Application of the Neural Network-based Machine Learning Method to Classify Scientific Articles

Masood Ghayoomi*
PhD in Computational Linguistics; Assistant Professor;
Institute for Humanities and Cultural Studies; Tehran, Iran;
Email: M.Ghayoomi@ihcs.ac.ir

Maryam Mousavian
M.Sc. in Artificial Intelligence; Department of Computer Engineering; Amirkabir University of Technology; Tehran, Iran;
Email: maryam.mousavian@aut.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 37 | No. 4 | pp. 1217-1244
Summer 2022
https://doi.org/10.35050/JIPM010.2022.008

Received: 03, Jul. 2021 **Accepted: 05, Dec. 2021**

Abstract: Since 2000s (1380s according to the Iran's solar calendar), the increasing rate of writing and publishing scientific articles in Iran has become very intense. In addition to the governmental organizations, such as Irandoc & the National Library and Archives of the Islamic Republic of Iran, this caused numerous other online systems, such as the General Portal of Humanities, Noormags, Magiran, Elmnet, Civilica, etc, to manage knowledge and to provide structured archives of the scientific documents. Each of these archives provides facilities to the user. One of these facilities is searching on the documents. An accurate search can greatly improve the usage of these online systems. To increase the accuracy of the search result, it is necessary to determine the scientific field of articles. Classifying large volumes of scientific resources in different fields is very time-consuming. Using machinery methods can be a solution to reduce the severity of the task.

The main contribution of this paper is to provide a classification model to classify Persian scientific articles. Although in previous studies, the classification task has been mainly used for simple texts, in this study, the neural network-based classification models, such as convolutional and perceptron neural networks, are used with the contextualized semantic representation, such as ParsBERT; and the results are compared with the other common method utilized for vectorization, namely Word2Vec. To this end, we use the data from the General Portal of Humanities, which includes various articles in the Humanities and each article contains the label of the field. One of the neural network characteristics is that a set of hidden features from the data in the vector space is created and used to train the model. According to the experimental results, the Perceptron classifier that utilized ParsBERT representation obtained the highest

* Corresponding Author

performance which is 74.71% based on the Micro F-score, and 72.55% based on the Macro F-score.

Keywords: Scientific Publications, Humanities, Classification, Neural Network, Vector Space, ParsBERT

کاربرد یادگیری ماشینی مبتنی بر شبکهٔ عصبی برای دسته‌بندی مستندات علمی

مسعود قیومی
دکتری زبان‌شناسی رایانشی؛ استادیار؛ پژوهشکده زبان‌شناسی؛ پژوهشگاه علوم انسانی و مطالعات فرهنگی؛ تهران، ایران؛
پدیدآور رابط M.Ghayoomi@ihcs.ac.ir

مریم موسویان
کارشناسی ارشد هوش مصنوعی؛ دانشکدهٔ مهندسی کامپیوتر؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران؛
maryam.mousavian@aut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۰۴/۱۲ **پذیرش:** ۱۴۰۰/۰۹/۱۴ مقاله برای اصلاح به مدت ۱۶ روز نزد پدیدآوران بوده است.

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 ۳۷ | شمارهٔ ٤ | ص‌ص ۱۲۱۷-۱۲٤٤
تابستان ۱٤۰۱
https://doi.org/10.35050/JIPM010.2022.008

چکیده: از دههٔ ۱۳۸۰ شمسی، نگارش و انتشار مقالات علمی در ایران سرعت بسیار زیادی یافته و سبب شده افزون بر سازمان‌های دولتی مانند «ایرانداک» و «سازمان اسناد و کتابخانهٔ ملی جمهوری اسلامی ایران»، سامانه‌های برخط متعدد دیگری چون «پرتال جامع علوم انسانی»، «نورمگز»، «مگ‌ایران»، «علم‌نت»، «سیویلیکا» و غیره اقدام به مدیریت دانش و تهیهٔ بایگانی‌های ساختارمند مستندات علمی کنند. هر کدام از این بایگانی‌ها امکاناتی را در اختیار کاربر قرار می‌دهد. یکی از این امکانات، قابلیت جست‌وجوست و جست‌وجوی دقیق می‌تواند بر کاربریِ این سامانه‌ها تأثیر به‌سزایی بگذارد. برای افزایش دقت جست‌وجو نیاز است حوزهٔ علمی مقالات مشخص شود. دسته‌بندی حجم زیاد منابع علمی در حوزه‌های مختلف بسیار زمان‌بر است و استفاده از روش‌های ماشینی به‌عنوان یک راه‌حل می‌تواند از این کار طاقت‌فرسا بکاهد.

هدف اصلی این مقاله ارائهٔ یک مدل دسته‌بندی برای تعیین حوزهٔ مقالات علمی است. اگرچه در پژوهش‌های پیشینِ دسته‌بندی، به‌طور عمده، الگوریتم‌های دسته‌بندیِ متداول برای متن ساده به‌کار رفته‌اند، در این پژوهش تلاش می‌شود افزون بر استفاده از این دسته‌بندی‌ها، از دسته‌بندهای مبتنی بر شبکهٔ عصبی، مانند شبکهٔ عصبی «پیچشی» و «پرسپترون»، به همراه بازنمایی معنایی مبتنی بر بافت، مانند «پارس‌برت» استفاده شود و نتایج آن با سایر روش‌های متداول در ساخت بردار مستندات، مانند «ورد۲وک» مقایسه شود. برای این هدف، از داده‌های «پرتال علوم انسانی» که دربرگیرندهٔ مقالات متنوع علوم انسانی است، استفاده می‌کنیم.

ویژگی این داده مشخص بودن حوزهٔ تخصصی هر مقاله است. یکی از ویژگی‌های شبکهٔ عصبی این است که برایندی از ویژگی‌های نهفته از داده در فضای برداریِ ساخته‌شده شکل می‌گیرد و برای آموزش مدل استفاده می‌شود. بر اساس نتایج عملی، دسته‌بند «پرسپترون» مبتنی بر «پارس‌برت» بالاترین کارایی ۷۴/۷۱ درصدی بر اساس امتیاز F میکرو و کارایی ۷۲/۵۵ درصدی بر اساس امتیاز F ماکرو را به‌دست آورده است.

کلیدواژه‌ها: مستندات علمی، علوم انسانی، دسته‌بندی، شبکه عصبی، فضای برداری، پارس‌برت، معناشناسی توزیعی

۱. مقدمه

از دههٔ ۱۳۸۰ شمسی، نگارش و انتشار مقالات علمی در ایران در مقایسه با دهه‌های گذشته سرعت بسیار بیشتری یافته و سبب شده سرانهٔ انتشار مقالات علمی توسط پژوهشگران اعم از دانشجو یا عضو هیئت علمی در مراکز آموزش عالی رشد چشمگیری داشته باشد. از جمله وظایف سازمان‌های دولتی مانند «ایرانداک» و «سازمان اسناد و کتابخانهٔ ملی جمهوری اسلامی ایران»، مدیریت دانش و بایگانی این دسته از محصولات علمی است. شایان ذکر است که در این میان، سامانه‌های برخط متعدد دیگری همچون «پرتال جامع علوم انسانی»، «نورمگز»، «مگ‌ایران»، «علم‌نت»، «سیویلیکا» و مانند آن با رویکرد پژوهشی، تجاری یا ترکیبی از این دو، اقدام به مدیریت دانش و تهیهٔ بایگانی‌های ساختارمند داده‌های علمی کرده‌اند. هر یک از این سامانه‌ها امکاناتی را در اختیار کاربر قرار می‌دهد. برای مثال، در سامانهٔ «علم‌نت» با جست‌وجوی یک مقاله، ضمن نمایش اطلاعات شناسنامه‌ای سند علمی اعم از مقاله یا پایان‌نامه، تعداد محدودی از مستندات علمی دیگر که از نظر محتوایی با مقالهٔ جست‌وجوشده مرتبط است، به کاربر نمایش داده می‌شود. مثال دیگر، «پرتال جامع علوم انسانی» است که مقالات علوم انسانی در آن مدیریت و بایگانی شده است. در این پرتال، مقولات مستندات علمی مانند زبان‌شناسی، روان‌شناسی و غیره، که بیانگر نوع حوزهٔ علوم انسانی است، مشخص شده است. طبقه‌بندی مستندات علمی در فرایند جست‌وجو کمک شایانی به کاربر می‌کند و مشخص‌بودن حوزهٔ علمی مقالات به افزایش دقت جست‌وجو در مستندات بسیار کمک می‌کند.

وجود حجم زیادی از مستندات علمی که سالانه توسط پژوهشگران منتشر می‌شود، سبب می‌شود فرایند طبقه‌بندی و تعیین حوزهٔ مقالات علمی به‌صورت دستی بسیار زمان‌بر و طاقت‌فرسا باشد. از این رو، استفاده از روش‌های ماشینی می‌تواند ضمن کاهش

میزان دخالت نیروی انسانی، به افزایش سرعت در مدیریت و ورودی داده به سامانه‌های بایگانی مستندات علمی و به‌روزرسانی سریع آن‌ها کمک نماید. هدف از انجام این پژوهش، ارائهٔ مدل یادگیری ماشینی مبتنی بر شبکهٔ عصبی در کنار روش‌های متداول یادگیری ماشینی برای دسته‌بندی و تعیین مقولهٔ مستندات علمی است تا فرایند مدیریت دانش و تهیهٔ بایگانی‌ها را تسهیل نماید. همچنین، در این پژوهش به تحلیل مواردی که به اشتباه دسته‌بندی شده‌اند، پرداخته می‌شود.

ساختار مقالهٔ حاضر به‌شرح زیر است: در بخش ۲، به‌کارگیری رویکردهای رایانشی نوین در مدیریت دانش معرفی می‌شود. در بخش ۳، مطالعات انجام‌شده در حوزهٔ دسته‌بندی و شیوه‌های بازنمایی اطلاعات توضیح داده می‌شود. در بخش ۴، مدل دسته‌بندی معرفی و توضیح داده می‌شود. ویژگی‌های مجموعهٔ دادهٔ استفاده‌شده در این پژوهش در بخش ۵ معرفی شده و فرایند انجام آزمایش‌ها بیان می‌گردد. و سرانجام، مقاله با نتیجه‌گیری در بخش ۶ به پایان می‌رسد.

۲. به‌کارگیری رویکردهای رایانشی نوین در مدیریت دانش

مدیریت دانش یک حوزهٔ علمی است که قدمتی بیش‌از ۳۰ سال دارد (Nonake 1991). اگرچه مدیریت دانش در دانشگاه‌ها شروع شد، این موضوع بخش جدایی‌ناپذیر زندگی امروزی است. منظور از «مدیریت دانش»، فرایند تولید، اشتراک‌گذاری، کاربرد و مدیریت دانش و اطلاعات مربوط به یک سازمان است تا بتوان با بهره‌گیری از این دانش بهتر به اهداف سازمانی رسید (Girard & Girard 2015). افزون بر این موارد، مجموعهٔ فعالیت‌های به‌دست آوردن، تعیین، سازماندهی، ذخیره، بازنمایی، انتقال و کاربرد مجدد دانش نیز جزء ویژگی‌های مدیریت دانش برشمرده شده است (Aharony 2011). هدف اصلی مدیریت دانش به‌اشتراک‌گذاری تجارب، اطلاعات و دانش است تا بتوان از آن‌ها در زمان و مکان مناسب برای تصمیم‌گیری و بازیابی اطلاعات استفاده نمود.

وجود حجم زیاد داده در محیط اطراف سبب شده که کار مدیریت دانش از حالت سنتی که استفاده از روش‌های دستی بود، تغییر کند و با کمک رایانه انجام پذیرد. در فرایند مدیریت دانش باید ضمن ساختارمندسازی داده‌های بدون ساختار، آن‌ها را قابل پردازش نمود تا امکان به‌دست آوردن دانش میسر شود. در اصل، فرایند مدیریت دانش با علم داده همراستاست؛ چرا که هدف نهایی علم داده نیز استخراج دانش از داده و اطلاعات

است. بنابراین، این دانش یک ارزش افزوده‌ای است که می‌تواند در تصمیم‌گیری‌ها جنبهٔ کاربردی به خود بگیرد. مفاهیم داده، اطلاعات و دانش به‌واسطهٔ کاربرد رایانه تخصصی شده و رابطهٔ آن‌ها در شکل ۱، در قالب هرم دانش نمایش داده شده است (Landauer 1998; Boisot & Canals 2004; Rowley 2007; Sharma 2008).

شکل ۱. هرم دانش

دانش که به تبیین علم می‌پردازد، در اصل به داده و اطلاعات وابسته است و امروزه حجم زیاد داده به‌صورت الکترونیکی در دسترس است. این ویژگی سبب شده است که روش‌های مدیریت دانش دستخوش تغییر باشد. همان‌طور که اشاره شد، سازماندهی دانش یکی از فعالیت‌های مدیریت دانش است که سبب می‌شود با کمک رایانه و الگوریتم‌های پردازشی انجام پذیرد. در این مقاله تلاش می‌شود با ارائهٔ مدلی برای دسته‌بندی مقالات علمی در حوزهٔ علوم انسانی، کار دسته‌بندی مستندات علمی به‌صورت ماشینی انجام شده و در جهت نیل به اهداف مدیریت دانش قدم برداشته شود.

۳. پیشینهٔ مطالعاتی

پژوهش‌های انجام‌شده در حوزهٔ یادگیری ماشینی، افزون بر روش تحلیل ماشینی، داده‌محور بوده و به داده نیاز دارد. بر این اساس، پژوهش‌های انجام‌شدهٔ مرتبط با موضوع این پژوهش را به دو دسته تقسیم می‌کنیم.

۳-۱. تهیهٔ پیکرهٔ زبانی از مستندات علمی

در زبان انگلیسی، «بِرد» و همکارانش پیکره‌ای به‌نام «پیکرهٔ مرجع هستان‌شناسی انجمن

زبان‌شناسی رایانشی»[1] تهیه کرده‌اند (Bird et al. 2008). این پیکره حاوی ۱۰۹۲۱ مقاله در حوزهٔ زبان‌شناسی رایانشی است. پردازش‌های انجام‌شده بر روی این پیکره برچسب‌گذاری مقولات دستوری واژه‌ها و بن‌واژه‌سازی است. «دگاتانواورتلیب» و همکارانش از مستندات علمی یک پیکرهٔ زبانی به‌نام «سایتکس»[2] تهیه کرده‌اند (Degaetano-Ortlieb et al. 2013). این پیکره شامل ۳۴ میلیون واژه و حاوی ۹ رشتهٔ علمی است. «کواری» پیکره‌ای با ۸۹۵ مقاله از پایگاه «الزویر» در چهار حوزهٔ علوم پزشکی، علوم زیست‌شناسی، علوم فیزیک و علوم اجتماعی تهیه کرده و واژه‌های کلیدی را با توجه به بافت استخراج نموده است (Kwary 2018).

پژوهش «کامیابی‌گل» و همکاران از جمله معدود پژوهش‌های متمرکز بر مستندات علمی فارسی است که در آن افزون بر گردآوری مستندات علمی برای تشکیل یک پیکرهٔ زبانی، به نشانه‌گذاری زبان‌شناختی مستندات علمی نیز پرداخته شده است (۱۳۹۷). آن‌ها مجموعه‌ای متشکل از ۱۱۰۰ مقاله با حجم حدود ۷ میلیون واژه را جمع‌آوری کرده و با کمک ابزارهای پردازش زبان فارسی، کار بهنجارسازی[3]، واحدسازی[4]، بن‌واژه‌سازی[5] واژه‌های به‌کاررفته در این متون، برچسب‌گذاری مقولهٔ دستوری واژه‌ها و تجزیهٔ نحوی وابستگی متون را انجام داده‌اند. همچنین، به‌تازگی، پیکره‌ای حاصل از مستندات علمی در سازمان «ایرانداک» به‌نام «پیکرهٔ پژوهشنامه» (علایی ابوذر و همکاران ۱۴۰۰) ته‌یه شده و قابل دسترس است.[6] این پیکره تقریباً حاوی بیش‌از ۵/۴ میلیون واژهٔ تخصصی و میان‌رشته‌ای در رشته‌های علمی مختلف است. از جمله ویژگی‌های این پیکره، تخصیص اطلاعات زبان‌شناسی، مانند برچسب مقولهٔ دستوری به واژه‌هاست.

۳-۲. تحلیل محتوایی مستندات علمی

تحلیل محتوایی مستندات علمی در دو حوزهٔ مستندات علمی و خبری قابل مقایسه است؛ چرا که اولاً هر دو داده متنی است و از نظر الگوریتمی نیز به یکدیگر بسیار شبیه است. ثانیاً پژوهش‌های انجام‌شده در حوزهٔ پردازش مستندات علمی در فارسی از تنوع چندانی برخوردار نیست.

1. The ACL Anthology Reference Corpus
2. SciTex
3. normalization
4. tokenization
5. lemmatization
6. https://sapa.irandoc.ac.ir/

۳-۲-۱. دسته‌بندی مستندات علمی

در زبان انگلیسی، «کیم و جیل» از بسامد واژه-معکوس بسامد سند»[1] (1975) Salton که به اختصار TF-IDF می‌نامیم و همچنین، تخصیص «دریشلهٔ پنهان»[2] در مدل‌سازی موضوع (Blei et al. 2003) برای خوشه‌بندی مستندات علمی بر اساس تشابه موضوعی مقالات در یک خوشه استفاده کرده‌اند (Kim and Gil 2019).

«چاودری و شوئن» با استفاده از چندین دسته‌بند یادگیری ماشینی متداول، مانند ماشین بردار پشتیبان[3]، «بیز ساده»[4]، درخت تصمیم[5] و کا نزدیکترین همسایه[6]، برای دسته‌بندی مقالات علمی استفاده کرده‌اند (Chowdhury & Schoen 2020). بر اساس نتایج این پژوهش، درخت تصمیم بالاترین کارایی را به‌دست آورده است.

«ریوست، ویگنولا-گان و آرکامبالت» و همکاران در پژوهش خود به دسته‌بندی ۴۰ میلیون مقالهٔ علمی با استفاده از شبکه‌های عصبی پرداخته‌اند. از جمله ویژگی‌هایی که آن‌ها در مدل خود استفاده کرده‌اند، بهره‌گیری از اطلاعات منابع، ارجاعات مستقیم و دسته‌بندی دستی است. بر اساس نتایج این پژوهش، استفاده از ویژگی‌های ساده، مانند ارجاع مستقیم، بیشترین تأثیر را بر الگوریتم‌های متداول دسته‌بند گذاشته و استفاده از اطلاعات کتابنامه بیشترین تأثیر را در شبکهٔ عصبی گذاشته است (Rivest, Vignola-Gagné & Archambault 2021).

۳-۲-۲. دسته‌بندی متون خبری

تا جایی که می‌دانیم مطالعات انجام‌شده در حوزهٔ دسته‌بندی متون فارسی بیشتر بر روی متن خبری متمرکز بوده و مطالعات محدودی در حوزهٔ تحلیل و دسته‌بندی مستندات علمی با کمک پردازش زبان طبیعی و هوش مصنوعی انجام پذیرفته است که در ادامه، بررسی می‌گردد. «امامی آزادی و الماس‌گنج» با استفاده از روش «تحلیل معنایی پنهان احتمالاتی»[7] (Hofmann 1999) کار دسته‌بندی موضوعی، شش موضوع از مقالات فارسی موجود در پیکرهٔ متنی «فارس‌دات»[8] (Bijankhan, Sheikhzadegan & Roohani (1994 را انجام داده‌اند. آنان برای بهبود مدل از مشخصات نویسندگان استفاده کرده‌اند (۱۳۸۵).

«تیمورپور، سپهری و پزشک» با استفاده از یک مدل بی‌نظارت مبتنی بر بازیابی

1. term frequency-inverse document frequency (TF-IDF)
2. Latent Dirichlet allocation
3. support vector machine (SVM)
4. naive Bayes
5. decision tree
6. k nearest neighbor
7. probabilistic latent semantic analysis
8. FARSDAT

اطلاعـات کار خوشـه‌بندی و سـپس، دسـته‌بندی مقـالات نمایه‌شـدهٔ ISI در حـوزهٔ نانـو را انجـام داده‌انـد. در ایـن مطالعـه از پیکـره‌ای متشـکل از ۱۹۹۰ مقالـه طـی سـال‌های ۱۳۸۲ تـا ۱۳۸۸ اسـتفاده شـده اسـت (۱۳۸۸).

«ربیعـی، حسـینی مطلـق و مینایـی بیدگلـی» بـا اسـتفاده از ماشـین بـردار پشـتیبان بـه رده‌بنـدی پژوهش‌هـای حـوزهٔ محیـط زیسـت پرداختـه و سـپس، پارامترهـای تأثیرگـذار در کیفیـت ایـن رده‌بنـدی را ارزیابـی کرده‌انـد. ایـن پژوهـش در زمینـهٔ اسـتخراج کلیدواژه‌هـا و نمایه‌سـازی متـون کاربـرد دارد. در ایـن پژوهـش، آنـان روش جدیـدی بـرای وزن‌دهـی در هنـگام سـاخت بردارهـا بـا نـام NG-TF معرفـی کرده‌انـد کـه حاصـل تضریـب بسـامد واژه و تعـداد دفعـات تکـرار توالـی اجـزای یـک عبـارت اسـت. داده‌هـای اسـتفاده شـده در ایـن پژوهـش، ۱۶۶۲۶ سـند مرتبـط بـا حـوزهٔ محیـط زیسـت اسـت کـه از میـان پایان‌نامه‌هـای کارشناسـی ارشـد و رسـاله‌های دکتـری بایگانی‌شـده در پایـگاه اطلاعـات علمـی« ایرانـداک» به‌دسـت آمـده اسـت (۱۳۹۸).

«شـکوهیان» و همـکاران از دو مـدل یادگیـری ماشـینی بانظـارت و بی‌نظـارت بـرای دسـته‌بندی موضوعـی مسـتندات علمـی حـوزهٔ سـلامت اسـتفاده کرده‌انـد. بـرای انجـام ایـن پژوهـش، آنـان بـا روش جسـت‌وجوی کلیـدواژه، پیکـره‌ای از متـون علمـی بـه زبـان انگلیسـی در حـوزهٔ سـلامت در بـازهٔ زمانـی سـال‌های ۲۰۰۹ تـا ۲۰۱۹ در پایـگاه «پابمـد»[1] تهیـه کرده‌انـد. تعـداد مسـتندات در ایـن پیکـره ۶۸۷۰ متـن بـوده اسـت کـه ابتـدا بـا اسـتفاده از تخصیـص «دریشـلهٔ پنهـان» کار دسـته‌بندی و تحلیـل محتوایـی مسـتندات انجـام شـده و پـس از برچسـب‌گذاری دادهٔ خوشه‌بندی‌شـده، بـردار حاصـل از تخصیـص «دریشـلهٔ پنهـان»، بـرای آمـوزش دسـته‌بند ماشـین بـردار پشـتیبان اسـتفاده شـده اسـت (۱۳۹۸).

«کرمـی، گانگوپادهیـای و خـرازی» یـک مـدل فـازی بـرای اسـتخراج موضوعـات مسـتندات علمـی در حـوزهٔ بهداشـت و درمـان معرفـی، و مـدل خـود را بـا مـدل تخصیـص «دریشـلهٔ پنهـان» مقایسـه کرده‌انـد. داده‌ای کـه آنـان بـرای پژوهـش خـود اسـتفاده کرده‌انـد، داده «پابمـد» بـوده اسـت (Karami, Gangopadhyay & Kharrazi 2018).

از آنجـا کـه الگوریتم‌هـای داده‌هـای متنـی مسـتندات علمـی و خبـری شـبیه بـه یکدیگـر اسـت، در ادامـه، خلاصـه‌ای از پژوهش‌هـای انجام‌شـده در زمینـهٔ دسـته‌بندی متـون فارسـی خبـری بررسـی می‌گـردد. «باقـری» و همـکاران بـا اسـتفاده از دسـته‌بند «بیـز» سـاده بـه دسـته‌بندی متـون خبـری فارسـی پرداخته‌انـد (۱۳۸۷).

1. Pubmed

«یعقوبی» به بررسی و تهیهٔ سامانهٔ خودکاری پرداخته است که بتواند اخبار را در بستر وب به‌طور خودکار طبقه‌بندی کند. در این سامانه از روش‌های وب‌کاوی و روش‌های دسته‌بندی متون استفاده شده است. روش به‌کاررفته در این سامانه استخراج واژه‌های کلیدی از متن و یافتن گروه خبری آن متن بر اساس آن کلیدواژه‌هاست. ویژگی پژوهش «یعقوبی» این است که کارهای پردازشی و تحلیلی به‌صورت برخط و در زمان بسیار محدود انجام می‌گیرد (۱۳۹۱).

«منفرد» با کمک هوش مصنوعی کار دسته‌بندی اخبار وب را به‌طور خودکار انجام داده است و سپس، به افزایش دقت و کارایی دسته‌بند پرداخته است (۱۳۹۳). به نظر می‌رسد روش مطرح‌شده مکمل مدل ارائه‌شده در «یعقوبی» (۱۳۹۱) است.

در پژوهش «رباطی» به این نکته پرداخته شده است که به هنگام دسته‌بندی اخبار از چه ویژگی‌های مؤثری در فرایند دسته‌بندی می‌توان استفاده کرد تا میزان کارایی الگوریتم دسته‌بند افزایش یابد. برای این هدف از معیار ارزش‌گذاری ویژگی بنام «ای-دومیننس»[1] استفاده شده است. این معیار بر روش وزن‌دهی TF-IDF و هم‌رخداد[2] TF-IDF استوار است (۱۳۹۳).

«نوریان و زاده‌طبری میثم» از شبکه‌های عصبی با الگوریتم پس‌انتشار و شبکه‌های باور عمیق مبتنی بر یادگیری عمیق برای دسته‌بندی اخبار پیکرهٔ روزنامهٔ «همشهری» (AleAhmad et al. 2009) استفاده کرده‌اند. در این روش از بردار وزنی TF-IDF استفاده شده است (۱۳۹۴).

«هاشمی و حورعلی» بر روی اخبار حوزهٔ دفاعی متمرکز شده و افزون بر استفاده از واژه‌ها، از یک هستان‌شناسی به‌عنوان پایگاه دانش خارجی به‌منظور تعمیم ارتباط بین واژه‌ها استفاده کرده‌اند (۱۳۹۶).

«جمالی، میرعابدینی و هارون‌آبادی» تلاش کرده‌اند از ترکیب روش‌های دسته‌بندی، به دسته‌بندی متون فارسی بپردازند. در این پژوهش بردار وزنی TF-IDF با استفاده از مربع خی[3] و بسامد سند انتخاب می‌شود (۱۳۹۹).

«ممتازی و قیومی» در پژوهش خود از مدل‌سازی موضوع برای برچسب‌گذاری پیکرهٔ روزنامهٔ «همشهری» با حداقل دخالت انسان برای دسته‌بندی و برچسب‌زنی مقولهٔ متون

1. E-Dominance 2. co-occurrence 3. Chi-square test

خبری استفاده کرده‌اند. آن‌ها برای رسیدن به این هدف، از تخصیص «دیریشلهٔ پنهان» استفاده کرده‌اند (Momtazi & Ghayoomi 2014).

«ایمای، ناکامورا و اومامودا» تلاش کرده‌اند که تصویر تجسمی از تحلیل اطلاعات روزنامه‌ای به‌دست آورند. برای این تصویرسازی نیاز به دسته‌بندی متون است. راهکاری که آن‌ها برای دسته‌بندی متون و اتصال آن‌ها به یکدیگر استفاده کرده‌اند، کاربرد روش وزن‌دهی TF-IDF و چندواژه‌ای[1] بوده است (Imai, Nakamura & Ohmameuda 2015).

«احمدی، تابنده و غلامپور» تلاش کرده‌اند که با استفاده از مدل‌سازی موضوع به دسته‌بندی متون فارسی بپردازند. در این پژوهش آنان مدل خود را با مدل کیسه‌واژه[2] مقایسه کرده و به‌صورت عملی کارایی بالاتری در حدود ۹ درصد با به‌کارگیری مدل‌سازی موضوع در دسته‌بندی به‌دست آورده‌اند (Ahmadi, Tabandeh & Gholampour 2016).

«دادگر، عراقی و فراهانی» روشی را با استفاده از روش وزن‌دهی TF-IDF و ماشین بردار پشتیبان پیشنهاد کرده‌اند که برای دسته‌بندی اخبار در وبلاگ‌ها و شبکه‌های اجتماعی به‌کار می‌رود (Dadgar, Araghi & Farahani 2016).

«جهان‌تیغ، دانش‌پور و اوروجلو» تلاش کرده‌اند که با استفاده از الگوریتم کا نزدیک‌ترین همسایه، دسته‌بندی «بیز» ساده و ترکیب این دو دسته‌بند، تعداد ۵۳۳۰ متن خبری روزنامهٔ «همشهری» را به‌طور صحیح دسته‌بندی نمایند. آن‌ها برای استخراج ویژگی، از بسامد واژه و روش وزن‌دهی TF-IDF استفاده کرده‌اند. در این پژوهش برای کاهش تأثیر ایست‌واژه[3]، از آنتروپی[4] واژه استفاده شده است (Jahantigh, Daneshpour & Orojlou 2016).

«یاکوبی، ون‌آته‌ولت و ولبرس» به چالش حجم زیاد دادهٔ خبری و کاربرد عملی آن به‌عنوان ابزار کمکی در تحلیل برای خبرنگاران پرداخته‌اند. برای این هدف، اخبار مربوط به فناوری انرژی هسته‌ای از سال ۱۹۴۵ تا ۲۰۱۵ در روزنامهٔ «نیویورک‌تایمز» را جمع‌آوری کرده و با استفاده از روش تخصیص «دیریشلهٔ پنهان» تحلیل کرده‌اند (Jacobi, van Atteveldt & Welbers 2016).

1. *n*-gram
2. bag of word
3. stop word
4. entropy

٤. معرفی مدل دسته‌بندی مستندات علمی

هدف از انجام این پژوهش، ارائهٔ مدل‌های مبتنی بر یادگیری ماشینی از جمله دسته‌بندهای یادگیری ماشینی متداول و شبکه‌های عصبی برای تعیین حوزهٔ مقالات علمی علوم انسانی و مقایسهٔ نتایج آن‌ها با یکدیگر است. برای رسیدن به این هدف، نیاز به بازنمایی متون مقالات علمی در قالب بردار است که از دو مدل «ورد۲وک»[1] (Mikolov et al. 2013) و «پارس‌برت»[2] (Farahani et al. 2020) استفاده می‌شود. بردارهای «ورد۲وک» برای دسته‌بندهایی مانند «بیز» ساده، رگرسیون لجستیک[3]، درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، کا نزدیک‌ترین همسایه و دو شبکهٔ عصبی «پرسپترون»[4] و «پیچشی»[5] مورد استفاده قرار می‌گیرد. سپس، برای بهره‌گیری از بازنمایی‌های مبتنی بر بافت و دانش موجود در مدل‌های زبانی از پیش آموزش‌دیده، از مدل «پارس‌برت» به‌عنوان ورودی شبکهٔ عصبی «پرسپترون» و «پیچشی» استفاده می‌شود.

بازنمایی معنایی متن یکی از مسائل اساسی در پردازش زبان طبیعی است که در آن محتوای زبانی به بردارهایی که قابل استفاده در الگوریتم‌های پردازشی باشد، تبدیل می‌شود. این بازنمایی می‌تواند در سطح واژه، جمله و یا متن باشد. هدف از ایجاد بازنمایی معنایی به‌صورت یک مدل فضای برداری متون این است که اسناد متنی حتی بدون ساختار به‌صورت عددی نشان داده شود و با رویکرد ریاضی‌گونه قابل محاسبه باشد. از این رو، تبدیل متن به مدل فضای برداری پایه و اساس فرایند پردازش زبان طبیعی را شکل می‌دهد و این موضوع نشان‌دهندهٔ اهمیت زیاد آن است.

یکی از ابتدایی‌ترین روش‌های بازنمایی معنایی، استفاده از یک ماتریس $n \times m$ است که n تعداد مستندات در ردیف‌ها و m تعداد واژه‌های موجود در واژگان است که از تمام داده‌ها استخراج شده و تعداد ابعاد بردار را مشخص می‌نماید. خروجی این ماتریس، تعریف یک بردار به ازای هر سند و به ابعاد تعداد واژه‌های موجود در واژگان است. در این شیوه بردارسازی، حضور هر یک از واژه‌ها در یک سند با عدد یک و بقیه ابعاد با عدد صفر نشان داده می‌شود. یکی‌از نقاط ضعف این روش وجود بردارهای با ابعاد بزرگ و تُنُک به‌منظور بازنمایی واژه‌هاست. همچنین، در این روش، شباهت معنایی میان واژه‌ها در نظر گرفته نمی‌شود و در این فضای برداری هیچ ارتباطی میان واژه‌ها وجود ندارد.

---

1. Word2Vec
2. ParsBERT
3. logistic regression
4. perceptron
5. convolutional

«هریس» معتقد است که معنای واژه از بافتی که در آن به‌کار رفته است، به‌دست می‌آید (Harris 1954). این نظر پایه‌گذار معناشناسی توزیعی[1] بوده است. برای نمایش اطلاعات بافتی در چارچوب نظریهٔ معناشناسی توزیعی، «سونگ» و همکارانش دو روش کلی را معرفی کرده‌اند. یک روش مبتنی بر روش «بِیز» است که بر اساس رویکردهای مربوط به «مدل‌سازی موضوع»[2] Blei et al. (2003) استوار است و روش دیگر مبتنی بر ویژگی است که بر اساس بازنمایی اطلاعات بافتی به‌صورت بردار پایه‌ریزی شده است (Song et al. 2016).

رویکردهای مدل‌سازی بافت در بازنمایی معنایی می‌تواند مبتنی بر تجزیهٔ ماتریس حاصل از هم‌رخدادی واژه‌ها و یا روش‌های مبتنی بر شبکهٔ عصبی باشد. هدف رویکرد تجزیهٔ ماتریس، تبدیل یک ماتریس تُنُک به یک ماتریس چگال[3] با ابعاد پایین‌تر است (Pennington, Socher & Manning 2014). یکی‌از ویژگی‌های مدل‌های زبانی عصبی این است که افزون بر آموزش یک مدل زبانی بر مبنای شبکهٔ عصبی، ساختاری را برای نگاشت[4] واژه‌ها به فضای برداری فراهم می‌نماید که به تعبیه‌سازی واژه[5] معروف است. مدل‌های پیشرفتهٔ تعبیه‌سازی واژه بر مبنای بافت مطرح شده است و به مدل‌های از پیش آموزش‌دیده، از جمله مدل‌های خانواده «برت»[6]، منتهی شده است. این پژوهش، از دو شیوهٔ تعبیه‌سازی «ورد۲وک» (Mikolov et al. 2013) و «پارس‌برت» ParsBERT (Farahani et al. 2020) برای استخراج ویژگی‌های غنی مبتنی بر بافت جایگاهی واژه‌ها استفاده می‌کند که در ادامه توضیح داده می‌شود.

٤-١-١. «ورد۲وک»

در بردارسازی با «ورد۲وک» که توسط Mikolov et al. (2013) معرفی شده، ارتباط معنایی واژه‌ها در یک مدل فضای برداری بر اساس همجواری واژه‌ها در متون بازنمایی می‌گردد. در این شیوه مدل‌سازی، واژه‌ها با بردارهایی با ابعاد کمتر بازنمایی می‌شود. این روش بردارسازی کاربردهای فراوانی در مدل‌های شبکهٔ عصبی در پردازش زبان طبیعی دارد.

هنگام تهیهٔ یک مدل زبانی می‌توان با آموزش یک شبکهٔ عصبی یک فضای برداری متراکم ساخت. هر یک از ابعاد فضای برداری بیانگر یک ویژگی است که توسط شبکه

1. distributional semantics
2. topic modeling
3. dense
4. mapping
5. word embedding
6. bidirectional encoder representations from transformers (BERT)

از حجم زیادی از داده استخراج می‌شود. این شبکه به دو روش مختلف کیسه‌واژهٔ پیوسته[1] و پرش‌نگاشت پیوسته[2] آموزش داده می‌شود. مزیت این شیوه مدل‌سازی این است که به دادهٔ برچسب‌خورده نیاز ندارد (Rong 2014).

٤-١-٢. «پارس‌برت»

افزایش مدل‌های زبانی از پیش آموزش‌دیده با هدف ساخت مدل‌های غنی زبانی سبب شده است که فصل نوینی در زمینهٔ پردازش زبان طبیعی آغاز شود. در میان این مدل‌ها، مدل‌های مبتنی بر انتقال‌دهنده[3]، مانند «برت» که توسط Devlin et al. (2019)، معرفی شده، به‌دلیل عملکرد عالی محبوبیت بیشتری پیدا کرده است. با این حال، این مدل‌ها به‌طور معمول، بر روی زبان انگلیسی متمرکز بوده و زبان‌های دیگری که منابع محدودی دارند، از مدل‌های چندزبانه بهره می‌برند. «پارس‌برت» یک مدل «برت» تک‌زبانه برای زبان فارسی است که عملکرد بهتری در مقایسه با معماری‌ها و مدل‌های چندزبانه، مانند مدل «برت» چندزبانه، از خود نشان داده است. با وجود مجموعه داده‌های بسیار محدود در زبان فارسی در حوزهٔ پردازش زبان طبیعی، یک پیکرهٔ زبانی پرحجم با بیش‌از ۳ میلیارد واژه که از منابع مختلف جمع‌آوری شده، برای آموزش این مدل استفاده شده است (Farahani et al. 2020).

مدل «برت» یک مدل زبانی یادگیری عمیق است. «برت» و مدل‌های مشابه آن، مانند «المو»[4] (Peters et al. 2018) و «جی‌پی‌تی»[5] (Radford et al. 2018)، مدل‌های تعبیه‌سازیِ مبتنی بر بافتِ از پیش آموزش‌دیده است. در این مدل‌ها، تعبیهٔ واژه‌ها با توجه به جمله‌ای که در آن قرار گرفته، مشخص می‌شود. بنابراین، بازنمایی‌های متفاوتی از یک واژه با توجه به بافت جایگاهی واژه در جمله‌های متفاوت به‌دست خواهد آمد. از این رو، ویژگی ابهام واژگانی چالشی برای این‌گونه بردارسازی نخواهد بود؛ چرا که هر واژه با توجه به بافت شامل بردارهای مختلف است؛ در حالی که در شیوهٔ بردارسازی با «ورد۲وک» هر واژه فقط یک بردار دارد و ابهام واژگانی در بردار واژه همچنان مستتر است. مدل‌های تعبیه‌ساز مبتنی بر بافت برای وظیفه‌هایی مانند مدل‌سازی زبان بر روی مجموعه داده‌های متنی کلان، آموزش داده می‌شوند. غنی بودن مدل «برت» به این است که از دوازده لایهٔ

1. continuous bag of word
2. continuous skip-gram
3. transformer
4. Embeddings from Language Models (ELMO)
5. Generative Pre-trained Transformer (GPT)

کدگذار انتقال‌دهنده ساخته شده است. این مدل با دو تابع مختلف آموزش می‌بیند که عبارت است از تهیهٔ مدل زبانی پوشانده‌شده (مسک) و پیش‌بینی جملهٔ بعدی.

در تابع مدل زبانی پوشانده‌شده تعدادی از واحدهای زبانی در جملهٔ ورودی با «مسک» پوشانده می‌شود و سپس، توسط مدل حدس زده می‌شود. در تابع پیش‌بینی جملهٔ بعدی، دو جمله الحاق‌شده با نماد «سپ»[1] که بیانگر مرزنمای جمله است، به‌عنوان ورودی به مدل داده می‌شود و سپس، این مسئله که کدام جمله می‌تواند ادامه جمله اول باشد، توسط مدل مشخص می‌شود. کارکردهای مختلف این مدل در پردازش زبان طبیعی سبب می‌شود که مقدار دقیق وزن‌های مدل تنظیم شود. در واقع، بازنمایی مبتنی بر بافت هر واژه با تجمیع تعبیهٔ واژه، تعبیهٔ بخش و تعبیهٔ مکانی ایجاد می‌شود. مدل «برت» با استفاده از انتقال‌دهنده توانسته است محدودیت ترتیب پردازش دنباله‌های ورودی را از سر راه بردارد. همچنین، استفاده از انتقال‌دهنده سرعت آموزش مدل را نسبت به سایر مدل‌ها افزایش داده و امکان آموزش الگوریتم با پیکره‌های بزرگ‌تر را میسر ساخته است (Devlin et al. 2019).

٤-٢. مدل‌های یادگیری ماشینی

در پژوهش حاضر، دو دسته از الگوریتم‌های دسته‌بندی برای برچسب‌زنی حوزهٔ علمی مستندات علوم انسانی استفاده می‌شود. در دستهٔ اول، از دسته‌بندهای متداول، مانند ماشین بردار پشتیبان (Cortes and Vapnik 1995)، «بیز» ساده (Xu 2018)، رگرسیون لجستیک، جنگل تصادفی، کا نزدیک‌ترین همسایه (Shah et al. 2020) و درخت تصمیم (Charbuty & Abdulazeez 2021) استفاده می‌شود، و در دستهٔ دوم، دسته‌بندهای مبتنی بر شبکهٔ عصبی، مانند شبکهٔ عصبی «پیچشی» و «پرسپترون» در پردازش زبان طبیعی (Kalchbrenner, Grefenstette & Blunsom 2014)، به‌کار برده می‌شود. در فرایند دسته‌بندی، از دو شیوهٔ بردارسازی «ورد۲وک» و «پارس‌برت» برای آموزش دسته‌بندها استفاده می‌شود. شایان ذکر است که با استفاده از مدل «ورد۲وک» میانگین بازنمایی واژه‌ها به‌عنوان ورودی الگوریتم‌های دسته‌بند برای پیش‌بینی حوزهٔ مقالات علمی استفاده می‌گردد.

1. SEPerator (SEP)

۵. تنظیمات آزمایش‌ها

۵-۱. داده

به‌طور کلی، زبان فارسی یکی از زبان‌هایی است که در حوزهٔ پردازش زبان طبیعی از منابع زبانی متنوعی برخوردار نیست. اگر چه تلاش‌هایی برای توسعهٔ منابع زبانی برای این زبان انجام شده است، داده‌های تهیه‌شده برای حوزه‌های پژوهشی مختلف دارای تنوع نبوده و روزآیندسازی منظم در آن‌ها اتفاق نمی‌افتد. سالانه حجم زیادی از مستندات علمی توسط پژوهشگران در زبان فارسی منتشر می‌شود. گردآوری این مستندات در حوزه‌های مختلف علمی می‌تواند یک پیکرهٔ زبانی را شکل دهد که برای رعایت اصول حق انتشار، پیکره می‌تواند حاوی فقط چکیدهٔ مستندات باشد. تعیین حوزهٔ علمی هر یک از این مستندات می‌تواند در فرایند جست‌وجو مفید باشد. با توجه به این‌که سالانه تعداد زیادی مستندات علمی در حال انتشار است، برای تسریع در فرایند بایگانی انتشارات می‌توان از روش‌های ماشینی جهت تعیین حوزهٔ علمی مستندات استفاده نمود.

ازمیان بایگانی‌های موجود، مانند «پرتال جامع علوم انسانی»[1]، نسخهٔ پیشین «پایگاه اطلاعات علمی ایران (گنج)»[2]، «نورمگز»[3]، «مگیران»[4]، «پایگاه مرکز اطلاعات علمی جهاد دانشگاهی»[5]، سامانهٔ نشریات علمی ایران در «کتابخانهٔ ملی ایران»[6] و «مرکز منطقه‌ای اطلاع‌رسانی علوم و فناوری»[7]، که دسترسی به چکیدهٔ مقالات از طریق وب وجود دارد، «پرتال جامع علوم انسانی» را به دو دلیل انتخاب کردیم و به تهیهٔ پیکرهٔ مورد نظر از طریق خزش آن پرتال اقدام نمودیم. دلیل اول صدور مجوز و سهولت برای گردآوری مستندات از این پرتال بود، و دلیل دوم که مهم‌تر بود، وجود برچسب مقولهٔ حوزهٔ علمی برای هر یک از مستندات بود. ویژگی دیگر این منبع اطلاعاتی این بود که به داده‌های حوزهٔ علوم انسانی محدود بود و شامل مستندات علوم پزشکی و فنی-مهندسی نمی‌شد. از خزش «پرتال جامع علوم انسانی» تعداد چکیدهٔ ۳۲۶،۳۴۵ مقاله در بازهٔ زمانی ۱۲۸۵ تا ۱۳۹۹ به‌دست آمد. پس از بررسی داده‌های به‌دست‌آمده از فرایند خزش با دو نکتهٔ قابل توجه مواجه شدیم:

الف) بعضی از اسناد چکیده نداشت. برای نداشتن چکیده دو دلیل عمده یافت شد. یکی این‌که ساختار مقالات قدیمی با ساختار مقالات امروزی متفاوت بوده و در این دسته از

1. https://www.ensani.ir
2. https://ganj-old.irandoc.ac.ir
3. https://www.noormags.ir
4. https://www.magiran.com
5. https://www.sid.ir/fa/journal/AdvanceJournal.aspx
6. https://iranjournals.nlai.ir
7. https://search.ricest.ac.ir/Inventory/index_10702.htm

مقالات، چکیده بخشی از اجزای اصلی مقاله نبوده است. دلیل دیگر اینکه اگرچه برای بعضی از مقالات در فایل «پی‌دی‌اف» چکیده موجود بود، به‌دلیل نبودِ چکیده در صفحهٔ وبِ مربوط به سند علمی به‌صورت متن حروفچینی‌شده، امکان خزش داده میسر نشد.

ب) در داده‌های خزش‌شده از منابع مختلف، اسناد علمی به زبان‌هایی به‌جز فارسی، مانند عربی و انگلیسی، موجود بود که باید از میان مستندات علمی حذف می‌شد و پیکره‌ای تمام‌فارسی به‌دست می‌آمد. برای این منظور، از کتابخانهٔ «پلی‌گلات»[1] که به زبان برنامه‌نویسی «پایتون»[2] موجود است، استفاده شد و با مجموعهٔ دادگان wili-2018[3] آموزش دیده و ارزیابی شد.[4]

پس از پالایش داده‌های موجود در «پرتال جامع علوم انسانی»، تعداد ۱۱۴،۱۷۲ سند به‌دست آمد. این داده که آن را «پیکرهٔ مستندات علمی علوم انسانی» می‌نامیم، حاوی فقط مقالات فارسی بوده و برای تمامی مستندات، چهار اطلاعاتِ عنوان و چکیدهٔ مقاله، تاریخ انتشار، مقولهٔ حوزهٔ علمی و نام نویسنده موجود است. در پژوهش حاضر از این تعداد مقاله برای ساخت مدل پردازشی و ارزیابی استفاده می‌گردد. در راستای این هدف، این پیکره به سه دسته تقسیم شده است: الف) دادهٔ آموزش (۷۰ درصد)، ب) دادهٔ اعتبارسنجی (۱۰ درصد)، و ج) دادهٔ آزمون نهایی (۲۰ درصد). در جدول ۱، اطلاعات آماری این مجموعه داده ارائه شده است.

جدول ۱. اطلاعات آماری تقسیم‌بندی مستندات از «پیکرهٔ مستندات علمی علوم انسانی»

داده	تعداد
آموزش	۸۲۲۰۲
اعتبارسنجی	۹۱۳۴
آزمون	۲۲۸۳۴

مقالات علمی این مجموعه داده در ۱۶ حوزهٔ مختلف مربوط به علوم انسانی جمع‌آوری شده است. جزئیات مربوط به توزیع آماری دادهٔ آموزش، اعتبارسنجی و آزمون برای هر حوزه در جدول ۲، گزارش شده است.

1. polyglot 2. Python 3. https://martin-thoma.com/wili
4. https://polyglot.readthedocs.io/en/latest/Detection.html

جدول ۲. اطلاعات آماری توزیع داده‌ها برای حوزه‌های مختلف علوم انسانی در «پیکرهٔ مستندات علمی علوم انسانی»

موضوع	تعداد کل مستندات	دادهٔ آموزش	دادهٔ اعتبارسنجی	دادهٔ آزمون
ادبیات	۸۸۲۶	۶۳۵۷	۶۹۸	۱۷۷۱
اقتصاد	۹۹۶۵	۷۱۳۴	۸۳۹	۱۹۹۲
تاریخ	۵۰۰۴	۳۵۵۱	۴۳۱	۱۰۲۲
تربیت بدنی	۴۴۵۴	۳۱۴۹	۳۸۱	۹۲۴
جغرافیا	۱۰۳۷۰	۷۴۸۱	۸۲۴	۲۰۶۵
حقوق	۵۲۴۰	۳۷۸۱	۴۳۰	۱۰۲۹
روان‌شناسی و علوم تربیتی	۱۳۵۴۳	۹۸۴۶	۱۰۱۸	۲۶۷۹
زبان‌شناسی	۳۱۹۹	۲۳۱۰	۲۳۴	۶۵۵
علوم اجتماعی و ارتباطات	۸۸۷۲	۶۳۴۹	۷۰۵	۱۸۱۸
علوم اسلامی	۱۱۴۰۱	۸۲۱۹	۹۱۴	۲۲۶۸
علوم سیاسی و روابط بین‌الملل	۶۵۹۶	۴۶۷۹	۵۵۱	۱۳۶۶
علوم کتابداری	۲۶۰۰	۱۹۰۶	۲۰۶	۴۸۸
فلسفه و منطق	۴۸۶۳	۳۵۱۴	۴۲۰	۹۲۹
مدیریت و حسابداری	۱۴۵۶۹	۱۰۵۴۵	۱۱۰۴	۲۹۲۰
مطالعات زنان	۲۰۷۷	۱۴۹۸	۱۷۳	۴۰۶
مطالعات هنر	۳۱۹۹	۲۳۱۰	۲۶۶	۶۲۳

۵-۲. نتایج به‌دست‌آمده

در این پژوهش نتایج دسته‌بندی متون علمی با دو رویکرد یادگیری ماشینی و یادگیری عمیق بررسی شده است. افزون بر آن، تأثیر استفاده از مدل بازنمایی مبتنی بر بافت «پارس‌برت» در مقایسه با مدل بازنمایی قدیمی‌تر، مانند «ورد۲وک»، بررسی شده است.

ابتدا با آموزش مدل «ورد۲وک» بر روی متون مقالات علمی، بازنمایی واژه‌های موجود در هر مقاله استخراج شد. سپس، میانگین بازنمایی واژه‌ها برای آموزش الگوریتم‌های

یادگیری ماشینی «بیز» ساده، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان و کا نزدیک‌ترین همسایه استفاده شد. نتایج استفاده از الگوریتم‌های یادگیری ماشین و مدل بازنمایی «ورد۲وک» در جدول ۳، نشان داده شده است. همان‌طور که در این جدول قابل مشاهده است، الگوریتم رگرسیون لجستیک بهترین نتیجه را با استفاده از مدل بازنمایی «ورد۲وک» به‌دست آورده است. الگوریتم‌های درخت تصمیم و «بیز» ساده به‌ترتیب، ضعیف‌ترین عملکرد را در دسته‌بندی متون علمی داشته است.

جدول ۳. کارایی الگوریتم‌های یادگیری ماشین مبتنی بر مدل بازنمایی «ورد۲وک»

مدل	معیار F (درصد)	
	میکرو	ماکرو
«ورد۲وک»- درخت تصمیم	۵۲/۵۵	۴۷/۶۹
«ورد۲وک» - بیز ساده	۶۲/۱۴	۵۹/۳۳
«ورد۲وک» - ماشین بردار پشتیبان	۶۹/۷۲	۶۶/۳۱
«ورد۲وک» - جنگل تصادفی	۷۰/۳۸	۶۶/۵۳
«ورد۲وک» - کا نزدیکترین همسایه	۷۰/۴۶	۶۶/۰۸
«ورد۲وک» - رگرسیون لجستیک	۷۰/۷۲	۶۷/۰۴

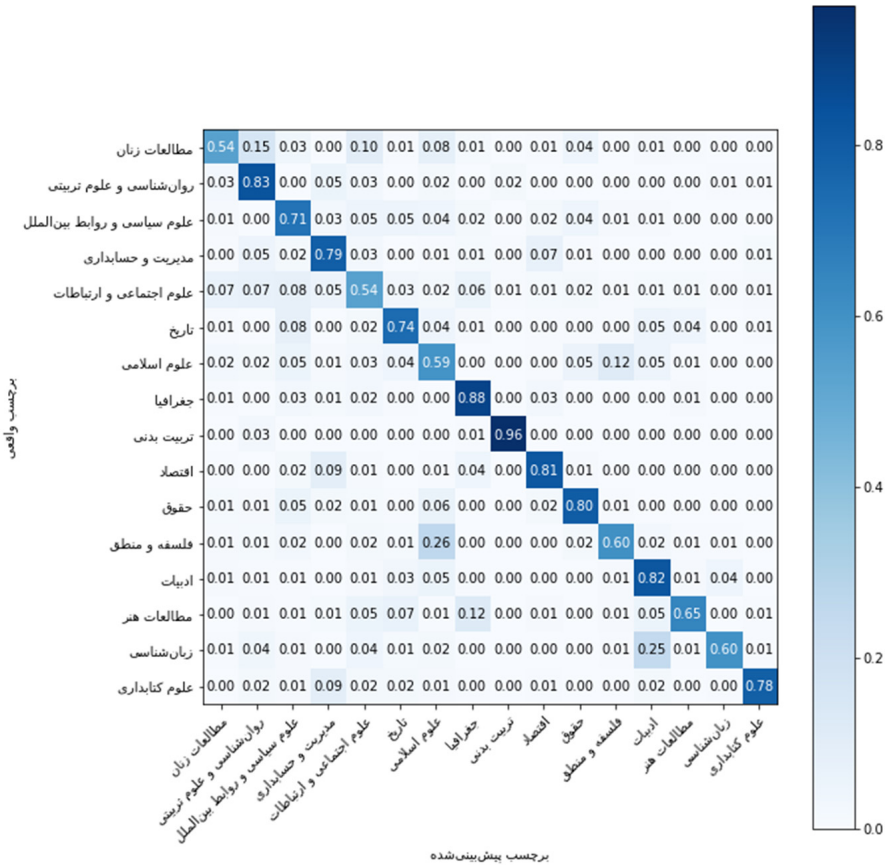
در مرحلهٔ بعد، برای بهبود نتایج به‌دست‌آمده، از مدل‌های مبتنی بر یادگیری عمیق و مدل‌های بازنمایی از پیش‌آموزش‌دیدهٔ مبتنی بر بافت کمک گرفته شد. در این پژوهش از دو مدل شبکهٔ عصبی «پرسپترون» و «پیچشی» برای دسته‌بندی استفاده شد و در دو آزمایش جداگانه، بازنمایی میانگین واژه‌ها با مدل «ورد۲وک» به شبکهٔ عصبی «پرسپترون» و بازنمایی تک‌تک واژه‌ها با مدل «ورد۲وک» به شبکهٔ عصبی «پیچشی» داده شد. طبق نتایج آزمایش‌ها در جدول ۴، شبکهٔ عصبی «پرسپترون» در مقایسه با شبکهٔ عصبی «پیچشی» عملکرد بهتری برای دسته‌بندی مقالات علمی به‌دست آورد. در نهایت، مدل «ورد۲وک»–«پرسپترون» باعث بهبود ۲/۸۴ درصدی در معیار F (میکرو) نسبت به الگوریتم یادگیری ماشینی رگرسیون لجستیک شد. برای آزمایش‌های بعدی، مدل «پارس‌برت» که حاوی بازنمایی واژه‌ها مبتنی بر بافت بود، به‌کار برده شد. در این آزمایش‌ها نیز شبکهٔ عصبی «پرسپترون» نتایج بهتری را برای دسته‌بندی مقالات علمی به‌دست آورد؛ به این

صورت که استفاده از مدل «پارس‌برت»–«پرسپترون» به بهترین دقت رسید و در مقایسه با مدل «ورد۲وک»–«پرسپترون» باعث بهبود ۱/۱۵ درصدی در معیار F (میکرو) شد.

جدول ٤. کارایی الگوریتم‌های یادگیری ماشینی مبتنی بر مدل بازنمایی«ورد۲وک» و «پارس‌برت»

مدل	معیار F (درصد)	
	میکرو	ماکرو
«ورد۲وک» - رگرسیون لجستیک	۷۰/۷۲	۶۷/۰۴
«ورد۲وک» - «پیچشی»	۶۷/۷۹	۶۴/۰۶
«ورد۲وک» - «پرسپترون»	۷۳/۵۶	۷۰/۵۷
«پارس‌برت» - «پیچشی»	۷۳/۴۳	۷۰/۹۳
«پارس‌برت» - «پرسپترون»	۷۴/۷۱	۷۲/۵۵

همان‌طور که در ماتریس درهم‌ریختگی نرمال‌سازی‌شدهٔ به‌دست‌آمده از نتایج مدل «پارس‌برت»–«پرسپترون» در شکل ۲، مشاهده می‌شود، تعیین برچسب مستندات در حوزه‌های «مطالعات زنان»، «علوم اجتماعی و ارتباطات»، «علوم اسلامی»، «فلسفه و منطق»، و «زبان‌شناسی» به‌ترتیب با بیشترین خطای مدل همراه بوده است. اما تعیین برچسب مستندات در حوزه‌های «تربیت‌بدنی»، «جغرافیا»، «روان‌شناسی»، و «علوم تربیتی» کمترین خطای مدل را داشته است.



شکل ۲. ماتریس درهم‌ریختگی نرمال‌سازی‌شده حاصل از نتایج مدل «پارس‌برت» – «پرسپترون»

۵-۳. تحلیل خطا

در این بخش، تعدادی از داده‌های به غلط دسته‌بندی‌شده در دادهٔ آزمون توسط مدل «پارس‌برت»–«پرسپترون» مورد بررسی و تحلیل قرارگرفته است تا با رفع کاستی‌ها، نتایج مدل در پژوهش‌های آینده بهبود یابد. نمونهٔ دادهٔ (۱) را در نظر بگیرید:

(۱) *مطالعات میان‌رشته‌ای و افزایش سطح تعامل میان رشته‌های مختلف علوم انسانی در بررسی‌ها و مطالعات اخیر جایگاه ویژه‌ای یافته است. علل گرایش به الحاد نیز موضوعی است که امکان دارد از زوایای مختلف روان‌شناسانه، جامعه‌شناسانه و دین‌شناسانه بررسی شود. در این مقاله چرایی و چگونگی تأثیرگذاری علل غیرفلسفی مانند آسیب‌های عاطفی، روانی، تربیتی و فضای فکری و فرهنگی*

جامعه (که اغلب توسط روشنفکران رهبری می‌شود) بررسی خواهد شد تا مشخص شود که مداخلهٔ عوامل غیرفلسفی تا چه اندازه در برگزیدن جهان‌بینی الحادی مؤثر است. «پاسکوبینی» با استفاده از نظریهٔ «پدر ناکارآمد» پروفسور «ویتز»، نقش عوامل روان‌شناسانه و اجتماعی را در گرایش به الحاد بررسی کرده است.

برچسب این نمونهٔ دادهٔ در واقعیت «فلسفه و منطق» است. اما مدل «پارس‌برت»—«پرسپترون» برچسب سند را «علوم اجتماعی و ارتباطات» پیش‌بینی کرده است. همان‌طور که از چکیدهٔ مقاله استنباط می‌شود، در این مقاله مطالعه‌ای میان‌رشته‌ای دربارهٔ چند حوزهٔ علوم انسانی شامل «فلسفه و منطق»، «علوم اجتماعی و ارتباطات» و «روان‌شناسی و علوم تربیتی» انجام شده است. بنابراین، به‌دلیل ابهام، تعیین درست مقولهٔ حوزهٔ این مقاله توسط مدل با دشواری همراه شده است. کاربرد واژه‌هایی چون *انسانی*، *روان‌شناسانه*، *جامعه‌شناسانه*، *عاطفی*، *روانی*، *تربیتی*، *فرهنگی*، *اجتماعی*، و *تعامل* باعث تعیین حوزهٔ مقاله در حوزهٔ «علوم اجتماعی و ارتباطات» شده است. تنها واژه‌ای که می‌تواند وزن زیادی در حوزهٔ «فلسفه» داشته باشد، خود واژهٔ *فلسفه* است. برای بررسی دقیق‌تر این مسئله، شباهت کسینوسی این چکیده با تمامی داده‌های دادهٔ آموزش را بر اساس بازنمایی به‌دست‌آمده از مدل زبانی «پارس‌برت» محاسبه کردیم. از میان ۱۰ سندی که بیشترین شباهت را به این نمونهٔ داده داشت، ۶ چکیده در حوزهٔ «مطالعات زنان»، ۲ چکیده در حوزهٔ «علوم اسلامی»، ۱ چکیده در حوزهٔ «علوم اجتماعی و ارتباطات» و ۱ چکیده در حوزهٔ «تاریخ» وجود داشت. بنابراین، نتایج به‌دست‌آمده از شباهت کسینوسی وجود ابهام در نمونهٔ دادهٔ ذکرشده را تأیید می‌کند. به همین جهت، تعیین حوزهٔ این مقاله برای مدل پیشنهادی دشوار بوده و به غلط پیش‌بینی و تعیین شده است.

در نمونهٔ قبلی، شبیه‌ترین چکیده به چکیدهٔ مورد نظر، به حوزهٔ «مطالعات زنان» تعلق داشت. تعداد زیادی از مقاله‌های موجود در مجموعهٔ داده در زمینهٔ «مطالعات زنان» خود نیز دارای ابهام بود. بنابراین، این وابستگی موضوعات مختلف در مقالات و وجود مقاله‌های میان‌رشته‌ای در حوزه‌های مختلف علوم انسانی در مجموعهٔ داده، دسته‌بندی اسناد را با مشکل روبه‌رو می‌کند. به‌عنوان مثال، نمونهٔ دادهٔ (۲) دارای برچسب «مطالعات زنان» است. اما مدل پیشنهادی برچسب «علوم اجتماعی و ارتباطات» را برای این چکیده پیش‌بینی کرده است:

(۲) *مسئلهٔ فقر به‌طور کلی، و فقر زنان به‌طور خاص، از جمله مسائل اجتماعی است که در سال‌های اخیر تلاش گردیده است راهبردهایی جهت کاهش آن اتخاذ گردد. نتایج بررسی‌های آماری نشان می‌دهد که فقر و نابرابری در مناطق روستایی بیش از نواحی شهری و در بین زنان بیش از مردان است. این تحقیق در صدد است در یک مطالعهٔ موردی با روش کیفی و بر اساس رهیافت «گراندد تئوری» علل اجتماعی و فرهنگی فقر زنان روستایی را مورد مطالعه قرار دهد. نمونه‌گیری تحقیق، به شیوهٔ نمونه‌گیری مبتنی بر هدف بوده است و از زنان متأهل و سرپرست خانوار و نیز دختران جوان تحت پوشش «کمیتهٔ امداد امام خمینی (ره)» مصاحبهٔ گروهی انجام شده است. سؤالات محوری این پژوهش عبارت‌اند از: عوامل تعیین‌کنندهٔ فقر از دیدگاه زنان و دختران روستای مورد مطالعه چه بوده است؟ راهبردهای فردی آن‌ها در مواجههٔ مسئلهٔ فقر چه بوده است؟ پیامدهای فقر از نظر آن‌ها چه بوده است؟ یافته‌های تحقیق نشان می‌دهد که تبیین‌کننده‌های فقر از دیدگاه زنان و دختران روستایی، ناآگاهی اجتماعی، تعصبات طایفه‌ای، اولویت رسیدگی به افراد ذکور در خانواده، تقدیرگرایی، فقر بین نسلی، جمعیت زیاد خانواده، محدودیت‌های تحصیلی، بیکاری و آثار جنگ تحمیلی است. نتایج تحقیق نشان می‌دهد که بارزترین علل فقر در منطقهٔ مورد مطالعه، عوامل اجتماعی و فرهنگی است.*

در این چکیده به یکی از موضوعات اجتماعی یعنی فقر و به‌طور خاص، به فقر زنان پرداخته شده است. بنابراین، می‌توان این چکیده را متعلق به دو حوزهٔ «علوم اجتماعی و ارتباطات» و همچنین «مطالعات زنان» دانست که سبب سخت‌شدن پیش‌بینی برچسب توسط مدل می‌شود. برای بررسی بیشتر این مسئله، شباهت کسینوسی این چکیده نسبت به کل مجموعهٔ دادهٔ آموزش مورد بررسی قرار گرفت. از میان ۱۰ سند اول مشابه به این نمونهٔ داده، ۴ چکیده به حوزهٔ «علوم اسلامی»، ۲ چکیده به حوزهٔ «مطالعات زنان»، ۲ چکیده به حوزهٔ «علوم اجتماعی و ارتباطات»، ۱ چکیده به حوزهٔ «زبان‌شناسی» و ۱ چکیده به حوزهٔ «ادبیات» تعلق داشت. همان‌طور که مشخص است، چکیدهٔ مقاله‌ها در حوزه‌های «مطالعات زنان» و «علوم اجتماعی و ارتباطات» به نسبت یکسان مشابه نمونهٔ دادهٔ هدف شده است که بیانگر وجود ابهام و دشواری تشخیص حوزهٔ این مقاله توسط مدل پیشنهادی است.

نمونهٔ (۳) را با برچسب واقعی «علوم اجتماعی و ارتباطات» و برچسب پیش‌بینی‌شدهٔ «علوم سیاسی و روابط بین‌الملل» توسط مدل بررسی کرده‌ایم:

(۳) *این پژوهش به فرهنگ سیاسی (که یکی از موضوعات مهم در حوزه‌های علوم اجتماعی و سیاسی است) دانشجویان دانشگاه تهران و عوامل مؤثر بر آن می‌پردازد. برای تحلیل فرهنگ سیاسی از نظریه‌های کسانی چون «آلموند»، «وربا»، «پای»، «راش»، «لیپست» و «دال» استفاده شده است. روش تحقیق در این پژوهش از نوع پیمایش و تکنیک جمع‌آوری داده‌ها و اطلاعات از طریق پرسشنامهٔ استاندارد و برای تجزیه و تحلیل از نرم‌افزار «اس‌پی‌اس‌اس» استفاده گردیده و از آماره‌های تحلیلی مانند رگرسیون سود جسته تا به ارائهٔ تحلیل مسیر و در نهایت، رسم مدل نهایی دست یابیم. شیوهٔ نمونه‌گیری تصادفی ساده و حجم نمونه با استفاده از فرمول «کوکران» برابر با ۳۰۰ نفر بوده است. نتیجهٔ تحقیق چنین نشان می‌دهد که سطح نظری فرهنگ سیاسی دانشجویان بیشتر از نوع تبعی است و بین سطح نظری و سطح رفتاری فرهنگ سیاسی دانشجویان شکاف وجود دارد. همچنین، متغیرهایی مانند منافع سیاسی، مهارت‌های سیاسی، پایگاه اجتماعی–اقتصادی، سن بر این شکاف تأثیر دارد و تأثیر متغیرهایی همچون موانع اجتماعی و خانوادگی، اعتماد سیاسی بر شکاف فوق‌الذکر تأیید نگردید.*

در این چکیده، به موضوع فرهنگ سیاسی دانشجویان در دو سطح نظری و رفتاری پرداخته شده است. بنابراین، می‌توان این مقاله را نیز با دو حوزهٔ علوم انسانی، یعنی «علوم اجتماعی و ارتباطات» و «علوم سیاسی و روابط بین‌الملل»، مرتبط دانست. با محاسبهٔ شباهت کسینوسی برای این سند، همانند دو نمونهٔ قبلی، ۱۰ سندی که بیشترین شباهت به این نمونهٔ داده را داشته است، فهرست کرده‌ایم که عبارت است از: ۴ چکیده در حوزهٔ «تاریخ»، ۳ چکیده در حوزهٔ «مطالعات زنان»، ۲ چکیده در حوزهٔ «علوم سیاسی و روابط بین‌الملل» و ۱ چکیده در حوزهٔ «روان‌شناسی و علوم تربیتی. از میان این ۱۰ سند مشابه، هیچ مقاله‌ای به حوزهٔ «علوم اجتماعی و ارتباطات» تعلق نداشت و ۲ مقاله به حوزهٔ «علوم سیاسی و روابط بین‌الملل» مربوط بود. بنابراین، وجود مقالات بین رشته‌ای در حوزه‌های مختلف علوم انسانی و واژه‌های مشترک میان آن‌ها سبب ایجاد ارتباط بین حوزه‌های مختلف شده است. این ارتباط و وابستگی باعث شناسایی و کشف الگوهایی برای تعیین حوزهٔ مقالات

شده است. همین الگوها می‌توانند مدل را برای برچسب‌زنی یک مقالهٔ میان‌رشته‌ای جدید گمراه کنند و دادهٔ جدید به اشتباه برچسب‌گذاری گردد.

سه نمونه دادهٔ ذکرشده به حوزه‌های «مطالعات زنان»، «علوم اجتماعی و ارتباطات» و «علوم سیاسی و روابط بین‌الملل» تعلق داشت. طبق ماتریس درهم‌ریختگی در بخش قبل، این سه حوزه چندین بار به اشتباه به‌جای یکدیگر برچسب‌گذاری شده است. به‌طور کلی، تشخیص یک حوزهٔ مشخص برای تعدادی از مقالات بین رشته‌ای می‌تواند برای نیروی انسانی نیز دشوار باشد. به همین دلیل، مدل پیشنهادی برای پیش‌بینی برچسب‌ها با مشکل روبه‌رو شده است. یک راهکار پیشنهادی این است که به‌جای تخصیص یک برچسب به یک سند، چند برچسب بر اساس ترتیب احتمالاتی برچسب به یک سند تخصیص داده شود.

٦. بحث و نتیجه‌گیری

در این پژوهش، ابتدا مجموعهٔ داده‌ای از چکیدهٔ مقالات علمی به زبان فارسی در حوزهٔ علوم انسانی که حاوی ۱۶ حوزهٔ مختلف بود، به‌منظور دسته‌بندی جمع‌آوری گردید. سپس، دسته‌بندی متون علمی از دو جهت بررسی شد: مقایسهٔ انواع مدل‌های یادگیری ماشینی و مدل‌های مبتنی بر یادگیری عمیق در کنار مقایسهٔ میزان اثرگذاری استفاده از مدل‌های بازنمایی مبتنی بر بافت، مانند «پارس‌برت» و مدل‌های بازنمایی قدیمی‌تر، مانند «ورد۲وک». ابتدا، بازنمایی واژه‌های متون علمی با استفاده از مدل بازنمایی «ورد۲وک» استخراج شد. سپس، میانگین بازنمایی واژه‌های هر متن به‌عنوان ورودی الگوریتم‌های یادگیری ماشینی استفاده گردید. الگوریتم رگرسیون لجستیک بهترین نتیجه را در معیار F به‌دست آورد. از بین الگوریتم‌های استفاده‌شده در آزمایش‌ها، الگوریتم‌های درخت تصمیم و «بیز» ساده ضعیف‌ترین عملکرد را داشت. همان‌طور که انتظار می‌رفت، عملکرد شبکه‌های عصبی در مجموع در مقایسه با الگوریتم‌های یادگیری ماشینی بهتر بود. از بین شبکه‌های عصبی «پیچشی» و «پرسپترون»، استفاده از شبکهٔ عصبی «پرسپترون» نتیجهٔ بهتری برای دسته‌بندی متون مقاله‌های علمی به همراه داشت. در نهایت، استفاده از مدل بازنمایی مبتنی بر بافت «پارس‌برت» در مقایسه با مدل «ورد۲وک» با استفاده از شبکهٔ عصبی «پرسپترون» با تفاوت معناداری باعث بهبود ۱/۱۵ درصدی در معیار F (میکرو) شد. وجود ابهام و مقالات بین رشته‌ای در مجموعهٔ داده از عوامل مؤثر در وجود خطا در خروجی مدل‌های دسته‌بندی اسناد علمی بود.

تقدیر و تشکر

این پژوهش در چارچوب طرح پژوهشی شماره ۲۸۱۱۱ در مجموعهٔ طرح‌های «طرح جامع اعتلای علوم انسانی معطوف به پیشرفت کشور» در پژوهشگاه علوم انسانی و مطالعات فرهنگی انجام پذیرفته است.

فهرست منابع

امامی آزادی، طاهره، و فرشاد الماس‌گنج. ۱۳۸۵. «دسته‌بندی موضوعی متون فارسی بر اساس روش آنالیز معنایی پنهان احتمالاتی بهبود یافته،» در *مجموعه‌مقالات دوازدهمین کنفرانس سالانه انجمن کامپیوتر ایران*. تهران.

باقری، ایوب، حامد فرزانه‌فر، محمدحسین سرایی، و محمدرضا احمدزاده. ۱۳۸۷. «دسته‌بندی متون خبری فارسی با استفاده از الگوریتم Naïve Bayes،» *دومین کنفرانس داده کاوی ایران*. دانشگاه صنعتی امیرکبیر، تهران.

تیمورپور، بابک، محمدمهدی سپهری، و لیلا پزشک. ۱۳۸۸. روشی نوین برای دسته‌بندی هوشمند متون علمی (مطالعه موردی مقالات فناوری نانو متخصصان ایران). *سیاست علم و فناوری*، ۲ (۲): ۱-۱۵.

رباطی، زهرا .۱۳۹۳. *دسته‌بندی اخبار فارسی با استفاده از تکنیک‌های هوش مصنوعی*. پایان‌نامه کارشناسی ارشد. دانشگاه صنعتی شاهرود. دانشکده کامپیوتر و فناوری ارتباطات، شاهرود. ایران.

ربیعی، محمد، سید مهدی حسینی مطلق، و بهروز مینایی بیدگلی. ۱۳۹۸. ارائه روش رده‌بندی تک‌رده‌ای برای شناسایی متون پژوهشی حوزه محیط زیست ایران با استفاده از ماشین بردار پشتیبان. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۴ (۳): ۱۲۱۱-۱۲۳۴.

جمالی، ایمان، سید جواد میرعابدینی، و علی هارون‌آبادی. ۱۳۹۹. ارائهٔ یک مدل جهت دسته‌بندی متون فارسی با استفاده از ترکیب روش‌های دسته‌بندی،» *فصلنامه تخصصی مهندسی مخابرات* ۱۰ (۳۸): ۶۱-۷۲.

شکوهیان، محبوبه، عاصفه عاصمی، احمد شعبانی، و مظفر چشمه‌سهرابی. ۱۳۹۸. ارائهٔ مدل دسته‌بندی موضوعی تولیدات علمی حوزه سلامت با استفاده از روش‌های متن‌کاوی. *پژوهشنامه پردازش و مدیریت اطلاعات* ۳۵ (۲): ۵۵۳-۵۷۴.

علایی ابوذر، الهام، نصرالله پاک‌نیت، علی‌اصغر حجت‌پناه، و مجتبی زالی و محمدهادی آقالویی‌آغمیونی. ۱۴۰۰. «معرفی یک پیکره متنی تخصصی: پیکره پژوهشنامه،» مجله *پژوهش‌های زبان‌شناسی تطبیقی*، ۱۱ (۲۲): ۲۷۱-۲۸۹. https://rjhll.basu.ac.ir/article_4226.html

کامیابی گل، عطیه، الهام اخلاقی باقوجری، احسان عسگری، و هانیه حبیبی. ۱۳۹۷. استخراج اطلاعات از پیکره زبانی: معرفی پیکره مقاله‌های علمی-پژوهشی دانشگاه فردوسی مشهد. *کتابداری و اطلاع‌رسانی* ۲۱ (۲): ۳-۲۵.

منفرد، زینت. ۱۳۹۳. *توسعه راهکارهایی هوشمند جهت پردازش خبرهای فارسی*. پایان‌نامه کارشناسی ارشد. دانشگاه شیراز. دانشکده مهندسی برق و الکترونیک. شیراز، ایران.

نوریان، زهرا، و یداله زاده‌طبری میثم. ۱۳۹۴. «دسته‌بندی اسناد فارسی با استفاده از شبکه‌های عصبی،» در مجموعه‌مقالات *کنفرانس بین‌المللی دستاوردهای نوین در علوم مهندسی و پایه*. مرکز پژوهشی زمین کاو با همکاری انجمن علوم مهندسی لندن، اودسا، اوکراین.

هاشمی، سیامک، و مریم حورعلی. ۱۳۹۶. «دسته‌بندی اخبار فارسی حوزه دفاعی با استفاده از هستان‌شناسی،» در *مجموعه‌مقالات دومین کنفرانس بین‌المللی پژوهش‌های دانش‌بنیان در مهندسی کامپیوتر و فناوری اطلاعات*، تهران، ایران.

یعقوبی، ملیکا. ۱۳۹۱. *سیستم مکانیزه طبقه‌بندی اخبار در بستر وب*. پایان‌نامه کارشناسی ارشد. دانشگاه صنعتی شاهرود. دانشکده کامپیوتر و فناوری ارتباطات. شاهرود، ایران.

References

Aharony, N. 2011. Librarians' attitudes toward knowledge management. *College & Research Libraries* 72 (2): 111-126.

Ahmadi, P., M. Tabandeh, & I. Gholampour. 2016. "Persian text classification based on topic models," In *Proceedings of the 24th Iranian Conference on Electrical Engineering*, pp: 86-p1, IEEE Computer Society. Shiraz, Iran.

AleAhmad, A., H Amiri, E. Darrudi, M. Rahgozar, & F. Oroumchian. 2009. "Hamshahri: A standard Persian text correction. *Knowledge-based Systems* 22 (5): 382-387.

Bijankhan, M., J. Sheikhzadegan, & M.R. Roohani. 1994. "FARSDAT-The speech database of Farsi spoken language," In *Proceedings of the Australasian Conference in Speech Science & Technology*, Vol.2, pp: 826-830. Perth, Australia.

Bird, S., R. Dale, B. Dorr, B. Gibson, M. Joseph, M.Y. Kan, D. Lee, B. Powley, D. Radev & Y. Fan Tan. 2008. "The ACL Anthology Reference Corpus: A reference dataset for bibliographic research in Computational Linguistics," In *Proceedings of the 6th International Conference on Language Resources & Evaluation*, Marrakech, Morocco, pp: 1755-1759.

Blei, D. M.; A. Y. Ng, M. I. Jordan, & J. Lafferty. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Boisot, M. & A. Canals. 2004. Data, information, & knowledge: Have we got it right. *Journal of Evolutionary Economics* 14: 43–67.

Cortes, C., & V. Vapnik. 1995. Support-vector networks. *Machine Learning* 20 (3): 273-297.

Charbuty, B., & A. Abdulazeez. 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science & Technology Trends*. 2 (1): 20-28.

Chowdhury, S., & M. P. Schoen. 2020. "Research paper classification using supervised machine learning techniques," In *Proceedings of the Intermountain Engineering, Technology & Computing*, pp. 1-6.

Dadgar, S. M. H., M. S. Araghi, & M. M. Farahani. 2016. “A novel text mining approach based on TF-IDF & support vector machine for news classification,” In *Proceedings of 2016 IEEE International Conference on Engineering & Technology*, pp. 112–116. Wuhan, China.

Degaetano-Ortlieb, S., H. Kermes, E. Lapshinova-Koltunski, & E. Teich. 2013. “SciTex - A diachronic corpus for analyzing the development of scientific registers,” In P. Bennett, M. Durrell, S. Scheible, & R. J. Whitt (eds.), *New Methods in Historical Corpus Linguistics* 3: 93–104.

Devlin, J., M. W. Chang, K. Lee, & K. Toutanova. 2019. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp: 4171–4186, Minneapolis: Association for Computational Linguistics.

Farahani, M., M. Gharachorloo, M. Farahani, & M. Manthouri. 2020. “ParsBERT: Transformer-based model for Persian language understanding,” arXiv preprint arXiv: 2005.12515.

Girard, J. P., & J. L. Girard. 2015. Defining knowledge management: Toward an applied compendium. *Online Journal of Applied Knowledge Management* 3 (1): 14.

Harris, Z. S. 1954. Distributional Structure. *Word* 23: 146–162.

Hofmann, T. 1999. “Probabilistic latent semantic indexing,” In *Proceedings of the Twenty-Second Annual International SIGIR Conference on Research & Development in Information Retrieval*, pp: 211-218. California, Berkeley, USA.

Imai, T., K. Nakamura, & T. Ohmameuda. 2015. “Visualization of similar news articles with network analysis & text mining,” In *Proceedings of 2015 IEEE 4th Global Conference on Consumer Electronics*, Osaka, Japan, pp: 151-152.

Jacobi, C., W. van Atteveldt, & K. Welbers. 2016. Quantitative analysis of large amounts of journalistic texts using topic modelin. *Digital Journalism* 4 (1): 89–106.

Jahantigh, M., N. Daneshpour, & M. E. N. Orojlou. 2016. “Presenting an improved combination for classification of Persian texts,” In *Proceedings of 2016 Eighth International Conference on Information & Knowledge Technology* (IKT), pp. 234–240. Bu-Ali Sina University, Hamedan, Iran.

Kalchbrenner, N., E. Grefenstette, & P. Blunsom. 2014. “A convolutional neural network for modelling sentences,” In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 655--665, June, Baltimore, Maryland: Association for Computational Linguistics.

Karami, A., Aryya Gangopadhyay, B. Z., & H. Kharrazi. 2018. Fuzzy approach topic discovery in health & medical corpora. *International Journal of Fuzzy Systems* 20: 1334-1345.

Kim, S. W., & J. M. Gil. 2019. Research paper classification systems based on TF-IDF & LDA schemes. *Human-centric Computing & Information Sciences* 9: 30.

Kwary, D. A. 2018. A corpus & a concordancer of academic journal articles. *Data in Brief* 16: 94-100.

Landauer, C. 1998. “Data, information, knowledge, understanding: Computing up the meaning hierarchy,” In *Proceedings of the 1998 IEEE International Conference on Systems, Man, & Cybernetics*, San Diego, California, pp. 2255-2260.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, & J. Dean. 2013. “Distributed representations of words & phrases & their compositionality,” In *Advances in Neural Information Processing Systems* 26, eds. Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., & Weinberger, K. Q., Curran Associates, Inc., pp. 3111–3119.

Momtazi, S. & M. Ghayoomi. 2014. “Weekly supervised text categorization using topic modeling,” In *Proceedings of the 3rd Conference on Computational Linguistics.* Tehran, Iran.

Nonaka, I. 1991. *Harvard Business Review* 69 (6): 96–104.

Pennington, J., R. Socher, & C.D. Manning. 2014. “Glove: Global Vectors for word representation,” In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, vol. 14, pp. 1532–1543. Doha, Qatar.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee & L. Zettlemoyer. 2018. “Deep contextualized word representations,” In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp: 2227–2237.

Radford, A., K. Narasimhan, T. Salimans, & I. Sutskever. 2018. Improving language understanding by generative pre-training. https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf (accessed May 6, 2022)

Rivest, M., E. Vignola-Gagné, & É. Archambault. 2021. “Article-level classification of scientific publications: A comparison of deep learning, direct citation & bibliographic coupling”. *PLoS ONE* 16 (5): e0251493.

Rong, X. 2014. word2vec parameter learning explained. arXiv preprint arXiv: 1411. 2738.

Rowley, J. 2007. The wisdom hierarchy: Representations of the DIKW hierarchy. *Journal of Information Science* 33 (2): 163–180.

Salton, G. 1971. *The SMART Retrieval System---Experiments in Automatic Document Processing*. Upper Saddle River, NJ: Prentice-Hall, Inc.

Shah, K., H. Patel, D. Sanghvi, & M. Shah. 2020. A comparative analysis of logistic regression, random forest & KNN models for the text classification. *Augmented Human Research* 5 (1): 1-16.

Sharma, N. 2008. The origin of the data information knowledge wisdom hierarchy. https://www.researchgate.net/publication/292335202_The_Origin_of_Data_Information_Knowledge_Wisdom_DIKW_Hierarchy. (accessed May 6, 2022)

Song, L., Z. Wang, H. Mi, & D. Gildea. 2016. “Sense embedding learning for word sense induction,” In *Proceedings of the 5th Joint Conference on Lexical & Computational Semantics*, The *SEM 2016 Organizing Committee, pp. 85–90. Berlin, Germany.

Xu, S. 2018. Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science* 44 (1): 48-59.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell Publishing Ltd.

مسعود قیومی

متولد سال ۱۳۵۸، دارای مدرک تحصیلی دکتری در رشتهٔ زبان‌شناسی رایانشی از دانشگاه آزاد برلین، آلمان، است. ایشان هم‌اکنون استادیار پژوهشکده زبان‌شناسی در پژوهشگاه علوم انسانی و مطالعات فرهنگی است.

زبان‌شناسی رایانشی و پردازش زبان طبیعی، مدل‌سازی زبانی، یادگیری ماشینی، نحو و معناشناسی واژگانی از جمله علایق پژوهشی وی است.

مریم موسویان

متولد سال ۱۳۷۳، دارای مدرک تحصیلی کارشناسی ارشد در رشتهٔ هوش مصنوعی از دانشگاه صنعتی امیرکبیر است.

پردازش زبان طبیعی و یادگیری ماشین از جمله علایق پژوهشی وی است.