Presenting a Topic Classification Model of Health Scientific Productions Using Text-Mining Methods

Mahboobeh Shokouhian
PhD Candidate in Knowledge and Information Science; University of Isfahan Email: mahboobeshokohian@gmail.com

Asefeh Asemi*
PhD in Knowledge and Information Science; Associate Professor; University of Isfahan; and PhD in Business Informatics; Corvinus University of Budapest; Email: asemi@edu.ui.ac.ir

Ahmad Shabani
PhD in Knowledge and Information Science; Professor; University of Isfahan Email: shabania@edu.ui.ac.ir

Mozaffar Cheshmehsohrabi
PhD in Information and Communication Sciences; Associate Professor; University of Isfahan Email: mo.sohrabi@edu.ui.ac.ir

Iranian Journal of Information Processing and Management

Received: 31, Jul. 2019 **Accepted: 08, Jan. 2020**

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 35 | No. 2 | 553-574
Winter 2020

Abstract: With the proliferation of the Internet and the rapid growth of electronic articles, text classification has become one of the key and important tools for data organization and management. In text classification a set of basic knowledge is provided to the system by learning. Then, new input documents enter to one of the subject groups. In health literature due to wide variety of topics, preparing such a set of early education is a very time consuming and costly task. The purpose of this article is to present a hybrid model of learning (supervised and unsupervised) for the subject classification of health scientific products that performs the classification operation without the need for an initial labeled set. To extract the thematic model of health science texts from 2009 to 2019 at PubMed database, data mining and text mining were performed using machine learning. Based on Latent Dirichlet Allocation model, the data were analyzed and then the Support Vector Machine was used to classify the texts. In the findings of this study, the model was introduced in three main steps. In data preprocessing, the unnecessary words were eliminated from the data set and the accuracy of the proposed model increased. In the second step, the themes in the texts were extracted using the Latent Dirichlet Allocation method, and as a basic training set in step 3, the data were backed

* Corresponding Author

up by the Support Vector Machine algorithm and the classifier learning was performed with the help of these topics. Finally, with the help of the classification, the subject of each document was identified. The results showed that the proposed model can build a better classification by combining unsupervised clustering properties and prior knowledge of the samples. Clustering on labeled samples with a specific similarity criterion merges related texts with prior knowledge, and the learning algorithm teaches classification by supervisory method. Combining classification and clustering can increase the accuracy of classification of health texts.

Keywords: Health, Latent Dirichlet Allocation, Machine Learning, Scientific Production, Support Vector Machine Algorithm, Text Classification, Text Mining, Topic Model

ارائهٔ مدل دسته‌بندی موضوعی تولیدات علمی حوزهٔ سلامت با استفاده از روش‌های متن‌کاوی

محبوبه شکوهیان
دانشجوی دکتری علم اطلاعات و دانش‌شناسی؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه اصفهان؛
mahboobeshokohian@gmail.com

عاصفه عاصمی
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه اصفهان؛ دکتری؛ بیزینس اینفورماتیک، دانشگاه کروینوس بوداپست؛ پدیدآور رابط asemi@edu.ui.ac.ir

احمد شعبانی
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه اصفهان؛
shabania@edu.ui.ac.ir

مظفر چشمه‌سهرابی
دکتری علوم اطلاعات و ارتباطات؛ دانشیار؛ گروه علم اطلاعات و دانش‌شناسی؛ دانشگاه اصفهان؛
mo.sohrabi@edu.ui.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۸/۰۵/۰۹ **پذیرش:** ۱۳۹۸/۱۰/۱۸ مقاله برای اصلاح به مدت ۳۶ روز نزد پدیدآوران بوده است.

چکیده: با گسترش اینترنت و رشد سریع و روزافزون مقالات الکترونیکی، دسته‌بندی متون به یکی از ابزارهای کلیدی و مهم برای سازماندهی و مدیریت داده تبدیل شده است. در دسته‌بندی متون، یک مجموعهٔ دانش اولیه در اختیار سامانه قرار می‌گیرد تا با یادگیری از این مجموعه، اسناد جدید ورودی به یکی از گروه‌های موضوعی ملحق گردد. در متون سلامت به‌علت تنوع زیاد موضوعات، آماده‌کردن چنین مجموعهٔ آموزش اولیه عملی بسیار زمان‌بر و هزینه‌بر است. هدف از این مقاله ارائهٔ مدلی ترکیبی از یادگیری (با نظارت و بدون نظارت) برای دسته‌بندی موضوعی تولیدات علمی حوزهٔ سلامت است که بدون نیاز به مجموعهٔ برچسب خوردهٔ اولیه عمل دسته‌بندی را انجام دهد. برای استخراج مدل موضوعی، متون تولیدات علمی سلامت طی سال‌های ۲۰۰۹ تا ۲۰۱۹ در پایگاه «پابمد» با استفاده از روش آمیختهٔ داده‌کاوی، شامل متن‌کاوی و یادگیری ماشینی انجام

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۵ | شمارهٔ ۲ | ص‌ص ۵۵۳-۵۷۴
زمستان ۱۳۹۸

گرفت. بر اساس مدل موضوعی تخصیص پنهان «دیریکله»، داده‌ها تحلیل و سپس برای دسته‌بندی متون، از مدل ماشین بردار پشتیبان استفاده شد. در یافته‌های این پژوهش، مدل دسته‌بندی متون سلامت در سه گام اصلی معرفی شد. در گام اول، پیش‌پردازش‌های لازم بر روی مجموعهٔ داده برای حذف کلمات کم‌تکرار و غیرضروری از مجموعهٔ داده و افزایش دقت مدل پیشنهادی انجام گرفت. در گام دوم، موضوعات موجود در متون به کمک روش احتمالاتی تخصیص پنهان «دیریکله» استخراج شد. سپس، با استفاده از الگوریتم‌های ماشین بردار پشتیبان و با معیار قرار دادن بردارهای پشتیبان، دسته‌بندی موضوعی انجام شد. در نهایت، بر این اساس، موضوع هر سند مشخص گردید. نتایج نشان داد که مدل پیشنهادی می‌تواند با استفاده از ترکیب کردن خواص و بدون نظارت خوشه‌بندی و دانش پیشین نمونه‌ها یک دسته‌بندی بهتر بسازد. انجام دادن خوشه‌بندی روی نمونه‌های برچسب‌دار با یک معیار شباهت مشخص، متن‌های مرتبط را با هم ادغام و یک دانش پیشین ایجاد کرده، سپس الگوریتم یادگیری، دسته‌بندی را با روشی نظارتی آموزش می‌دهد. ترکیب دسته‌بندی و خوشه‌بندی می‌تواند دقت دسته‌بندی متون سلامت را افزایش دهد.

کلیدواژه‌ها: تولیدات علمی، دسته‌بندی متون، سلامت، متن‌کاوی، مدل تخصیص پنهان دیریکله، مدل موضوعی، ماشین بردار پشتیبان، یادگیری ماشینی

۱. مقدمه

امروزه با رشد چشمگیر اینترنت و گسترش روزافزون تولیدات علمی برای مدیریت اطلاعات سلامت، اهمیت دسته‌بندی متون سلامت افزایش یافته است. دسته‌بندی متون[1] را می‌توان به‌عنوان یکی از روش‌های متن‌کاوی[2] دانست که دسته‌بندی صفحات وب، و دسته‌بندی موضوعی مقالات از جمله کاربردهای آن است (Dang and Hamid Ahmad 2014). دسته‌بندی شامل فرایند اختصاص هر سند به دستهٔ مرتبط به آن است. اکثر روش‌های معتبر و دقیق در دسته‌بندی متون نیازمند مجموعهٔ آموزش اولیهٔ وسیع برچسب‌خورده هستند تا بر مبنای این آموزش اولیه اقدام به تشخیص دستهٔ متون جدید نمایند (Zhai & Massung 2016).

با توجه به این‌که رسالت سازمان‌های مراقبت بهداشت و سلامت حفظ و ارتقای سلامت مردم است، فراهم‌آوری سریع اطلاعات سلامت مورد نیاز افراد به بهبود کیفیت زندگی آنان کمک می‌کند. تشخیص موضوع و مرتب‌کردن فوری اطلاعات سلامت افراد در سلسله‌مراتبی موضوعی، جست‌وجوی ساخت‌یافته و پیدا کردن اسناد مورد نیاز فرد،

1. text classification
2. text mining

صرفه‌جویی در وقت و هزینه و افزایش رضایت و کاهش استرس افراد از جمله کاربردهای استفاده از موضوع برای دسته‌بندی متون سلامت هستند. تولید دانش موضوعی و کسب اطلاعات هدفمند با استفاده از روش‌های علمی برای پاسخگویی به نیازهای خاص، موجب ارتقای وضعیت سلامت جامعه می‌گردد و افراد می‌توانند وضعیت سلامت خود را به‌طور مؤثرتر مدیریت کنند.

مسئلهٔ اصلی پژوهش حاضر از آنجا ناشی می‌شود که با رشد گستردهٔ داده‌ها و اطلاعات سلامت، افراد با انبوهی از اطلاعات مواجه هستند و برای استفادهٔ بهینه از آن‌ها به مدیریت پیشرفتهٔ این اطلاعات نیاز دارند. نیاز به ابزارهای جدید مبتنی بر تکنولوژی وبی برای کشف و مرور مجموعه‌های بزرگ اطلاعات سلامت به روش خودکار توسط کاربران بسیار ضروری است. بدین سبب، دغدغهٔ اصلی محقق یافتن راهی برای حل مشکل دسته‌بندی موضوعی اطلاعات سلامت است. استفاده از تکنولوژی در مدل‌های موضوعی به‌عنوان یک روش جدید قدرتمند برای کشف ساختار یک مجموعه اطلاعات، بر کشف الگوهای استفاده از کلمه و اتصال اسناد مبتنی است. بدین‌گونه می‌توان به‌طور خودکار پایگاه‌های اطلاعاتی سلامت را سازمان‌دهی نمود تا مرور و کاوش اطلاعات در آن‌ها تسهیل شود. دسترسی افراد به اطلاعات سلامت بر اساس یک دسته‌بندی موضوعی مشکلی است که افراد با آن مواجه هستند و هنوز ابزارهای لازم برای ایجاد این امکان فراهم نشده است. دسته‌بندی آنلاین موضوعی به‌صورت پیوسته یک تکنولوژی بسیار عالی و کارآمد برای افراد است که به موجب آن مدیریت و کنترل بسیار مؤثرتری بر پیامدهای سلامت شخصی و ارتقای کیفیت سلامت امکان‌پذیر می‌شود. دسترسی به چنین اطلاعات موضوعی به این معناست که افراد و ارائه‌دهندگان خدمات بهداشتی تصمیمات را بر اساس اطلاعات در موقعیت بهتری اتخاذ می‌کنند. اگرچه در ایران کاربرد فناوری‌های نوین همواره مد نظر پژوهشگران حوزهٔ سلامت است، اما تاکنون مطالعه‌ای در دسته‌بندی موضوعی تولیدات این حوزه اجرا نشده است. در این مقاله سعی شده است که یک مدل ترکیبی از یادگیری (با نظارت و بدون نظارت) برای دسته‌بندی موضوعی تولیدات حوزهٔ سلامت بدون نیاز به مجموعهٔ برچسب‌خوردهٔ اولیهٔ وسیع طراحی شود، تا عمل دسته‌بندی اطلاعات به‌طور خودکار انجام گیرد. با توجه به این مدل موضوعی می‌توان پیشنهادهایی به محققان و طراحان سیستم ارائه داد.

۲. روش‌های متن‌کاوی و یادگیری ماشینی

متن‌کاوی را اولین بار در سال ۱۹۹۵ «فلدمن و داگان»[1] مطرح کردند. متن‌کاوی به فرایند استخراج اطلاعات مفید برای کشف دانش جدید از متن اشاره دارد و یک زمینهٔ چندرشته‌ای از بازیابی اطلاعات، پردازش زبان طبیعی و استخراج اطلاعات، آمار و یادگیری ماشین است. از جمله کاربردهای متن‌کاوی می‌توان به دسته‌بندی، خوشه‌بندی، خلاصه‌سازی و یافتن روابط میان مفاهیم در متون اشاره کرد. متن‌کاوی دارای دو روش یادگیری «با نظارت»[2] و «بدون نظارت»[3] است (Allahyari et al. 2017).

روش‌های «با نظارت» با این‌که در مواردی نتایج خوبی در دسته‌بندی موضوعی دارند، ولی نیازمند مجموعهٔ آموزش برچسب‌خوردهٔ وسیعی هستند و آماده‌کردن چنین مجموعهٔ داده‌ای برای حوزه‌های خاص مثل سلامت بسیار دشوار است. الگوریتم ماشین بُردار پشتیبان[4] یکی از محبوب‌ترین الگوریتم‌های دسته‌بندی است که در سال‌های اخیر کارایی نسبتاً خوبی داشته است. ماشین‌های بُردار پشتیبان اولین بار توسط «وپنیک»[5] در سال ۱۹۹۰ برای حل مسئلهٔ بازشناسی الگوهای دو کلاسی ارائه شد. مبنای کار این الگوریتم دسته‌بندی خطی داده‌هاست. در تقسیم خطی داده‌ها، سعی بر انتخاب خطی است که حاشیهٔ اطمینان بیشتری نسبت به سایر خطوط دارد (Klinkenberg and Joachims 2000). ماشین بُردار پشتیبان بر پایهٔ نظریهٔ یادگیری محاسباتی توسعه‌یافته و بر اصل «حداقل‌سازی خطای ساختاری» تکیه دارد. ایدهٔ حداقل‌سازی خطای ساختاری این است که مقدار h به‌دست بیاید، به‌طوری که بتوان برای آن کمترین «خطای مطلق»[6] را تضمین کرد. در دسته‌بندی ماشین بُردار پشتیبان، دسته‌بند از داده‌های جدایی‌پذیر به‌صورت خطی، قوانین تصمیم خطی را یاد می‌گیرد. این قوانین در رابطهٔ ۱ آمده است.

$$\vec{h}(\vec{x}) = sign\{\vec{w}.\vec{x} + b\} = \begin{cases} +1 & if\ (\vec{w}.\vec{x} + b > 0) \\ -1 & else \end{cases} \qquad \text{رابطهٔ ۱}$$

در رابطهٔ ۱، $\vec{w}$ یک بردار وزن و b یک حد آستانه است. نمونه‌های زیادی از کاربرد ماشین بُردار پشتیبان در دسته‌بندی متون وجود دارد. این روش یک فضای بُرداری تعریف می‌کند و ایدهٔ اصلی آن است که یک جداکنندهٔ مناسب، بیشترین فاصله را با نقاط همسایه

1. Feldman & Dagan
2. supervised
3. unsupervised
4. support vector machine (SVM)
5. Vipnic
6. true error

از هر دو طبقه دارد (Srivastava, Singh and Suri 2019).

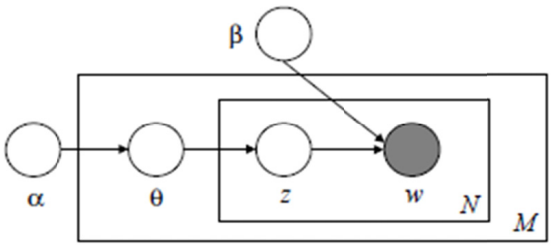
یادگیری «بدون نظارت» روش‌هایی برای پیدا کردن ساختار پنهان از داده‌های بدون برچسب هستند. خوشه‌بندی[1] و مدل موضوعی[2] دو الگوریتم یادگیری بدون نظارت مورد استفاده در متن‌کاوی هستند. در خوشه‌بندی، مجموعه‌هایی از اسناد که بیشترین شباهت به یکدیگر را دارند، در یک بخش با هم قرار می‌گیرند. در مدل موضوعی، اسناد در قالب ترکیبی از موضوعات مخفی تعریف می‌شوند و هر موضوع نیز در قالب توزیعی بر روی کلمات ایجاد می‌شود (Allahyari et al. 2017). اکثر پژوهش‌ها بر پایهٔ مدل‌های موضوعی شامل «آنالیز معنایی پنهان احتمالاتی»[3] (Hofman 1999) و «تخصیص دیریکله پنهان»[4] (Blei, Ng and Jordan 2003) هستند.

در مدل «دیریکله» پنهان برای استخراج موضوعات فرض می‌شود زمانی که یک کاربر قصد نوشتن متنی را دارد، ابتدا در ذهن خود موضوعاتی را که می‌خواهد در مورد آن‌ها بنویسد، انتخاب می‌کند. سپس، بر حسب هر موضوع و توزیع کلمات در آن موضوع، یک کلمهٔ مشخص را برای نوشتن انتخاب می‌کند. این عمل به‌صورت تکراری تا تکمیل‌شدن متن به‌صورت زیر انجام می‌شود: فرض کنید هر نظر از n کلمه به شکل $d = (w_1, w_2, \ldots, w_n)$ تشکیل شده است و مجموعهٔ داده از m سند مختلف به شکل $D = (d_1, d_2, \ldots, d_m)$ شکل گرفته است. روند انتخاب هر کلمه در هر نوشته به‌صورت زیر است:

۱. انتخاب توزیع موضوع: ابتدا نویسنده بر طبق توزیع $\theta \sim Dir(\alpha)$ یک توزیع از موضوعات موجود را انتخاب می‌کند (شبیه انتخاب توزیعی برای برداشتن یک تاس).
۲. اجرای دو گام زیر برای هر یک از کلمات موجود در هر سند
   - ◇ انتخاب یک موضوع برحسب $z_n \sim Multinomial(\theta)$
   - ◇ انتخاب یک کلمه از توزیع $p(w_n | z_n, \beta)$ (Blei, Ng and Jordan 2003)

روند معرفی‌شده با استفاده از دو توزیع «دیریکله» شکل می‌گیرد. با نمونه‌گیری از توزیع اول، موضوع مورد نظر کاربر مشخص می‌شود و توزیع دوم احتمال رخداد هر کلمه به شرط موضوع را تعیین می‌کند. در شکل ۱، نحوهٔ استفاده از این دو توزیع آمده است.

1. clustering 2. topic model 3. probabilistic latent semantic analysis (PLSA)
4. latent Dirichlet allocation (LDA)



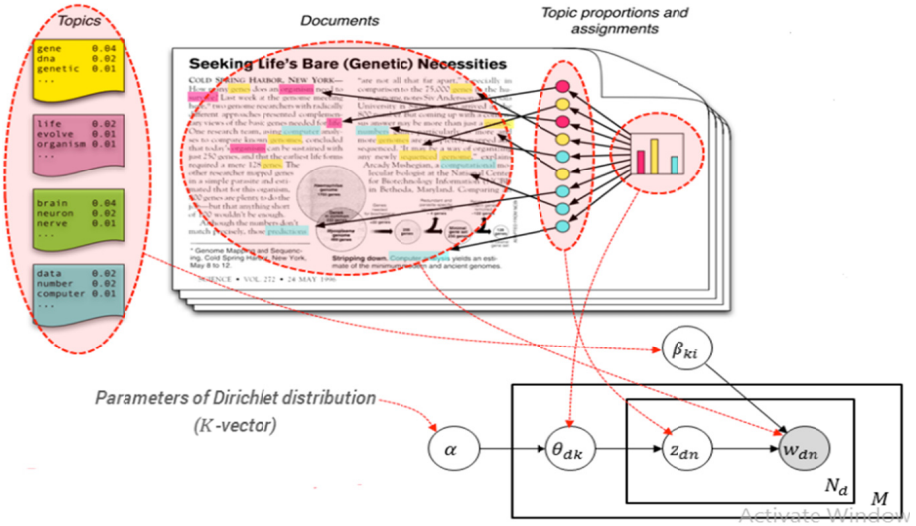
شکل ۱. شمای گرافیکی مدل LDA (Blei, Ng and Jordan 2003)

طبق شکل ۱، برای انتخاب هر کلمه ابتدا توزیع «دیریکله» با پارامتر α، یک توزیع موضوع انتخاب شده و سپس، طبق توزیع استخراجی و یک توزیع «دیریکله» دیگر با پارامتر β، کلمهٔ مورد نظر نمونه‌برداری می‌شود. برای محاسبهٔ احتمالات شرح داده‌شده، معمولاً از نمونه‌برداری «گیبس»[1] استفاده می‌گردد (Griffiths and Steyvers 2004). به این منظور، احتمال هر موضوع به‌شرط هر کلمه از دو احتمال مستقل به‌صورت رابطهٔ ۲، تعیین می‌شود. این دو احتمال به‌ترتیب انتخاب یک موضوع و سپس از روی آن، انتخاب یک کلمه است. مقدار دو عبارت $p(z|\alpha)$ و $p(w|z,\beta)$ بر طبق توزیع «دیریکله» محاسبه می‌شود (Frigyik, Kapila and Gupta 2010 و Blei, Ng and Jordan 2003). خروجی این گام مجموعه کلماتی است که هر مجموعه معرف یک موضوع در مجموعهٔ داده است.

$$p(z|w) \sim p(z|\alpha)p(w|z,\beta) \qquad \text{رابطهٔ ۲}$$

به‌طور کلی، می‌توان گفت که طبق شکل ۲، روش مدل‌سازی موضوعی LDA یک مدل احتمالی تولیدی بوده و قابل اجرا بر روی دادگان گسسته، همانند پیکره‌های متنی است؛ به‌نحوی که اسناد را به‌صورت مجموعه‌ای از کلمات در نظر می‌گیرد که ترتیب کلمات در آن مهم نیست. در این روش یک مدل بیزی بر روی پیکره‌ای از اسناد ایجاد می‌شود و فرض می‌شود که هر سند ترکیبی از موضوعات مختلف و هر موضوع توزیعی روی کلمات است که با توجه به کلمات تشکیل‌دهندهٔ هر سند می‌توان سهم هر موضوع را مشخص کرد. از این رو، در این روش، مقدار معلوم، توزیع کلمات در هر سند است و مقدار مجهول، توزیع موضوع‌ها در هر یک از اسناد و توزیع کلمات در هر یک از موضوعات است (Blei 2012).

1. Gibbs



شکل ۲. توزیع موضوع در متن (Blei 2012)

۳. پیشینهٔ پژوهش

در این بخش به تعدادی از پژوهش‌های ارائه‌شده برای دسته‌بندی متون اشاره می‌شود. مطالعات اولیه در حوزهٔ دسته‌بندی متون بیشتر دربارهٔ روش‌های یادگیری ماشین هستند (Sun, Lim and Liu 2009; Joachims 1998). «بلی، ان‌جی و جوردن» روش «تخصیص دیریکلهٔ پنهان» را برای مدل‌سازی متون پیشنهاد دادند و از آن در دسته‌بندی متون استفاده کردند (Blei, Ng and Jordan 2003). «گاندانگ، زانگ و زو» روش «آنالیز معنایی پنهان احتمالاتی» را برای دسته‌بندی صفحات وب به‌کار گرفتند (Guandong, Zhang & Zhou 2005). «لی» و همکاران یک روش دسته‌بندی ترکیبی با استفاده از دو دسته‌بند نزدیک‌ترین همسایه و ماشین بُردار پشتیبان ارائه کردند. آن‌ها در جهت کاهش زمان آموزش به‌ازای هر دسته، ابتدا از دسته‌بند ماشین بُردار پشتیبان استفاده کرده و سپس، دسته‌های مختلف به‌دست‌آمده به‌عنوان «داده‌های آموزشی نزدیک‌ترین همسایهٔ داده» به دسته‌بند ارائه شدند. برای محاسبهٔ فاصلهٔ میانگین ما بین دادهٔ آزمون و بُردار پشتیبان هر دسته، تابع فاصلهٔ اقلیدسی مورد استفاده قرارگرفته است. تصمیم نهایی بر مبنای دسته‌ای است که بُردار پشتیبانش کمترین فاصله را با دادهٔ آزمون دارد. نتایج استخراج‌شده نشان‌دهندهٔ کارایی مناسب این روش است (Lee et al. 2012).

«کرمی» و همکاران در پژوهش خود مدل موضوعی LDA را به‌صورت فازی[1] ارائه داده و دقت FLDA را در دسته‌بندی متون سنجیده‌اند تا بتوانند موضوعات اصلی را استخراج کنند (Karami et al. 2018). روش FLDA برای دسته‌بندی متون نیاز به مجموعهٔ برچسب‌خورده دارد. ولی، مدل پیشنهادی این پژوهش بدون وابستگی به هیچ مجموعهٔ برچسب‌خورده‌ای قادر است دسته‌بندی متون را انجام دهد. «وانگ» و همکاران یک روش دسته‌بندی متن بالینی با استفاده از نظارت کم ارائه کردند. آن‌ها ابتدا به‌طور خودکار برچسب‌ها را برای داده‌های آموزشی آماده و سپس، از کلمات پیش‌آموزش‌دیده به‌عنوان ویژگی‌های نمایشی برای آموزش مدل‌های یادگیری ماشین استفاده کردند (Wang et al. 2019).

«سریواستاوا، سینگ و سوری» بررسی خود را بر اساس دو مجموعه مراقبت‌های بهداشتی و غیربهداشتی ارائه کرده و با استفاده از «ماشین بُردار پشتیبانی»، شبکهٔ عصبی و درخت تصمیم‌گیری به دسته‌بندی متون پرداختند. آن‌ها از روش‌های انتخاب ویژگی کیسهٔ کلمات، TF-IDF و دسته‌بندی **معنایی** نهفتهٔ کلمات[2] استفاده کردند که عاملی در افزایش سازگاری و غنی‌سازی ویژگی‌ها بود (Srivastava, Singh & Suri 2019).

تعدادی از پژوهش‌ها، برای نمونه، در زمینهٔ دسته‌بندی متون برای زبان فارسی معرفی می‌گردد. «عرب سرخی و فیلی» یک روش دسته‌بندی با استفاده از بُردارهای فراوانی ریشهٔ کلمات و الگوریتم بیزین ساده ارائه داده و روش خود را با ترکیب روش بیزین با ایدهٔ نگهداری کلمات همنشین بهبود بخشیدند (۱۳۸۵). «امامی آزادی و الماس گنج» یک روش بدون نظارت برای دسته‌بندی متون فارسی با استفاده از «آنالیز معنایی پنهان احتمالاتی» و روش «کی-میانگین»[3] برای تعیین دستهٔ مربوط به متون جدید استفاده کردند (۱۳۸۵). «شمس و بارانی دستجردی» در پژوهش خود تلاش کردند که مدل موضوعی LDA را با اضافه‌کردن هم‌رخدادی[4] کلمات از موضوعات مشابه غنی کنند (Shams & Baraani Dastjerdi 2017).

با توجه به گستردگی حجم اطلاعات متون الکترونیکی سلامت، در صورت عدم نمایه‌گذاری و دسته‌بندی مناسب، کار بازیابی و پردازش اطلاعات سلامت با مشکل روبه‌رو می‌گردد. برخی از کارهای انجام‌گرفته در حوزهٔ دسته‌بندی متون در پی اِعمال

1. fuzzy
2. latent semantic indexing (LSI)
3. k-means
4. co-occurrence

روشی جدید یا تجربه‌نشده در این حوزه بوده‌اند و برخی از روش‌ها در پی ایجاد بهبود از منظرهای مختلف و به‌واسطهٔ پارامترها و دیدگاه‌های مختلف در روش‌های مزبور هستند. مطالعات نشان می‌دهد که روش‌های دسته‌بندی متون به‌تنهایی نمی‌توانند کیفیت دسته‌بندی را از یک حدی بیشتر افزایش دهند، اما با بهره‌گیری از روش‌های ترکیبی می‌توان کیفیت دسته‌بندی را ارتقا داد. در این پژوهش به ارائهٔ مدلی ترکیبی از یادگیری (با نظارت و بدون نظارت) برای دسته‌بندی موضوعی خودکار تولیدات حوزهٔ سلامت می‌پردازیم که بدون نیاز به مجموعهٔ برچسب‌خوردهٔ اولیه عمل دسته‌بندی را انجام دهد. این مزیت کمک زیادی در زمینه‌های گسترده و وسیع نظیر سلامت دارد.

٤. روش پژوهش

پژوهش حاضر از نوع پژوهش‌های کاربردی است و با استفاده از روش آمیختهٔ داده‌کاوی شامل متن‌کاوی و یادگیری ماشینی انجام گرفت. بر اساس مدل موضوعی تخصیص پنهان «دیریکله» داده‌ها تحلیل و سپس برای دسته‌بندی از مدل ماشین بُردار پشتیبان استفاده شد.

با استفاده از فرمول جست‌وجوی ترکیبی کلمات کلیدی «مش»[1] در پایگاه «پابمد»[2] در تاریخ ۱۲ ماه می ۲۰۱۹، تولیدات علمی حوزهٔ سلامت در فاصلهٔ سال‌های ۲۰۰۹-۲۰۱۹ به زبان انگلیسی شناسایی و مجموعهٔ داده به‌دست آمد (جدول ۱). روایی و پایایی ابزار «پابمد» به این دلیل که از معتبرترین و پرکاربردترین ابزارها در این حوزه است، نیاز به بررسی ندارد.

جدول ۱. فرمول جست‌وجو در پایگاه «پابمد»

فرمول جست‌وجو
(“electronic health record”[All Fields] OR “health care information system”[All Fields]) OR “personal health record”[All Fields] AND ((“2009/07/02”[PDAT]: “2019/06/29”[PDAT]) AND English [Lang])

پس از استخراج مجموعهٔ داده، هر سند از آن به یک فایل «ایکس‌ام‌ال»[3] تبدیل شد تا روال نمایه‌گذاری مجموعهٔ داده بهتر و ساده‌تر انجام پذیرد. در زیر یک نمونه از متون مجموعهٔ داده آمده است.

1. Mesh 2. Pubmed 3. XML

```
<book-part>
<title>Power (Social sciences)</title>
<p> There's a noble tradition among social scientists of trying to clarify how power works:
who gets what, when, where, and why. </p>
</book-part>
```

همان‌گونه که در شکل مشخص است، هر فایل از دو تگ مجزا تشکیل شده است: تگ <title> که عنوان سند در آن آمده و تگ <p> که بیانگر متن سند است. لازم به ذکر است که در مثال ذکرشده به‌علت اختصار، تنها بخشی از متن آورده شده است. اطلاعات آماری مجموعهٔ دادهٔ مورد استفاده در جدول ۲، آمده است.

جدول ۲. اطلاعات آماری مجموعهٔ داده

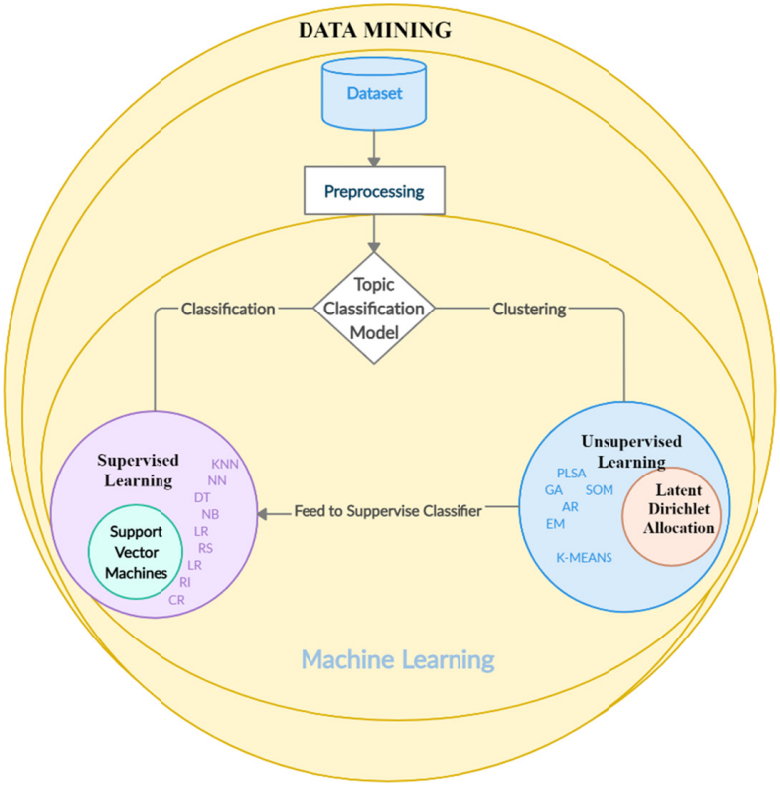
تعداد واژه‌ها پس از حذف کلمات توقف و کم تکرار و نشانه‌ها	تعداد کل واژگان یکتا	تعداد اسناد فاقد برچسب	تعداد اسناد دارای برچسب	تعداد کل اسناد
۹۷۳۲	۳۱۱۰۸	۳۹۷۶	۲۵۹۴	۶۸۷۰

برای مدل‌سازی موضوعی از نرم‌افزار «ایکلیپس»[1] و کتابخانهٔ «لوسین»[2]، کتابخانهٔ تخصیص پنهان «دیریکله» و کتابخانهٔ ماشین بُردار پشتیبان برای اجرای مدل و برای ترسیم جداول و نمودارها از نرم‌افزار «اکسل» استفاده شد.

۵. یافته‌ها

مدل موضوعی پیشنهادی همان‌گونه که در شکل ۳، مشخص است، در سه گام اصلی پیش‌پردازش، خوشه‌بندی و دسته‌بندی ارائه گردید. هر کدام از این گام‌های اصلی شامل چندین مرحله است. گام اول، پیش‌پردازش شامل دو مرحلهٔ نمایه‌گذاری کلمات و فیلترکردن واژگان، گام دوم، شامل سه مرحلهٔ خوشه‌بندی و استخراج موضوعات موجود در مجموعهٔ داده با استفاده از مدل موضوعی تخصیص پنهان «دیریکله»، وزن‌دهی ویژگی‌ها، برچسب‌گذاری موضوعات استخراجی و گام سوم، دسته‌بندی متون است. هدف این مدل موضوعی، خودکار کردن دسته‌بندی موضوعی اطلاعات است تا بدین‌وسیله پردازش و بازیابی داده‌های متنی سلامت غیرساخت‌یافته تسهیل گردد. در زیر، به شرح گام‌ها پرداخته شده است.

1. Eclipse Java 2019
2. Lucene



شکل ۳. مدل دسته‌بندی موضوعی متون

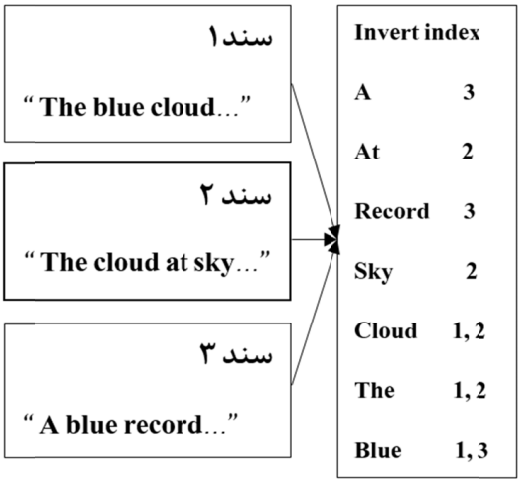
۵-۱. شرح مدل

۵-۱-۱. گام اول پیش‌پردازش

برای بررسی مجموعهٔ بزرگ متون اسناد بایستی که پیش‌پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره شوند. منظور از پیش‌پردازش انجام فعالیت‌هایی به‌منظور آماده‌نمودن مجموعهٔ دادهٔ اولیه برای به‌کارگیری در روش پیشنهادی است. این عملیات شامل نمایه‌گذاری و فیلترنمودن کلمات است که در زیر به معرفی هر یک پرداخته می‌شود.

۵-۱-۱-۱. نمایه‌گذاری

در مرحلهٔ اول پیش‌پردازش برای آماده‌سازی متون ابتدا نمایهٔ معکوس[1] و فشرده‌سازی انجام گرفت. این شیوه سبب می‌شود که سرعت اجرای الگوریتم و توانایی بررسی حالات مختلف بسیار بهبود یابد. نمایه‌گذاری معکوس با استفاده از کتابخانهٔ «لوسین» در برنامهٔ «ایکلیپس» بدین‌صورت انجام گرفت که ابتدا متن شکسته شد و کلمات داخل متن به تعداد ۳۱ هزار به‌دست آمد. در نمایهٔ معکوس مشخص می‌شود که هر کلمه در چه اسنادی و به چه تعداد تکرار شده است. هر کلمه در متن با لیست واژگان به‌دست‌آمده یک‌به‌یک مقایسه گردید و در صورت تطابق شمارندهٔ فراوانی کلمه، یکی اضافه می‌شود. در شکل ۴، عمل نمایه‌گذاری نشان داده شده است. مثلاً کلمهٔ ابر در سند شماره ۱، یک بار و در سند شماره ۲، یک بار تکرار شده است. تمامی اسناد شکسته‌شده و نمایه‌ای از تمام کلمات موجود در همهٔ اسناد به همراه لیست رخداد مکانی آن‌ها به‌وجود آمد. در ادامه، سایر گام‌های مدل پیشنهادی بر روی این نمایهٔ به‌دست‌آمده، انجام گرفت.



شکل ٤. نمایه‌گذاری معکوس

۵-۱-۱-۲. فیلترکردن کلمات

در گام پیشین، مجموعهٔ داده به یک نمایهٔ معکوس تبدیل شد. در این مرحله کلماتی

1. invert index

از مجموعهٔ داده که کاربرد چندانی در نتایج ندارند و تنها سبب کند شدن روال اجرای الگوریتم شده‌اند، کنار گذاشته می‌شوند. با این عمل پاک‌سازی مجموعهٔ داده انجام‌گرفته و ماتریس کلمات موجود کاهش می‌یابد. این کلمات شامل سه دستهٔ کلیِ کلمات توقف و ربط، نشانه‌ها و اعداد، و کلمات کم‌تکرار هستند.

با حذف سه دسته کلمات اشاره‌شده، باقی‌ماندهٔ مجموعهٔ نمایهٔ معکوس به تعداد حدود ده هزار کلمه به‌عنوان نمایندهٔ اسناد مجموعهٔ داده به گام دوم، یعنی خوشه‌بندی فرستاده می‌شوند. عملاً در پایان گام اول اسناد در قالب یک نمایهٔ پاک‌سازی‌شده که کلمات غیرکاربردی آن حذف شده است، برای گام بعدی آماده می‌گردد.

گام دوم و سوم این مدل تحت یادگیری ماشینی انجام پذیرفت. در گام دوم، تحت یادگیری بدون نظارت، عمل خوشه‌بندی توسط مدل تخصیص پنهان «دیریکله» انجام گرفت و مجموعهٔ موضوعات به‌دست آمد. در گام سوم که تحت یادگیری با نظارت توسط الگوریتم ماشین بُردار پشتیبان انجام گرفت، مجموعهٔ دادهٔ به‌دست‌آمده از گام دوم به‌عنوان برچسب اولیه در اختیار دسته‌بند قرار گرفت تا دسته‌بندی موضوعی به‌صورت خودکار و تحت یادگیری ماشینی انجام گیرد.

۵-۱-۲. گام دوم خوشه‌بندی

گام بعدی استفاده از LDA برای مدل‌سازی موضوعی است. بعد از انجام پیش‌پردازش داده‌ها، نوبت به کشف موضوعات پنهان با مدل‌سازی موضوعی می‌رسد. این مدل‌سازی موضوعی با استفاده از روش یادگیری ماشینی بدون نظارت انجام می‌گیرد که حاصل آن ایجاد خوشه‌بندی موضوعی است.

۵-۱-۲-۱. استخراج موضوعات موجود در مجموعهٔ داده با استفاده از مدل موضوعی تخصیص پنهان «دیریکله»

در این پژوهش با استفاده از مدل موضوعی تخصیص پنهان «دیریکله» به خوشه‌بندی داده‌ها پرداخته شد. این مدل موضوعی بر مبنای توزیع آماری «دیریکله» ایجاد شده که برای استخراج موضوعات فرض می‌گردد وقتی دو کلمه با هم در اسناد زیاد تکرار شده‌اند، احتمالاً با یکدیگر ارتباط معنایی دارند. این کلمات باهم، هم‌معنا نیستند ولی ارتباط موضوعی زیادی با یکدیگر دارند. مثلاً کلمات health و care با یکدیگر در اسناد زیاد تکرار شده‌اند. در نتیجه، با هم بار معنایی زیادی دارند. نتیجهٔ خروجی مجموعه

شامل کلمات مرتبط به‌همی است که هر کلمه به همراه احتمال تخصیص آن به موضوع، مشخص شده است.

در این مرحله، ابتدا یک فایل متنی با استفاده از فهرست کلمات نمایه‌گذاری‌شده به‌دست می‌آید که هر سطر آن شامل دنباله‌ای از کلمات هر متن است. مجموع سطور ماتریس کلمه/ متن، کلیهٔ اسنادی است که مجموعهٔ متن نامیده می‌شود. سپس، تعیین می‌گردد که چه تعداد موضوع از طریق مدل موضوعی «دیریکله» به‌دست آید. مثلاً ۱۰ موضوع به‌عنوان مثال، مشخص گردید. فهرست کلمات نمایه‌گذاری‌شده، مجموعهٔ متن و تعداد موضوع به‌عنوان ورودی به الگوریتم LDA در نرم‌افزار «ایکلیپس» داده شد. عملکرد الگوریتم LDA بدین‌صورت است که به تمامی کلمات یک احتمال یکسان در هر موضوع داده می‌شود. در اینجا مثلاً فرض می‌شود که هر موضوع با یک تاس شناخته شده و مجموع تاس‌ها در یک کیسهٔ کلمات قرار دارند. در ابتدا، کلمات در وجوه مختلف تاس به‌صورت متقارن هستند. با انجام عمل نمونه‌برداری، احتمال رخداد بعضی از کلمات بیشتر می‌شود. با تکرار زیاد آن، تاس در بعضی از وجوه خود شروع به تغییر و نامتقارن‌شدن می‌کند. در نتیجه، سند به سمت یک موضوع خاص کشیده می‌شود. سپس، بقیهٔ اسناد نیز که مشابه هستند، به سمت آن موضوع خاص گرایش می‌یابند. وقتی تاس نامتقارن می‌شود، یادگیری ماشینی صورت می‌گیرد؛ یعنی سیستم متوجه می‌شود که مثلاً کلمات «سلامت»، «پروندهٔ الکترونیک سلامت»، و «مراقبت بهداشتی» به هم مربوط‌اند، زیرا در یک‌سری از اسناد با هم زیاد تکرار شده‌اند. در نتیجه، این کلمات ارتباط معنایی بیشتری با هم دارند و با هم تشکیل یک موضوع را می‌دهند. در پایان این گام تعداد ۱۰ موضوع تولید شد که هر موضوع توزیعی روی کلمات به هم مربوط است.

احتمال رخداد هر کدام از کلمات متفاوت است، زیرا این تاس نامتقارن بر اساس احتمال بیشتر یا کمتر کلمات آن دسته موضوع است؛ اما به هر حال، شانس آمدن برای تمامی کلمات وجود دارد. مجموع احتمالات همهٔ کلمات در یک تاس موضوعی برابر یک است. مثلاً تاس موضوعی «مراقبت بهداشتی» شامل موضوعات با احتمال رخداد داخل پرانتز، سیستم (۰/۲)، امنیت (۰/۱)، مراقبت بهداشتی (۰/۵)، سلامت (۰/۲) باشد. احتمال آمدن کلمه‌ای که مرتبط‌تر است، مثلاً «مراقبت بهداشتی» بیشتر است. مدل تخصیص پنهان «دیریکله» احتمال رخداد هر موضوع به هر کلمه را مشخص می‌کند. مثلاً احتمال رخداد کلمهٔ «مراقبت بهداشتی» در موضوع اول ۰/۰۵، در موضوع دوم ۰/۱، در موضوع سوم ۰/۶۵،

در موضوع چهارم ۰/۱ و در موضوع پنجم ۰/۱ است که مجموع احتمالات آن‌ها یک است.

جدول ۳. موضوعات

موضوع ۱	موضوع ۲	موضوع ۳	موضوع ٤
library	health	information	national
academic	care	technology	security
students	medical	social	policy
libraries	system	digital	international
resources	healthcare	political	cloud
collection	patient	economic	military
access	patients	role	legal
university	safety	communications	space
faculty	quality	global	privacy
research	systems	public	identity

در جدول ۳، ۱۰ کلمهٔ اول تعدادی از موضوعات استخراج‌شده توسط مدل موضوعی LDA آمده است. هر ستون نماینده یک موضوع و کلمات هر ستون دارای ارتباط زیادی با یکدیگر هستند. برای مثال، کلمات موضوع ۱، در زمینهٔ کتابخانه و مباحث دانشگاهی است یا کلمات موضوع ۲، بیشتر در حوزهٔ سیستم‌های سلامت است و به همین‌ترتیب، در سایر موضوعات نیز این همبستگی موضوعی بین کلمات دیده می‌شود.

۵-۱-۲-۲. وزن‌دهی ویژگی‌ها

در این گام، از احتمالات استخراجی LDA به‌عنوان وزن اختصاص‌یافته به هر کلمه استفاده شد. این احتمال چون مربوط به هر موضوع است، می‌تواند به دقت بهتر الگوریتم کمک کند. روش LDA احتمال اختصاص هر کلمه به شرط هر موضوع را تعیین می‌کند و احتمال بیشتر به معنای مرتبط‌بودن کلمه به موضوع است. واضح است که کلمه‌های موجود در هر موضوع از نظر نزدیکی به آن موضوع یکسان نیستند و عملاً بعضی از کلمه‌ها مفهوم بیشتری در حیطهٔ آن موضوع را در خود دارند. مثلاً در کلمه‌های موضوع سوم، کلماتی نظیر health، care و patient بار موضوعی بیشتری نسبت به quality در این موضوع دارند.

۵-۱-۲-۳. برچسب‌گذاری موضوعات استخراجی

موضوعاتی که از مدل موضوعی تخصیص پنهان «دیریکله» حاصل می‌شوند، به‌صورت مجموعه‌ای از کلمات هستند و برچسب موضوع برای آن‌ها مشخص نیست. برای مثال، در جدول ۳، موضوع ۲ شامل کلماتی نظیر، care، system، healthcare، medical، systems، patient است، ولی برچسبی برای این دسته تعیین نشده و عملاً LDA توانایی تعیین برچسب برای دسته را ندارد؛ اما این مسئله در دسته‌بندی متون مشکل‌زاست، چون هدف نهایی در دسته‌بندی اسناد، تعیین دستهٔ آن‌هاست و در این عمل بهتر است یک برچسب به‌عنوان خروجی تعیین شود؛ مثلاً مشخص شود که سند در دستهٔ healthcare systems قرار دارد.

به این منظور در این گام برای هر یک از موضوعات استخراجی یک برچسب به‌عنوان معرف موضوع انتخاب گردید. برای مثال ذکرشده، healthcare یا health care systems برچسب مناسبی است. برای انتخاب برچسب از احتمال اختصاص‌یافته LDA به هر کلمه استفاده می‌گردد. همان‌گونه که بیان شد، در LDA احتمال هر کلمه به‌شرط هر موضوع تعیین می‌شود. با استفاده از این احتمالات، می‌توان n کلمه با بیشترین احتمال را به‌عنوان برچسب موضوع انتخاب نمود. برای مثال، اگر در جدول بالا n=2 در نظر گرفته شود، برچسب موضوع اول library academic خواهد بود و برچسب موضوعات بعدی نیز به‌ترتیب health care، information technology و national security خواهد بود.

۵-۱-۳. گام سوم مدل‌سازی دسته‌بندی متون

هر سند صرفاً یک موضوع ندارد، بلکه ترکیبی از چند موضوع است. برای تعیین موضوعات یک سند وقتی اطلاعاتی از آن در دست نیست، باید بر اساس یادگیری ماشینی، سیستم به‌طور خودکار، کار دسته‌بندی کلمات و تعیین موضوع را انجام دهد. در مجموعه‌های بزرگ اطلاعات که برچسب اولیه وجود ندارد، می‌توان با استفاده از روش خوشه‌بندی، این برچسب اولیه را فراهم و در اختیار دسته‌بند قرارداد.

در این پژوهش ماشین بُردار پشتیبانی به‌عنوان الگوریتم دسته‌بندی در مدل پیشنهادی استفاده شده است. از موضوعات استخراج‌شده توسط مدل موضوعی «دیریکله» به‌عنوان مجموعهٔ آموزش اولیه استفاده گردید. به این ترتیب که هر مجموعهٔ کلمه‌های استخراجی از LDA که معرف یک موضوع است، به همراه وزن‌های اختصاصی به هر کلمه به‌عنوان

یک آموزش اولیه به ماشین بُردار پشتیبانی داده شده و سپس، عمل آموزش و یادگیری با نظارت به‌وسیلهٔ این مجموعهٔ کلمه‌ها شکل می‌گیرد. در ادامه، هر سند بر مبنای شباهت به مجموعهٔ کلمه‌های موضوعی به یک موضوع نسبت داده شده و در انتها می‌توان از برچسب تعیین‌شده برای موضوع نیز به‌عنوان برچسب موضوعی سند استفاده کرد. سیستم توسط این مرحله از مدل یاد می‌گیرد چگونه به‌صورت خودکار و بدون داشتن مجموعهٔ اولیهٔ دانش، عمل دسته‌بندی را انجام دهد.

ماشین بُردار پشتیبان بر مبنای مجموعهٔ آموزش داده‌های به‌دست‌آمده از گام قبل، برای دسته‌بندی متون بر اساس تحلیلی که انجام می‌دهد، احتمال می‌دهد که این سند مثلاً به موضوع یک یا دو یا سه مربوط است و به وزن‌دهی کلمات نگاه می‌کند. هر کدام از دستهٔ کلمات که جمع وزنی بیشتری داشته باشد، سند برای تعیین موضوع به آن سمت گرایش می‌یابد؛ یعنی در سند ورودی اگر وزن مجموعهٔ کلمات موضوع اول (health، care و patient) بیشتر از موضوعات دیگر باشد، احتمالاً موضوع سند ورودی health care است.

٦. نتیجه‌گیری

در این پژوهش روشی ترکیبی (با نظارت و بدون نظارت) برای دسته‌بندی تولیدات علمی پایگاه «پابمد» در حوزهٔ سلامت معرفی گردید. مدل پیشنهادی شامل سه گام اصلی بود که در گام اول پیش‌پردازش‌های لازم بر روی مجموعه داده انجام شد. اجرای این پیش‌پردازش‌ها سبب می‌شود کلمات کم‌تکرار و غیرضروری از مجموعهٔ داده حذف شده، متون پاک‌سازی شده و دقت روش پیشنهادی افزایش یابد. در گام دوم موضوعات موجود در متن به کمک روش احتمالاتی LDA به‌صورت یادگیری بدون نظارت استخراج شدند. این موضوعات به‌عنوان یک مجموعهٔ آموزش اولیه در گام سوم به الگوریتم دسته‌بندی ماشین بُردار پشتیبان داده شد تا عمل یادگیری با نظارت دسته‌بند و به کمک این موضوعات انجام گیرد. در نهایت، به کمک دسته‌بند ماشین بُردار پشتیبان، موضوع هر سند مشخص گردید.

«کرمی» و همکاران در پژوهش خود همانند مدل پیشنهادی این پژوهش، از مدل موضوعی LDA استفاده کردند تا بتوانند موضوعات اصلی را استخراج کنند (Karami et al. 2018). این مدل برای دسته‌بندی متون نیاز به مجموعهٔ برچسب‌خورده دارد، ولی مدل

پیشنهادی این پژوهش بدون وابستگی به هیچ مجموعهٔ برچسب‌خورده‌ای قادر است دسته‌بندی متون را انجام دهد. مدل پیشنهادی این پژوهش همانند پژوهش Wang et al. (2019) و Srivastava, Singh & Suri (2019) با استفاده از ماشین بُردار پشتیبانی، به دسته‌بندی متون پرداختند، با این تفاوت که این پژوهش از مدل LDA برای تعیین موضوعات و داده‌های اولیه استفاده نموده است.

مهم‌ترین مزیت مدل ارائه‌شده برای دسته‌بندی متون آن است که این مدل به‌صورت یک چارچوب کلی ارائه گردیده است. در هر یک از مراحل آن می‌توان الگوریتم‌ها و شیوه‌های دیگری را جایگزین نمود؛ برای مثال، به‌جای روش دسته‌بندی ماشین بُردار پشتیبان در گام سوم از روش نزدیک‌ترین همسایه استفاده نمود یا به‌جای مدل تخصیص پنهان «دیریکله» از مدل آنالیز معنایی پنهان احتمالاتی بهره برد. این مدل قابلیت کارکرد در مجموعهٔ داده‌های مختلف به زبان‌های مختلف را دارد. هرچند مدل پیشنهادی در این پژوهش فقط بر روی یک مجموعه دادهٔ انگلیسی مورد بررسی قرار گرفت، ولی از آنجا که مدل مبتنی بر یک روش احتمالاتی است و وابستگی خاصی به نوع و زبان نوشته ندارد، بر روی مجموعهٔ داده‌های گوناگون به‌سادگی قابل اجراست. این مدل برخلاف بسیاری از روش‌های معتبر در حوزهٔ دسته‌بندی که نیاز به مجموعهٔ برچسب‌خوردهٔ اولیهٔ وسیعی دارند، نیازی به مجموعهٔ دانش اولیهٔ برچسب‌خورده ندارد و قادر به دسته‌بندی اسناد است. این مزیت می‌تواند در زمینه‌های گسترده و وسیع نظیر پزشکی و سلامت مفید واقع شود.

مدل پیشنهادی به دسته‌بندی متون سلامت با گستردگی زیاد کمک زیادی می‌کند، زیرا در ابتدا مجموعهٔ دادهٔ برچسب‌خوردهٔ اولیه برای آموزش دسته‌بندها وجود ندارد و همین مسئله سبب دشواری و کاهش دقت دسته‌بندی در متون سلامت می‌شود. مدل پیشنهادی با بهره‌گیری از دانش موجود در موضوعات استخراجی از LDA این مشکل را تا حد زیادی برطرف می‌سازد. این مدل می‌تواند اطلاعات بدون ساختار را با روشی سریع و مقرون‌به‌صرفه ساختار دهد. این امر موجب می‌شود که در زمان و هزینهٔ تجزیه و تحلیل داده‌های متنی صرفه‌جویی شده و اطلاعات دسترس‌پذیرتر گردد. با انجام دسته‌بندی متون خطاهای انسانی کاهش‌یافته و در شرایط بحرانی، بازیابی اطلاعات خاص ممکن می‌شود. استفاده از این مدل می‌تواند به‌عنوان نقشهٔ راه برای متخصصان و طراحان پایگاه داده باشد.

به‌عنوان یکی از مطالعات پیشِ روی این مقاله می‌توان از روش‌ها و الگوریتم‌های مختلف در هر یک از گام‌های مدل پیشنهادی بهره برد و مناسب‌ترین نتیجه را به‌دست آورد. هرچند مدل پیشنهادی در این مقاله بر روی یک مجموعهٔ دادهٔ انگلیسی انجام گرفت، پیشنهاد می‌گردد که بر روی مجموعهٔ متون فارسی نیز انجام گیرد.

قدردانی

این مطالعه حاصل پایان‌نامهٔ مقطع دکتری رشتهٔ علم اطلاعات و دانش‌شناسی در دانشگاه اصفهان بود. دست‌اندرکاران این پژوهش از دانشگاه اصفهان به‌دلیل حمایت‌های مالی و معنوی از این پایان‌نامه تقدیر و تشکر می‌نمایند.

فهرست منابع

امامی آزادی، طاهره، و فرشاد الماس گنج. ۱۳۸۵. دسته‌بندی موضوعی متون فارسی بر اساس روش آنالیز معنایی پنهان احتمالاتی بهبودیافته. دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران، تهران، دانشگاه شهید بهشتی. https://www.civilica.com/Paper-ACCSI12-ACCSI12_283.html (دستیابی در ۹۸/۶/۱۸)

عرب‌سرخی، محسن، و هشام فیلی. ۱۳۸۵. ارائهٔ یک سیستم دسته‌بندی موضوعی متون فارسی بر اساس روش‌های احتمالاتی. در *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*. تهران: دانشگاه صنعتی شریف. ۱۵۱-۱۶۱.

References

Allahyari, M., P. Pouriyeh, M. Assefi, A. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. e-prints. arXiv: 1707.02919

Beil, F., M. Ester & X. Xu. 2002. Frequent term-based text clustering. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. https://doi.org/10.1145 / 775047.775110 (accessed July 2, 2019).

Blei, D. 2012. Probabilistic topic models. *Communications of the acm* 55 (4), 77-84. https://doi.org/10.1145/2133806.2133826 (accessed 10 July. 2019).

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3: 993-1022. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf (accessed July 5, 2019).

Dang, Sh., P. Hamid Ahmad. 2014. Text Mining: Techniques and its application. *International Journal of Engineering & Technology Innovations* 1 (4): 22-25.

Frigyik, B. A., A. Kapila, & M. R. Gupta. 2010. Introduction to the Dirichlet Distribution and Related Processes. Technical Report, Department of Electrical Engineering, University of Washignton. https://pdfs.semanticscholar.org/775e/5727f5df0cb9bf834af2ea2548a696c27a38.pdf?_ga=2.58651690.1018243960.1564408728-575202613.1556829791 (accessed July 10, 2019).

Griffiths, T. L., & M. Steyvers. 2004. Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101: 5228–5235. https://doi.org/ 10.1073/pnas.0307752101 (accessed July 10, 2019).

Guandong, X., Y. Zhang, & Z. Zhou. 2005. Using Probabilistic Latent Semantic Analysis for Web Page Grouping. Proceedings of Research Issues in Data Engineering: Stream Data Mining and Applications, 29-36. http://staff.itee.uq.edu.au/zxf/_papers/RIDE05.pdf (accessed July 8, 2019).

Han, J., M. Kamber, & J. Pei. 2011. Data Mining: Concepts and Techniques. Elsevier. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf (accessed July 10, 2019).

Hofmann, T. 1999. Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. 50–57. https://doi.org/10.1145/312624.312649 (accessed 15 July. 2019).

Joachims, T. 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features in Machine Learning. *10th European Conference on Machine Learning.* 137-142. https://doi.org/10.1007/BFb0026683 (accessed July 12, 2019).

Karami, A., A. Gangopadhyay, B. Zhou, & H. Kharrazi. 2018. Fuzzy Approach Topic Discovery in Health and Medical Corpora. *International Journal of Fuzzy Systems* 20 (4): 1334–1345. https://doi.org/10.1007/s40815-017-0327-9 (accessed July 14, 2019).

Klinkenberg, R. & T. Joachims. 2000. Detecting Concept Drift with Support Vector Machines. Proceedings of the Seventeenth International Conference on Machine Learning (ICML). San Francisco, CA., USA.

Lee, L. H., C. H. Wan, R. Rajkumar, D. Isa. 2012. An enhanced Support Vector Machine classification framework by using Euclidean distance function for text document categorization. Appl Intell, 37, 80-99. https://doi.org/10.1007/s10489-011-0314-z (accessed July 14, 2019).

Lewis D. D. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. *In Proceedings of ECML-98, 10th European Conference on Machine Learning*. 4–15. https://doi.org/10.1007/BFb0026666 (accessed July 11, 2019).

Michael, M., H. Erik, & G. Otis. 2010. Lucene in Action, Second Edition. USA: Manning Publication.

Shams, M., & A. Baraani Dastjerdi. 2017. Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction. *Expert Systems with Applications.* 80. https://doi.org/10.1016/j.eswa.2017.02.038. (accessed July 10 2019).

Srivastava, S. K., S. K. Singh, & J. S. Suri. 2019. Effect of incremental feature enrichment on healthcare text classification system: A machine learning paradigm. *Computer Methods and Programs in Biomedicine* 172: 35–51. https://doi.org/10.1016/j.cmpb.2019.01.011 (accessed July 22, 2019).

Sun, A., E.-P. Lim, & Y. Liu. 2009. On strategies for imbalanced text classification using SVM: A comparative study. *Decision Support Systems* 48 (1): 191-201.https://ink.library.smu.edu.sg/sis_research/757 (accessed July 1, 2019).

Vapnik, V. N. 1990. *Statistical Learning Theory*. New York: J. Wiley.

Wang, Y., S. Sohn, S. Liu, F. Shen, L. Wang, E. J. Atkinson, & H. Liu. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC medical informatics and decision making* 19 (1): 1. https://doi.org/10.1186/s12911-018-0723-6 (accessed July 10, 2019).

Yao, L., C. Mao, & Y. Luo. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC medical informatics and decision making* 19 (71). https://doi.org/10.1186/s12911-019-0781-4 (accessed July 20, 2019).

Zhai, C., & S. Massung. 2016. Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining. https://doi.org/10.1145/2915031 (accessed July10, 2019).

محبوبه شکوهیان

متولد ۱۳۶۳، دانشجوی دکتری علم اطلاعات و دانش‌شناسی دانشگاه اصفهان و کتابدار نهاد کتابخانه‌های عمومی استان اصفهان است.

سازماندهی اطلاعات و سیستم‌های اطلاعاتی از جمله علایق پژوهشی وی است.

عاصفه عاصمی

متولد ۱۳۴۹، دارای مدرک دکتری بیزینس اینفورماتیک از دانشگاه کوروینوس بوداپست است. ایشان هم‌اکنون محقق دانشگاه کوروینوس بوداپست و استاد بازنشستهٔ دانشگاه اصفهان است.

تولید هوشمند، محصول هوشمند، یادگیری عمیق و سیستم‌های توصیه‌گر از جمله علایق پژوهشی وی است.

احمد شعبانی

متولد ۱۳۳۵، دارای مدرک دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون استاد گروه علم اطلاعات و دانش‌شناسی دانشگاه اصفهان است.

مدیریت دانش و روش تحقیق از جمله علایق پژوهشی وی است.

مظفر چشمه‌سهرابی

متولد سال ۱۳۵۳، دارای مدرک تحصیلی دکتری در رشتهٔ علوم اطلاعات و ارتباطات از دانشگاه استاندال فرانسه است. ایشان هم‌اکنون دانشیار گروه علم اطلاعات و دانش‌شناسی دانشگاه اصفهان است.

وب معنایی، ذخیره و بازیابی اطلاعات، داده‌کاوی، سنجش و ارزیابی علم و پژوهش، علم‌شناسی و اخلاق علمی از جمله علایق پژوهشی وی است.