی، بهبود عادات کیفیت زندگی و کاهش هزینه‌های پزشکی مفید واقع شد. نتایج این پژوهش نیز در تأیید نتایج پژوهش حاضر و کارایی مدل‌سازی موضوعی و همچنین الگوریتم تخصیص پنهان دریکله است.

موضوعاتی که بیشتر مورد تمرکز پژوهش‌گران ایرانی در حوزه غدد بوده است بر این اساس در ۱۰ دسته موضوعی علائم و درمان، بیماری‌های قلبی و عروقی، عوارض دیابت، سندرم متابولیک، دیابت نوع ۲، بیماری‌های متابولیک استخوان، سرطان‌ها در حوزه غدد، دیابت، سبک زندگی و رژیم غذایی و بیماری‌های مزمن در حوزه غدد و دسته‌های ۲۰تایی از واژه‌های مرتبط ارائه شد. با دقت نظر در نتایج به‌دست‌آمده و با توجه به واژه‌های سرعنوان موضوعی پزشکی موضوعاتی مانند بیماری آدرنال، اختلالات جنسی، تیروئید و پاراتیروئید، بیماری‌های هیپوفیز و سل-غدد کمتر مورد توجه قرار گرفته‌اند.

بررسی پژوهش‌های خارجی و داخلی نشان می‌دهند که تاکنون بررسی در حوزه موضوعی غدد درون‌ریز و متابولیسم با استفاده از مدل‌سازی موضوعی انجام نشده است. همچنین پژوهش‌هایی که در حوزه پزشکی به مدل‌سازی موضوعی پرداخته‌اند از ابزارها و روش‌های مختلفی استفاده کرده‌اند، تاکنون هیچ پژوهشی در حوزه پزشکی از نرم‌افزار متلب استفاده نکرده‌اند. با توجه به اینکه این یک ابزار مجزا و مختص به مدل‌سازی موضوعی دارد به نظر می‌رسد به‌طور مؤثری بتواند در این زمینه مفید واقع شود.

با توجه به اهمیت بیماری‌های حوزه غدد درون‌ریز و متابولیسم شناخت موضوعاتی که روی آنها پژوهش بیشتری صورت گرفته و موضوعاتی که در این حوزه روی آنها پژوهشی صورت نگرفته یا کمتر مورد پژوهش قرار گرفته‌اند ضروری است. نتایج بیانگر این است که اجرای مدل تخصیص پنهان دریکله سبب تولید موضوعاتی می‌شود که اساساً توزیع احتمالاتی کلماتی هستند که به خوبی می‌توانند یک موضوع یا محتوای خاص را توصیف کنند. بررسی چندین‌باره مقالات بیانگر این است که ممکن است واژه‌ها در هر عنوان موضوعی کاملاً شبیه به هم نباشند اما به‌طور قطع با هم مرتبط هستند. تعدادی از موضوعات که از آموزش بدون نظارت ایجاد شده‌اند خنثی هستند بدین معنی که

1 . Wang
2 . Song

نمی‌توان آنها در خوشه‌بندی خاصی ارائه کرد. همچنین در برخی خوشه‌های موضوعی کلمه‌های غیرمرتبط به زمینه وجود دارند اما اگر به‌درستی و با توجه به مقالاتی که از آن برآمده‌اند تجزیه و تحلیل شوند تا حدودی متوجه ارتباط آنها خواهیم شد. به‌عنوان مثال واژه mortality در دسته موضوعی بیماری‌های قلبی و عروقی احتمالاً در خصوص آن دسته از بیماری‌های قلبی و عروقی اشاره دارد که سبب مرگ‌ومیر در این حوزه می‌شوند و یا واژه‌های male, gender, men به درصد و تعداد ابتلا زنان و مردان و مسائل جنسیتی در حوزه‌های مختلف اشاره دارند.

همان‌طور که مشاهده شد الگوریتم تخصیص پنهان دریکله عملکرد قابل قبولی در این پژوهش و دسته‌بندی موضوعات حوزه غدد داشته است. دسته‌های موضوعی استخراج‌شده دارای تجانس و ارتباط موضوعی خوبی با یکدیگر هستند. نتایج نشان‌دهنده فعالیت متخصصین حوزه غدد است و بیانگر توجه و تمرکز آنها در بخش‌هایی از این حوزه موضوعی می‌باشد. بنابراین این الگوریتم می‌تواند جهت تحلیل و بررسی موضوعات سایر حوزه‌های پزشکی و غیرپزشکی مورد استفاده قرار گیرد.

پیشنهادهای اجرایی

موضوعات اختلالات جنسی، بیماری سل-غدد، بیماری‌های آدرنال و بیماری‌های خود ایمنی در بیماری‌های غدد درون‌ریز و متابولیسم از دید پژوهشگران ایرانی مورد اغفال قرار گرفته‌اند که جای کار و پژوهش بیشتری دارند.

از نتایج پژوهش حاضر در سیاست‌گذاری‌های رشته غدد استفاده شود.

پیشنهاد برای پژوهش‌های آتی

از این روش برای پژوهش در سایر حوزه‌ها جهت بهبود و ارتقای جایگاه این حوزه‌ها صورت پذیرد.

تشکر و قدردانی

از آقای روح‌الله احسانی بابت همکاری در تهیه کدهای مربوطه و راهنمایی‌های بی‌دریغ‌شان در اجرای مدل در نرم‌افزار متلب نهایت تشکر را داریم.

از خانم دکتر ناهید هاشمی مدنی، دکتر ملیحه قدیر، دکتر آتوسا نجم‌الدین و دکتر هدا طاهری (فوق تخصص غدد درون‌ریز و متابولیسم) بابت همکاری‌های صمیمانه ایشان در بررسی‌های تخصصی سپاسگزاریم.

فهرست منابع

حشمتی، هاشم، بهنام‌پور، ناصر، خراسانی، فرشته، و مقدم، زهرا. (۱۳۹۲). شیوع عوارض مزمن دیابت و برخی عوامل مرتبط آن در بیماران دیابتی نوع دو مراجعه کننده به مرکز دیابت شهرستان فریدون‌کنار. مجله دانشکده علوم پزشکی نیشابور، ۱ (۲).

شکرچیان چالشتری، رضا. (۱۳۹۵). مدل‌سازی موضوعی با استفاده از خوشه‌بندی برای اسناد دامنه خاص. پایان‌نامه کارشناسی ارشد. دانشگاه تهران.

صابری، محمدکریم، و اسفندیاری مقدم، علیرضا. (۱۳۹۰). بررسی میزان دسترس‌پذیری و زوال استنادهای وبی مقالات نمایه‌شده در مؤسسه اطلاعات علمی (ISI) در حوزه اطلاعات سلامت و کتابداری پزشکی. مدیریت اطلاعات سلامت، ۸ (۲)، ۱۸۹-۱۹۷.

قاضی میرسعید، جواد، و صنیعی، نادیا. (۱۳۹۴). ارزیابی جایگاه علمی مراکز تحقیقاتی غدد درون‌ریز، دیابت و متابولیسم دانشگاه‌های علوم پزشکی کشور به روش Exergy. مجله علمی دانشگاه علوم پزشکی کردستان، ۲۰ (۵)، ۱۱۰- ۱۱۹.

لاریجانی، باقر، و دیگران. (۱۳۹۵). چارچوب ملی ارائه خدمت در بیماری دیابت در راستای سند ملی پیشگیری و کنترل بیماری‌های غیرواگیر. تهران: کمیته ملی پیشگیری و کنترل بیماری‌های غیرواگیر.

مسعودی، بابک، و راحتی، سعید. (۱۳۹۴). رفع ابهام معنایی واژگان مبهم فارسی با مدل موضوعی LDA. فصل‌نامه علمی پژوهشی پردازش علائم و داده‌ها، ۱۲ (۴)، ۱۱۷-۱۲۵.

Bastani, K., Namavari, H., & Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints. 127, 256-271.

Blei, M. D. (2017). Latent dirichlet allocation. J Mach Learn Res. 3:993–1922.

Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003). Latent Dirichlet Allocation. Journal of machine Learning research, 3,993–1922.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. 41(6), 391-407.

Golden, S.H., Robinson, K.A., Saldanha, I. Anton, B. & Ladenson, W. (2009). Prevalence and Incidence of Endocrine & Metabolic Disorders in the United States: A Comprehensive Review. J Clin Endocrinol Metab,94(6),1853–1878.

Hidayatullah, A. F., Aditya, S. K., & Karimah, S. T. (2019). Topic modeling of weather and climate condition on twitter using latent dirichlet allocation (LDA). Paper presented at the IOP Conference Series: Materials Science and Engineering. IOP Publishing.

Hofmann, T. (1999). Probabilistic latent semantic indexing. Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. Mach Learn,42(1–2),177–196.

Howes, C., Purver, M. & McCabe, R. (2013). Using conversation topics for predicting therapy outcomes in schizophrenia. Biomed Inf Insights, 6, BII. S11661.

Kandula, S., Curtis, D., Hill, B., & Zeng-Treitler, Q. (2011). Use of topic modeling for recommending relevant education material to diabetic patients. Paper presented at the AMIA annual symposium proceedings.

Liu, L., Tang, L., Dong, W, Shaowen, Y. & Zhoucorresponding W. (2016). An overview of topic modeling and its current applications in bioinformatics. SpringerPlus,5(1),1608.

Mehler, A. & Waltinger, U. (2009). Enhancing document modeling by means of open topic models Crossing the frontier of classification schemes in digital libraries by example of the DDC. Library Hi Tech,27(4),520-539.

Özmutlu, S. & Çavdur, F. (2005) Neural network applications for automatic new topic identification. Online Information Review,29(1),34-53.

Park, S., Choi, D., Lee, W., Jung, D., Kim, M., & Moon, C. (2014). Disease-medicine topic model for prescription record mining. Paper presented at the 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC).

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. Paper presented at the Proceedings of the 2009 conference on empirical methods in natural language processing.

Robinson, S.D. (2019). Temporal topic modeling applied to aviation safety reports: A subject matter expert review. Safety Science. 116, 275-286.

Sendhilkumar, S., Nachiyar, S.N., & Mahalakshmi, G.S. (2013). Novelty Detection via Topic Modeling in Research Articles. Paper presented at the Proceedings of international conference ICCSEA.

Song, C.W., Jung, H., & Chung, K. (2017). Cluster Comput. Development of a medical big-data mining process using topic modeling. Cluster Computing, (22),1949–1958.

Taghva, K., Russell, B. and Sadeh, M. (2003). A list of farsi stopwords. Retrieved Sept,7(2).

Wang, L., Zhang, Y., Zhang, Y., Xu, X., & Cao, S. (2017). Prescription function prediction using topic model and multilabel classifiers. Evidence-Based Complementary and Alternative Medicine, 2017.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. Paper presented at the BMC bioinformatics.

Verheggen, K., Ræder, H., Berven, Frode S., Martens, L., & Barsnes, H. (2020). Anatomy and evolution of database search engines—a central component of mass spectrometry based proteomic workflows. mass spectrometry review,39(3),292-306.

پیوست‌ها

پیوست ۱. دسته موضوعی استخراج‌شده

موضوع ۱. علائم و درمان

Disease and treatment		
کلمه	احتمال	تکرار
Patient	0.213422	12282
Treatment	0.04204	3026
Disease	0.036613	4492
Infection	0.028564	621
Therapy	0.017157	1210
Test	0.016053	1591
Diagnosis	0.013891	978
disorder	0.012925	1123
symptom	0.009659	407
Diagnose	0.009475	600
Leukemia	0.008969	195
Regimen	0.008601	187
Hsct	0.008509	185
Treat	0.007589	977
Chronic	0.007083	805
Gvhd	0.006761	147
Mortality	0.006531	618
Marrow	0.006117	512
sexual	0.005336	116
thalassemia	0.005014	122

موضوع ۲. بیماری‌های قلبی و عروقی

Cardiovascular disease		
کلمه	احتمال	تکرار
patient	0.134784	12282
diabetes	0.064576	7145
Diabetic	0.051643	4359
Disease	0.051462	4492
Age	0.036366	3057
Coronary	0.029607	657
mellitus	0.02821	2274
Artery	0.022667	503
Heart	0.021721	565
Mortality	0.019963	618
Hypertension	0.015952	748
Retinopathy	0.012978	288
Cardiovascular	0.012888	838
Stroke	0.011176	248
Male	0.01068	1125
Death	0.01032	363
Eye	0.010184	226
gender	0.009553	441
Old	0.009508	581
Man	0.009193	966

موضوع ۳. عوارض دیابت

Complications of Diabetes		
کلمه	احتمال	تکرار
treatment	0.059828	3026
Test	0.020501	1591
medicine	0.012923	258
chronic	0.012923	805
therapeutic	0.010131	384
global	0.008376	105
therapy	0.008137	1210
endocrine	0.007817	366
gastric	0.0071	89
periodontal	0.00702	88
secretion	0.006621	304
Adult	0.006222	672
rural	0.006222	78
pathogenesis	0.006142	296
melatonin	0.005903	74
relate	0.005743	176
diseases	0.005664	71
periodontitis	0.005504	69
inflammatory	0.004946	409
basic	0.004866	61

موضوع ٤. سندرم متابولیک

Metabolic Syndrome		
کلمه	احتمال	تکرار
metabolic	0.059163	1652
Mets	0.047891	922
Age	0.04571	3057
Bmi	0.043528	893
obesity	0.043164	831
Man	0.034542	966
cardiovascular	0.020881	838
Cvd	0.020569	396
hypertension	0.020465	748
Adult	0.019375	672
weight	0.019271	996
Obese	0.018907	397
diabetes	0.01818	7145
predict	0.014856	316
systolic	0.014076	271
Sex	0.013297	617
predictor	0.013297	381
anthropometric	0.01231	328
diastolic	0.012051	232
criterion	0.011427	499

موضوع ۵. دیابت نوع ۲

Type2 Diabetes		
کلمه	احتمال	تکرار
glucose	0.080416	2794
Serum	0.070608	3712
diabetes	0.05408	7145
Fast	0.049494	1142
patient	0.046996	12282
Lipid	0.042501	1180
resistance	0.041684	1129
Cholesterol	0.035418	907
plasma	0.030695	1204
t2dm	0.029832	904
hba1c	0.028016	617
Lipoprotein	0.026336	580
diabetic	0.025156	4359
Triglyceride	0.021932	566
Adiponectin	0.020978	462
mellitus	0.018753	2274
glycemic	0.017346	382
Sugar	0.015121	375
Hdl	0.014076	310
Leptin	0.013441	296

موضوع ۶. بیماری متابولیک استخوان

Metabolic bone disease		
کلمه	احتمال	تکرار
bone	0.03816	2329
tumor	0.035429	608
disease	0.028427	4492
diagnosis	0.023106	978
lesion	0.019185	302
tnfalpha	0.015614	223
adrenal	0.014844	212
lung	0.011833	169
injury	0.010153	251
brain	0.009873	259
tissue	0.009172	836
diagnostic	0.009032	243
biopsy	0.008192	117
inflammatory	0.007982	409
diagnose	0.007842	600
mri	0.007492	107
drug	0.006862	640
tooth	0.006652	95
symptom	0.006582	407
pathological	0.006512	93

موضوع ۷. سرطان‌ها در حوزه غدد

Endocrine cancers		
کلمه	احتمال	تکرار
cancer	0.067159	837
disease	0.040359	4492
mscs	0.027281	340
immune	0.024874	310
endothelial	0.019417	242
cytokine	0.019257	286
breast	0.015566	194
growth	0.012678	555
chronic	0.010993	805
oxide	0.010912	207
nitric	0.01003	176
hepatitis	0.009709	121
tumor	0.008184	608
organ	0.007623	208
prostate	0.007542	94
lymphocyte	0.007221	90
antigen	0.007221	117
vegf	0.007221	90
death	0.007061	363
malignant	0.006981	131

موضوع ۸. دیابت

Diabetes		
کلمه	احتمال	تکرار
diabetic	0.137459	4359
diabetes	0.072167	7145
glucose	0.030035	2794
treatment	0.02957	3026
animal	0.02714	645
Treat	0.023056	977
mellitus	0.022591	2274
Male	0.019955	1125
injection	0.017111	517
Hyperglycemia	0.014268	418
Streptozotocin	0.013906	269
pancreatic	0.013803	267
Stz	0.012614	244
Oil	0.01251	242
weight	0.012148	996
enzyme	0.011735	399
Liver	0.009874	761
Islets	0.009667	187
Test	0.009202	1591
Acid	0.00884	847

موضوع ۹. سبک زندگی و رژیم غذایی

Lifestyle and diet		
کلمه	احتمال	تکرار
child	0.082503	977
dietary	0.055565	658
diet	0.051005	604
age	0.034031	3057
food	0.033609	398
adolescent	0.021365	253
t1dm	0.018494	219
fat	0.017902	344
girl	0.016889	200
healthy	0.014862	1087
nutritional	0.012413	147
weight	0.01216	996
lifestyle	0.011316	191
disease	0.011062	4492
adult	0.010725	672
selfcare	0.010556	125
nutrition	0.010218	121
milk	0.010049	119
gender	0.009458	441
Vegetable	0.009205	109

موضوع ۱۰. بیماری‌های مزمن در حوزه غدد

Endocrine Cronic disease		
کلمه	احتمال	تکرار
disease	0.061458	4492
disorder	0.041926	1123
cardiac	0.026931	293
ckd	0.019335	196
skin	0.018349	219
diagnosis	0.015192	978
pulmonary	0.015192	154
congenital	0.015093	200
respiratory	0.01243	126
limb	0.010654	108
symptom	0.010161	407
chronic	0.009174	805
organ	0.008385	208
male	0.008286	1125
cah	0.007991	81
tuberculosis	0.007991	81
curcumin	0.007497	76
pon1	0.007399	75
hyperplasia	0.0073	74
bmmscs	0.006905	70