The Role of Ontology and Knowledge Graph in Text Document Classification: A Review of Studies

Saiede Khalilian[1], Mitra Pashootanizadeh[2], Ali Mansori [3], Hamidreza Baradaran Kashani [4]

Abstract

Purpose: With the increasing use of the internet and the growing volume of electronically accessible documents on the web, automatic text classification has become a critical method for enhancing information retrieval and managing digital text collections. Text classification allows individuals to search for and retrieve information more accurately and quickly. The significance of automatic document classification lies in labeling documents into predefined classes so that documents within a class exhibit the highest similarity and the most remarkable dissimilarity with documents from other classes while utilizing semantic relationships. This study investigates the application of ontology and knowledge graphs in automatic text document classification.

Method: This study reviewed research and documents related to applying semantic tools such as ontologies and knowledge graphs in text document classification. To collect texts, three domestic databases, including the "National Journal Database," the "Scientific Information Database of Jihad University," and "Marefate Danesh," along with three internal databases "Magiran," "SID" and "Civilica" and three external citation databases, such as "Web of Science", "Scopus" and "Google Scholar" It has been examined in both categories, regardless of the period.

Findings: Results of text exploration show that the vector space model does not consider the semantic relationships between words and disregards the word order in sentences. Neglecting the semantic and syntactic relationships between words in natural language provides a different representation of documents. However, ontologies and knowledge graphs help strengthen machine learning models by capturing the meaning of entities and classes. These tools act as an external reference during the classification process and provide domain knowledge for classification models. Using these tools generally allows machines to comprehend the meaning of the data they work with.

Conclusion: The application of ontologies and knowledge graphs in classifying textual documents can strengthen the results of machine learning algorithms through background knowledge. These tools can free the meanings of words from ambiguous sentences and solve problems related to natural language. Using ontology and knowledge graphs can effectively help classify textual documents and improve the accuracy and efficiency of classification models. However, constructing and integrating ontologies and knowledge graphs is a tedious, time-consuming, and complex task that limits the feasibility and practical application of these tools. In the Persian language, in addition to the problems raised in the application of ontologies and knowledge graphs in the classification of documents, there are limitations such as the specific features of the language in writing and technical limitations. Therefore, the use of ontology and knowledge graphs in discussing the classification of textual documents requires attention to linguistic limitations and technical complexity, and the need for further development and efforts is felt, especially in Persian.

Keywords

Automatic Classification, Text Documents, Knowledge Graph, Ontology, Domain Knowledge

Citation: Khalilian, S., Pashootanizadeh, M., Mansori, A., & Baradaran Kashani, H. (2024). The Role of Ontology and Knowledge Graph in Text Document Classification: A Review of Studies. *Librarianship and Information Organization Studies*, 35(2): 167-196.

Doi: 10.30484/nastinfo.2024.3548.2264

Article Type: Review Article
Article history:
Received: 11 Jan. 2024
Accepted: 20 Apr. 2024

1.Ph.D. Candidate, Knowledge and Information Science, University of Isfahan, Isfahan, Iran
skhalilian71@gmail.com

2. Associate Professor, Knowledge and Information Science Group, University of Isfahan, Isfahan, Iran; (Corresponding Author)
m.pashootanizade@edu.ui.ac.ir

3. Associate Professor, Knowledge and Information Science Group, University of Isfahan, Isfahan, Iran
a.mansouri@edu.ui.ac.ir

4. Assistant Professor, Computer Engineering Group, University of Isfahan, Isfahan, Iran
hrb.kashani@eng.ui.ac.ir

Publisher: National Library and Archives of I.R. of Iran

بررسی نقش هستی‌شناسی و نمودار دانش در طبقه‌بندی اسناد متنی: مروری بر مطالعات

سعیده خلیلیان[۱] | میترا پشوتنی‌زاده[۲] | علی منصوری[۳] | حمیدرضا برادران کاشانی[۴]

۱. دانشجوی دکتری، علم اطلاعات و دانش‌شناسی، دانشگاه اصفهان، اصفهان، ایران
Skhalilian71@gmail.com

۲. دانشیار، گروه علم اطلاعات و دانش‌شناسی، دانشگاه اصفهان، اصفهان، ایران؛ (نویسندهٔ مسئول)
m.pashootanizade@edu.ui.ac.ir

۳. دانشیار، گروه علم اطلاعات و دانش‌شناسی، دانشگاه اصفهان، اصفهان، ایران
a.mansouri@edu.ui.ac.ir

۴. استادیار، گروه مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران
hrb.kashani@ eng.ui.ac.ir

چکیده

هدف: باتوجه‌به افزایش نرخ استفاده از اینترنت و افزایش حجم اسناد الکترونیکی قابل‌مشاهده در وب، طبقه‌بندی خودکار متن تبدیل به یکی از روش‌های کلیدی برای ارتقای بازیابی اطلاعات و مدیریت دانش مجموعه‌های متنی دیجیتالی شده است. افراد با طبقه‌بندی متون می‌توانند اطلاعات موردنیاز خود را با دقت بیشتر و سرعت بالاتر جستجو و بازیابی کنند. آن چیزی که در بحث طبقه‌بندی خودکار اسناد حائز اهمیت است، برچسب‌گذاری اسناد به کلاس‌های از پیش تعریف شده است، به‌گونه‌ای که اسنادی که در یک طبقه جای می‌گیرند بیشترین شباهت و با اسناد سایر طبقه‌ها بیشترین تفاوت را داشته باشند و قابلیت استفاده از روابط معنایی را داشته باشد. در این راستا، پژوهش حاضر به بررسی نقش هستی‌شناسی و نمودار دانش در طبقه‌بندی خودکار اسناد متنی می‌پردازد.

روش: این مطالعه به‌مرور پژوهش‌ها و اسناد مرتبط با کاربرد ابزارهای معنایی مانند هستی‌شناسی‌ها و نمودار دانش در طبقه‌بندی اسناد متنی پرداخته است. به‌منظور جمع‌آوری متون، سه پایگاه اطلاعاتی داخلی شامل «بانک اطلاعات نشریات کشور»، «پایگاه مرکز اطلاعات علمی جهاد دانشگاهی» و «مرجع دانش» و سه پایگاه استنادی خارجی یعنی «وب آو ساینس»، «اسکوپوس» و «گوگل اسکالر» بدون درنظرگرفتن بازهٔ زمانی در هر دو دسته بررسی شده است.

یافته‌ها: نتایج واکاوی متون نشان داد در مدل فضای برداری ارتباط معنایی بین کلمات در نظر گرفته نمی‌شود و ترتیب کلمات در جملات از بین می‌رود. با نادیده‌گرفتن روابط معنایی و نحوی مختلف بین کلمات در زبان طبیعی، بازنمایی متفاوتی از اسناد فراهم می‌شود؛ اما هستی‌شناسی‌ها و نمودار دانش با دریافت معنای موجودیت‌ها و کلاس‌ها به تقویت مدل‌های یادگیری ماشینی کمک می‌نمایند. استفاده از این ابزارها به‌عنوان یک مرجع خارجی در حین فرایند طبقه‌بندی عمل می‌کند و دانش زمینه را برای مدل‌های طبقه‌بندی فراهم می‌نماید. به‌طورکلی استفاده از این ابزارها به ماشین‌ها اجازه می‌دهند معنای داده‌هایی را که با آن‌ها کار می‌کنند، درک کنند.

نتیجه‌گیری: کاربست هستی‌شناسی‌ها و نمودار دانش در طبقه‌بندی اسناد متنی می‌تواند موجب تقویت نتایج الگوریتم‌های یادگیری ماشین از طریق بهره‌برداری از دانش زمینه شود. این ابزارها می‌توانند معانی کلمات را از جملات دارای ابهام آزاد نموده و مشکلات مرتبط با زبان طبیعی را حل کند. استفاده از هستی‌شناسی و نمودار دانش می‌تواند به‌طور مؤثری در طبقه‌بندی اسناد متنی کمک کند و باعث ارتقای دقت و کارایی مدل‌های طبقه‌بندی شود؛ اما ساخت و ادغام هستی‌شناسی و نمودار دانش امری خسته‌کننده، زمان‌بر و پیچیده است که امکان‌پذیری و ارزش عملی آن‌ها را محدود می‌کند. در زبان فارسی علاوه بر مشکل مطرح‌شده در به‌کارگیری هستی‌شناسی‌ها و نمودار دانش در طبقه‌بندی اسناد، محدودیت‌هایی مانند ویژگی‌های خاص زبان فارسی در نگارش و محدودیت فنی وجود دارد؛ لذا استفاده از هستی‌شناسی و نمودارهای دانش عمومی و یا دامنه در بحث طبقه‌بندی اسناد نیازمند توجه به این محدودیت‌ها و پیچیدگی‌های فنی است و علاوه بر این مستلزم توسعه و تلاش‌های بیشتری بالأخص در زبان فارسی است.

کلیدواژه‌ها

طبقه‌بندی خودکار، اسناد متنی، نمودار دانش، هستی‌شناسی، دانش دامنه

استناد: خلیلیان، سعیده، پشوتنی‌زاده، میترا، منصوری، علی و برادران کاشانی، حمیدرضا (۱۴۰۳). بررسی نقش هستی‌شناسی و نمودار دانش در طبقه‌بندی اسناد متنی: مروری بر مطالعات. *مطالعات کتابداری و سازماندهی اطلاعات*، ۳۵(۲): ۱۶۷-۱۹۶.

Doi: 10.30484/nastinfo.2024.3548.2264

فصلنامهٔ مطالعات کتابداری و سازماندهی اطلاعات، ۳۵ (۲)، تابستان ۱۴۰۳

نوع مقاله: مروری

تاریخ دریافت: ۱۴۰۲/۱۰/۲۱

تاریخ پذیرش: ۱۴۰۳/۰۲/۰۱

مقدمه

اطلاعات در قالب‌ها و فرمت‌های متنوعی در بستر وب تولید می‌شود. بیش از ۸۰ درصد این اطلاعات را مستندات و داده‌های متنی در بر می‌گیرند. در میان انواع مختلف اطلاعات، داده‌های متنی[1] ازنظر کمیت و کیفیت مهم‌ترین نوع داده‌ها هستند؛ زیرا که انواع مختلف اطلاعات و دانش را دربر می‌گیرند. به دلیل اینکه داده‌های متنی ماهیت پیچیده‌ای دارند ساختار معنایی محتوای آن‌ها به‌خوبی توصیف نمی‌شود، همین امر باعث می‌شود اطلاعات موجود در این منابع به‌صورت دانش پنهان باقی بماند و پردازش خودکار، سازمان‌دهی و مدیریت آن‌ها نسبت به سایر داده‌ها چالش‌برانگیز باشد (Fkih & Omri, 2020؛ سلیمانی‌نژاد و همکاران ،۱۳۹۷). ازاین‌رو در سال‌های اخیر به‌دلیل تولید حجم عظیمی از داده‌ها به شکل متن، متن‌کاوی به‌طور پیوسته اهمیت پیدا کرده است. درواقع متن‌کاوی آشکار کردن اطلاعات پنهان است که با استخراج اطلاعات ضمنی، بدون ساختار یا نیمه ساختاریافته و مفید از حجم عظیمی از داده‌های متنی سروکار دارد. این روش به دنبال اطلاعات باارزشی مانند روابط، روندها و الگوها در داده‌های متنی بوده و به‌طور گسترده برای کشف روابط پیچیده در متون و اسناد علمی به کار می‌رود (Song et al., 2005 ؛ رمضانی و همکاران،۱۳۹۳). متن‌کاوی کاربردهای مهمی مانند طبقه‌بندی اسناد، پالایه کردن اسناد، خلاصه‌سازی، و تجزیه‌وتحلیل احساسات/طبقه‌بندی نظرات دارد (Altınel & Ganiz, 2018). با توجه به افزایش نرخ استفاده از اینترنت و افزایش حجم اسناد الکترونیکی

1. Textual data

قابل‌مشاهده در وب، طبقه‌بندی داده‌های متنی تبدیل به یکی از روش‌های کلیدی برای ارتقای بازیابی اطلاعات و مدیریت دانش مجموعه‌های متنی دیجیتالی شده است (Song et al., 2005; Joorabchi & Mahdi, 2011)).

مفهوم طبقه‌بندی یا دسته‌بندی متن در سال ۱۹۵۰ بیان شد و بر روی نمایه‌سازی مجلات علمی با استفاده از واژگان متمرکز شد (ریحانی آرانی و لاجوردی، ۱۳۹۵). طبقه‌بندی متن به عمل برچسب‌گذاری موضوعی متون زبان طبیعی براساس مجموعه از پیش تعریف‌شده گفته می‌شود. به عبارت رسمی‌تر اگر d_i سندی از کل مجموعه اسناد D باشد و $\{c_1,c_2,...,n\}$ مجموعه‌ای از همه دسته‌ها باشد، سپس طبقه‌بندی متن یک دسته c_j را به سند d_i اختصاص می‌دهد (Zhang & Xu, 2020).

در ارتباط با روش‌های طبقه‌بندی اسناد متنی روش‌های مختلفی از دستی تا کاملاً خودکار وجود دارد. تا دهه ۹۰ طبقه‌بندی متون عمدتاً دستی بودند، یا از فنون دستی (طبقه‌بندی توسط افراد خبره) استفاده می‌شد. روش‌های طبقه‌بندی دستی محدودیت‌هایی دارد مانند نیاز داشتن به نیروهای متخصص در هر حوزه، بالا بودن هزینه و زمان‌بر بودن که میزان کارایی را کاهش می‌دهد، همچنین برچسب‌گذاری متون براساس دانش و تجربه‌ی فردی بدون خطا نخواهد بود و نیز در صورت اعمال تصمیمات چند فرد خبره در فرآیند طبقه‌بندی، سیستم به دلیل تفاوت تصمیمات افراد مختلف دچار ناسازگاری می‌گردد (مدنی ۱۳۹۱؛ Qian et al., 2021) اما با توجه به گستردگی و تنوع متون و نیز ظهور حوزه‌های جدید، شناسایی و یا تولید طبقه‌بندی‌هایی که بتواند با جامعیت کامل اسناد موجود را در طبقه‌های مناسب جای دهد بسیار مشکل است و در برخی موارد غیرممکن است. ازاین‌رو مطلوب است که از روش‌های خودکار جهت طبقه‌بندی متون برای به دست آوردن نتایج مطمئن‌تر و کمتر ذهنی استفاده شود (Dalal & Zaveri, 2011). استفاده از رویکردهای مبتنی بر یادگیری ماشینی به روش اصلی جهت طبقه‌بندی خودکار اسناد تبدیل شده است. درواقع ماشین از طریق داده‌های آموزشی برچسب‌گذاری شده، ارتباط ذاتی بین متون و برچسب‌ها را یاد می‌گیرد و می‌تواند برای داده‌های بدون برچسب به کار گیرد (Suneera & Prakash, 2020؛ Minaee et al., 2020). به‌عبارت‌دیگر در طبقه‌بندی خودکار به یک مجموعه داده اولیه نیاز است تا ماشین بتواند از طریق آن بیاموزد و برای داده‌های بدون برچسب استفاده نماید(Suneera & Prakash, 2020). فرآیند انجام طبقه‌بندی خودکار به این صورت است که ابتدا باید اسناد پیش‌پردازش شوند، در این مرحله میزان داده‌ها اسناد متنی

ورودی به میزان قابل‌توجهی کاهش می‌یابد(Mumivand et al., 2021)؛ زیرا اسناد از هر داده که ممکن است تأثیر منفی بگذارد، پاک‌سازی می‌شود. این مرحله شامل چندین گام است که عبارت‌اند از: نرمال‌سازی[1]، تقسیم‌بندی جملات[2]، تبدیل جملات به مجموعه‌ای از کلمات (تک‌واژه سازی[3])، حذف کلمات بی‌اثر[4]، ریشه‌یابی[5]. سپس باید ویژگی‌هایی که دارای اهمیت بیشتری هستند، انتخاب و استخراج گردند (Uysal & Gunal, 2014). به این منظور یک سند متنی به بردار سند تبدیل می‌شود به‌گونه‌ای که اسناد به‌صورت برداری از کلمات نمایش داده می‌شوند، و ویژگی‌ها با توجه به میزان اهمیت آن‌ها نسبت به سند متنی و دسته آن وزن‌دهی می‌شوند که به آن مدل فضای برداری[6] یا مدل کیسه لغات[7] گفته می‌شود (Kilimci & Akyokus, 2019؛ بهروزیان‌نژاد و همکاران، ۱۳۹۳). پس از گذشت از مراحل قبلی اسناد به فرمتی تبدیل می‌شوند که به‌راحتی می‌توان به‌وسیله الگوریتم‌های مختلف یادگیری ماشینی مدل‌سازی شوند. درنهایت مدل طبقه‌بندی ایجادشده باید با استفاده از مجموعه آزمایشی اسناد متنی آزمایش شود. اگر دقت طبقه‌بندی مدل‌سازی شده برای مجموعه آزمایشی قابل‌قبول باشد، از این مدل می‌توان برای طبقه‌بندی نمونه‌های جدید اسناد متنی استفاده نمود (Novaković et al., 2017; Burgueño et al., 2017).

معمولاً در فرآیند کلی طبقه‌بندی اسناد هر سند در مدل فضای برداری نشان داده می‌شود. در مدل فضای برداری (فرمت نمایش مبتنی بر سند) هر سند به‌عنوان یک بردار نمایش داده می‌شود که شامل مجموعه‌ای از عبارت (کلمات) است که در سند ظاهر می‌شود. مجموعه اسناد با نمایش برداری آن‌ها ماتریس سند–عبارت[8] را تشکیل می‌دهد و اهمیت هر عبارت در یک سند با محاسبه وزن آن‌ها مشخص می‌شود (Kilimci & Akyokus, 2019). در این مدل از کلمات یا ریشه کلمات به‌عنوان ویژگی برای نمایش محتوای سند استفاده می‌کند. با انجام این کار، الگوریتم‌های یادگیری انتخاب‌شده تنها به تشخیص الگوها در اصطلاحات استفاده‌شده

1. Normalization
2. Sentence Segmentation
3. Tokenization
4. Stop-words
5. Stemming
6. Vector Space Model (VSM)
7. Bag Of Words model
8. Document-term matrix

محدود می‌شوند، درحالی‌که الگوهای مفهومی نادیده گرفته می‌شوند. در مدل فضای برداری اسناد به‌عنوان کیسه‌هایی از اصطلاحات متداول در نظر گرفته می‌شوند که هر اصطلاح برای خود یک ویژگی مستقل است (Bloehdorn et al., 2005). آنچه در فرمت نمایش مبتنی بر سند مهم است، نحوه‌ی نمایش بهتر برای مستندات است مانند تبدیل اسناد به یک فرمت میانی و نیمه ساخت‌یافته یا بکار بردن برچسب بر روی آن‌ها یا هر نوع نمایش دیگری که کار کردن با اسناد را کاراتر می‌کند. این مدل منجر به بردارهای سند با ابعاد بالا (تعداد زیادی از ویژگی‌ها) می‌شود که برای مدیریت تعداد زیادی از ویژگی‌ها باید از فن‌های کاهش ویژگی استفاده نمود(Dalal & Zaveri, 2011). این نوع نمایش ارتباط معنایی بین کلمات را در نظر نمی‌گیرد و ترتیب کلمات در جملات از بین می‌رود. نادیده گرفتن روابط معنایی و نحوی مختلف بین کلمات در زبان طبیعی، بازنمایی اسناد را ساده‌تر می‌سازد. به این دلیل که عبارات چندکلمه‌ای را با جدا کردن آن‌ها به اصطلاحات مستقل نادیده می‌گیرد و کلمات چندمعنایی (کلمات دارای معانی متعدد) را به‌عنوان یک موجود واحد در نظر می‌گیرد، زیرا کلمه از کلمات مجاور خود که معنای آن را تعیین می‌کنند جدا می‌شود. همچنین کلمات مترادف را به اصطلاحات مجزا تبدیل می‌کند(Altınel & Ganiz, 2018). در این روش رابطه و زمینه کلمات موجود در متون در نظر گرفته نمی‌شود. درواقع عملکرد این نوع طبقه‌بندی تا حد زیادی به کیفیت ویژگی‌های دست‌ساز[1] بستگی دارد و نیاز به تحلیل ویژگی‌های خسته‌کننده برای به دست آوردن عملکرد خوب دارد و به دلیل وابستگی شدید به دانش دامنه برای طراحی ویژگی‌ها، تعمیم روش به وظایف جدید را دشوار می‌کند. درنهایت، این مدل‌ها نمی‌توانند از حجم زیادی از داده‌های آموزشی استفاده کامل کنند، زیرا ویژگی‌ها (یا قالب‌های ویژگی) از پیش تعریف شده‌اند(Minaee et al., 2020). بنابراین طبقه‌بندی خودکار اسناد به روش ذکرشده در بالا زمانی مناسب است که داده‌های با کیفیت بالا و ابعاد پایین داشته باشد. همچنین این روش برای استخراج الگوهای پیچیده از داده‌ها با ابعاد بالا دقت کافی را ندارند(Patterson & Gibson, 2017). مباحث مطرح‌شده در قسمت قبل نشان‌دهنده معضل بزرگ در طبقه‌بندی اسناد متنی است به‌ویژه زمانی که برچسب‌های متعددی از اسناد وجود داشته باشد اما داده‌های آموزشی کافی برای هریک از طبقه‌ها در دسترس نباشد. همچنین

1. Hand-crafted features

دسترسی به داده‌های برچسب‌گذاری شده با کیفیت برای آموزش معمولاً در دنیای واقعی بسیار گران است (Altınel & Ganiz, 2018). لذا وجود یک مدل دقیق که بتواند اسناد را با توجه به معنای آن‌ها در طبقه‌های مناسب جای دهد ضرورت دارد. مرور سوابق نظری و تحقیقاتی موضوع پژوهش نشان داد که تاکنون پژوهش‌های بسیاری در خصوص طبقه‌بندی اسناد متنی و تقویت مدل‌های یادگیری با استفاده از هستی‌شناسی‌ها و نمودار دانش انجام پذیرفته است، اما پژوهش جامعی در مورد مسئله پژوهشی حاضر صورت نگرفته است و تمامی پژوهش‌ها به‌صورت عملیاتی و یا به‌صورت محدود یا پراکنده به اهمیت استفاده از هستی‌شناسی‌ها و نمودار دانش در طبقه‌بندی اسناد پرداخته‌اند. لذا این پژوهش در پی دستیابی به پاسخ برای پرسش‌های ذیل است:

- تعاریف و ساخت هستی‌شناسی و نمودار دانش چه هستند؟
- کاربرد و نقش هستی‌شناسی و نمودار دانش در طبقه‌بندی خودکار اسناد چیست؟
- چالش‌های کاربرد هستی‌شناسی و نمودار دانش در طبقه‌بندی خودکار اسناد چه هستند؟

روش پژوهش

روش این پژوهش رویکرد توصیفی و واکاوی به روش اسنادی و کتابخانه‌ای است. در این پژوهش سعی شده است مرور جامعی بر کاربرد هستی‌شناسی و نمودار دانش در بحث طبقه‌بندی اسناد متنی در بروندادهای علمی صورت گیرد.

در مطالعه حاضر به‌منظور جست‌وجوی متون از کلیدواژه‌های هم‌راستا با نظر متخصصان استفاده شد. در متون انگلیسی اصطلاحات ontology knowledge graph, domain semantic tools, semantic, knowledge, knowledge resource, domain ontology,, vocabulary resources, thesaurus, semantic text classification, wordnet, wikidata, Wiktionary, FarsNet, FarsBase, smantic relations, yago, DBPedia به‌صورت ترکیبی (AND) با اصطلاحات text classification, documents classification, text documents classification در سه پایگاه استنادی خارجی یعنی «وب آو ساینس[1]»، «اسکوپوس[2]» و «گوگل اسکالر[3]» مورد جست‌وجو قرار گرفت. همچنین در ادامه به‌منظور جمع‌آوری متون فارسی، با

1. Mjl.clarivate.com
2. Scopus.com
3. Scholar.google.com

کلیدواژه‌های «هستی‌شناسی»، «آنتولوژی»، «معنا»، «نمودار دانش»، «گراف دانش»، «روابط معنایی»، «دانش معنایی»، «منابع دانشی»، «فارس‌نت»، «دانش دامنه»، «اصطلاح‌نامه»، «وردنت»، «ویکی دیتا»، «یاگو»، به‌صورت ترکیبی با اصطلاحات «طبقه‌بندی اسناد»، «طبقه‌بندی اسناد متنی»، «طبقه‌بندی متن» در سه پایگاه اطلاعاتی داخلی شامل «بانک اطلاعات نشریات کشور[1]»، «پایگاه مرکز اطلاعات علمی جهاد دانشگاهی[2]» و «مرجع دانش[3]» انجام شد. به‌منظور حفظ جامعیت نتایج در جست‌وجوی هر دودسته منابع داخلی و خارجی بازه زمانی لحاظ نشد.

یافته‌ها

- بررسی سؤال اول پژوهش

تعاریف و ساخت هستی‌شناسی و نمودار دانش چیست؟

هستی‌شناسی یکی از لایه‌های وب معنایی است که برای اتصال مفاهیم از طریق موجودیت‌های نام‌گذاری شده بالقوه آن (به‌عنوان‌مثال، کلاس‌ها و روابط) استفاده می‌شوند. یک هستی‌شناسی به‌صورت محاسباتی مجموعه‌ای از واژگان مشترک را برای نمایش رسمی دانش تعریف می‌کند و درنتیجه اشتراک و استفاده مجدد دانش را تسهیل می‌کند (Fensel et al., 2000). همان‌طور که گروبر[4] (۱۹۹۳) توضیح می‌دهد یک هستی‌شناسی مشخصات واژگان نماینده را برای حوزه مشترک موردعلاقه توصیف می‌کند. به‌طورمعمول، هستی‌شناسی شامل مفاهیم اساسی، روابط، مصادیق و بدیهیات است که به‌طور رسمی در قالب ماشین‌خوان نمایش داده می‌شوند (Gruber, 1993). مفاهیم نشان‌دهنده مجموعه‌ای از موجودیت‌ها یا کلاس‌ها در یک دامنه هستند که می‌توانند به‌عنوان مفاهیم ابتدایی یا تعریف‌شده مشخص شوند. روابط بیانگر تعامل بین مفاهیم یا ویژگی‌های یک مفهوم است و بدیهیات موضوعات اصلی هستند که مقدار یک کلاس یا یک نمونه را محدود می‌کند و ویژگی‌های یک رابطه را مشخص می‌سازند. دو عملکرد اصلی هستی‌شناسی‌ها عبارت‌اند از: الف. هستی‌شناسی‌ها اطلاعات را در قالب معنای رسمی تعریف می‌کنند که امکان پردازش اطلاعات به‌وسیله ماشین فراهم می‌شود؛ ب. هستی‌شناسی‌ها یک معنا از دنیای واقعی را تعریف می‌کنند که از این طریق امکان اتصال میان

1. Magiran.com
2. Sid.ir
3. Civilica.com
4. Gruber

محتوای قابل‌پردازش ماشینی با معانی اصطلاحات رایج و موردتوافق افراد فراهم می‌شود. هستی‌شناسی‌ها مدل‌های داده‌ای هستند که برای نمایش معنایی مفاهیم دامنه از طریق اصطلاح هستی‌شناسی، یعنی کلاس‌ها (موجودات) و روابط (ویژگی‌ها) استفاده می‌شوند. هستی‌شناسی دامنه، واژگان یک دامنه خاص را به روشی رسمی نشان می‌دهد، بنابراین باید با اطلاعات موجود در یک سند در آن حوزه مطابقت داشته باشد (Lee et al., 2021). یکی از کاربردهای هستی‌شناسی نشانه‌گذاری موجودیت‌ها در متن است. درواقع هستی‌شناسی شامل موجودیت‌هایی است که می‌توان به دنبال همان موجودیت‌ها در متون بود و آن‌ها را به‌عنوان موجودیت واحد علامت‌گذاری کرد. همچنین می‌توان از دانش موجود در هستی‌شناسی برای ایجاد قوانین تعمیم استفاده کرد (Malik & Jain, 2021).

هستی‌شناسی زیرمجموعه‌ای از نمودار دانش است و برای توسعه آن موردنیاز است. مفهوم نمودار دانش اولین بار توسط گوگل در سال ۲۰۱۲ ارائه شد که به‌عنوان یک پایگاه دانش در مقیاس بزرگ متشکل از تعداد زیادی موجودیت‌ها و روابط بین آن‌ها تعریف می‌شود (Chen et al., 2020). ارلینگر و ووس[1] (۲۰۱۶) اصطلاح «نمودار دانش» را این‌گونه تعریف می‌نماید: "نمودار دانش، اطلاعات را از یک هستی‌شناسی به دست آورده، سپس یکپارچه می‌سازد و درنهایت با استفاده از یک استدلال برای استخراج دانش جدید استفاده می‌نماید". بنابراین نمودار دانش را می‌توان مجموعه‌ای بزرگ از موجودیت‌های مرتبط به هم دانست که به‌وسیله برچسب‌های معنایی غنی شده‌اند (Gomez-Perez et al., 2017). همچنین می‌توان یک نمودار دانش را به‌عنوان هستی‌شناسی بسیار بزرگ توصیف کرد اگرچه آن‌ها برتر از هستی‌شناسی‌ها هستند؛ زیرا ویژگی‌های اضافی را ارائه می‌کنند. نمودار دانش متشکل از گره‌ها و لبه‌ها هستند، که در آن گره‌ها به هر موجودیت در پایگاه دانش اشاره می‌کنند که بر اساس دسته‌بندی متمایز می‌شوند و هر لبه به ارتباط منطقی بین موجودیت‌ها اشاره دارد. درواقع نمودار دانش را می‌توان نوعی شبکه رابطه‌ای دانست که ترکیبی از دانش بین موجودات و نقشه‌برداری از دانش موجود در دنیای عینی است (Lai et al., 2015 ؛ Wei et al., 2019). نمودار دانش ابزاری فنی برای استخراج مکرر دانش ساختاریافته از مقدار زیادی داده در ساختارهای مختلف است. مزیت نمودار دانش در توانایی آن در یکپارچه‌سازی کارآمد

1. Ehrlinger & Wöß

داده‌های متعدد، زائد و ساختار متفاوت و استدلال دانش کامل است (Chen et al., 2018; Lili et al., 2020) . نمودارهای دانش دیدگاه متمرکز بر روی اشیاء دارند، درواقع نه‌تنها بروی نمایش ساختارمند دانش معنایی تأکید دارند بلکه به چگونگی اتصال و تفسیر آن‌ها هم می‌پردازد که همین امر به بررسی صحت، اتصال، و سازگاری اطلاعات کمک می‌کند (Guo et al., 2022; Issa et al., 2021; Jung et al., 2010; Ehrlinger & Wöß, 2016; Shin et al., 2015).

هستی‌شناسی‌ها و نمودار دانش نقش مهمی در مقوله‌بندی و تصویرسازی دانش دامنه دارند ازاین‌رو مطالعات بسیاری در رابطه با روش و اصول طراحی، مراحل، ابزارهای ساخت و ارزیابی هستی‌شناسی‌ها و نمودار دانش در حوزه‌های مختلف صورت پذیرفته است. بر همین اساس روش‌های متعددی برای ساخت هستی‌شناسی و نمودار دانش وجود دارند که هرکدام دارای چرخه و مراحلی متفاوتی هستند (هماوندی و همکاران، ۱۳۹۹) و رویکردهای مختلف از دستی تا کاملاً خودکار را دربر می‌گیرند (محمدی استانی و همکاران، ۱۳۹۷). در روش دستی دانش حوزه شامل مفاهیم و روابط توسط کارشناسان انسانی در نرم‌افزارهای طراحی هستی‌شناسی تعریف می‌شود و عمدتاً از منابع دانشی مانند اصطلاح‌نامه‌های چاپی بهره گرفته می‌شود و یا کارشناسان با توجه به دانش و تخصص خود اطلاعات و مفاهیم موردنیاز را جمع‌آوری و طبقه‌بندی می‌کنند تا ساختار و محتوای هستی‌شناسی شکل بگیرد. فرآیند ایجاد یک هستی‌شناسی به‌صورت دستی زمان‌بر، پرهزینه و چالش‌برانگیز است که منجر به کشف روش‌های خودکار شده است (Alobaidi et al., 2018). در روش خودکار برای ساخت هستی‌شناسی از روش‌های پردازش زبان طبیعی و یادگیری ماشینی برای استخراج مفاهیم و روابط بین آن‌ها استفاده می‌شود. تولید هستی‌شناسی خودکار به‌طور قابل‌توجهی هزینه کار و زمان موردنیاز برای ساخت هستی‌شناسی‌ها را کاهش می‌دهد (Ma et al., 2019). در روش نیمه‌خودکار از ترکیب روش‌های دستی و خودکار برای ساخت هستی‌شناسی استفاده می‌شود. در این روش به‌طورمعمول از ابزارهای یادگیری ماشینی برای استخراج مفاهیم و روابط استفاده می‌شود (محمدی استانی و همکاران، ۱۳۹۷) سپس کارشناسان باتوجه به دانش خود هستی‌شناسی را بهبود می‌بخشند. به‌طورکلی ساخت هستی‌شناسی از هر روشی که انجام گیرد، یک فرایند زمان‌بر و هزینه‌بر است اما به‌منظور صرفه‌جویی در زمان و پرهیز از دوباره‌کاری در ساخت هستی‌شناسی دامنه می‌توان از اصطلاح‌نامه‌های موضوعی بهره گرفت. درواقع اصطلاح‌نامه‌ها با انعکاس درجه‌ای از اجماع کلی و دارا بودن مجموعهٔ به نسبت کاملی از

اصطلاحات هر حوزه موضوعی، قابلیت دسترسی آسان به مجموعه‌ای از اصطلاحات یک حوزه و روابط سلسله مراتبی که تعریف شده است را فراهم می‌آورند(دمرچی لو و حسینی بهشتی، ۱۴۰۰).

وجود یک هستی‌شناسی برای ساخت نمودار دانش امری مهم تلقی می‌شود؛ زیرا هستی‌شناسی می‌تواند تمام اطلاعات مورد نیاز برای ساخت نمودار دانش را فراهم کند و تأثیر حوزه و رفتار موجودیت‌ها را معرفی کند. وجود هستی‌شناسی همچنین به درک قوانین و محدودیت‌های خاص یک حوزه کمک می‌کند و باعث بهبود استنباط از آن می‌شود (Ehrlinger & Wöß, 2016). یک هستی‌شناسی به یک نمودار دانش اجازه می‌دهد تا یکپارچه‌سازی، استدلال و ارائه دانش را به روشی ساختاریافته و معنادار ارائه دهد (Hurlburt, 2021; Αντωνίου, 2020). با درج نمونه‌های داده در هستی‌شناسی، امکان تبدیل به یک نمودار دانش فراهم می‌شود. درواقع با اضافه شدن داده‌ها به هستی‌شناسی روابط جدید بین موجودیت‌های هستی‌شناسی تسهیل می‌شود و امکان استنتاج روابط یا مسیرهای جدید بین موجودیت‌ها (پیوندهای قابل استنتاج) را فراهم می‌کند(Yahya et al., 2021). باید در نظر داشت که امکان ساخت نمودار دانش بدون هستی‌شناسی نیز امکان‌پذیر است. در این روش می‌توان از داده‌های بدون ساختار حوزه موردنظر مانند مقالات علمی آن حوزه استفاده نمود. اما ساخت نمودار دانش براساس متون بدون ساختار برای یک دامنه جدید در غیاب یک هستی‌شناسی طولانی و پیچیده است. برای ساخت خودکار نمودار دانش براساس داده‌های بدون ساختار می‌توان از روش‌های یادگیری عمیق استفاده کرد(Shin et Jung et al., 2010 ; al.,2015). اما ایجاد نمودارهای دانش منحصراً با استفاده از روش‌های مبتنی بر یادگیری ماشین یک معضل تحقیقاتی بوده است، زیرا کیفیت نمودارهای دانش و سودمندی آن‌ها موردتردید خواهد بود(Issa et al., 2021). لذا استفاده از روش‌های یادگیری ماشینی در کنار نظارت متخصصان انسانی مطلوب‌تر خواهد بود(Chaudhri et al., 2022). در این روش باید موجودیت‌های مختلف (اشخاص، شرکت‌ها، محصولات، مکان‌ها و غیره) اسناد و روابط بین آن‌ها شناسایی شود که یکی از وظایف کلیدی در اجرای نمودار دانش است. موجودیت‌های مختلف و روابط بین آن‌ها باید طبق یک طرح کلی شناسایی شود(Agrawal et al., 2022 ;Sun et al., 2016). سپس باید برای هر موجودیت شناسایی‌شده، ویژگی‌ها و صفات مرتبط مشخص شود. برای مثال، در موجودیت شخص، ویژگی‌های مرتبط می‌تواند شامل نام، سن، آدرس و شماره تماس باشد. در مرحله بعد باید نوع روابط بین موجودیت‌ها بررسی و تعیین

شود، عموماً روابط در نمودار دانش می‌توانند یک‌به‌یک، یک به چند یا چند به چند باشند. در مرحله بعدی باید خصوصیت‌ها (نوع داده، محدودیت‌ها، ارتباطات با سایر موجودیت‌ها و غیره) و عملکردهای (نحوه برخورداری و استفاده از خصوصیت) هر رابطه تعیین شود. همچنین در این مرحله باید شروط، محدودیت‌ها، و اولویت‌بندی هر خصوصیت و عملکردها مشخص شود. در آخرین مرحله با ابزارهایی مانندNeo4j نمودار دانش رسم شود (Li et al., 2021; Sun et al., 2022; Song et al., 2022; Zhang et al., 2022).

- بررسی سؤال دوم پژوهش

کاربرد و نقش هستی‌شناسی و نمودار دانش در طبقه‌بندی خودکار اسناد چیست؟

همان‌طور که بیان شد معمولاً نمایش‌های فضای برداری از اسناد قادر به ثبت ویژگی‌ها فراتر از ویژگی‌های آماری ساده نیستند و عمدتاً اطلاعات معنایی را که اغلب به‌طور صریح با داده‌های ورودی مرتبط هستند نادیده می‌گیرند. درصورتی‌که دانش معنایی یک حوزه به‌راحتی از طریق نمودارهای دانش و هستی‌شناسی‌ها در دسترس است؛ زیرا آن‌ها حاوی مفاهیم معنایی، مقولات و روابط میان آن‌ها هستند. نادیده گرفتن این دانش معنایی می‌تواند مزایا یا معایبی را در الگوریتم‌های یادگیری ماشینی ایجاد نماید. مزیت ازاین‌جهت است که اگر دانش معنایی در یادگیری لحاظ نشود می‌تواند نمایش‌های بالقوه متفاوتی را از داده‌های ورودی ایجاد کند و معایب آن این است که با نادیده گرفتن غنای دانش موجود هر ویژگی مفیدی باید از ابتدا دوباره یاد گرفت که این امر به منابع آموزشی زیادی نیاز دارد که این امر هم دشواری‌های خاص خود را به وجود می‌آورد (Denecke, 2022). در پژوهش‌های بسیاری از هستی‌شناسی و نمودار دانش جهت طبقه‌بندی خودکار اطلاعات و متون استفاده شده است(Zhou & El-Mali & Jain, Chicaiza & Reátegui, 2020; ; Kastrati et al., 2019 Gohary, 2016 ;Li et al., 2024 ; Lan et al., 2021 ;2021).

هستی‌شناسی‌ها و نمودار دانش ابزاری ارزشمند در طبقه‌بندی اسناد متنی هستند (Lee et al., 2009;Wijewickrema, 2015). این ابزارها می‌توانند برای تقویت نتایج الگوریتم‌های یادگیری ماشین از طریق بهره‌برداری از دانش زمینه استفاده شود. این ابزارها برای دریافت معنای موجودیت‌ها و کلاس‌ها به‌منظور تغذیه مدل‌های یادگیری ماشینی استفاده می‌شوند. استفاده از این ابزارها به‌عنوان یک مرجع خارجی در حین فرآیند طبقه‌بندی عمل می‌کند و دانش زمینه را فراهم می‌نماید (مدنی، ۱۳۹۱).

درواقع این ابزار می‌تواند به‌طور مؤثر معانی کلمات را از جملات متن دارای ابهام آزاد نماید و بر مشکلی که در زبان طبیعی با آن مواجه است، غلبه کند که در آن یک کلمه ممکن است بسته به زمینه کاربردی معانی متعددی داشته باشد (Pan, 2015 ; Lee et al., 2009;Wijewickrema, 2015 Gruber,1993;). افزون بر موارد بیان‌شده در جدول ۱ براساس واکاوی پژوهش‌های صورت گرفته در این زمینه به تفکیک به برخی از کاربردهای هستی‌شناسی و نمودار دانش اشاره شده است.

جدول ۱- تنظیمات کاربرد هستی‌شناسی و نمودار دانش در طبقه‌بندی اسناد

کاربرد		توضیحات	مطالعات
افزایش بازنمایی اسناد	استخراج و انتخاب ویژگی‌های مرتبط و متمایز	هستی‌شناسی‌ها و نمودار دانش درک عمیق‌تری از اصطلاحات، مفاهیم و روابط خاص دامنه ارائه می‌دهند. استخراج ویژگی‌های مرتبط و آموزنده از متن موجب بهبود نمایش اسناد برای وظایف طبقه‌بندی می‌شود و با درک مفاهیم و موجودیت‌های مهم در حوزه، شناسایی متمایزترین ویژگی‌هایی که به طبقه‌بندی دقیق کمک می‌کنند آسان‌تر می‌شود.	Khan & Bhatti, 2012; Lei et al., 2019;، Li et al., 2017; Shan et al., 2020
	درک متنی	هستی‌شناسی و نمودار دانش به درک زمینه و تفاوت‌های ظریف زبان خاص دامنه مورداستفاده در اسناد کمک می‌کند. این درک، تفسیر دقیق‌تر متن و بازنمایی بهتر معنای آن را ممکن می‌سازد.	Khan & Bhatti, 2012; Lei et al., 2019; Varga, 2014; Xu & Sarikaya, 2014
	نمایش معنایی پیشرفته	هستی‌شناسی و نمودار دانش دامنه امکان ترکیب اطلاعات معنایی را در بازنمایی سند فراهم می‌کند. با استفاده از این ابزارها می‌توان روابط سلسله‌مراتبی، مترادف‌ها و مفاهیم مرتبط را به دست آورد که می‌تواند درک معنایی و بازنمایی اسناد را بهبود بخشید.	Varga, 2014; Li et al., 2017; Sinoara et al., 2019
	پیش‌پردازش اختصاصی دامنه	هستی‌شناسی و نمودار دانش دامنه می‌توانند مراحل پیش‌پردازش مختص به حوزه موضوعی را راهنمایی کنند. به‌عنوان‌مثال، درک اختصارات یا کلمات اختصاری خاص دامنه می‌تواند به ارائه بهتر این عبارات در طول فرایندهای توکن‌سازی یا نرمال‌سازی کمک کند.	Varga, 2014; Li et al., 2017; Lu et al., 2008; Mali & Atique, 2021; Rozeva, 2012

کاربرد		توضیحات	مطالعات
استانداردسازی متن	استانداردسازی اصطلاحات	هستی‌شناسی و نمودار دانش دامنه می‌توانند فهرستی از اصطلاحات استانداردشده، اختصارات یا کلمات اختصاری را که معمولاً در دامنه استفاده می‌شود، ارائه دهند. با اطمینان از استفاده مداوم از این اصطلاحات، مدل‌های طبقه‌بندی متن می‌توانند به‌طور دقیق اسناد را بر اساس وجود یا عدم وجود اصطلاحات خاص دامنه شناسایی و طبقه‌بندی کنند.	Brscic et al., 2021; Perez et al., 2022
	شناسایی و نرمال‌سازی موجودیت‌ها	هستی‌شناسی و نمودار دانش دامنه می‌تواند حاوی فهرستی از موجودیت‌های مرتبط با دامنه باشد، مانند نام سازمان‌ها، مکان‌ها یا محصولات افراد. با استفاده از این دانش، مدل‌های طبقه‌بندی متن می‌توانند این موجودیت‌ها را با دقت شناسایی کنند.	Nguyen et al., 2019; Ren et al., 2015
	پیش‌پردازش اسناد خاص	هستی‌شناسی و نمودار دانش دامنه می‌تواند شامل قوانین پیش‌پردازش باشد که نحوه استفاده از انواع خاصی از متن را مشخص می‌کند. برای مثال، آن‌ها ممکن است نحوه کار با کاراکترهای خاص، علائم نگارشی یا عبارات عددی را مشخص کنند. با پیروی از این قوانین در طول پیش‌پردازش، مدل‌های طبقه‌بندی متن می‌توانند نمایش متن را استاندارد کرده و دقت طبقه‌بندی را بهبود بخشند.	HaCohen-Kerner et al., 2020

به‌طورکلی هدف اصلی استفاده از هستی‌شناسی‌ها و نمودار دانش در طبقه‌بندی اسناد بهبود دقت و افزایش قابلیت اطمینان مدل‌های به‌دست‌آمده است که بتواند اطلاعات را به‌صورت مناسب و دقیق سازمان‌دهی کند. درواقع هستی‌شناسی‌ها و نمودار دانش به ماشین‌ها اجازه می‌دهند معنای داده‌هایی را که با آن‌ها کار می‌کنند، درک کنند (Galkin et al., 2017). درنتیجه موجب بهبود طبقه‌بندی اسناد می‌شوند. طبقه‌بندی اسناد متنی با استفاده از این ابزارها می‌تواند مزایای زیر را فراهم سازد: کشف رابطه ضمنی یا صریح بین کلمات، استخراج و استفاده از روابط نهفته بین کلمات و اسناد، قابلیت ایجاد کلمات کلیدی نماینده برای رده‌های موجود، درک معنایی متون که باعث افزایش دقت طبقه‌بندی می‌شود. همچنین مدیریت مترادف‌ها و چند جمله در مقایسه با الگوریتم‌های طبقه‌بندی متنی فرکانسی بهبود می‌یابد، زیرا از معنا استفاده می‌کنند (Altınel & Ganiz, 2018).

- بررسی سؤال سوم پژوهش

چالش‌های کاربرد هستی‌شناسی و نمودار دانش در طبقه‌بندی اسناد متنی چه هستند؟

هستی‌شناسی و نمودار دانش در مطالعات بسیاری جهت افزایش دقت مدل‌های طبقه‌بندی اسناد به کار گرفته شده است اما استفاده از این ابزارها با چالش‌هایی همراه است. مهم‌ترین چالش مطرح‌شده ساخت و ادغام هستی‌شناسی یا نمودار دانش است. تولید هستی‌شناسی‌ها و نمودار دانش امری خسته‌کننده، زمان‌بر و پیچیده است زیرا به مهندسی دانش فراوان نیاز دارد که می‌تواند امکان‌پذیری و ارزش عملی آن را برای بسیاری از کاربردها محدود کند (Al-Arfaj & Al-Salman, 2015Hashemi et al., 2018;). لذا برای استفاده از قابلیت‌های این ابزارها، تعدادی هستی‌شناسی و نمودار دانش عمومی وجود دارد که می‌توان از آن‌ها بهره گرفت. برخی از مهم‌ترین این ابزارها که روابط معنایی میان مفاهیم را نشان می‌دهد عبارت‌اند از: وردنت[1]، دی بی پدیا[2]، یاگو[3]، ویکی داده[4]، فارس نت[5]، و ویکشنری[6]. هرکدام از موارد بیان‌شده دامنه‌های متعددی را پوشش می‌دهد که نشان‌دهنده تنوع گسترده‌ای از موجودیت‌ها و روابط است که در پژوهش‌های متعددی به‌منظور طبقه‌بندی اسناد متنی به کار گرفته شدند (Bouchiha et al., 2023; Yousif et al., 2019; Allahyari et al., 2014Wasi et al., 2020; مدنی، ۱۳۹۱). هستی‌شناسی‌ها و نمودار دانش معرفی‌شده جنبه عمومی دارند و برای حیطه تخصصی با محدودیت‌هایی روبه‌رو هستند و نمی‌توانند به‌طور کامل تمام مفاهیم و روابط معنایی یک حوزه خاص را پوشش دهند، اما باوجود دشواری (زمان‌بر بودن، هزینه‌بر بودن و تخصصی بودن) در ساخت و توسعه هستی‌شناسی‌ها و نمودار دانش تخصصی اغلب در رویکردهای مبتنی بر دانش برای طبقه‌بندی متن از منابع دانشی دامنه استفاده می‌شود. هستی‌شناسی و نمودار دانش دامنه هنگامی‌که روی مجموعه‌ای از متون در یک حوزه موضوعی خاص اعمال شود، بسیار مؤثرتر از منابع دانشی عمومی خواهد بود. بااین‌حال برای مجموعه‌های متنوع اسناد، استفاده از هستی‌شناسی یا نمودار دانش دامنه امکان‌پذیر نیست.

1. WordNet
2. DBPedia
3. Yago
4. Wikidata
5. FarsNet
6. Wiktionary

همچنین ساختار این ابزارها حاوی اطلاعاتی هستند که می‌تواند در قالب مفهوم نزدیکی، مترادف، ابرنام، انواع رابطه و غیره استفاده شود و با گذشت زمان، گستره منابع دانشی عمومی بزرگ‌تر می‌شوند، که موجب بهبود عملکرد آن‌ها می‌شود (Dumitrescu et al., 2013).

کاربرد هستی‌شناسی و نمودار دانش در طبقه‌بندی اسناد متنی فارسی علاوه بر چالش مطرح‌شده در قسمت قبل با چالش‌های دیگری هم مواجه است. به‌طورکلی چالش موجود در زبان فارسی به‌منظور طبقه‌بندی اسناد متنی با استفاده از هستی‌شناسی و یا نمودار دانش عبارت‌اند از:

- **ویژگی‌های خاص زبان فارسی در نگارش متون**

زبان فارسی دارای ویژگی‌های خاصی است که چالش‌هایی را در محیط دیجیتال و پردازش زبان طبیعی ایجاد می‌کند. به‌طورکلی۴۳ گروه چالش نگارشی در متون به زبان فارسی معرفی شده است (ستوده و هنرجویان، ۱۳۹۱). برخی از این چالش‌ها عبارت‌اند از: پیوسته یا جدانویسی، تنوع نشانه‌های جمع، تنوع دگرنوشته‌ها، فاصلهٔ بین حروف واژه. این ویژگی‌ها باعث می‌شوند پردازش زبان فارسی در مقایسه با زبان‌های دیگر چالش‌برانگیزتر باشد. چالش‌های زبان فارسی در پردازش زبان طبیعی در مطالعات متعددی موردتوجه قرار گرفته است. باقری و همکاران [1] (۲۰۱۳) مدلی را برای طبقه‌بندی احساسات در زبان فارسی با در نظر گرفتن موضوعاتی مانند پسوندهای عمودی، فاصله بین واژه‌ها و زبان غیررسمی ارائه کرد. حسینی پوزوه و همکاران [2] (۲۰۱۸) یک هستی‌شناسی برای بهبود وظایف مختلف زبان پردازش طبیعی فارسی ازجمله برچسب‌گذاری بخشی از گفتار و شناسایی موجودیت نام‌گذاری شده معرفی کرد. خشابی و همکاران [3] (۲۰۲۱) معیاری برای وظایف درک زبان فارسی را توسعه داد و آن را با عملکرد انسان مقایسه کرد. حبیب [4] (۲۰۲۱) از رویکردهای یادگیری ماشینی برای رسیدگی به چالش‌ها در طبقه‌بندی اسناد فارسی استفاده کرد و به نتایج امیدوارکننده‌ای دست یافت. مطالعات در این زمینه پتانسیل یادگیری ماشینی در غلبه بر چالش‌های زبان فارسی را برجسته می‌کند و تاکنون بسیاری از چالش موجود رفع شده است

1. Bagheri et al.
2. Hosseini Pozveh et al.
3. Khashabi et al.
4. Habib

اما درمجموع نیاز به پژوهش‌های بیشتری در این زمینه احساس می‌شود.

- **چالش‌های فنی**

علی‌رغم وجود چندین هستی‌شناسی و نمودار دانش چندزبانه شاهد هستیم که این ابزارها به منابع فارسی داده‌های پیوند کمی دارند یا اصلاً پیوندی وجود ندارند و از ابزارها برای استخراج دانش از منابع اطلاعاتی فارسی به‌طور کامل پشتیبانی نمی‌کنند (Asgari-Bidhendi et al., 2019). متفاوت بودن ساختارهای دستور زبانی و نحوی زبان فارسی با سایر زبان‌ها مانند انگلیسی یکی از دلایل عدم پشتیبانی کامل از زبان فارسی است. این تفاوت در ساختار زبانی می‌تواند موجب سختی در تبدیل هستی‌شناسی و نمودارهای دانش عمومی به فارسی شود. به‌عنوان‌مثال، برخی اصطلاحات و ساختارهای زبانی در فارسی به‌صورت ترکیبی و مضاعف وجود دارند که ممکن است در به‌کارگیری این ابزارهای برای متون فارسی مشکل ایجاد کند. لذا استفاده از زبان‌های هستی‌شناسی رایج مانند WWL در هستی‌شناسی‌ها و نمودارهای دانش برای زبان فارسی محدود است و استفاده از قوانین وجودی ممکن است منجر به استدلال غیرقابل تصمیم‌گیری شود(Krötzsch & Thost, 2016). به‌طورکلی مشکلات فنی در به‌کارگیری ابزارهای معنایی در یادگیری ماشینی ازجمله طبقه‌بندی اسناد متنی فارسی چندوجهی است که می‌توان به موارد زیر اشاره نمود:

- **کمبود منابع**: تحلیل متن فارسی از کمبود منابع رنج می‌برد. مجموعه داده جامع باعث بهبود وظایف پردازش زبان طبیعی می‌شود؛ اما ازآنجایی که تعداد منابع فارسی مانند کتاب‌ها، مقالات، روزنامه‌ها، نشریات و منابع دیگر نسبت به زبان‌هایی مانند انگلیسی محدود است، لذا شناخت متون فارسی برای یادگیری ماشینی دارای محدودیت است همین امر سبب ایجاد ناتوانی ابزارها برای تحلیل و پردازش متون فارسی می‌شود (Hosseini Pozveh et al., 2018; Habib, 2021).
- **ناهمخوانی و عدم تطابق ترجمه**: چالش عدم تطابق ترجمه در هستی‌شناسی و نمودارهای دانش چندزبانه یک نگرانی کلیدی در تطبیق آن‌ها در سایر زبان‌ها است. هستی‌شناسی‌ها و نمودارهای دانش عمومی با استفاده از منابع مختلفی ازجمله ویکی‌پدیا، جداول وب و متون بدون ساختار ایجاد شده‌اند که عمدتاً از زبان انگلیسی هستند و برای سایر زبان‌ها معادل آن استفاده می‌شود که در برخی موارد ناهماهنگی در ترجمه‌ها ایجاد می‌شود (Dos Santos et al., 2010).

به‌عبارت‌دیگر از اصطلاحاتی استفاده شود که در زبان فارسی معادل دقیقی وجود نداشته باشد یا بالعکس. همین امر می‌تواند منجر به سوءتفاهم و تفسیر نادرست شود.

- **جامع نبودن منابع دانشی فارسی**: علیرغم وجود منابع دانش فارسی مانند فارس بیس[1]، اف‌کی‌جی[2] و فارس نت (Shirmardi, 2022) اما آن‌ها به‌اندازه همتایان جهانی خود جامع نیستند. به‌عنوان‌مثال، فارس بیس اولین نمودار دانش چند منبعی فارسی است، اما پوشش آن در مقایسه با سایر نمودارهای دانش در مقیاس بزرگ محدود است (Asgari-Bidhendi et al., 2019). به‌طور مشابه اف‌کی‌جی که شامل بیش از ۵۰۰ هزار موجودیت و ۷ میلیون رابطه است، به‌اندازه سایر پایگاه‌های دانش جهانی گسترده نیست (Sajadi et al., 2020) و فارس نت، هستی‌شناسی واژگانی زبان فارسی نیز ازنظر جامعیت نسبت به همتای خود وردنت دارای محدودیت‌هایی است.

به‌طورکلی توسعه و استفاده از هستی‌شناسی و نمودارهای دانش عمومی و یا دامنه در بحث پردازش زبان طبیعی فارسی و همچنین طبقه‌بندی اسناد نیازمند تلاش‌های بیشتری است، بعلاوه مستلزم توجه به ویژگی‌های نحوی و دستور زبانی زبان فارسی است.

نتیجه‌گیری

اخیراً مطالعات مرتبط نشان داده‌اند ابزارهای دانش معنایی ازجمله هستی‌شناسی و نمودار دانش را می‌توان برای تقویت نتایج روش‌های یادگیری ماشین از طریق بهره‌برداری از دانش زمینه استفاده کرد. به‌ویژه در سال‌های گذشته، شاهد توسعه روزافزون هستی‌شناسی‌ها و نمودارهای دانش هستیم که تعداد زیادی موجودیت و روابط آن‌ها را توصیف می‌کنند؛ که از آن‌ها می‌توان برای دریافت معنای مفاهیم و برچسب‌ها برای تقویت مدل‌های یادگیری استفاده کرد. استفاده از هستی‌شناسی و نمودار دانش در مدل‌های طبقه‌بندی اسناد متنی می‌تواند باعث بهبود دقت و عملکرد مدل‌های طبقه‌بندی اسناد شود. این ابزارها جهت استاندارد کردن اصطلاحات موجود در اسناد و جملات

1. FarsBase
2. Farsi Knowledge Graph (FKG)

استفاده می‌شود، سپس برای طبقه‌بندی اسناد از شکل استانداردشده اسناد استفاده خواهد شد. درواقع با استفاده از این ابزارها اسناد و واژگان موجود آن به مفاهیم دامنه تبدیل می‌شوند که در این صورت مشکلات عدم تطابق کلمات یا ابهام را برطرف می‌نماید. به‌طورکلی با نگاشت اسناد به مفاهیم و توصیفگرهای نمودار دانش و هستی‌شناسی می‌توان ارتباط بین یک سند و مفهومی خاص را در نظر گرفت و برای ارزیابی مفاهیم نماینده هر دسته از اسناد استفاده شود. همچنین استفاده از مفاهیم معرف برای انعکاس اسناد می‌تواند ابعاد فضای بردار ویژگی را کاهش دهد، زیرا تعداد مفاهیم دامنه به‌طور قابل‌توجهی نسبت به واژگانی که در فن‌های طبقه‌بندی مبتنی بر ویژگی واژگانی ظاهر می‌شوند کمتر است. علاوه بر این، بردارهای سند نشان داده‌شده توسط مفاهیم دامنه قابل‌درک هستند که نتایج طبقه‌بندی را قابل تفسیرتر می‌کند (Nguyen et al.,2021 Lee et al., 2021;). با استفاده از دانش معنایی در حین طبقه‌بندی منجر به فراتر رفتن از معیارهای ظاهری و آماری می‌شود و با جایگزین کردن مفاهیم به‌جای کلمات مرتبط در متن، موجب افزایش وزن آن مفهوم غالب در متن می‌شود که همین امر در افزایش دقت طبقه‌بندی نقش بسزایی دارد (Song et al.,2005 Denecke, 2022؛ مدنی، ۱۳۹۱؛ هاشمی و حورعلی، ۱۳۹۶). همچنین معنای مفاهیم و نحوه ارتباط آن‌ها با اسناد و مفاهیم اسناد دیگر مشخص می‌شود و با تعبیه این اطلاعات در الگوریتم‌های یادگیری ماشین می‌تواند به معناشناسی بیشتری منجر شود و موجب یادگیری بهتر ارتباط بین مفاهیم متن و طبقه‌های هدف شود. که برای استفاده از این منابع دانشی تعدادی منابع دانش عمومی وجود دارد مانند وردنت، یاگو، دی بی پدیا، ویکی دیتا و فارس نت است که بسیاری از موجودیت‌ها و روابط موجود میان آن‌ها را توصیف می‌کند اما برای استفاده در حوزه‌های خاص محدودیت‌هایی دارند که نمی‌توانند تمامی مفاهیم موجود در یک حیطه موضوعی خاص را پوشش دهند، لذا برای استفاده از این ابزار جهت طبقه‌بندی اسناد یک حوزه لازم است که از نمودار دانش یا هستی‌شناسی خاص همان حوزه بهره گرفت.

منابع

بهروزیان نژاد، محمد، عطار زاده، ایمان، افتخار، شادی، کاظمی، احمد و شکیبا فخر، محسن (۱۳۹۳). *استفاده از تکنیک داده‌کاوی در دسته‌بندی خودکار اسناد متنی*. اولین همایش ملی مهندسی کامپیوتر و فناوری اطلاعات دانشگاه پیام نور. اصفهان: ۱۵-۲۳. https://civilica.com/doc/337433

دمرچی لو، منصوره و حسینی بهشتی، ملوک‌السادات (۱۴۰۰). قابلیت تبدیل اصطلاح‌نامه به هستی‌شناسی (مرور سیستماتیک). *پژوهشنامه کتابداری و اطلاع‌رسانی*، ۱۱(۲): ۱۰۵-۱۲۷.

رمضانی، هادی، علیپورحافظی، مهدی و مؤمنی، عصمت (۱۳۹۳). نقشه‌های علمی: فنون و روش‌ها. *ترویج علم*، ۵ (۶): ۵۳-۸۴.

ریحانی آرانی، احسان و لاجوردی، محمدرضا (۱۳۹۵). *بررسی روش‌های طبقه‌بندی خودکار اسناد متنی*. کنفرانس ملی برق و کامپیوتر سیستم‌های توزیع‌شده و شبکه‌های هوشمند، کاشان: https://civilica.com/doc/622157 ،۵۴۶-۵۴۰

ستوده، هاجر و هنرجویان، زهره (۱۳۹۱). مروری بر دشواری‌های زبان فارسی در محیط دیجیتال و تأثیرات آن‌ها بر اثربخشی پردازش خودکار متن و بازیابی اطلاعات. *کتابداری و اطلاع‌رسانی*، ۱۵(۴): ۵۹-۹۲.

سلیمانی نژاد، عادل، سلاجقه، مژده و طبیبی‌نیا، الهام (۱۳۹۷). خوشه‌بندی مقالات علمی بر پایهٔ الگوریتم k_mean مطالعه موردی: پایگاه پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک). *پژوهشنامه پردازش و مدیریت اطلاعات*، ۳۴(۲):۸۷۱-۸۹۶.

محمدی استانی، مرتضی، آذرگون، مریم و چشمه سهرابی، مظفر (۱۳۹۷). روش‌شناسی ساخت و طراحی هستی نگاشت‌ها: مورد پژوهی علم‌سنجی. *پردازش و مدیریت اطلاعات*،۳۳(۴): ۱۷۶۵-۱۷۹۲.

مدنی، صباالسادات (۱۳۹۱). *دسته‌بندی اسناد فارسی به کمک هستان‌شناسی فارس نت*. پایان‌نامه کارشناسی ارشد، دانشگاه صنعتی شاهرود، شاهرود.

هاشمی، سیامک و حور علی، مریم (۱۳۹۶). *دستهٔ اخبار فارسی حوزه دفاعی با استفاده از هستان‌شناسی*. دومین کنفرانس بین‌المللی پژوهش‌های دانش‌بنیان در مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه خوارزمی، تهران: ۱-۱۵.

هماوندی، هدی، فهیم نیا، فاطمه، ناخدا، مریم و حسینی بهشتی، ملوک‌السادات (۱۳۹۹). مطالعه روش‌های ایجاد هستی‌شناسی: شناسایی مؤلفه‌ها و ویژگی‌ها بر مبنای تحلیل پژوهش‌های انجام‌شده. *تحقیقات کتابداری و اطلاع‌رسانی دانشگاهی*، ۵۴(۱): ۱۳-۳۹.

References

Agrawal, G., Deng, Y., Park, J., Liu, H., & Chen, Y. C. (2022). Building Knowledge Graphs from Unstructured Texts: Applications and Impact Analyses in Cybersecurity Education. *Information*, 13(11): 526. DOI: 10.3390/info13110526

Al-Arfaj, A., & Al-Salman, A. (2015). Ontology construction from text: challenges and trends. *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, 6(2): 15-26. URL:

https://www.cscjournals.org/library/manuscriptinfo.php?mc=IJAE-169

Allahyari, M., Kochut, K. J., & Janik, M. (2014). *Ontology-based text classification into dynamically defined topics.* In 2014 IEEE International Conference on Semantic Computing (pp.273-278). IEEE. DOI**:** 10.1109/ICSC.2014.51

Alobaidi, M., Malik, K. M., & Sabra, S. (2018). Linked open data-based framework for automatic biomedical ontology generation. *BMC bioinformatics*, 19(1): 1-13. DOI: 10.1186/s12859-018-2339-3

Altınel, B., & Ganiz, M. C. (2018). Semantic text classification: A survey of past and recent advances. *Information Processing and Management,* 54(6): 1129–1153. DOI: 10.1016/j.ipm.2018.08.001

Asgari-Bidhendi, M., Hadian, A., & Minaei-Bidgoli, B. (2019). Farsbase: The persian knowledge graph. *Semantic Web,* 10(6): 1169-1196. DOI: 10.3233/SW-190369

Αντωνίου, Τ. Ά. (2020). *Ontology-based application for knowledge management in ancient Greek mythology*, PhD Thesis, Αριοταστοτ\acute\varepsilonλειοταο Πανεπιοταστ\acute\etaμιοταο Θεσσαλον\ac. τειοτακης. https://ikee.lib.auth.gr/record/320900

Bagheri, A., Saraee, M., & de Jong, F. (2013, May). *Sentiment classification in Persian: Introducing a mutual information-based method for feature selection*. In 2013 21st Iranian conference on electrical engineering (pp.1-6). DOI: 10.1109/IranianCEE.2013.6599671

Behrouziannejad, M., Attarzadeh, I., Eftekhar, S., Kazemi, A., & Shakibafakhr, M. (2015, March). *Using data mining techniques in the automatic classification of text documents*. The first national conference of computer engineering and information technology of Payam Noor University, Isfahan.15-23. https://civilica.com/doc/337433 [In Persian]

Bloehdorn, S., Cimiano, P., Hotho, A., & Staab, S. (2005). An Ontology-based Framework for Text Mining. *LDV Forum - GLDV Journal for Computational Linguistics and Language Technology*, *20*(1): 87–112. DOI: 10.21248/jlcl.20.2005.70

Bouchiha, D., Bouzianae, A., & Doumi, N. (2023). Ontology based Feature Selection and Weighting for Text classification using Machine Learning. *Journal of Information Technology and Computing,* 4(1): 1-14. DOI: 10.48185/jitc.v4i1.612

Brscic, M., Contiero, B., Magrin, L., Riuzzi, G., & Gottardo, F. (2021). The use of the general animal-based measures codified terms in the scientific literature on farm animal welfare. *Frontiers in Veterinary Science*, 8, 634498. https://doi.org/10.3389/fvets.2021.634498

Burgueño, L., Hilken, F., Vallecillo, A., & Gogolla, M. (2017). Testing Transformation Models Using Classifying Terms. In E. Guerra and M. Van Den Brand (Eds.), *Theory and Practice of Model Transformation*, 10374, 69–85. Springer International Publishing. DOI:10.1007/978-3-319-61473-1_5

Chaudhri, V., Baru, C., Chittar, N., Dong, X., Genesereth, M., Hendler, J., Kalyanpur, A., Lenat, D., Sequeda, J., Vrandečić, D., & Wang, K. (2022). Knowledge Graphs: Introduction, History and, Perspectives. *AI Magazine*, 43(1): 17-29. https://doi.org/10.1002/aaai.12033

Chen, X., Jia, S., & Xiang, Y. (2020). A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications, 141*, 112948. DOI: https://doi.org/10.1016/j.eswa.2019.112948

Chen, Z. Y., Shang, Y., & Qian, D. M. (2018). Research on intelligent question answering system based on knowledge graph. *Computer Applications and Software*, 35(2): 178–182.

Chicaiza, J., & Reátegui, R. (2020). *Using domain ontologies for text classification. A use case to classify computer science papers*. In Knowledge Graphs and Semantic Web: Second Iberoamerican Conference and First Indo-American Conference, KGSWC 2020, Mérida, Mexico, November 26–27, 2020, Proceedings 2 (pp.166-180). Springer International Publishing. DOI:10.1007/978-3-030-65384-2_13

Dalal, M. K., & Zaveri, M. A. (2011). Automatic text classification: A technical review. *International Journal of Computer Applications, 28*(2): 37–40. DOI:10.5120/3358-4633

Damerchiloo, M., & Hosseini Beheshti, M. S. (2021). Converting Thesaurus to Ontology (a Systematic Review). *Library and Information Science Research*, 11(2): 105-127. DOI: 10.22067/infosci.2021.23662.0. [In Persian]

Denecke, K. (2022). Does Enrichment of Clinical Texts by Ontology Concepts Increases Classification Accuracy? *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation, 290*, 602–606. DOI: https://doi.org/10.3233/SHTI220148

Dos Santos, C. T., Quaresma, P., & Vieira, R. (2010). *An API for multilingual ontology matching*. In Proc. 7th conference on Language Resources and Evaluation Conference (LREC) (pp. 3830-3835). No commercial editor.URL: http://www.lrec-conf.org/proceedings/lrec2010/pdf/691_Paper.pdf

Dumitrescu, S. D., Trausan-Matu, S., Brut, M., & Sedes, F. (2013*). Ontology-based flexible topic classification of crowdsourcing textual resources. Proceedings of the Fifth International*

Conference on Management of Emergent Digital EcoSystems (pp.145–151). DOI: https://doi.org/10.1145/2536146.2536172

Ehrlinger, L., & Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTICS (Posters, Demos, SuCCESS)*, 48(1-4): 2. URL: https://ceur-ws.org/Vol-1695/paper4.pdf

Fensel, D., Horrocks, I., Van Harmelen, F., Decker, S., Erdmann, M., & Klein, M. (2000). *OIL in a Nutshell*, 1–16. DOI: https://doi.org/10.1007/3-540-39967-4_1

Fkih, F., & Omri, M. N. (2020). Hidden data states-based complex terminology extraction from textual web data model. *Applied Intelligence*, *50*(6): 1813–1831. DOI: https://doi.org/10.1007/s10489-019-01568-4

Galkin, M., Auer, S., Vidal, M. E., & Scerri, S. (2017, April). Enterprise Knowledge Graphs: A Semantic Approach for Knowledge Management in the Next Generation of Enterprise Information Systems. *Proceedings of the 19th International Conference on Enterprise Information Systems* (pp.88–98). DOI: https://doi.org/10.5220/0006325200880098

Gomez-Perez, J. M., Pan, J. Z., Vetere, G., & Wu, H. (2017). Enterprise Knowledge Graph: An Introduction. In *Exploiting Linked Data and Knowledge Graphs in Large Organisations*, 1–14. Springer International Publishing. Doi: https://doi.org/10.1007/978-3-319-45654-6_1

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Creation Diffusion Utilization*, *5*(April): 199–220.URL: https://tomgruber.org/writing/ontolingua-kaj-1993.pdf

Guo, L., Yan, F., Li, T., Yang, T., & Lu, Y. (2022). An automatic method for constructing machining process knowledge base from knowledge graph. *Robotics and Computer-Integrated Manufacturing*, *73*1, 02222. DOI:10.1016/j.rcim.2021.102222

Habib, M. K. (2021). The challenges of Persian user-generated textual content: A machine learning-based approach. *arXiv preprint*, 2101.08087. Doi: https://doi.org/10.48550/arXiv.2101.08087

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PloS one*, 15(5): e0232525. Doi: 10.1371/journal.pone.0232525

Hashemi, P., Khadivar, A., & Shamizanjani, M. (2018). Developing a domain ontology for knowledge management technologies. *Online Information Review*, 42(1): 28-44. DOI:10.1108/OIR-07-2016-0177

Hashemi, S., & Horali, M. (2016). *Category of Persian news in the field*

of defense using ontology. Second International Conference on Knowledge-Based Research in Computer Engineering and Information Technology: 1-15. [In Persian]

Homavandi, H., Fahimnia, F., Nakhoda, M., & Hoseini Beheshti, M. (2021). A study on ontology building methods: understanding of the features and requirements. *Academic Librarianship and Information Research,*54(1): 13-39. [In Persian]

Hosseini Pozveh, Z., Monadjemi, A., & Ahmadi, A. (2018). FNLP ONT: A feasible ontology for improving NLP tasks in Persian. *Expert Systems*, 35(4): e12282. DOI:10.1111/exsy.12282

Hurlburt, G. F. (2021). The Knowledge Graph as an Ontological Framework. *IT Professional*, 23(4): 14-18. DOI: 10.1109/ MITP.2021.3086918

Issa, S., Adekunle, O., Hamdi, F., Cherfi, S.S.S., Dumontier, M., & Zaveri, A. (2021). Knowledge graph completeness: Asystematic literature review. *IEEE Access*, 9, 31322–31339. Doi: https://DOI.org/10.1109/ACCESS.2021.3056622

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science*, 37(5): 499–514. DOI: 10.1177/0165551511417785

Jung, Y., Ryu, J., Kim, K. M., & Myaeng, S. H. (2010). Automatic construction of a large-scale situation ontology by mining how-to instructions from the web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 8(2-3): 110-124. DOI: http://dx.doi.org/10.2139/ssrn.3199480

Kastrati, Z., Imran, A. S., & Yayilgan, S. Y. (2019). The impact of deep learning on document classification using semantically rich representations. *Information Processing & Management*, 56(5): 1618-1632. DOI: https://doi.org/10.1016/j.ipm.2019.05.003

Khan, S. A., & Bhatti, R. (2012). Application of social media in marketing of library and information services: A case study from Pakistan. *Webology*, 9(1): 1-8. URL: http://www.webology.org/2012/v9n1/a93.html

Khashabi, D., Cohan, A., Shakeri, S., Hosseini, P., Pezeshkpour, P., Alikhani, M., & Yaghoobzadeh, Y. (2021). Parsinlu: a suite of language understanding challenges for persian. *Transactions of the Association for Computational Linguistics*, 9,1147-1162. DOI:10.1162/tacl_a_00419

Kilimci, Z. H., & Akyokus, S. (2019). *The Evaluation of Word Embedding Models and Deep Learning Algorithms for Turkish Text Classification*. 2019 4th International Conference on Computer

Science and Engineering (UBMK) (pp.548–553). DOI: https://doi.org/10.1109/UBMK.2019.8907027

Krötzsch, M., & Thost, V. (2016). *Ontologies for knowledge graphs: Breaking the rules*. In The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, Proceedings, Part I, 15, 76-392. Springer International Publishing. DOI:10.1007/978-3-319-46523-4_23

Lai, S., Xu, L., Liu, K., & Zhao, J. (2015). Recurrent Convolutional Neural Networks for Text Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1). DOI: https://doi.org/10.1609/aaai.v29i1.9513

Lan, G., Li, Y., Hu, M., Sun, Y., & Zhang, Y. (2021*). Knowledge Graph Integrated Graph Neural Networks for Chinese Medical Text Classification*. In 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 682-687, IEEE. DOI: 10.1109/BIBM52615.2021.9669286

Lee, Y. H., Tsao, W. J., & Chu, T. H. (2009). *Use of ontology to support concept-based text categorization. In Designing E-Business Systems*. Markets, Services, and Networks: 7th Workshop on E-Business, WEB 2008, Paris, France. Revised Selected Papers 7, 201-213, Springer Berlin Heidelberg. DOI:10.1007/978-3-642-01256-3_17

Lee, Y. H., Hu, P. J. H., Tsao, W. J., & Li, L. (2021). Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. *Expert Systems with Applications, 174*, 114681. DOI: https://doi.org/10.1016/j.eswa.2021.114681

Lei, X., Cai, Y., Xu, J., Ren, D., Li, Q., & Leung, H. F. (2019). *Incorporating task-oriented representation in text classification*. In Database Systems for Advanced Applications: 24th International Conference, DASFAA 2019, Chiang Mai, Thailand, Proceedings, Part II 24, 401-415.Springer International Publishing. https://doi.org/10.1007/978-3-030-18579-4_24

Li, L., Zhang, Z., & Zhang, S. (2021). Knowledge graph entity similarity calculation under active learning. *Complexity*, 2021, 1-11. Doi: https://doi.org/10.1155/2021/3522609

Li, S., Chen, L., Song, C., & Liu, X. (2024). Text Classification Based on Knowledge Graphs and Improved Attention Mechanism. *arXiv preprint*, 2401.03591. DOI: https://doi.org/10.48550/arXiv.2401.03591

Li, Y., Wei, B., Yao, L., Chen, H., & Li, Z. (2017). Knowledge-based document embedding for cross-domain text classification. In *2017 International Joint Conference on Neural Networks (IJCNN)*,

1395-1402. IEEE. DOI: 10.1109/IJCNN.2017.7966016

Lili, D., Jiong, C., Xiang, Z., & Na, Y. E. (2020). Research on disease diagnosis method combining knowledge graph and deep learning. *Journal of Frontiers of Computer Science and Technology,* 14(5): 815. URL: https://arxiv.org/pdf/2305.00359.pdf

Lu, H., Zhengtao, Y., Jinhui, D., Cheng, Z., Cunli, M., & Jianyi, G.2008. *The effects of domain knowledge relations on domain text classification*. In 2008 27th Chinese Control Conference, 460-463. IEEE. DOI: 10.1109/CHICC.2008.4605079

Ma, Z., Cheng, H., & Yan, L. (2019). Automatic construction of OWL ontologies from Petri nets. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 15(1): 21-51. DOI: 10.4018/IJSWIS.2019010102

Madani, S. (2011). *Classification of Persian documents with the help of Fars Net ontology*. Master thesis, Shahrood University of Technology. [In Persian]

Mali, M., & Atique, M. (2021). *The Relevance of Preprocessing in Text Classification*. in Proceedings of Integrated Intelligence Enable Networks and Computing, in Algorithms for Intelligent Systems. Singapore: Springer, 553–559. DOI: 10.1007/978-981-33-6307-6_55.

Malik, S., & Jain, S. (2021). Semantic Ontology-Based Approach to Enhance Text Classification. *ISIC*, 85–98. URL: http://star.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-2786/Paper16.pdf

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., &d Gao, J. (2020). Deep Learning Based Text Classification: A Comprehensive Review. *ACM Computing Surveys (CSUR)*, 54(3): 1–40. DOI: https://doi.org/10.48550/arXiv.2004.03705

Mohammadi Ostani, M., Azargoon, M., & Cheshmesohrabi, M. (2018). Methodology of Construction and Design of Ontologies: a Case Study of Scientometrics Field. *Iranian Journal of Information Processing and Management*, 33(4): 1761-1788. DOI: 10.35050/JIPM010.2018.033. [In Persian]

Mumivand, H., Piri, R., S., & Kheiraei, F. (2021). A New Model for Automatic Text Classification. *Electrical Science and Engineering*, 3(1): 37–40. DOI: https://doi.org/10.30564/ese.v3i1.3170

Nguyen, D. N., Phan, T. T., & Do, P. (2021). Embedding knowledge on ontology into the corpus by topic to improve the performance of deep learning methods in sentiment analysis. *Scientific Reports*, 11(1): 23541. DOI: https://doi.org/10.1038/s41598-021-03011-6

Nguyen, N. T., Gabud, R. S., & Ananiadou, S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species

occurrence from biodiversity literature. *Biodiversity data journal*, (7). DOI: 10.3897/BDJ.7.e29626

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., and Tomović, M. (2017). Evaluation of classification models in machine learning. *Theory and Applications of Mathematics and Computer Science 7*(1): 39.URL: https://typeset.io/pdf/evaluation-of-classification-models-in-machine-learning-1u2pog86m5.pdf

Pan, X. (2015). *A context-based free text interpreter / A Context-Based Free Text Interpreter.* [PhD Thesis, California Polytechnic State University]. http://eil.stanford.edu/xpan/CFTI-Paper.pdf

Patterson, J., & Gibson, A. (2017). *Deep learning: A practitioner'. approach*. Sebastopol: O'Reilly Media.

Perez, Z. G., Zafar, M. A., Ziganshin, B. A., & Elefteriades, J. A. (2022). Toward standard abbreviations and acronyms for use in articles on aortic disease. *JTCVS open*, 10, 34-38. https://doi.org/10.1016/j.xjon.2022.04.010

Qian, L., Hao, P., Jianxin, L., Congying, X., Renyu, Y., Lichao, S., Philip, S. Y., & Lifang, H. (2021). A Survey on Text Classification: From Traditional to Deep Learning. *ACM Trans. Intell. Syst. Technol*, 37(4): 39. DOI: https://arxiv.org/pdf/2008.00364.pdf

Ramezani, H., Alipour-Hafezi, M., & Momeni, E. (2014). Scientific Maps: Methods and Techniques. *Popularization of Science*, 5(1): 53-84. [In Persian]

Ren, X., El-Kishky, A., Wang, C., & Han, J. (2015, August). Automatic Entity Recognition and Typing from Massive Text Corpora: A Phrase and Network Mining Approach. *In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2319-2320. https://doi.org/10.1145/2783258.2789988

Reyhani-Arani, E., & Lajardi, M. R. (2015, Aug). *Investigating methods of automatic classification of textual documents*. National Conference on Electricity and Computer, Distributed Systems and Smart Networks, 3.URL: https://www.sid.ir/fa/seminar/ViewPaper.aspx?ID=80521[In Persian]

Rozeva, A. (2012). Classification of text documents supervised by domain ontologies. *Applied Innovations and Technologies*, 8(3): 1-12. Doi:10.15208/ati.2012.11

Sajadi, M. B., & Minaei Bidgoli, B. (2020). The Architecture of Farsi Knowledge Graph System. *Iranian Journal of Information Processing and Management*, 35(2): 425-462. DOI: 10.35050/JIPM010.2020.057

Shan, G., Foulds, J., & Pan, S. (2020). Causal feature selection with dimension reduction for interpretable text classification. *arXiv preprint*, 2010.04609. https://doi.org/10.48550/arXiv.2010.04609

Shin, J., wu, S., Wang, feiran, De Sa, C., Zhang, C., & Re, C. (2015). Incremental knowledge base construction using deepdive. *Proceedins of the VLDB Endowment International Conference on Very Large Data Base, 8*, 1310. DOI:10.14778/2809974.2809991

Shirmardi, F., Hosseini, S. M. H., & Momtazi, S. (2021). FarsWikiKG: an Automatically Constructed Knowledge Graph for Persian. *International Journal of Web Research*, 4(2): 25-30. DOI: 10.22133/IJWR.2022.337760.1112

Sinoara, R. A., Camacho-Collados, J., Rossi, R. G., Navigli, R., & Rezende, S. O. (2019). Knowledge-enhanced document embeddings for text classification. *Knowledge-Based Systems*, 163, 955-971. https://doi.org/10.1016/j.knosys.2018.10.026

Soleimani Nezhad, A., Salajegheh, M., & Tayyebi Nia, E. (2019). Clustering scientific articles based on the k_means algorithm Case Study: Iranian Research Institute for information Science and Technology (IranDoc). *Iranian Journal of Information Processing and Management*, 34(2): 871-896. DOI: 10.35050/JIPM010.2019.060. [In Persian]

Song, M. H., Lim, S. Y., Kang, D. J., & Lee, S. J. (2005). Automatic classification of web pages based on the concept of domain ontology. *12th Asia-Pacific Software Engineering Conference*. DOI: 10.1109/APSEC.2005.46

Song, X., Bai, L., Liu, R., & Zhang, H. (2022). *Temporal Knowledge Graph Entity Alignment via Representation Learning*. In International Conference on Database Systems for Advanced Applications, 391-406, Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-031-00126-0_30

Sotoudeh, H. & Honarjoyan, Z. (2012). An overview of the difficulties of the Persian language in the digital environment and their effects on the effectiveness of automatic text processing and information retrieval. *Library and Information Sciences*, 15(4): 59-92. [In Persian]

Sun, K., Liu, Y., Guo, Z., & Wang, C. (2016). Visualization for knowledge graph based on education data. *International Journal of Software and Informatics*, 10(3): 1-13. DOI: 10.1145/2968220.2968227

Sun, M., Guo, Z., & Deng, X. (2021). Intelligent BERT-BiLSTM-CRF Based Legal Case Entity Recognition Method. *In Proceedings of the ACM Turing Award Celebration Conference-China*. 186-191. DOI: 10.1145/3472634.3474069

Suneera, C. M., & Prakash, J. (2020). *Performance Analysis of Machine Learning and Deep Learning Models for Text Classification*. In 2020 IEEE 17th India Council International Conference (INDICON), 1–6. DOI:https://doi.org/10.1109/INDICON49873.2020.9342208

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*, 50(1): 104–112. DOI: https://doi.org/10.1016/j.ipm.2013.08.006

Varga, A. (2014). Exploiting domain knowledge for cross-domain text classification in heterogeneous data sources. [Doctoral dissertation, University of Sheffield]

Wasi, S., Sachan, M., & Darbari, M. (2020). Document classification using wikidata properties. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD* 2018,729-737. Springer Singapore. DOI: 10.1007/978-981-13-7166-0_73

Wei, F., Qin, H., Ye, S., & Zhao, H. (2019). Empirical Study of Deep Learning for Text Classification in Legal Document Review. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 3317–3320. DOI: https://doi.org/10.1109/BigData.2018.8622157

Wijewickrema, C. M. (2015). Impact of an ontology for automatic text classification. *Annals of Library and Information Studies (ALIS)*, 61(4): 263-272. DOI: http://nopr.niscair.res.in/bitstream/123456789/30334/1/ALIS%2061(4)%20263-272.pdf

Xu, P., & Sarikaya, R. (2014, May). Contextual domain classification in spoken language understanding systems using recurrent neural network. *In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*,136-140. IEEE. DOI: 10.1109/ICASSP.2014.6853573

Yahya, M., Breslin, J. G., & Ali, M. I. (2021). Semantic web and knowledge graphs for industry 4.0. *Applied Sciences*, 11(11): 5110. DOI: https://doi.org/10.3390/app11115110

Yousif, S. A., Sultani, Z. N., & Samawi, V. W. (2019). Utilizing Arabic WordNet Relations in Arabic Text Classification: New Feature Selection Methods. *IAENG International Journal of Computer Science*, 46(4): 750-761.

Zhang, R., Trisedya, B. D., Li, M., Jiang, Y., & Qi, J. (2022). A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning. *The VLDB Journal*, 31(5): 1143-1168. DOI: https://doi.org/10.48550/arXiv.2103.15059

Zhang, W., & Xu, C. (2020). Microblog Text Classification System Based on TextCNN and LSA Model. *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, 469–474. DOI:https://doi.org/10.1109/ISCTT51595.2020.00090

Zhou, P., & El-Gohary, N. (2016). Ontology-based multilabel text classification of construction regulatory documents. *Journal of Computing in Civil Engineering*, 30(4): 04015058. DOI: https://doi.org/10.1061/(ASCE)CP.1943-5487.0000530