مجله علم سنجی کاسپین
Caspian Journal of Scientometrics
e-ISSN: 2423-4710 p-ISSN: 2383-157X
دانشگاه علوم پزشکی
و خدمات بهداشتی درمانی بابل

[DOR: 20.1001.1.24234710.1401.9.1.9.5]
[DOI: 10.22088/cjs.9.1.104]

Analytical Comparison of Iranian Scientific Documents in Text Mining

Mohadeseh Rafiee (MA) [1]*, Abdalsamad Keramatfar (PhD) [1]

1. SID, Academic Center for Education, Culture and Research (ACECR), Tehran, Iran.

ABSTRACT

Article Type:
Research Paper

Background and aim: Policymakers seek to evaluate their country's scientific performance and measure it in terms of effectiveness and problem-solving. The aim of this study was to make an analytical comparison of Iranian scientific documents in text mining based on domestic and foreign databases.

Materials and methods: The present study is descriptive survey with a bibliometric approach. In order to find scientific documents related to text mining in the Scopus database, related terms were searched, and then the results were limited to Iran. Scientific Information Database (SID) was used to search for Persian scientific documents. Bibexcel, VOSviewer, Python programming language, and Excel 2017 were used to analyze the data.

Findings: The total number of Iranian scientific documents in text mining in the Scopus citation database was 1082 and 284 (26.25%) of scientific documents indexed in Scopus were in Persian. Moreover, according to the Scientific Information Center, the number of scientific documents in this field was 89 and the number of scientific documents in Persian was 51 (57.30%). The Journal of Lecture Notes in Computer Science has published most international scientific papers in Iran, and the Journal of Signal and Data Processing has published most domestic scientific papers in Iran in text mining. A t-test was used to determine that there was a significant difference in the number of scientific documents in Persian between Scopus and SID databases ($p<0.0001$).

Conclusion: The average growth rate of Iranian scientific documents in text mining was higher than in other subject areas. The United States, Britain, and Australia have had the most collaboration with Iranian researchers in this field. It was also found that international scientific documents in English received more citations than scientific documents in Persian.

Received:
23 Aug. 2020
Revised:
26 June 2022
Accepted:
2 July 2022

Keywords: **Data mining, Text mining, Evaluating science, Bibliometrics, Natural language processing**

Cite this article: Rafiee M, Keramatfar A. Analytical Comparison of Iranian Scientific Documents in Text Mining. *Caspian Journal of Scientometrics*. 2022; 9(1): 104-116.

Publisher: Babol University of Medical Sciences

***Corresponding Author:** Mohadeseh Rafiee
Address: SID, Academic Center for Education, Culture and Research (ACECR), Tehran, Iran.
E-mail: mohadeseh.rafie2012@gmail.com

مجله علم‌سنجی کاسپین
Caspian Journal of Scientometrics
e-ISSN: 2423-4710 p-ISSN: 2383-157X
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی بابل

مقایسه تحلیلی مدارک علمی ایران در حوزه موضوعی متن‌کاوی

محدثه رفیعی خشنود (MA) [۱*]، عبدالصمد کرامت فر (PhD) [۱]

۱. مرکز اطلاعات علمی جهاد دانشگاهی، تهران، ایران.

چکیده

نوع مقاله:
مقاله پژوهشی

سابقه و هدف: سیاست‌گذاران تلاش می‌کنند تا عملکرد علمی کشور خود را مورد ارزیابی قرار داده و آن را از نظر اثربخشی و حل مشکلات مورد سنجش قرار دهند. این مقاله به مقایسه تحلیلی مدارک علمی ایران در حوزه موضوعی متن‌کاوی بر اساس پایگاه‌های داخلی و خارجی می‌پردازد.

مواد و روش‌ها: پژوهش حاضر از نوع توصیفی- پیمایشی و با رویکرد کتاب‌سنجی انجام شده است. برای بازیابی مدارک علمی مرتبط با متن‌کاوی در پایگاه اسکوپوس عبارات مرتبط با آن جستجو و سپس نتایج به ایران محدود شد. برای بازیابی مدارک علمی مجلات داخلی از پایگاه مرکز اطلاعات علمی جهاد دانشگاهی به شیوه مشابه استفاده شد. برای تجزیه و تحلیل داده‌ها از نرم‌افزارهای Bibexcel، Vosviewer، زبان برنامه‌نویسی Python و Excel استفاده شد.

یافته‌ها: تعداد کل مدارک علمی ایران در حوزه موضوعی متن‌کاوی در پایگاه استنادی اسکوپوس، برابر با ۱۰۸۲ است. ۲۸۴ مدرک علمی (۲۶/۲۵٪) از مدارک علمی نمایه‌شده در اسکوپوس، بر زبان فارسی متمرکز هستند. همچنین بر اساس داده‌های پایگاه مرکز اطلاعات علمی، تعداد مدارک علمی این حوزه موضوعی برابر با ۸۹ و مدارک علمی متمرکز بر زبان فارسی برابر با ۵۱ (۵۷/۳۰٪) است. مجله Lecture notes in computer science بیشترین تعداد مدارک علمی بین‌المللی ایران و مجله پردازش علایم و داده‌ها، بیشترین تعداد مدارک علمی داخلی ایران را در حوزه موضوعی متن‌کاوی منتشر کرده‌اند. با استفاده از آزمون تی مستقل مشخص شد بین تعداد مدارک علمی متمرکز بر زبان فارسی پایگاه اسکوپوس و مرکز اطلاعات علمی جهاد دانشگاهی، تفاوت معناداری وجود دارد ($p<۰/۰۰۰۱$).

نتیجه‌گیری: میانگین نرخ رشد مدارک علمی ایران در حوزه متن‌کاوی بالاتر از حوزه‌های موضوعی دیگر است. کشورهای آمریکا، انگلیس و استرالیا بیشترین میزان مشارکت را با محققان ایرانی در این حوزه موضوعی داشته‌اند. همچنین مشخص شد مدارک علمی بین‌المللی که بر زبان انگلیسی متمرکز هستند، استناد بیشتری نسبت به مدارک علمی متمرکز بر زبان فارسی دریافت می‌کنند.

دریافت:
۱۳۹۹/۶/۲
ویرایش:
۱۴۰۱/۴/۵
پذیرش:
۱۴۰۱/۴/۱۱

واژگان کلیدی: **داده‌کاوی، متن‌کاوی، ارزیابی علم، کتاب‌سنجی، پردازش زبان طبیعی**

استناد: محدثه رفیعی خشنود، عبدالصمد کرامت فر. مقایسه تحلیلی مدارک علمی ایران در حوزه موضوعی متن‌کاوی. مجله علم‌سنجی کاسپین. ۱۴۰۱؛ ۹(۱): ۱۱۶-۱۰۴.

Publisher: Babol University of Medical Sciences

---

* **مسئول مقاله:** محدثه رفیعی خشنود

آدرس: تهران، مرکز اطلاعات علمی جهاد دانشگاهی.

رایانامه: mohadeseh.rafie2012@gmail.com

مقدمه

اخیراً، حجم بالایی از داده‌های متنی، به دلیل تنوع شبکه‌های اجتماعی، وب ۲ و سایر نرم‌افزارهای مبتنی بر متن به وجود آمده است (۱). برای شناخت و کشف این الگوهای متنی، متن‌کاوی اولین بار توسط فلدمن و همکاران مطرح شد (۲). متن‌کاوی از حوزه‌های موضوعی مثل بازیابی اطلاعات، استخراج اطلاعات، خوشه‌بندی، مصورسازی، تکنولوژی پایگاه داده، یادگیری ماشین و داده‌کاوی تشکیل شده است (۳). برای شناخت و درک الگوهای پنهان در متن، از مجموعه داده‌های متنی مثل شبکه‌های اجتماعی، وب‌سایت‌های خرید و فروش کالا، مجموعه داده‌های (Data sets) گوگل (Google)، داده بانک جهانی (World Bank Data)، آمازون (Amazon) و غیره در زبان انگلیسی و در زبان فارسی شامل فارس‌تیل (FarsTail)، فارسی ان ال پی (Farsinlp)، سای تو لب (sci2lab) دیجی‌کالا و غیره استفاده می‌شود.

یکی از راه‌های سنجش پیشرفت هر کشور، سنجش پژوهش‌های منتشرشده آن در غالب مدارک علمی منتشرشده است (۴). مطالعات کتاب‌سنجی، به کسانی که وظیفه تصمیم‌گیری و ارائه خط مشی‌های تحقیقاتی را دارند، کمک می‌کند تا به بودجه‌ریزی مناسب در علوم و فناوری بپردازند (۵). کتاب‌سنجی یک حوزه مطالعاتی است که از داده‌های کتاب‌شناختی، مدارک علمی و روابط استنادی برای ارزیابی و آشکارسازی ساختار یک حوزه تحقیق استفاده می‌کند (۶)، دو شاخص مهم در کتاب‌سنجی، تعداد مدارک علمی منتشرشده و تعداد استنادات است (۷). مطالعات کتاب‌سنجی بر اساس پایگاه‌های استنادی در سطح بین‌الملل، نظیر اسکوپوس (Scopus) و وب‌آوساینس (Web of Science) انجام می‌شود.

در حوزه کتاب‌سنجی مطالعات متعددی در کشور انجام شده است (۸). در اینجا به بررسی برخی از این موارد پرداخته می‌شود. می‌توان مشاهده نمود که اکثر مطالعات به بررسی و مقایسه وضعیت کلان کشور تمرکز داشته‌اند. فرزین یزدی و رضایی شریف آبادی در بررسی تولیدات علمی حوزه موضوعی هوش مصنوعی در کشورهای خاورمیانه طی سال‌های ۱۹۹۶ تا ۲۰۱۴، به این نتیجه رسیدند که کشورهای خاورمیانه، تنها ۴/۰۳ درصد انتشارات جهان در حوزه هوش مصنوعی را به خود اختصاص داده‌اند. ایران از نظر تعداد تولیدات علمی و تعداد مدارک قابل استناد در رتبه ۱۷ جهانی و با تولید ۵۱۵۶ مدرک دارای رتبه اول در خاورمیانه است؛ اما از نظر وضعیت استنادی، مشارکت بین‌المللی در تولید علم و پیشرفت علمی نیاز به تقویت و توسعه دارد و رژیم صهیونیستی بر اساس شاخص‌های هرش، تعداد استنادات، متوسط استناد به هر سند و پیشرفت علمی در رتبه اول خاورمیانه قرار دارد (۹). کرامت‌فر و رفیعی خشنود در پژوهشی که به بررسی مدارک علمی پژوهشگاه رویان پرداختند به این نتیجه رسیدند که پژوهشگاه رویان یکی از موفق‌ترین موسسات پژوهشی ایران در حوزه پزشکی است و با شتاب بیشتری از کل کشور در حال انجام تحقیق و پژوهش است. همچنین، بررسی شاخص‌های کیفیت مدارک علمی این پژوهشگاه نیز نشان‌دهنده کیفیت نسبی مدارک علمی آن است (۱۰). ورزنه، بهمنی و قادری آزاد در مطالعه‌ای با بررسی وضعیت تولیدات علمی ایران در حوزه‌ی انرژی و سوخت و مقایسه‌ی آن با کشورهای خاورمیانه به این نتیجه رسیدند که روند مدارک علمی ایران در این حوزه رو به افزایش است، اگرچه ایران از لحاظ تعداد مدارک، جایگاه خوبی در خاورمیانه دارد؛ اما از لحاظ تعداد استنادات و شاخص هرش عملکرد ضعیفی داشته است. همچنین با بررسی موضوعات کار شده در مقالات پراستناد دنیا مشخص شد این مقالات بیشتر روی موضوعات مربوط به انرژی‌های تجدیدپذیر تمرکز داشته‌اند؛ در حالی که بیشترین تمرکز پژوهشی ایران روی موضوعات مربوط به انرژی‌های تجدیدناپذیر است (۱۱). عطایی آشتیانی با مطالعه مدارک علمی ایران و چین و با محاسبه نرخ رشد مدارک علمی در دو دوره ۱۹۹۴–۱۹۸۰ و ۲۰۰۹–۱۹۹۵ به این نتیجه رسید که بیشترین نرخ رشد در جهان مربوط به ایران است. سپس با بررسی نرخ مدارک علمی ریترکت شده نرمال شده (Normalized Ratio of Retracted Documents) به این نتیجه رسید که بالاترین این نرخ مربوط به چین و ایران است (۱۲).

برنگی و خاصه، با تحلیل جایگاه جهانی ایران در پژوهش‌های علوم کامپیوتر با به‌کارگیری فنون علم‌سنجی به این نتیجه رسیدند که پژوهش‌های علوم کامپیوتر ایران در پایگاه ISI طی دوره سی ساله که مورد بررسی قرار گرفته، رشدی نسبی داشته است که نشان از سهم ۰/۹۸۶ درصدی ایران و کسب رتبه ۲۴ در میان کلیه کشورهای جهان دارد که رتبه چندان مناسبی به شمار نمی‌آید. ازآنجا که تحقق اهدافی نظیر دولت الکترونیک و پیشرفت در دیگر علوم تا حدود زیادی به توسعه علوم کامپیوتر بستگی دارد، نیاز به سرمایه‌گذاری و برنامه‌ریزی جامع در این حوزه بیش از پیش محسوس است (۱۳).

همچنین مطالعات متن‌کاوی نیز در ادبیات پژوهشی کشور با استفاده از روش‌های آن به تحلیل‌های مختلفی نظیر خلاصه‌سازی متن و تحلیل کاربران پرداخته است. کشاورزیان و براردخت، در پژوهشی با عنوان جایگاه کتاب و کتاب‌خوانی در سایت تبیان با رویکرد متن‌کاوی و تحلیل شبکه‌های اجتماعی با استفاده از روش خوشه‌بندی موضوعی و متن‌کاوی به بررسی مطالعه در این سایت پرداخت. این پژوهش با استفاده از روش خوشه‌بندی موضوعی و متن‌کاوی، به همراه مشخص‌کردن خوشه‌های موضوعی برجسته و کم‌اهمیت در میان کاربران شبکه تبیان پرداختند و به این نتیجه رسیدند که با توجه به ماهیت مجازی و الکترونیکی بودن سایت تبیان، لازم است بخش آموزش‌های الکترونیکی این سایت با تولید محتوای آموزشی بیشتر این کمبود را پوشش دهد (۱۴) . قانع، سپیدنام و جعفری به خلاصه‌سازی متون فارسی با استفاده از الگوریتم یادگیری عمیق پرداختند. با توجه به رشد روزافزون اطلاعات و افزایش سریع حجم داده‌های دیجیتالی، نیاز به خلاصه‌سازی متون احساس شد. خلاصه‌سازی متون و اخبار یکی از نیازهای لازم در حوزه پردازش زبان طبیعی است. روش‌های مورد استفاده در این پژوهش، آمار، وردنت، روش‌های مبتنی بر گراف و غیره می‌باشند. خلاصه‌سازهای ماشینی، سیستم‌های تصمیم یار، سیستم‌های پاسخگو،

موتورهای جستجو و غیره هستند. بر این اساس، برای خلاصه‌سازی متون با استفاده از الگوریتم‌های یادگیری عمیق صورت گرفت. نمایش یک سند متن فارسی را عددی به دست آوردند، به طوری که معنای کلمات آنها حذف نشود. نتایج تحقیق نشان داد که مدل درخت، تصمیم با میزان دقت ۱۰۰٪ بهترین مدل اعمال شده است و مدل شبکه عصبی با دقت ۹۸٪ در رتبه دوم قرار دارد (۱۵).

با بررسی مقالات مشخص شد که بیشتر آنها بر موضوع تعداد مقالات علمی و استناد آن توجه داشته‌اند و کمتر به تحلیل متنی آن پرداختند، هم‌زمان با توجه به رشد سریع مدارک علمی ایران در سطح جهان (۱۶) و ماهیت حوزه موضوعی متن‌کاوی (۱۷) این مقاله به بررسی مدارک علمی ایران در این حوزه موضوعی و در پایگاه‌های استنادی اسکوپوس و مرکز اطلاعات علمی جهاد دانشگاهی می‌پردازد. در این پژوهش، مدارک علمی حوزه موضوعی متن‌کاوی را به دو دسته، استفاده از داده فارسی و داده انگلیسی تقسیم می‌کند تا مشخص شود آیا مدارک علمی ایران که در پایگاه استنادی اسکوپوس نمایه می‌شوند به زبان فارسی اهمیت بیشتری می‌دهند یا به زبان انگلیسی؟

مواد و روش‌ها

پژوهش حاضر از نوع توصیفی- پیمایشی و با رویکرد کتاب‌سنجی انجام شده است. برای انجام این مقاله از دو پایگاه استنادی اسکوپوس و مرکز اطلاعات علمی استفاده شده است. متن‌کاوی برای تجزیه و تحلیل متون می‌پردازد و با داده‌های متنی بزرگ سروکار دارد. به این منظور در این پژوهش، برای تحلیل متن مدارک علمی ایران، آنها را به دودسته تقسیم می‌کنیم: فارسی و انگلیسی. مدارک علمی که از داده‌های فارسی استفاده کرده‌اند، به‌عنوان مدرک علمی متمرکز بر زبان فارسی و آنهایی که از مجموعه داده زبان انگلیسی استفاده کرده‌اند، متمرکز بر زبان انگلیسی است. در واقع می‌توان چنین تحلیل کرد که مدارک علمی که از مجموعه داده‌های فارسی استفاده می‌کنند به‌صورت مستقیم بر نیازهای متن‌کاوی زبان فارسی تکیه دارند و مدارک علمی که از داده‌های زبان انگلیسی استفاده کرده‌اند به حل چالش‌های این زبان متمرکز هستند.

به‌منظور بررسی مدارک علمی بین‌المللی ایران در حوزه متن‌کاوی، از پایگاه استنادی اسکوپوس استفاده شد. اسکوپوس، بزرگ‌ترین پایگاه استنادی دنیا است که به نمایه مجلات علمی می‌پردازد. این پایگاه برای کمک به محققان فراهم شده است (۱۸). برای بازیابی داده‌ها در اسکوپوس از راهبرد جستجوی زیر استفاده شد:

"text mining" or "text processing" or "classification of text" or "natural language processing" or "text classification" or "text analysis" or "text clustering" or "text categorization or nlp" AND NOT "neuro-linguistic programming"

در قسمت عنوان، چکیده و کلمات کلیدی جستجو و سپس به کشور ایران محدود شد. برای بازیابی اطلاعات مدارک علمی که بر زبان فارسی متمرکز بودند، از روش زیر استفاده شد:

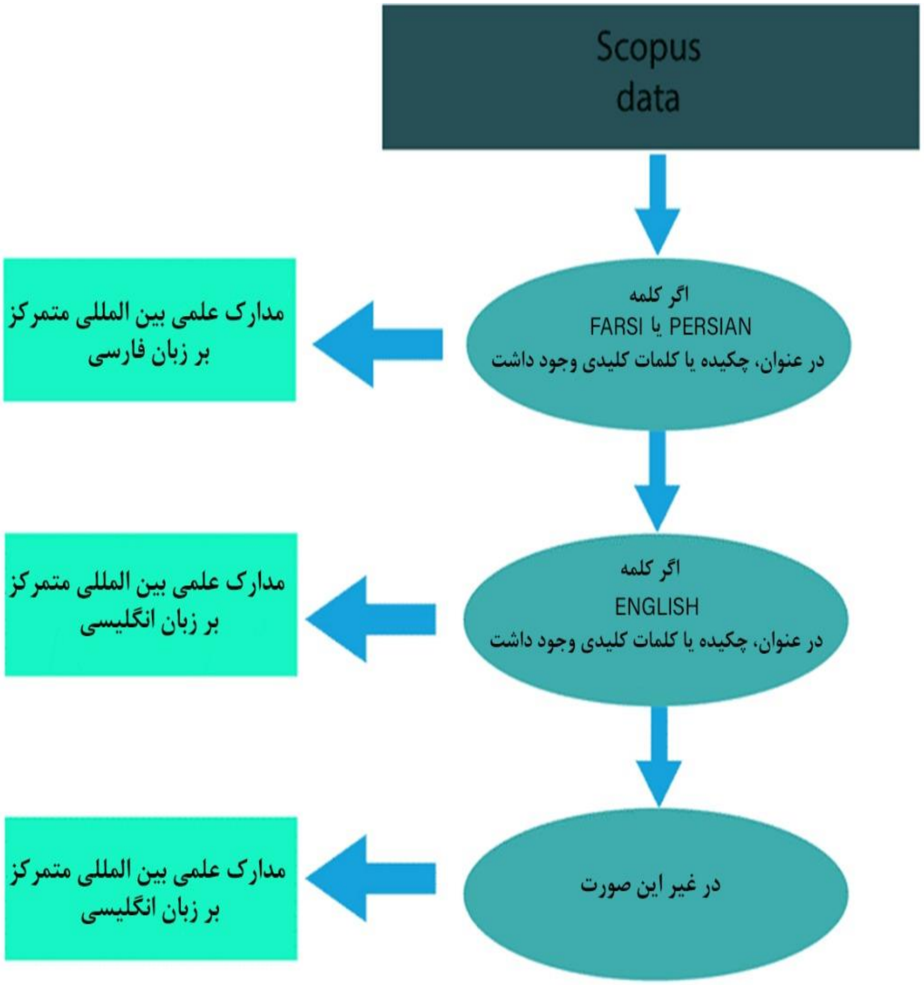
ابتدا کلمات Farsi و Persian جستجو و با دیتاست موردنظر and شد و برای بازیابی مدارک علمی که از زبان انگلیسی استفاده کرده‌اند، کلمه English جستجو و با دیتاست موردنظر and شد. سپس برای بازیابی مدارک علمی که برچسب فارسی یا انگلیسی نداشتند، از این روش استفاده شد:

اگر در مدرک علمی زبان ذکر نشده باشد، بر زبان انگلیسی تمرکز دارد. دلیل این امر این است که مدارک علمی متن‌کاوی به‌صورت معمول روی زبان انگلیسی کار می‌کنند و نیازی به ذکر آن در چکیده نیست؛ اما اگر زبان دیگری مورد بررسی باشد نویسندگان در چکیده آن را پررنگ می‌کنند. ۷۵ مدرک علمی به‌صورت تصادفی با این قاعده بررسی شد و هیچ‌کدام متمرکز بر زبان فارسی نبود؛ بنابراین این نتیجه حاصل شد که اگر در مدرک علمی عبارت زبان انگلیسی نبود، مربوط به زبان انگلیسی است.

پایگاه مرکز اطلاعات علمی جهاد دانشگاهی، تنها پایگاه علمی دسترسی آزاد ایران است که از سال ۱۳۸۳ به نمایه مجلات علمی‌پژوهشی، طرح‌های پژوهشی، کنفرانس‌ها و همایش‌های داخلی و بین‌المللی معتبر می‌پردازد (۱۹). کلمات کلیدی "متن کاوی" یا "پردازش متن" یا "طبقه بندی متن" یا "پردازش زبان طبیعی" یا "طبقه بندی متون" یا "تجزیه و تحلیل متن" یا "خوشه بندی متن" یا "دسته بندی متن" در پایگاه مرکز اطلاعات علمی جستجو شدند. سپس، همه مدارک علمی بازیابی شده، به صورت دستی بررسی شدند. تاریخ جستجو در هردو پایگاه اول شهریور ۱۴۰۰ بود.

برای تجزیه و تحلیل داده‌های مدارک علمی اسکوپوس از نرم‌افزار Bibexcel و Vosviewer استفاده شد، برای بررسی کلمات کلیدی نویسندگان از Python استفاده شد. همچنین برای تجزیه و تحلیل مدارک علمی فارسی از نرم‌افزار Excel 2017 استفاده شد. برای ترسیم ابر کلمات کلیدی نیز از لُب (http://vs1.sid.ir/cloud) استفاده شد. برای تجزیه و تحلیل داده‌ها جهت وجود رابطه معنادار، از روش‌های تحلیل آماری و آزمون تی تست (T-Test) استفاده شد.

[DOR: 20.1001.1.24234710.1401.9.1.9.5]
[DOI: 10.22088/cjs.9.1.104]



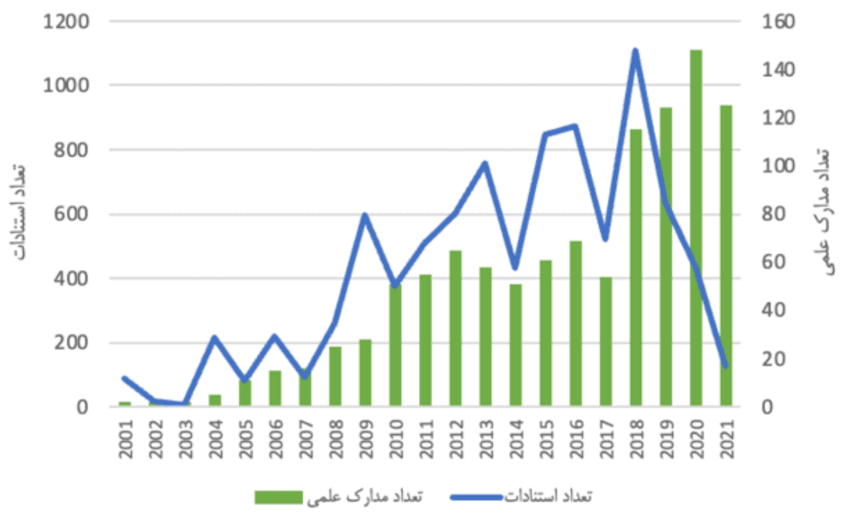
شکل ۱. دیاگرام دسته‌بندی مبتنی بر قاعده

یافته‌ها

نمودار ۱، روند مدارک علمی بین‌المللی ایران و استنادات آن در حوزه متن‌کاوی را نشان می‌دهد. بر این اساس، بیشترین تعداد مدارک علمی در این حوزه موضوعی در سال ۲۰۲۰ نگارش شده است و این حوزه موضوعی روند رو به رشدی را در سال‌های اخیر طی کرده است. تعداد کل استنادات این ۱۰۸۲ مدرک علمی حوزه موضوعی متن‌کاوی برابر با ۸۷۸۹ است. ۵۶ کشور در نگارش مدارک علمی ایران در حوزه متن‌کاوی مشارکت داشتند. شکل ۲، شبکه هم‌تالیفی کشورهای مشارکت‌کننده با محققان ایرانی در نگارش مدارک علمی مرتبط با متن‌کاوی را نشان می‌دهد. بر این اساس، محققان ایرانی بیشترین میزان هم‌تالیفی را به ترتیب با محققان کشورهای آمریکا، انگلیس و استرالیا در ۶۴، ۳۸ و ۳۵ مدرک علمی داشته‌اند. هشام فیلی دانشیار دانشکده مهندسی برق و کامپیوتر دانشگاه تهران، مهرنوش شمس فرد دانشیار دانشکده مهندسی و علوم کامپیوتر دانشگاه شهید بهشتی و مصطفی فخراحمد استادیار بخش مهندسی و علوم کامپیوتر دانشکده برق و کامپیوتر دانشگاه شیراز و بهروز مینایی بیدگلی دانشیار دانشکده علوم کامپیوتر دانشگاه علم و صنعت ایران به ترتیب با انتشار ۴۰، ۳۲، ۱۶ و ۱۶ مدرک علمی، نویسندگان ایرانی هستند که بیشترین تعداد مدارک علمی مرتبط با متن‌کاوی را منتشر کرده‌اند. محمد احسان بصیری از دانشگاه شهرکرد با دریافت ۴۰۲ استناد، موثرترین نویسنده ایرانی است. بالاترین میزان شاخص هرش نیز مربوط به مهرنوش شمس فرد با مقدار هرش هشت است. دانشگاه‌های تهران، صنعتی شریف و آزاد اسلامی، با انتشار ۱۱۳، ۴۷ و ۴۴ مدرک علمی، دانشگاه‌هایی هستند که بیشترین تعداد مدارک علمی در حوزه متن‌کاوی را منتشر کرده‌اند (جدول ۱).

[DOR: 20.1001.1.24234710.1401.9.1.9.5]

[DOI: 10.22088/cjs.9.1.104]



نمودار ۱. روند مدارک علمی ایران در حوزه متن‌کاوی در مدارک علمی بین‌المللی

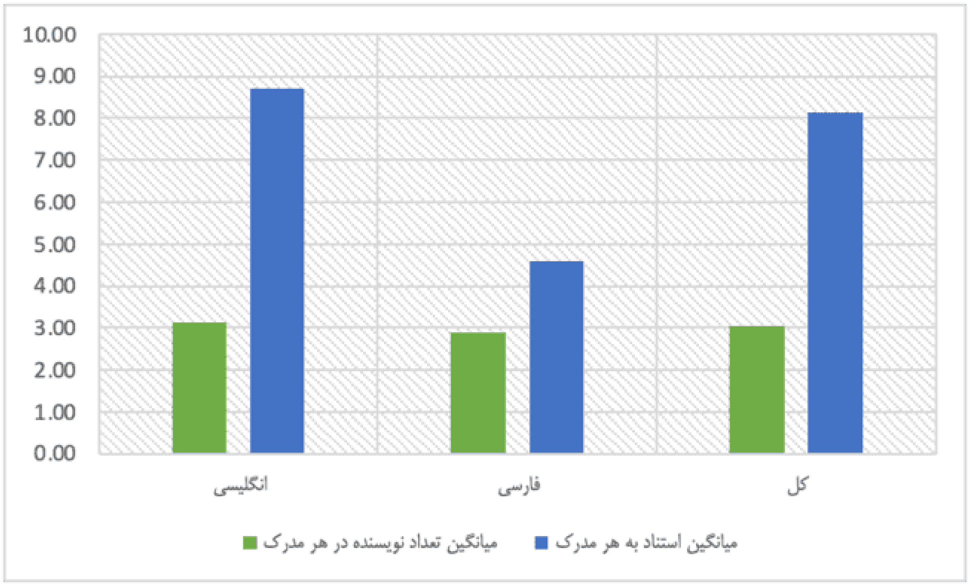
شکل ۲. شبکه هم‌تالیفی کشورهای مشارکت‌کننده با محققان ایرانی در نگارش مدارک علمی مرتبط با متن‌کاوی

جدول ۱. نویسندگان با بیشترین تعداد مدارک علمی ایران در حوزه متن‌کاوی

نویسنده	وابستگی سازمانی	تعداد مدارک علمی	تعداد استنادات	شاخص هرش
هشام فیلی	دانشگاه تهران	۴۰	۱۲۰	۶
مهرنوش شمس فرد	دانشگاه شهید بهشتی	۳۲	۲۹۲	۸
مصطفی فخراحمد	دانشگاه شیراز	۱۶	۶۴	۵
بهروز مینایی بیدگلی	دانشگاه علم و صنعت	۱۶	۵۲	۴
محمد احسان بصیری	دانشگاه شهرکرد	۱۲	۴۰۲	۷
غلامرضا قاسم خانی	دانشگاه صنعتی شریف	۱۲	۴۳	۴
سعیده ممتازی	دانشگاه صنعتی امیرکبیر	۱۲	۲۳	۳
حسین صامتی	دانشگاه صنعتی شریف	۱۲	۲۳	۲
هادی ویسی	دانشگاه تهران	۱۱	۱۲۳	۴
وب‌سایت	دانشگاه گیلان	۱۱	۳۴	۴

Lect (Lecture Notes in Computer Science Including Subseries Lecture Notes in Artificial Intelligence)، Expert Syst. Appl (Expert Systems with Applications) و RANLP (International Conference Recent Advances in Natural Language Processing) به ترتیب با انتشار ۴۶، ۱۳ و ۱۳ مدرک علمی، منابعی هستند که بیشترین تعداد مدارک علمی حوزه موضوعی متن‌کاوی ایران در آنها منتشر شده است، یک درصد از مدارک علمی ایران در این سه منبع منتشر شده است، که هیچ کدام ایرانی نیستند.

نمودار ۲، میانگین تعداد نویسنده و استناد را در هر سه نوع مدرک علمی، یعنی مدارک علمی متمرکز بر زبان انگلیسی، فارسی و کل نشان می‌دهد. میانگین استناد به هر سند مدارک علمی بین‌المللی متمرکز بر زبان فارسی برابر با ۴/۶۰ و برای مدارک علمی متمرکز بر زبان انگلیسی برابر با ۸/۷۳ است. بر این اساس، مدارک علمی متمرکز بر زبان انگلیسی، تعداد نویسنده بیشتری نسبت به سایرین داشته‌اند (۲۱). پژوهش‌ها نشان می‌دهد بین تعداد نویسنده و تعداد استنادات دریافتی رابطه وجود دارد و با افزایش تعداد نویسندگان، استنادات نیز افزایش می‌یابد (۲۲). چنانچه مشاهده می‌شود مدارک علمی بین‌المللی متمرکز بر زبان فارسی به صورت متوسط استناد کمتری دریافت کرده‌اند. این مورد خود می‌تواند یکی از دلایل توجه به زبان انگلیسی در میان محققین ایرانی باشد، چرا که محققین تمایل دارند کارهای آن‌ها بیشتر مورد توجه و استناد واقع شود.



نمودار ۲. میانگین تعداد نویسنده و استناد در هر مدرک علمی بین‌المللی

بررسی بسامد کلمات کلیدی مقیاسی مهم در روش تحلیل محتوا است. با استفاده از بسامد کلمات کلیدی می‌توان رویکردهای اصلی آن حوزه موضوعی را مورد سنجش قرار داد. کلمات کلیدی نمایی کلی از محتوای مدارک علمی هر حوزه موضوعی را نشان می‌دهد (۱۰). شکل ۳، کلمات کلیدی به کار رفته در مدارک علمی ایران در حوزه متن‌کاوی را نشان می‌دهد. بر این اساس، Natural Language Processing، Sentiment Analysis و Deep learning به ترتیب پربسامدترین کلمات کلیدی و Covid-19، Recommender systems وLatent Dirichlet Allocation جدیدترین کلمات کلیدی مدارک علمی ایران در حوزه متن‌کاوی هستند. Natural Language Processing حوزه‌ای تحقیقاتی است که به بررسی نحوه استفاده از رایانه‌ها برای درک و پردازش متن به منظور انجام کارهای مفید می‌پردازد (۲۰). منظور از پردازش زبان طبیعی این است که رایانه‌ای داشته باشیم که قادر باشد زبان انسان را تحلیل کند، بفهمد و به تولید زبان طبیعی برسد (۲۱). Sentiment analysis، دومین کلمه پربسامد در این حوزه موضوعی است. Sentiment analysis یا تحلیل احساس، به تجزیه و تحلیل نظرات، احساسات و نگرش‌ها در مورد موجودیت‌ها و جنبه‌های آنها که که غالبا به صورت متن بیان شده می‌پردازد (۲۲). تحلیل احساس می‌کوشد با استفاده از روش‌های یادگیری ماشین و پردازش زبان طبیعی، به استخراج، درک و تولید خودکار احساس در ماشین می‌پردازد (۲۵-۲۳). این حوزه از سال‌های ابتدایی قرن بیست و یکم شروع به فعالیت کرده است. یکی از دلایل عملکرد خوب بسیاری از روش‌های یادگیری ماشین، بازنمایی (Representation) و ویژگی‌های ورودی آن‌هاست که توسط انسان طراحی می‌شود. در چنین روش‌هایی، یادگیری ماشین به یافتن وزن‌هایی می‌پردازد که پیش‌بینی را بهینه کنند. در مقابل این رویکرد، یادگیری عمیق (Deep learning) وجود دارد. الگوریتم‌های یادگیری عمیق زیرمجموعه‌ای از یادگیری ماشین

[DOR: 20.1001.1.24234710.1401.9.1.9.5]
[DOI: 10.22088/cjs.9.1.104]

[DOR: 20.1001.1.24234710.1401.9.1.9.5]

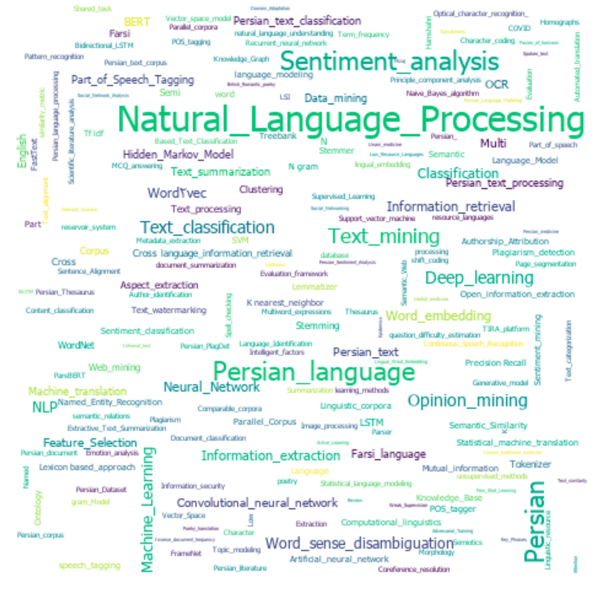
[DOI: 10.22088/cjs.9.1.104]

است که هدف آن کشف چندین سطح بازنمایی شده از داده ورودی است. استفاده از یادگیری عمیق در برخی از حوزه‌ها مثل دسته‌بندی تصاویر از قدرت انسان پیشی گرفته است (۲۶). سه ویژگی اصلی که روش‌های یادگیری عمیق که آن را نسبت به روش‌های سنتی متمایز می‌کند: اول، توانایی روش یادگیری عمیق در یادگیری مستقیم از داده‌های پردازش نشده، دوم، ساختار سلسله مراتبی و عمیق و سوم، بهینه بودن روش یادگیری عمیق نسبت به روش سنتی. در روش‌های یادگیری عمیق، ویژگی‌ها در ساختاری سلسله مراتبی و عمیق و همچنین در حالتی خودکار ایجاد می‌شوند (۲۷ و ۲۸).

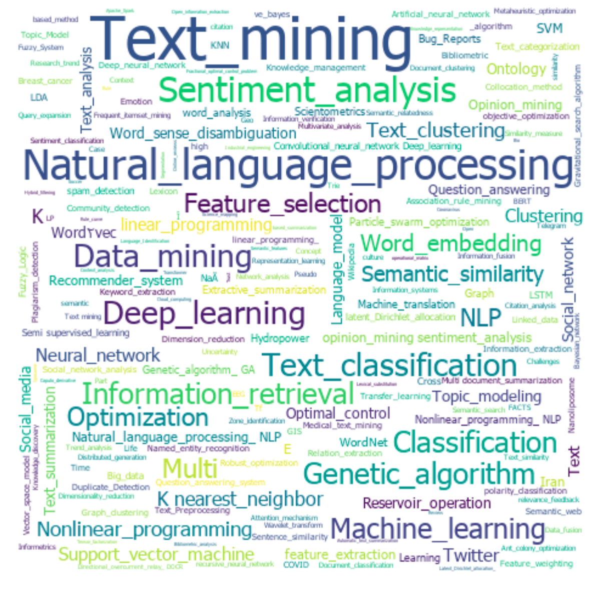
در میان کلمات کلیدی مدارک علمی متمرکز بر زبان فارسی کلمات کلیدی Natural language processing، Persian language و Sentiment analysis و در مدارک علمی متمرکز بر زبان انگلیسی نیز Natural language processing، Sentiment analysis وdeep learning پربسامدترین کلمات کلیدی هستند.

الف: کل

ب. متمرکز بر زبان فارسی



ج . متمرکز بر زبان انگلیسی

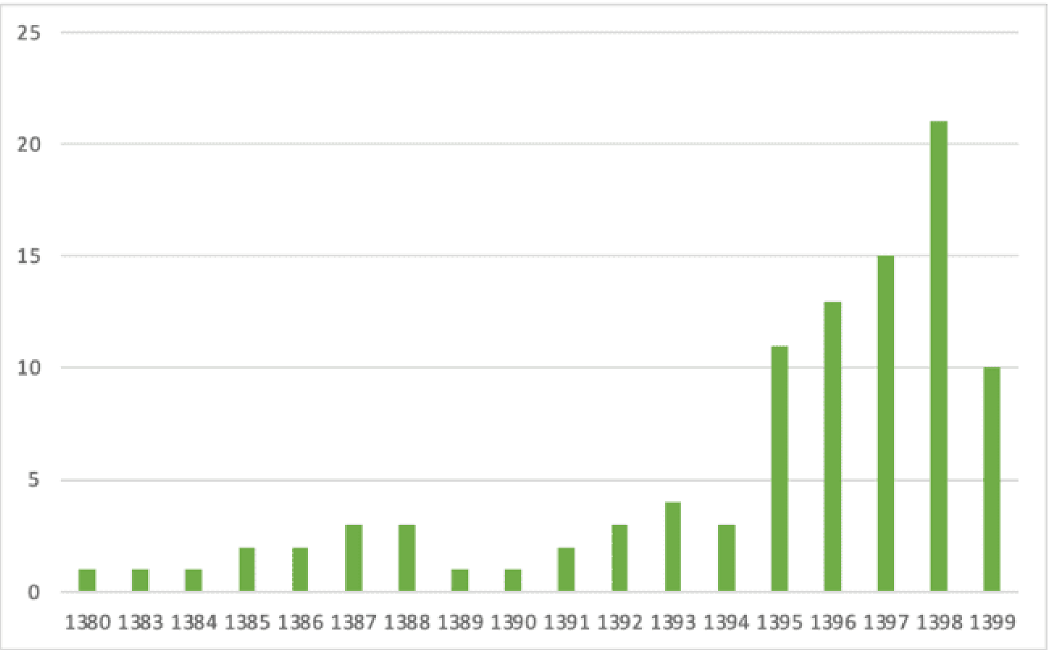


شکل ۳. کلمات کلیدی به کار رفته در مدارک علمی بین‌المللی در حوزه متن‌کاوی

[DOR: 20.1001.1.24234710.1401.9.1.9.5]

[DOI: 10.22088/cjs.9.1.104]

نمودار ۳، روند مدارک علمی داخلی مرتبط با متن‌کاوی را نشان می‌دهد. بر این اساس، بیشترین تعداد مدرک علمی در سال ۱۳۹۸ با تعداد ۲۱ مدرک علمی نگارش شده است. هشام فیلی، بهروز مینایی بیدگلی و حبیب زارع احمدآبادی به ترتیب با انتشار ۵، ۴ و ۳ مدرک علمی از دانشگاه تهران، دانشگاه علم و صنعت و دانشگاه یزد، نویسندگانی هستند که بیشترین تعداد مدارک علمی مرتبط با متن‌کاوی را منتشر کرده‌اند. پردازش علایم و داده‌ها، پردازش و مدیریت اطلاعات (علوم و فناوری اطلاعات) و مهندسی برق و مهندسی کامپیوتر ایران، مجلاتی هستند که بیشترین تعداد مدارک علمی متن‌کاوی داخلی را منتشر کرده‌اند.



نمودار ۳. روند مدارک علمی داخلی مرتبط با متن‌کاوی

جدول ۲، تعداد مدارک علمی داخلی و بین‌المللی ایران در حوزه موضوعی متن‌کاوی را نشان می‌دهد. بر این اساس، تعداد کل مدارک بین‌المللی ایران برابر با ۱۰۸۲ و تعداد مدارک علمی داخلی برابر با ۸۹ است. تعداد مدارک علمی بین‌المللی متمرکز بر زبان فارسی، برابر با ۲۸۴ (۲۶/۲۵٪) و تعداد مدارک علمی داخلی متمرکز بر زبان فارسی برابر با ۵۱ (۵۷/۳۰٪) است. بیشتر مدارک علمی بین‌المللی ۷۸۵ (۷۲/۵۵٪) بر زبان انگلیسی متمرکز هستند، در حالی که این تعداد برای مدارک علمی داخلی برابر با ۳۸ (۴۲/۶۹٪) است. نتایج آزمون تی مستقل (T-test) نشان داد که بین تعداد مدارک علمی متمرکز بر زبان فارسی پایگاه اسکوپوس و مرکز اطلاعات علمی جهاد دانشگاهی (SID) تفاوت معناداری وجود دارد ($P<0/001$). می‌توان چنین تفسیر کرد که مجلات بین‌المللی توجه بیشتری به زبان انگلیسی و مجلات داخلی توجه بیشتری به زبان فارسی دارند.

جدول ۲. تعداد مدارک علمی داخلی و بین‌المللی ایران در حوزه موضوعی متن‌کاوی

مدارک علمی داخلی	مدارک علمی بین‌المللی	نوع
۵۱ (۵۷/۳۰٪)	۲۸۴ (۲۶/۲۵٪)	متمرکز بر زبان فارسی
۳۸ (۴۲/۶۹٪)	۷۸۵ (۷۲/۵۵٪)	متمرکز بر زبان انگلیسی
۸۹ (۱۰۰٪)	۱۰۸۲ (۱۰۰٪)	مجموع

شکل ۴، کلمات کلیدی به کار رفته در مدارک علمی مرتبط با متن‌کاوی را در پایگاه مرکز اطلاعات علمی نشان می‌دهد. بر این اساس، کلمات کلیدی پردازش زبان طبیعی، زبان فارسی، خوشه‌بندی و بازیابی اطلاعات بیشترین بسامد را در میان مدارک علمی مرتبط با متن‌کاوی در پایگاه مرکز اطلاعات علمی داشته‌اند. همچنین، پیش‌بینی فناوری، دسته‌بندی متون و خوشه‌بندی متن پربسامدترین کلمات کلیدی مدارک علمی متمرکز بر زبان انگلیسی و پردازش زبان طبیعی، زبان فارسی و بازیابی اطلاعات پربسامدترین مدارک علمی متمرکز بر زبان فارسی هستند.

الف: کل

ج . متمرکز بر زبان انگلیسی

ب. متمرکز بر زبان فارسی

شکل ٤. کلمات کلیدی به کار رفته در مدارک علمی داخلی مرتبط با متن‌کاوی

بحث و نتیجه‌گیری

پژوهش حاضر به بررسی متنی مدارک علمی ایران در حوزه متن‌کاوی پرداخت و با استفاده از آزمون آماری مشخص شد که این تفاوت معنادار است و مجلات داخلی بیشتر به زبان فارسی توجه دارند. به این معنا که مدارک علمی که در مجلات بین‌المللی در حوزه متن‌کاوی نمایه می‌شوند، بیشتر از مجموعه داده‌های انگلیسی استفاده کرده‌اند و توجه کمی به زبان فارسی دارند. در مقابل مدارک علمی منتشر شده در مجلات داخلی، بیشتر بر زبان فارسی و چالش‌های آن تمرکز دارند. این مورد اهمیت توجه به مجلات داخلی و توسعه آنها را گوشزد می‌کند. باید توجه داشت که سوگیری محققین ایرانی به مجموعه داده‌های

انگلیسی مختص ایشان نیست و در سایر زبان‌ها نیز گزارش شده است (۲۹). با این وجود مطالعات کرامت فر و امیرخانی نشان می‌دهد که این سوگیری در حوزه تحلیل احساس در میان محققین ایرانی بیشتر است (۲۲). از دلایل دیگر این سوگیری می‌توان به گستردگی مجموعه داده‌ها در زبان انگلیسی و فقدان آن در زبان فارسی اشاره کرد. در این زمینه توسعه مجموعه داده‌های مختلف زبان فارسی می‌تواند حائز اهمیت باشد (۳۰). از جمله دیگر علل می‌تواند ذهنیت جامعه علمی در زمینه پذیرش مقالات علمی باشد. به عبارت دیگر ممکن است محققین ایرانی تصور کنند که اگر مقاله آنها با استفاده از مجموعه داده‌های فارسی نگارش شود، احتمال پذیرش مقاله در مجلات معتبر بین‌المللی کاهش یابد. بر اساس آمار ارائه شده در این مقاله مشخص می‌شود که مقالات متمرکز بر زبان انگلیسی بیشتر از مقالات متمرکز بر زبان فارسی در مجلات بین‌المللی منتشر شده؛ اما اینکه آیا این مقادیر نشان‌دهنده احتمال پذیرش مقالات است محل سوال است، چرا که نسبت مذکور در مورد ارسال مقالات برای ما روشن نیست. بنابراین صحت یا عدم صحت این مورد نیاز به پژوهش‌های آتی دارد. همچنین یافته‌های مقاله حاضر نشان می‌دهد که مقالات بین‌المللی متمرکز بر زبان انگلیسی استناد بیشتری نسبت به مقالات بین‌المللی متمرکز بر زبان فارسی دریافت می‌کنند. این مورد خود دلیل دیگری بر توجه جامعه علمی به مجموعه داده‌های انگلیسی است، چرا که نشان می‌دهد جامعه علمی بین‌المللی پذیرش بهتری نسبت به مقالات مبتنی بر مجموعه داده‌های انگلیسی دارد. همچنین مقایسه تعداد کلی مقالات منتشرشده در مجلات بین‌المللی در مقایسه با مجلات داخلی می‌تواند نشان‌دهنده رجحان انتشار در مجلات بین‌المللی در مقایسه با مجلات داخلی برای محققین ایرانی باشد. بر اساس نتایج پژوهش آزادی از آنجا که محققان ایرانی به شدت در انتشار مجلات علمی (خارجی) انگیزه دارند (به دلیل ارتقاء علمی و غیره)، اغلب توجه کم یا ناچیزی به نیازهای فکری و فنی کشور می‌شود. آنها تقریبا فقط بر قابل انتشار بودن آثار خود تمرکز می‌کنند (چنانچه پیشتر ذکر شد). همچنین، تعداد کمی از مقالات (در برخی از رشته‌ها) هیچ تحلیل جامع و دارای چشم‌انداز از برخی بحران‌های شدید کشور را ارائه نمی‌دهند (مثلا کمبود آب) (۳۱). در نهایت می‌توان چنین گفت، صرف چاپ مقالات در نمایه‌های بین‌المللی اگرچه سبب ارتقا دانش و مهارت محققین خواهد شد، نمی‌تواند به صورت مستقیم مشکلات کشور را بر طرف کند. یکی دیگر از دلایل عدم توجه به مجموعه داده‌های فارسی می‌تواند شکاف میان جامعه علمی و بخش صنعت باشد. بار اصلی تولید علم را دانشگاه‌های کشور به دوش می‌کشند و همانند بسیاری از کشورهای در حال توسعه تعامل بین پژوهشگران دانشگاه‌ها و سازمان‌های صنعتی ایران ضعیف و ناپایدار است (۳۲).

امروزه، انتشار مقالات علمی یکی از ملاک‌های سنجش افراد، سازمان‌ها و کشورها است. با بررسی یافته‌ها مشخص شد میانگین نرخ رشد مدارک علمی ایران در حوزه موضوعی متن‌کاوی برابر با ۳۲/۵ درصد و نرخ رشد در این حوزه موضوعی بالاتر از حوزه موضوعی کامپیوتر است. با بررسی میزان هم‌تالیفی پژوهشگران ایرانی در این حوزه موضوعی مشخص شد آمریکا بزرگ‌ترین شریک علمی ایران است و بعد از آن انگلیس و استرالیا قرار دارند. الگوی میانگین هم‌تالیفی نشان می‌دهد که نرخ آن برابر با ۳ است. با بررسی مجلات منتشرکننده مدارک بین‌المللی ایران مشخص شد که پراکندگی آن از قانون لوتکا پیروی نمی‌کند (۳۳). ۱۵ درصد از مدارک علمی بین‌المللی ایران در حوزه متن‌کاوی در سه مجله نگارش شده است. شناخت مجلات اصلی می‌تواند به محققان در هر حوزه موضوعی کمک کند تا از مهم‌ترین مدارک علمی حوزه موضوعی خود با خبر شوند. مجله Lecture Notes In Computer Science بیشترین تعداد مدارک علمی بین‌المللی ایران و مجله پردازش علایم و داده‌ها، بیشترین تعداد مدارک علمی داخلی ایران را در حوزه موضوعی متن‌کاوی منتشر کرده‌اند.

با بررسی کلمات کلیدی مشخص شد که Covid-19، Recommender systems و Latent Dirichlet Allocation جدیدترین کلمات کلیدی مدارک علمی بین‌المللی ایران هستند که با توجه به شیوع کرونا در جهان و به‌وجودآمدن داده‌های متنی در مورد این بیماری، مدارک علمی که به بررسی این موضوع پرداختند رو به افزایش است. پردازش زبان طبیعی، زبان فارسی، خوشه‌بندی و بازیابی اطلاعات پربسامدترین کلمات کلیدی مدارک علمی داخلی هستند. هشام فیلی محققی است که بیشترین تعداد مدارک علمی بین‌المللی و داخلی ایران را در حوزه موضوعی متن‌کاوی منتشر کرده است. بهروز مینایی بیدگلی جز محققان برتر ایران در موضوع متن‌کاوی است. علاوه بر این محمد احسان بصیری، مؤثرترین نویسنده و مهرنوش شمس فرد اثربخش‌ترین نویسنده ایرانی در مدارک علمی بین‌المللی ایران هستند. علاوه بر این دانشگاه‌های تهران، صنعتی شریف و آزاد اسلامی، بیشترین تعداد مدارک علمی بین‌المللی ایران در حوزه متن‌کاوی را منتشر کرده‌اند.

در نهایت می‌توان گفت میانگین نرخ رشد مدارک علمی ایران در حوزه متن‌کاوی بالاتر از حوزه‌های موضوعی دیگر است. همچنین در این حوزه موضوعی مانند سایر حوزه‌ها، مدارک علمی بین‌المللی که بر زبان انگلیسی متمرکز هستند، استناد بیشتری نسبت به مدارک علمی متمرکز بر زبان فارسی دریافت می‌کنند.

ملاحظات اخلاقی: در این پژوهش، مسائل اخلاقی از جمله سرقت ادبی، انتشار یا تسلیم دوگانه و همچنین اصول محرمانگی در ارائه‌ی داده‌های پژوهش به‌طور کامل رعایت شده است.

تضاد منافع: نویسندگان تصریح می‌نمایند که هیچ‌گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

References

1. Aggarwal CC, Zhai C. An Introduction to Text Mining. In: Aggarwal CC, Zhai C, editors. Mining Text Data. Boston, MA: Springer US; 2012. p. 1-10.

2. Keshavarzian S, Barardokht H. The position of books and reading on the site of tebyan based on text mining and analysis of social networks. Journal of Business Intelligence Management Studies. 2017; 6(21): 169-88. Available at: https://ims.atu.ac.ir/article_8516.html [In Persian]

3. Tan A-H, editor Text mining: The state of the art and the challenges. Proceedings of the pakdd 1999 workshop on knowledge disocovery from advanced databases; 1999: Citeseer.

4. Mojgani P, Jalali M, Keramatfar A. Bibliometric study of traumatic brain injury rehabilitation. Neuropsychol Rehabil. 2022; 32(1): 51-68.

5. Nourmohammadi H, Keramatfar M, Keramatfar A, Esparaein F. Research in which fields? Determining the Iranian research priorities based on their effects on economic growth. Caspian Journal of Scientometrics. 2014; 1(1): 48-53. Available at: http://cjs.mubabol.ac.ir/browse.php?a_id=31&sid=1&slc_lang=en [In Persian]

6. Keramatfar A, Rafiee M, Amirkhani H. Graph Neural Networks: a bibliometrics overview. arXiv: 2201.01188. 2021.

7. Habibi A ,Mousavi M, Jamali SM, Ale Ebrahim N. A bibliometric study of medical tourism. Anatolia. 2021: 1-11.

8. Imani H, Abdollahzadeh A, Pourezzat AA. Content analysis of the articles published in the journal of public administration in university of Tehran. Journal of Public Administration. 2018; 10(3): 387-414. Available at: https://jipa.ut.ac.ir/article_68037.html?lang=en [In Persian]

9. Farzin Yazdi M, Rezaei Sharifabadi S. Scientific publications in the subject area of Artificial Intelligence in Middle Eastern countries during 1996 to 2014. Scientometrics Research Journal. 2017; 3(6): 97-114. Available at: http://rsci.shahed.ac.ir/article_512.html [In Persian]

10. Keramatfar A, Rafiee Khoshnood M. Evaluation of scientific outputs of Royan Institute. Caspian Journal of Scientometrics. 2016; 3(1): 36-44. Available at: http://cjs.mubabol.ac.ir/browse.php?a_id=89&sid=1&slc_lang=en [In Persian]

11. Fazeli Varzaneh M, Bahmani M, Ghaderi Azad E. Iranian scientific outputs in the field of energy and fuel, and their comparison with those of the Middle East countries. Caspian Journal of Scientometrics. 2018; 5(1): 7-18. Available at: http://cjs.mubabol.ac.ir/article-1-137-en.html [In Persian]

12. Ataie-Ashtiani B. Chinese and Iranian Scientific Publications: Fast Growth and Poor Ethics. Science and Engineering Ethics. 2017; 23(1): 317-9.

13. Barangi H, Khasseh AA. An Investigation of Iran's Global Situation in Computer Science Using Scientometric Techniques. Journal of Knowledge and Information Management. 2017; 4(1): 59-74. Available at: https://lib.journals.pnu.ac.ir/article_4409.html?lang=en [In Persian]

14. Keshavarzian S, Barardokht H. The Position of Books and Reading on the Site of Tebyan Based on Text Mining and Analysis of Social networks. BI Management Studies. 2017; 6(21): 169-88. Available at: https://ims.atu.ac.ir/article_8516.html?lang=en [In Persian]

15. Ghaneh S, Sepidnam Gh, Jafari E. Persian text summarization using deep learning algorithm. The First National Conference on Modern and Smart Business Data Mining and Image Processing. Kerman: Technical and Vocational University; 2019. Available at: https://civilica.com/doc/990619/ [In Persian]

16. Kharabaf Sh, Abdollahi M. Science growth in Iran over the past 35 years. J Res Med Sci. 2012; 17(3): 275-9.
17. Hao T, Chen X, Li G, Yan J. A bibliometric analysis of text mining in medical research. Soft Computing. 2018; 22(23): 7875-92.
18. Surulinathi M, Amsaveni N, Maheswaran K, Srinivasaraghavan S. Scientometric Dimensions of Knowledge Management Research in India: A Study based on Scopus database. Sri Lankan Journal of Librarianship and Information Management. 2009; 2.
19. Wikipedia. Scientific Information Database: Wikipedia 2020.
20. Chowdhury GG. Natural language processing. Annual Review of Information Science and Technology. 2003; 37(1): 51-89.
21. Khoshian N, Mirzaeian V. The most widely used functions of natural language processing in the field of library science and information science. Knowledge Retrieval and Semantic Systems. 2020; 7(23):117-50. Available at: https://jks.atu.ac.ir/article_10929.html?lang=en [In Persian]
22. Keramatfar A, Amirkhani H. Bibliometrics of sentiment analysis literature. Journal of Information Science. 2019; 45(1): 3-15.
23. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment Analysis Is a Big Suitcase. IEEE Intelligent Systems. 2017; 32(6): 74-80.
24. Basiri ME, Kabiri A. Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining. ACM Trans Asian Low-Resour Lang Inf Process. 2018; 17(4): Article 26: 1-18.
25. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ. Sentiment analysis of twitter data. Proceedings of the Workshop on Language in Social Media (LSM 2011); 2011: 30-8.
26. Vogt M. An overview of deep learning techniques. at - Automatisierungstechnik. 2018; 66(9): 690-703.
27. Mousavi SM, Ebadi H, Kiani A. Provide an Optimal Deep-network Method for Spectral-spatial Classifying of High Resolution Images. Journal of Geomatics Science and Technology. 2019; 9(2): 151-70. Available at: http://jgst.issge.ir/article-1-814-en.html [In Persian]
28. Socher R. Recursive deep learning for natural language processing and computer vision. [PhD Thesis] California: Stanford University; 2014.
29. Ruder S. Why You Should Do NLP Beyond English. 2020. Available at: http://ruder.io/nlp-beyond-english.
30. Amirkhani H, AzariJafari M, Pourjafari Z, Faridan-Jahromi S, Kouhkan Z, Amirak A. Farstail: A persian natural language inference dataset. arXiv preprint arXiv:200908820. 2020.
31. Azadi P. The Structure of Corruption in Iran. Stanford Iran 2040 Project; 2020.
32. Erfanmanesh M, Moghiseh Z, Forouzandeh Shahraki M. Comparing the share of scholarly output published through the collaboration between academic and corporates in Iran, Middle East, and the World. Rahyaft. 2018; 28(69): 65-80. Available at: https://rahyaft.nrisp.ac.ir/article_13643.html?lang=en [In Persian]
33. Lotka AJ. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 1926; 16(12): 317-23.