نورواژه: یک دادگان معیار برای استخراج کلیدواژه از مقالات علمی فارسی

محمدامین طاهری[۱]*، محمدابراهیم شناسا[۲]، بهروز مینایی بیدگلی[۳]، سیدعلی حسینی[۴]

دانشجوی کارشناسی‌ارشد، دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم‌وصنعت ایران، تهران، ایران[۱*]

مربی، دانشکدهٔ مهندسی برق و کامپیوتر، دانشگاه آزاد اسلامی واحد تهران-شمال، تهران، ایران[۲]

دانشیار، دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم‌وصنعت ایران، تهران، ایران[۳]

دانشجوی پسادکترا، دانشکدهٔ مهندسی کامپیوتر، دانشگاه علم‌وصنعت ایران، تهران، ایران[۴]

چکیده

کلیدواژه‌ها، مهم‌ترین واژه‌های متن هستند که ایدهٔ بنیادین آن را در عباراتی کوتاه بیان می‌کنند. استخراج کلیدواژه یکی از کاربردهای پردازش زبان طبیعی است که پایهٔ بسیاری از عملیات‌ها نظیر طبقه‌بندی، خوشه‌بندی و خلاصه‌سازی متون است. تاکنون، دادگان‌های متعدّدی برای ارزیابی استخراج کلیدواژه در فارسی ارائه شده‌اند که اغلب آن‌ها به واژگان کلیدی نویسندگان مقالات اکتفا کرده و به سایر کلیدواژه‌های بالقوهٔ متن بی‌توجه‌اند. استفاده از چنین دادگانی، باعث ارزیابی نادرست روش‌های استخراج کلیدواژه می‌شود و دقت آن‌ها ناخواسته کاهش می‌یابد. در این پژوهش، ابتدا دادگان معیار نورواژه که از حدود ۱۴۰۰ مقالهٔ علمی جمع‌آوری شده‌است، برای ارزیابی روش‌های استخراج کلیدواژه معرفی می‌شود. در این دادگان علاوه‌بر واژه‌های کلیدی نویسندگان، سایر کلیدواژه‌ها توسط افراد خبره استخراج شده‌اند. برای اثبات قابلیت استفادهٔ این دادگان به‌عنوان معیار، روش‌های بی‌ناظر مختلفی روی آن آزمایش شده‌است. نتایج حاصل از این دادگان، مؤیّد نتایجی است که از سایر دادگان‌های معیار به‌دست می‌آید.

واژگان کلیدی: استخراج کلیدواژه، دادگان فارسی، یادگیری بی‌ناظر، روش‌های مبتنی بر گراف، بازیابی اطلاعات.

Noor-Vajeh: A Benchmark Dataset for Keyword Extraction from Persian Papers

Mohammad Amin Taheri [1*], Mohammad Ebrahim Shenassa[2], Behrouz Minaei Bidgoli[3], Seyyed Ali Hosseini[4]

Master Student, Faculty of Computer Engineering, Technology and Science of University, Tehran, Iran[1*]

Instructor, Electrical and Computer Engineering School, Tehran-North Branch of Azad University, Tehran, Iran[2]

Associate Professor, Faculty of Computer Engineering, Iran University of Science and Technology, Tehran, Iran[3]

Postdoctoral Student, Faculty of Computer Engineering, Technology and Science of University, Tehran, Iran[4]

Abstract

There are various ways to express the overall intention and focal points of a text, and keywords might be the most appropriate choice. Keywords are defined as the most prominent phrases in a document that convey its main message. By extracting relevant words and phrases from a text, keyword extraction can help to uncover meaningful patterns in the text and provide an overview of the content. It can also help highlight the most significant concepts in a text and focus the attention of a machine learning algorithm on them.

Keyword extraction is an imperative subtask of natural language processing. By reducing the complexity of the text and making it easier to process, keyword extraction can be used as the basis for

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۱/۶/۱۴ • تاریخ پذیرش: ۱۴۰۳/۹/۱۴ • تاریخ انتشار: ۱۴۰۳/۱۲/۳۰ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.21.4.113]

many other processing tasks such as text classification, clustering, and summarization. By extracting keywords from text, a machine can better understand the meaning and context of the text. This enables it to better analyze the text, recognize patterns, and make more accurate decisions. It can also reduce the amount of time it takes to process the text by eliminating unnecessary words and focusing on the most important words. Many datasets are proposed for evaluating keyword extraction methods in Persian, most of which only contain authors' keywords and do not cover all potential ones. Thus, using such datasets leads to incorrect judgments about the accuracy of the suggested supervised and unsupervised methods.

In this paper, we introduce Noor-Vajeh, a Persian keyword extraction dataset of about 1400 scientific papers. We asked experts to extract potential keywords besides the authors' keywords to complete the keywords set for each article. The resulting dataset is a valuable resource for ongoing research into Persian keyword extraction. To evaluate the dataset to be used as a benchmark, we tested several unsupervised keyword extraction methods. We used these methods because, compared to supervised methods, they take less time to execute and require minimal to no manual tuning. Moreover, they are able to extract keywords with a high degree of accuracy and generalize well to different articles. Furthermore, due to the wide variety of categories of unsupervised learning methods, graph-based methods have been regularly applied in different projects, so we describe and use some of their most famous ones, such as TextRank, SingleRank, and PositionRank. These algorithms can identify important words and phrases in a given text, as well as identify relationships between them. Doing so can provide insights into the overall structure and meaning of the text. This makes them especially useful for finding patterns and making predictions in a variety of tasks, such as machine translation, text summarization, and sentiment analysis. Furthermore, graph-based methods are highly versatile and can be adapted to different datasets and tasks. This makes them ideal for use in benchmark datasets. The results inferred from these methods confirm the comparisons made between the methods employed in other papers.

Keywords: Keyword Extraction, Persian Dataset, Unsupervised Learning, Graph-Based Methods, Information Retrieval.

۱-مقدمه

با افزایش روزافزون حجم اطلاعات غیرساخت‌یافته به ویژه در مقالات و نشریات علمی، نیاز به راهی برای توصیف سادهٔ محتوای این مستندات است. کلیدواژه‌ها عباراتی از متن هستند که این کار را انجام می‌دهند؛ علاوه‌براین می‌توانند به‌عنوان سرنخی جهت دستیابی موتورهای جست‌وجو به این اسناد باشند. استخراج عبارات کلیدی پایهٔ انجام بسیاری از عملیات‌های پردازش زبان طبیعی است، به گونه‌ای که بهبود عملکرد آن گاهی به بهبود عملیات‌های سطح بالا منجر می‌شود. گاهی اوقات برخی کلیدواژه‌ها به‌صراحت در متن وجود ندارند و نیاز به روش‌هایی برای تولید آن‌هاست.

روش‌های سنتی، فراوانی تکرار واژگان را محور استخراج کلیدواژه قرار می‌دهند، در این حالت برخی از کلیدواژه‌ها به‌دلیل تکرار کم از گزینه‌ها حذف می‌شوند. رایج‌ترین روش جایگزین، استفاده از روش‌های مبتنی بر گراف است که اهمیت یک عبارت را بر اساس ارتباط آن با سایر واژه‌های متن در نظر می‌گیرند.

با ظهور روش‌های یادگیری ماشین، اقسام متنوعی از ویژگی‌ها می‌توانند برای تشخیص کلیدواژه‌ها به‌کار گرفته شوند. این ویژگی‌ها در روش‌های بی‌ناظر به صورت دستی وزن‌دهی می‌شوند، اما در روش‌های باناظر این کار خودکار انجام می‌شود. تعیین کلیدواژه به روش باناظر به صورت یک طبقه‌بندی دودویی تعریف می‌شود که در آن به یک عبارت برچسبی نسبت داده می‌شود که نمایانگر بودن یا نبودن آن به‌عنوان کلیدواژه است.

بسیاری از روش‌ها از الگوهای زبانی برای استخراج کلیدواژه استفاده می‌کنند که روش‌ها را وابسته به زبان می‌سازد. در این صورت نیاز به دادگانی برای ارزیابی روش‌های استخراج کلیدواژه در آن زبان است. بر طبق اطلاعات ما تنها دادگان معیار موجود در فارسی، دادگان **Perkey** است که حاوی حدود ۴۵۰۰۰ متن خبری با ده کلیدواژه برای هر یک از آن‌هاست[۱]، اما مقالات و اسناد علمی یکی از منابعی هستند که جست‌وجو و دسترسی به آن‌ها به دفعات مورد نیاز است و به همین دلیل نیاز به دادگانی برای ارزیابی روش‌های استخراج کلیدواژه برای رسیدن به این مهم است. از طرفی مرسوم است که نویسندگان برای هر مقاله حداکثر پنج کلیدواژه تعیین می‌کنند که این کلیدواژه‌ها برای بازیابی اطلاعات مورد نیاز کافی نیست.

در این مقاله دادگان نورواژه معرفی می‌شود که متشکل از حدود ۱۴۰۰ مقالهٔ علمی است. برای هر مقاله ده کلیدواژه تولید شده‌است که پنج کلیدواژهٔ آن متعلق به خود نویسنده و پنج کلیدواژهٔ دیگر توسط افراد خبره تولید شده‌است. ما روش‌های متعدد بی‌ناظر را بر روی دادگان آزمایش کردیم تا نشان دهیم از این دادگان می‌توان

[DOI: 10.61186/jsdp.21.4.113]

به‌عنوان یک دادگان معیار برای استخراج کلیدواژه از مقالات علمی فارسی استفاده کرد.

در ادامهٔ مقاله، پژوهش‌های در زمینهٔ استخراج کلیدواژه و دادگان‌هایی که در این زمینه وجود دارند را مورد بررسی قرار می‌دهیم؛ سپس به معرفی روش‌های شاخص استخراج کلیدواژه و ارزیابی آن‌ها روی دادگان نوروژه می‌پردازیم، در مورد نتایج به‌دست‌آمده بحث می‌کنیم و نشان می‌دهیم نتایج به‌دست‌آمده مشابه نتایجی است که بر روی سایر دادگان‌های موجود حاصل شده‌است و در انتها جمع‌بندی کرده و پیشنهادهای بیشتر برای به‌کارگیری دادگان موردنظر را مطرح می‌کنیم.

۲- پژوهش‌ها و دادگان‌های موجود

در این بخش به معرفی کارهای شاخص در زمینهٔ استخراج کلیدواژه می‌پردازیم. جهت ارزیابی این پژوهش‌ها دادگان‌های متنوعی ارائه شده‌است که در ادامه آن‌ها را معرفی و ویژگی‌های آن‌ها را بررسی می‌کنیم.

۲-۱-کارهای پیشین

با نگاهی کلی به روش‌های استخراج کلیدواژه می‌توان آن‌ها را به دو دستهٔ بی‌ناظر و باناظر تقسیم‌بندی کرد که دسته‌بندی دقیق آن‌ها در شکل‌های (۱ و ۲) مشاهده می‌شود. روش‌های پایه از ویژگی‌های آماری نظیر فراوانی تکرار واژه برای استخراج کلیدواژه استفاده می‌کردند؛ برای مثال در مقالهٔ [۲] واژه‌های پرتکرار بر اساس میزان واگرایی و اطلاعات متقابل[1] بین هر جفت واژه خوشه‌بندی می‌شوند؛ سپس مربع کای برای هر واژه محاسبه می‌شود تا میزان هم‌رخدادی[2] آن با خوشهٔ مورد نظر مشخص شود و به این ترتیب واژه‌های با امتیاز بالاتر به‌عنوان واژه‌های کلیدی انتخاب می‌شوند. با آزمایش روی حدود بیست مقالهٔ علمی مشخص شد این روش امتیاز بالاتری نسبت به روش‌های فراوانی تکرار ساده و گراف کلیدی[3] دارد، اما نسبت به روش **TFIDF**[4] امتیاز پایین‌تری داشت.

روش [۳] برای بهبود نتایج از یک پیکرهٔ وابسته به زبان برای امتیازدهی به عبارات استفاده می‌کند؛ به این صورت که امتیاز هر عبارت به صورت نسبت فراوانی آن در متن به فراوانی آن در پیکرهٔ مورد نظر محاسبه می‌شود. هر قدر این نسبت کوچک‌تر باشد، شانس آن عبارت برای کلیدی‌بودن بیشتر خواهد بود. بر حسب سنجهٔ اف[5]، این روش نتایج بهتری از روش **TFIDF** بر روی دادگان **SemEval-2010** داشت.

بیشتر روش‌های بدون ناظر، از روش‌های مبتنی بر گراف برای استخراج کلیدواژه استفاده می‌کنند؛ برای مثال مقالهٔ [۴] ابتدا متن ورودی را به گرافی از واژه‌ها تبدیل می‌کند؛ سپس از مجموعه‌ای از ویژگی‌ها مانند انحطاط گراف، چگالی زیرگراف‌ها و انعطاف آن‌ها، معیار شاخص **TextRank** و انسجام هستهٔ مرکزی برای استخراج کلیدواژه‌ها بهره می‌گیرد؛ همچنین در مقالهٔ [۵] سامانه‌ای برای استخراج کلیدواژه از مکالمات برخط ارائه شده‌است که در آن ابتدا یک گراف هم‌رخدادی از واژه‌ها تولید می‌شود و سپس با استفاده از الگوریتم تجزیهٔ k بخشی به زیرگراف‌هایی تقسیم می‌شود. مقالهٔ دیگری [۶] متن ورودی را به گرافی از واژه‌ها تبدیل می‌کند و از شاخص‌های مرکزیت گراف و حذف واژه‌های مشابه زاید برای تشخیص ابرگره‌ها و درنهایت کلیدواژه‌ها استفاده می‌کند. این روش منجر به نتایج بهتری نسبت به روش‌های مبتنی بر گراف مشابه شده‌است.

مقالهٔ [۷] ابتدا با استفاده از مجموعه‌ای از قوانین اکتشافی نحوی، عبارات کلیدی نامزد را می‌یابد، سپس آن‌ها را در قالب بردارهای تعبیهٔ جملات[6] نمایش داده و رتبه‌بندی می‌کند. این روش بر روی دادگان‌های معیار موجود بهتر از روش‌های مبتنی بر گراف عمل کرده‌است.

یکی از روش‌های شاخص اخیر، روش **Yake** است که به‌دلیل نتایج خوب با وجود مستقل‌بودن از زبان و موضوع، معروف است[۸]. در این روش پس از پیش‌پردازش متن و تعیین عبارات نامزد کلیدی، از ویژگی‌هایی مانند موردبندی[7]، مکان عبارت در متن، فراوانی عادی‌شدهٔ عبارت، ارتباط عبارت با بافت متن و وجود آن در جملات مختلف برای امتیازدهی به عبارات استفاده می‌شود.

مقالهٔ [۹] برای استخراج کلیدواژه‌ها از وب‌نوشت‌های فارسی، تمامی ایست‌واژه‌ها، شکلک‌ها، پیوندها و فواصل زاید را از متن حذف می‌کند؛ سپس تمامی ترکیبات اسمی و وصفی را با برچسب‌گذار ادات سخن[8] استخراج کرده و به‌عنوان عبارات کلیدی نامزد در نظر می‌گیرد. در ادامه گرافی از کلیدواژه‌ها ساخته می‌شود که گره‌ها نمایانگر واژه‌ها و یال‌ها نمایانگر هم‌رخدادی بین آن‌هاست. بعد از رتبه‌بندی واژه‌ها، از روش پنجرهٔ لغزان برای تشخیص یال‌های نهایی استفاده می‌شود. این گراف به الگوریتم **TextRank** داده می‌شود تا کلیدواژه‌های نهایی را مشخص کند. این روش بر روی چهار دادگاه آزمون شد و

[1] Mututal Information
[2] Co-occurence
[3] Key-graph
[4] Term Frequency - Inverse Document Frequency
[5] F-measure
[6] Sentence Word Embedding
[7] Casing
[8] Part-of-Speech Tagger

[DOI: 10.61186/jsdp.21.4.113]

نتایج آن از روش‌های **TFIDF**، **Gensim** و تخصیص پنهان دیریکله بهتر بود.

مقالهٔ [۱۰] با استفاده از حالت بهبودیافتهٔ روش **RAKE** به استخراج عبارات کلیدی از رساله‌های علمی فارسی می‌پردازد. در حالت پایه پس از پردازش‌های لازم نظیر حذف ایست‌واژه‌ها و فواصل و علائم زاید، ماتریسی از هم‌رخدادی واژه‌های باقی‌مانده تشکیل می‌شود. این ماتریس که بر اساس پیکره‌ای از کلیدواژه‌ها به‌دست می‌آید، نشان می‌دهد فراوانی هم‌رخدادی یک واژه با واژهٔ دیگر در پیکرهٔ مورد نظر چقدر است و بر اساس این ماتریس، درجهٔ هر واژه محاسبه می‌شود. این معیار به همراه تکرار واژه‌ها در پیکرهٔ مورد نظر برای امتیازدهی نهایی به کلیدواژه‌های حاصل از آن‌ها استفاده می‌شود. در حالت بهبودیافته، درجهٔ عبارات نسبت به طولشان عادی‌سازی می‌شوند. این روش نسبت به الگوریتم پایه **RAKE**، **TFIDF**، **TextRank** و **Yake** سنجهٔ اف بالاتری حاصل کرده‌است.

مقالهٔ [۱۱] با استفاده از روش باناظر و مستقل از زبان و موضوع به استخراج کلیدواژه‌ها می‌پردازد. ابتدا متن به صورت شبکه‌ای از گره‌ها حاوی واژه‌ها تبدیل می‌شود؛ سپس از یک طبقه‌بند دودویی برای آموزش انتخاب گره‌ها استفاده می‌شود که در آن ویژگی هر گره از روش‌های بی‌ناظر و مبتنی بر گراف جمع‌آوری شده و برچسب هر گره بر حسب اینکه در کلیدواژه‌های دادگان آموزش وجود دارد یا خیر، تعیین می‌شود. برای طبقه‌بندی از دسته‌بندهای بیز ساده (رویهٔ گردایه‌سازی[1] و تقویت تطبیقی[2]) و تقویت **XG**[3] استفاده شده‌است. بهترین نتیجه بر مبنای سنجهٔ اف بر روی شش دادگان معیار توسط روش بیز ساده با تقویت **XG** حاصل شده‌است.

در مقالهٔ [۱۲] پس از مراحل اولیهٔ پیش‌پردازش روی متن و حذف ایست‌واژها و علائم زاید از متن، ویژگی‌هایی مانند **TFIDF**، درجهٔ ترکیب، سرواژه، آشفتگی و تبدیلات برای هر یک از واژگان محاسبه می‌شود و به یک الگوریتم یادگیری ماشین بردار پشتیبان داده می‌شود تا بتواند کلیدواژه‌ها را شناسایی و برچسب‌زنی کند.

مقالهٔ [۱۳] ابتدا پیش‌پردازش‌های اولیه، تقطیع واژگان و برچسب‌گذاری دستوری آن‌ها و ریشه‌یابی را بر روی واژه‌ها انجام می‌دهد و عبارات اسمی را استخراج می‌کند؛ سپس برای هر یک از این عبارات با استفاده از هستان‌شناسی فارس‌نت، زنجیرهٔ لغوی مرتبط ایجاد می‌شود. پس از آن، واژگان حاصل‌شده به دسته‌بندهای مبتنی بر یادگیری ماشین داده می‌شود تا کلیدی‌بودن یا نبودن آن‌ها مشخص شود.

مقالهٔ [۹] به مقایسهٔ روش‌های آماری در زمینهٔ استخراج واژه‌های کلیدی از متون فارسی و انگلیسی می‌پردازد. بهترین نتیجه به‌وسیلهٔ روش «نسبت واریانس واژه به بسامد لغوی آن» بر روی متن فارسی دادگان معرفی‌شده در این مقاله به‌دست می‌آید که واریانس واژه به تعداد تمام اسناد شامل واژهٔ مدّنظر، میانگین بسامد واژه در همهٔ اسناد و تعداد رخداد واژه در سند وابسته است.

با ظهور یادگیری عمیق، برخی مقالات از این فناوری برای استخراج کلیدواژه استفاده کرده‌اند؛ برای مثال مقالهٔ [۱۵] در مرحلهٔ آموزش از یک شبکهٔ رمزگذار-رمزگشا مبتنی بر شبکه‌های بازگشتی عمیق[4] یاد می‌گیرد که عنوان و متن اخبار ورودی را چگونه به عبارات کلیدی تبدیل کند. در این فرایند با استفاده از سازوکار توجه[5] به هر یک از واژگان ورودی وزنی تخصیص داده می‌شود. این روش نسبت به سایر روش‌های باناظر و بی‌ناظر مشابه خود نتیجهٔ مناسب‌تری به‌دست آورده‌است.

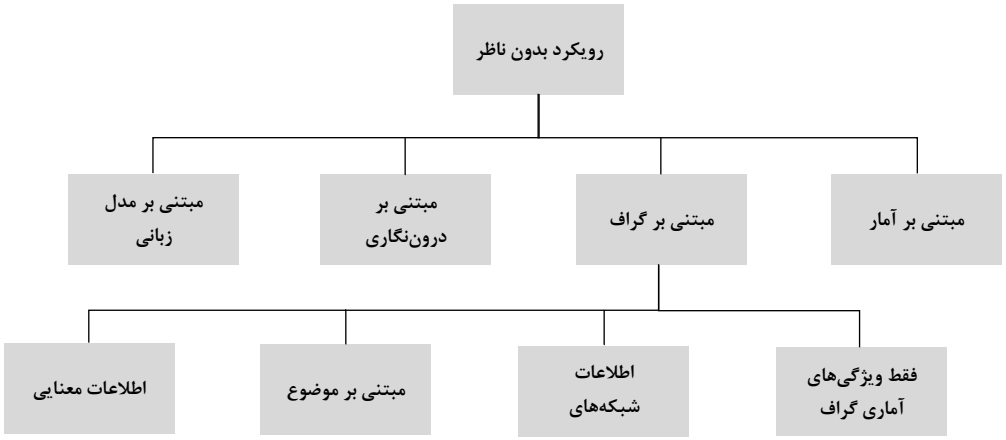
مقاله با رویکرد بدون ناظر و بهره‌گیری از روش بردار تعبیه **Word2Vec** به استخراج کلیدواژه‌ها پرداخته‌است. به این ترتیب که برای هر واژه، بردار تعبیه سیصد بعدی آن به‌دست می‌آید و فاصلهٔ برداری هر واژه با سایر واژه‌ها و احتمال رخداد آن واژه در متن محاسبه می‌شود و درنهایت واژه‌های با امتیاز کمتر به‌عنوان کلیدواژه انتخاب می‌شوند. این روش بر حسب سنجهٔ اف، امتیاز بالاتری نسبت به روش‌های **TFIDF**، **TextRank** و **YAKE** داشته‌است.

در مقالهٔ [۱۶] به مقایسهٔ روش‌های بیز ساده، رگرسیون ترابری[6]، ماشین‌های بردار پشتیبان، شبکه‌های عصبی کانولوشنی[7] و شبکه‌های عصبی طویل با حافظهٔ کوتاه مبتنی بر توجه[8] با رویکرد با ناظر در استخراج واژه‌های کلیدی پرداخته شده‌است. بهترین نتیجه با سنجهٔ اف، به‌وسیلهٔ روش شبکه‌های عصبی طویل با حافظهٔ کوتاه مبتنی بر توجه بر روی دادگان معرفی‌شده در این مقاله به‌دست می‌آید.

[1] Bagging
[2] Ada Boost
[3] XG Boost

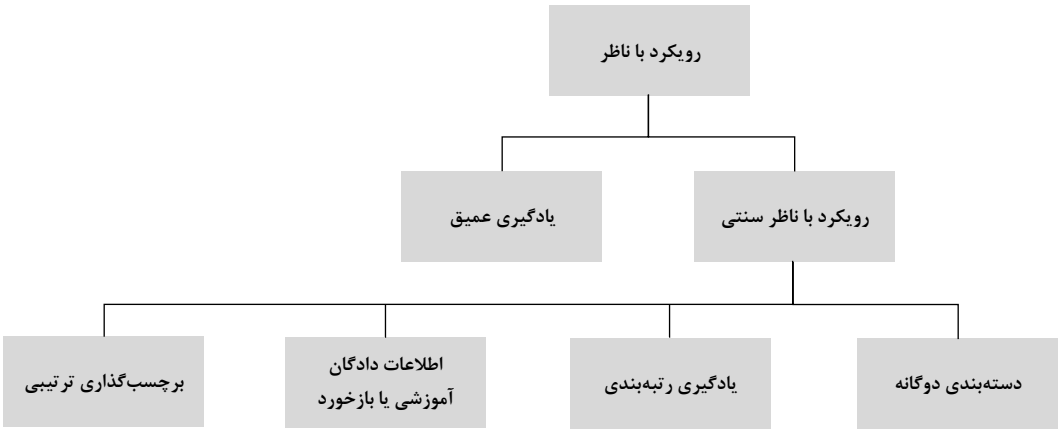
[4] Recurrent Neural Network
[5] Attention Mechanism
[6] Logistic Regression
[7] Convolutional Neural Networks
[8] Long Short-Term Memory Neural Networks with Attention

[DOI: 10.61186/jsdp.21.4.113]



(شکل-۱): اقسام رویکرد بدون‌ناظر استخراج کلیدواژه

(Figure-1): Types of unsupervised approach to keyword extraction



(شکل-۲): اقسام رویکرد باناظر استخراج کلیدواژه

(Figure-2): Types of supervised approach to keyword extraction

مقالهٔ [۱۷] از یک شبکهٔ **LSTM** دوجهته[1] برای استخراج کلیدواژه از نظرات برخط کاربران در مورد محصولات استفاده می‌کند. ابتدا از یک شبکهٔ **LSTM** برای پالایهٔ جملات غیر مرتبط استفاده می‌شود؛ سپس عبارات دو، سه و چهار گرمی از جملات مرتبط انتخاب می‌شوند و پرتکرارترین آن‌ها به‌عنوان کلیدواژه‌های نامزد انتخاب می‌شوند. دوباره یک شبکهٔ **LSTM** برای برچسب‌زنی جملات مرتبط آموزش می‌بینند که در اینجا هر یک از این کلیدواژه‌ها یک برچسب محسوب می‌شوند.

برچسب‌هایی که تعداد بیشتری جمله به آن‌ها منتسب می‌شوند به‌عنوان کلیدواژه‌های نهایی انتخاب خواهند شد. دقت این روش بر روی دادگان معرفی‌شده، برای هر دو بخش پالایه و استخراج واژگان کلیدی، از روش‌های مشابه مناسب‌تر است.

مقالهٔ [۱۸] با استفاده از مدل تعبیهٔ بافتاری **BERT**[2] به استخراج واژگان کلیدی می‌پردازد. پس از پردازش اولیهٔ متون، دنبالهٔ واژه‌ها به مدل برت داده می‌شود تا بردار تعبیهٔ آن‌ها تولید شود. در لایهٔ بعدی از مکانیزم توجه محلی یا سراسری برای تعیین اهمیت واژه‌ها استفاده می‌شود و در ادامه به یک لایهٔ رده‌بند داده می‌شوند تا در مورد کلیدواژه‌بودن یا نبودن آن‌ها تصمیم‌گیری کند. دقت این روش با ریزتنظیم[3] مدل برت با واژه‌های متن، از تمامی روش‌های یادشده بهتر بوده‌است.

۲-۲-دادگان‌های موجود

بر اساس موضوع، نوع کلیدواژه‌های استخراج‌شده و حجم آن‌ها دادگان‌های متنوعی ارائه شده‌است که هر یک مناسب کاربردی خاص‌اند. در ادامه به معرفی آن‌ها می‌پردازیم.

۲-۲-۱- دادگان SemEval-2010

توسط *کیم و همکاران* [۱۹] ارائه شده‌است و شامل ۲۸۴ مقالهٔ علمی به زبان انگلیسی، جمع‌آوری‌شده از کتابخانهٔ دیجیتال **ACM** است. ورودی نمونه‌ها، کل متن مقاله است و متوسط طول ورودی حدود ۵۲۰۰ واژه است و متوسط تعداد کلیدواژه‌ها دوازده عدد است که متشکل از کلیدواژه‌های نویسندگان و

[1] Bi-Directional Long-Short-Term-Memory
[2] Bidirectional Encoder Representation of Transformers
[3] Fine-Tune

[DOI: 10.61186/jsdp.21.4.113]

خوانندگان مقاله است. به طور معمول حدود صد سند از آن توسط روش‌ها مورد آزمون قرار می‌گیرد. از ویژگی‌های متمایزکنندهٔ این دادگان، کم‌بودن نرخ تطبیق کلیدواژه‌ها با متن مقاله است که این موضوع البته سبب تنوع بالای کلیدواژه‌های مورد ارزیابی است.

۲-۲-۲-دادگان Inspec

توسط *هلث و آنت*[۲۰] ارائه شده‌است و حاوی دوهزار چکیدهٔ مقاله علمی به انگلیسی در حوزه‌های کامپیوتر، کنترل و فناوری اطلاعات است. متوسط تعداد کلیدواژه‌های این دادگان ۲۱ عدد است که در میان دیگر دادگان‌ها بیشترین مقدار است و توسط خبرگان انتخاب و جمع‌آوری شده‌است. متوسط طول ورودی ۱۳۶ واژه و متوسط طول کلیدواژه‌ها ده نویسه است. این دادگان مناسب ارزیابی استخراج کلیدواژه به‌صورت خاص‌منظوره در حوزهٔ کامپیوتر است، اما در پژوهش‌های بیشتری نسبت به سایر پژوهش‌ها مورد استفاده قرار گرفته‌است.

۲-۲-۳-دادگان DUC-2001

توسط *وان و همکاران* [۲۱] جمع‌آوری و ارائه شده‌است، حاوی ۳۰۸ متن خبری به زبان انگلیسی است که توسط خوانندگان خبر کلیدواژه‌های آن استخراج شده‌است. هر سند به‌طور متوسط دارای نهصد واژه و به‌طور متوسط حاوی هشت کلیدواژه است. این دادگان در میان دادگان‌های خبری بیشترین استفاده را در کارهای پژوهشی داشته‌است.

۲-۲-۴-دادگان NUS

این دادگان که توسط *انگوین و همکاران* [۲۲] ارائه شده‌است شامل حدود ۲۱۱ مقالهٔ کنفرانس به طول چهار تا دوازده صفحه است. کلیدواژه‌های این دادگان در دو مجموعه کلیدواژهٔ نویسندگان و خبرگان گردآوری شده‌است و در مجموع نسبت به مابقی دادگان‌های موجود، تعداد کلیدواژه‌های بیشتری دارد و به طور طبیعی رسیدن به دقت بالا در چنین دادگان دارای چالش بیشتری خواهد بود.

۲-۲-۵-دادگان Perkey

توسط دوست‌محمدی و همکاران[۱] تهیه و جمع‌آوری شده‌است، حاوی حدود پانصدهزار متن خبری به همراه کلیدواژه است، اما از این تعداد حدود ۵۵ هزار متن حاوی بیش از نه کلیدواژه است. این دادگان از حدود شش خبرگزاری معروف جمع‌آوری شده‌است و کمابیش حدود ۶۵ درصد از کلیدواژه‌ها در متن خبر حضور دارند. حدود ۴۶ درصد از واژه‌های کلیدی، تک واژه‌ای و مابقی بیشتر از یک واژه‌اند و به نظر می‌رسد این دادگان مناسب زمانی است که واژه‌های کلیدی به صورت تک‌واژه‌ای تولید شوند؛ همچنین، دادگان‌های مورد استفاده در پژوهش‌های سال‌های اخیر مانند [۲۳]و [۲۷] دچار چالش‌هایی نظیر موارد زیر هستند:

- بسیاری از آن‌ها محدود به دامنه‌ای خاص مثل مقالات خبری یا موضوعات تخصصی علمی هستند.
- تنوع در ویژگی‌های متن مورد استفاده جهت استخراج کلیدواژه نظیر طول، پیچیدگی و لحن نوشتار به چشم نمی‌خورد.
- نرخ تطابق کلیدواژه‌ها با متن متناظر در مواردی بسیار کم است.
- تمام کلیدواژه‌های شاخص موجود در متن استخراج نشده‌اند.
- کلیدواژه‌ها بیشتر به‌صورت لغوی از متن استخراج شده‌اند و بیشتر توجّهی به کلیدواژه‌هایی که به صورت معنایی از متن قابل استخراج هستند، نشده‌است.

نوآوری دادگان نورواژه، بهبود این موارد با تکیه بر موضوعات متنوّع متون و استفاده از کلیدواژه‌های استخراج‌شده توسّط خبرگان به‌منظور افزایش نرخ تطابق با متن و همچنین اشاره به تمامی کلیدواژه‌های شاخص اعم از معنایی و لغوی است.

۳-معرفی دادگان نورواژه

همان‌طور که پیش‌تر اشاره شد، کلیدواژه‌های نویسندگان مقالات به‌تنهایی برای ارزیابی روش‌های استخراج کلیدواژه کافی نیست و لازم است مابقی کلیدواژه‌های بالقوه نیز استخراج شود. به همین دلیل ما دادگان نورواژه را ارائه کرده‌ایم که برای هر مقاله حدود پنج کلیدواژه متعلق به نویسنده و حدود هفت کلیدواژهٔ استخراج‌شده توسط خبرگان وجود دارد. وجود کلیدواژه‌های استخراج‌شده توسط خبرگان، وجه تمایز دادگان نورواژه با دادگان‌هایی است که تنها شامل کلیدواژه‌های نویسندگان بوده و از طریق خزش[1] سایت‌های علمی قابل جمع‌آوری‌اند.

این دادگان متشکل از چکیدهٔ ۱۳۶۴ مقاله به همراه کلیدواژه‌های آن‌هاست که با معیار رعایت تنوّع و گوناگونی در موضوعات، از زمینه‌هایی نظیر علوم انسانی، اسلامی، مهندسی، پزشکی، تربیت بدنی، اقتصاد، تاریخ، ادبیات، مدیریت، حقوق و... به‌وسیلهٔ مؤسسهٔ تحقیقات کامپیوتری علوم اسلامی نور[2] گردآوری شده و از طریق این پیوند[3] قابل بارگیری است. فرایند استخراج کلیدواژه توسط خبرگان به این صورت بوده‌است که در ابتدا مقالات به‌صورت تصادفی بین سه خبره توزیع و پس از استخراج، کلیدواژه‌ها توسّط دو فرد خبرهٔ دیگر تأیید شده‌اند. جدول (۱) ویژگی‌های آماری این دادگان را نمایش می‌دهد.

[1] Crawling
[2] https://noorsoft.org
[3] https://ai.inoor.ir/dataset/NoorManualKeywords.rar

[DOI: 10.61186/jsdp.21.4.113]

(جدول-۱): ویژگی‌های آماری دادگان نوروژه
(Table-1): Statistics of the Noorvajeh Dataset

شماره	آماره (واحد)	مقدار
۱	متوسط طول متن هر چکیده (واژه)	۱۷۲
۲	متوسط تعداد عبارات کلیدی نویسندگان (عبارت)	۴/۷
۳	متوسط تعداد عبارات کلیدی انسانی (عبارت)	۶/۶
۴	متوسط طول هر عبارت کلیدی نویسندگان (واژه)	۲/۶
۵	متوسط طول هر عبارت کلیدی انسانی (واژه)	۲/۸
۶	متوسط نرخ تطابق عبارات کلیدی نویسندگان با متن چکیده (٪)	۲۶/۴
۷	متوسط نرخ تطابق همهٔ عبارات کلیدی با متن چکیده (٪)	۵۶/۶

گفتنی است که در موارد شش و هفت، برای محاسبهٔ تطبیق عبارات کلیدی با چکیده، با استفاده از ابزار پردازش فارسی پارسی‌ور [۲۸] مراحلی از قبیل عادی‌سازی متن، ریشه‌یابی واژه‌ها انجام شده‌است؛ علاوه‌براین اگر عبارات کلیدی نویسندگان و خبرگان، عیناً در متون نیامده باشد و عبارات یا واژه‌هایی مترادف با آن‌ها در متن ظاهر شده باشد، با استفاده از فهرست واژگان مترادف که در [۲۹] ارائه شده‌است آن دسته از واژگانی که مترادف آن‌ها در متون ظاهر شده‌است با واژهٔ اصلی جای‌گذاری می‌شوند. همان طور که در مورد شش اشاره شده، میانگین، تنها ۲۶.۴٪ از کلیدواژه‌های نویسندگان در متن چکیده ظاهر شده‌است. در حالی که اگر کلیدواژه‌های استخراج‌شده توسط خبرگان را با این مجموعه اجتماع بگیریم (مطابق مورد هفت)، این عدد به ۵۶.۶٪ می‌رسد؛ این امر، حاکی از تأثیر کلیدواژه‌های استخراج‌شده توسط خبرگان در افزایش نرخ تطابق با واژگان متون چکیده است.

۴- روش‌های ارزیابی

برای تأیید قابلیت استفاده از نوروژه به‌عنوان یک دادگان معیار نیاز است روش‌های شاخص استخراج کلیدواژه روی این دادگان مورد آزمون و ارزیابی قرار بگیرند. به همین دلیل مجموعه‌ای از روش‌های بی‌ناظر انتخاب شد و از آنجا که مطابق شکل (۱)، قسم مبتنی بر گراف بیشترین استفاده را داشته‌است، ما نیز بیشتر از روش‌های مبتنی بر گراف برای ارزیابی استفاده کردیم. در ادامه به معرفی روش‌های ارزیابی می‌پردازیم.

۴-۱-رتبه‌بندی متن[1]

این روش، در اصل از روش رتبه‌بندی صفحات[2] که مربوط به صفحات وب است، برگرفته شده‌است؛ ابتدا جملات متن از یکدیگر جدا شده و نقش‌های دستوری واژگان هر یک، برچسب‌گذاری می‌شوند تا فقط واژه‌هایی از این مرحله عبور کنند که نقش‌های دستوری خاصی دارند (رویهٔ برچسب‌گذاری نقش‌های دستوری تنها به منظور پرداختن به واژه‌هایی است که احتمال می‌رود اهمیت بیشتری داشته باشند و شانس بیشتری برای انتخاب‌شدن به‌عنوان کلیدواژه دارند. به همین سبب، می‌توان این مرحله را انجام نداد و به بررسی تمامی واژگان جمله پرداخت). هر واژهٔ باقی‌مانده در این مرحله، به‌عنوان یکی از گره‌های گراف درنظر گرفته می‌شود؛ سپس، با اندازهٔ پنجرهٔ مشخصی از نخستین واژه تا آخرین واژه پیمایش شده و بین هر جفت واژه‌هایی که در یک پنجره قرار می‌گیرند، یک یال بدون جهت در گراف منظور می‌شود (هر یال بدون جهت، هم به‌عنوان یال ورودی و هم به‌عنوان یال خروجی در فرمول محاسبهٔ وزن گره‌ها درنظر گرفته می‌شود)؛ برای مثال اگر واژگان استخراج‌شده از مرحلهٔ قبل، {رایانه، پردازش، درون‌نگاری، هم‌رخدادی، زبان، گفتار} باشد و اندازهٔ پنجره را سه درنظر بگیریم، با پیمایش از ابتدا تا انتهای فهرست چهار پنجره خواهیم داشت:

۱) {رایانه، پردازش، درون‌نگاری}
۲) {پردازش، درون‌نگاری، هم‌رخدادی}
۳) {درون‌نگاری، هم‌رخدادی، زبان}
۴) {هم‌رخدادی، زبان، گفتار}

در مثال بالا برای پنجرهٔ اوّل، یک یال بدون جهت بین هر جفت گره {رایانه، پردازش}، {رایانه، درون‌نگاری} و {پردازش، درون‌نگاری} لحاظ می‌شود.

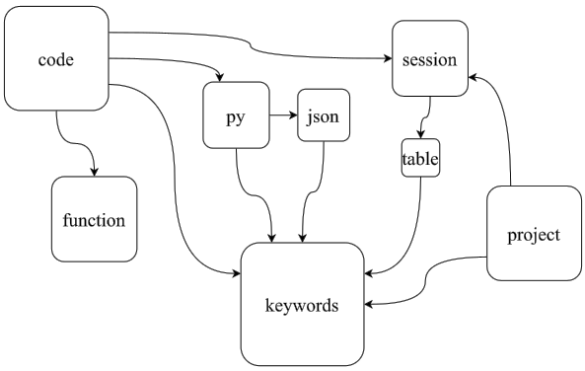
در مرحلهٔ بعد، برای هر گره مقدار وزن محاسبه می‌شود (این مقدار با فرمول‌های مختلفی محاسبه می‌شود که اشاره به آن‌ها در این مجال نمی‌گنجد) و گره‌هایی که وزنی بیشتر از یک حد آستانهٔ مشخص داشته باشند و یا تعداد از پیش تعیین‌شده‌ای از نامزدها با بیشترین وزن‌ها، به‌عنوان واژگان کلیدی شناخته می‌شوند[۳۰]. نمایی مفهومی از فرایند تشکیل گراف و امتیازدهی به گره‌ها در شکل (۳) مشاهده می‌شود. گره‌های با اندازهٔ بزرگ‌تر، امتیازات بیشتری دریافت کرده‌اند[۲۶].

۴-۲-رتبه‌بندی تکی[3]

در این روش، ابتدا تعدادی از اسناد مرتبط و مشابه با سند اصلی از طریق پیکرهٔ مربوطه و یا از سطح اینترنت جمع‌آوری می‌شود. مجموعهٔ سند اصلی و اسناد مرتبط به‌دست‌آمده، مجموعهٔ بسط‌یافتهٔ اسناد[4] را تشکیل می‌دهند (در این پژوهش، از مشابهت کسینوسی برای یافتن پنج مقالهٔ مشابه در دادگان به‌ازای هر مقاله استفاده شده‌است).

[1] TextRank
[2] PageRank
[3] SingleRank
[4] Expanded Document Set

[DOI: 10.61186/jsdp.21.4.113]



(شکل-۳): نمایی از گراف ایجادشده با الگوریتمTextRank
(Figure-3): A view of the graph created by the TextRank algorithm

در مرحلهٔ بعد، واژگان نامزد از طریق پالایه‌کردن نحوی واژه‌های موجود در مجموعهٔ اسناد، انتخاب شده و به‌عنوان گره‌های گراف درنظر گرفته می‌شوند. دو گره، تنها در صورتی به یکدیگر متصل می‌شوند که حداقل یک بار در یک پنجره قرار بگیرند؛ هر یال، یک وزن وابستگی[1] به خود می‌گیرد که بر اساس رابطهٔ هم‌رخدادی دو کلمهٔ متناظر به‌دست می‌آید و با فاصلهٔ بین آن دو واژه کنترل می‌شود. رابطهٔ هم‌رخدادی، در حقیقت انسجام بین واژگان را بیان می‌کند. گراف حاصل، یک گراف بدون جهت و بر اساس کل مجموعهٔ اسناد است و گراف وابستگی سراسری[2] نام دارد. بر اساس این گراف، امتیاز برتری[3] هر واژه از امتیاز واژگان دیگر به صورت بازگشتی محاسبه می‌شود (مقدار اولیهٔ امتیاز برتری تمام واژه‌ها، برابر یک در نظر گرفته می‌شود).

پس از مشخص‌شدن امتیاز برتری واژه‌ها، موقعیت واژگانی که در سند اصلی قرار دارند، مشخص می‌شود؛ اگر این واژه‌ها، تشکیل یک عبارت دهند که به نقش دستوری اسم ختم شود (البته در زبان فارسی می‌بایست با اسم آغاز شود) آن عبارت، مجاز درنظر گرفته شده و امتیاز آن، برابر با حاصل جمع امتیازهای برتری واژه‌های تشکیل‌دهنده‌اش خواهد بود؛ سپس، آن دسته از عباراتی که امتیازی بیشتر از یک حد آستانهٔ مشخص داشته باشند و یا تعداد از پیش تعیین‌شده‌ای از نامزدها با بیشترین امتیازات، به‌عنوان عبارات کلیدی انتخاب می‌شوند[۲۱].

۳-۴-رتبه‌بندی موقعیت[4]

در این روش، ابتدا بر روی متن ورودی، فرایند برچسب‌گذاری نقش‌های دستوری انجام می‌شود تا نقش‌های اسم و صفت در متن مشخص شوند. این نقش‌ها گره‌های گراف را تشکیل می‌دهند. دو گره در صورتی به یکدیگر یال دارند که حداقل در یک پنجره، هم‌رخداد شده باشند. وزن یک یال نیز بر اساس تعداد هم‌رخدادی واژه‌های متناظر با گره‌های ابتدا و انتهای آن در پنجره محاسبه می‌شود.

در ادامه، امتیاز هر گره با روش قدم‌زدن تصادفی[5] یا رتبه‌بندی صفحات [۳۱] محاسبه می‌شود و برای جلوگیری از گیرکردن این الگوریتم در حلقه‌های گراف، یک عامل تعدیل[6] به محاسبات اضافه می‌شود تا عملیات را به گره دیگری منتقل کند. در حقیقت، ایدهٔ اصلی این روش، لحاظ‌کردن وزن‌های بیشتر برای واژگانی است که زودتر در متن ظاهر شده‌اند و همچنین، به تعداد مناسبی تکرار شده‌اند. همین ایده سبب می‌شود تا روال وزن‌دهی با موقعیت ظاهرشدن واژه در متن نسبت عکس داشته باشد؛ اگر یک واژه، بارها در متن ظاهر شده باشد، این وزن‌ها با یکدیگر جمع می‌شوند و سپس رویهٔ عادی‌سازی انجام می‌شود. در آخرین گام این مرحله، امتیازات از طریق بازگشتی برای هر گره محاسبه می‌شوند.

در مرحلهٔ نهایی، ابتدا جایگاه واژگان در متن مشخص می‌شود و اگر تشکیل یک عبارت متشکل از نقش‌های دستوری اسم و صفت به صورت (صفت)*(اسم)+ بدهند، امتیازات تمام اجزاء با یکدیگر جمع می‌شود؛ سپس آن دسته از عباراتی که امتیازی بیشتر از یک حد آستانهٔ مشخص داشته باشند و یا تعداد از پیش تعیین‌شده‌ای از نامزدها با بیشترین امتیازات، به‌عنوان عبارات کلیدی درنظر گرفته می‌شوند [۳۲].

۵- آزمایش‌ها و بحث روی نتایج

روش‌های یادشده، در اصل برای زبان انگلیسی در [۲۹] معرفی شده‌اند و برای بهره‌برداری در زبان فارسی تغییر یافته‌اند و در [۳۰] ارائه شده‌اند. این روش‌ها برای عمل‌کردن بر روی زبان مورد نظر، به برچسب‌زن نقش‌های دستوری مربوط به همان زبان نیاز دارند؛ به همین منظور، در این پژوهش از یک برچسب‌زن نقش‌های دستوری زبان فارسی [۳۰] برای تکمیل عملکرد روش‌های پیش‌گفته استفاده شده‌است.

(جدول-۲): میانگین دقت متوسط روش‌ها
(Table-2): Average of Mean Accuracy of the Methods

روش	میانگین دقت متوسط (٪)
رتبه‌بندی موقعیت	**۲۱/۱۲**
رتبه‌بندی متن	۱۵/۶۱
رتبه‌بندی تکی	۹/۸۶

[1] Affinity Weight
[2] Global Affinity Graph
[3] Saliency Score
[4] PositionRank
[5] Random Walk
[6] Damping Factor

[DOI: 10.61186/jsdp.21.4.113]

(جدول-۳): دقت، بازخوانی و امتیاز F1 با میانگین ریزدانه
(Table-3): Micro Averages of Precision, Recall, and F1 Score

روش	@1			@3			@5			@10		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
رتبه‌بندی موقعیت	۸/۴۴	۰/۹	۱/۶	**۱۰/۶۳**	۳/۳۳	۵/۰۷	۱۰/۲۸	۵/۳۷	۷/۰۶	۹/۲۵	**۹/۶۷**	**۹/۴۵**
رتبه‌بندی متن	۳/۱۵	۰/۳۳	۰/۶	۶/۶۱	۲/۰۷	۳/۱۶	۷/۷۴	۴/۰۴	۵/۳۱	۸/۱	۸/۴۶	۸/۲۸
رتبه‌بندی تکی	۰/۹۵	۰/۱	۰/۱۸	۲/۶۴	۰/۸۳	۱/۲۶	۴/۱۸	۲/۱۸	۲/۸۶	۶/۱۳	۶/۴۱	۶/۲۶

(جدول-۴): دقت، بازخوانی و امتیاز F1 با میانگین درشت‌دانه
(Table-4): Macro Averages of Precision, Recall, and F1 Score

روش	@1			@3			@5			@10		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
رتبه‌بندی موقعیت	۸/۴۴	۰/۹۳	۱/۶۷	**۱۰/۶۳**	۳/۵۳	۵/۲۴	۱۰/۲۸	۵/۶۹	۷/۲۴	۹/۲۵	**۱۰/۰۷**	**۹/۵۲**
رتبه‌بندی متن	۳/۱۶	۰/۳۳	۰/۶	۶/۶۱	۲/۲۲	۳/۲۹	۷/۷۴	۴/۲۹	۵/۴۵	۸/۱	۸/۸۵	۸/۳۵
رتبه‌بندی تکی	۰/۹۵	۰/۱۱	۰/۱۹	۲/۶۴	۰/۸۵	۱/۲۸	۴/۱۸	۲/۲۷	۲/۹	۶/۱۳	۶/۶۷	۶/۳

پس از اجتماع مجموعه کلیدواژه‌های نویسندگان و خبرگان و حذف کلیدواژه‌های هم‌پوشان (که میانگین ۹٪ کلیدواژگان را به خود اختصاص می‌دادند)، نتایج زیر از اجرای روش‌های ارزیابی مطرح‌شده بر روی دادگان نوروژه به‌دست آمده‌اند. جدول (۲) نشان‌دهندهٔ میانگین دقت متوسط برای تمامی کلیدواژگان استخراج‌شدهٔ مقالات است.

همچنین جدول (۳) و جدول (۴)، معیارهای دقت[1]، بازخوانی[2] و سنجهٔ اف را برای یک کلیدواژه، سه کلیدواژه، پنج کلیدواژه و ده کلیدواژهٔ برتر استخراج‌شدهٔ مقالات را به ترتیب به‌صورت میانگین ریزدانه[3] و میانگین درشت‌دانه[4] نمایش می‌دهند. در زمان محاسبهٔ میانگین ریزدانه، امتیاز تک‌تک اسناد دادهٔ آزمون جداگانه محاسبه می‌شود؛ سپس میانگین این امتیازات اعلام می‌شود، اما در میانگین درشت‌دانه، امتیاز نهایی بر اساس دقت بازیابی واژگان کلیدی در تمام اسناد یکجا محاسبه می‌شود.

همان‌گونه که در بخش معرفی دادگان بیان شد، متوسط نرخ تطابق کلیدواژگان ارائه‌شده با متون چکیده ٪۵۶.۶ است که البته امکان اصلاح عبارات کلیدی به‌منظور افزایش نرخ تطابق وجود دارد، اما در پژوهش پیش‌رو، به جهت رعایت جانب امانت و حفظ اصالت دادگان، از انجام این عمل خودداری به عمل آمده‌است؛ همچنین، شایان توجه است که روش‌های پیش‌گفته، تنها برای معرفی دادگان مورد بررسی قرار گرفته‌اند و علت اساسی کاهش دقت و بازخوانی در آزمون‌های انجام‌شده، تمرکز این روش‌ها در استخراج کلیدواژه به صورت لغوی و همچنین، کمبود نرخ تطابق کلیدواژگان ارائه‌شده با متون چکیده است.

مطابق جداول (۳ و ۴) مشاهده می‌شود که روش «رتبه‌بندی موقعیت» توانسته است نسبت به هر دو روش «رتبه‌بندی متن» و «رتبه‌بندی تکی» به نتایج بهتری دست پیدا کند.

۶-نتیجه‌گیری

در این مقاله دلایل نیاز به دادگانی برای ارزیابی روش‌های استخراج کلیدواژه از متون علمی مورد بررسی قرار گرفت و دادگان نوروژه برای رفع مشکلات موجود معرفی شد؛ سپس دسته‌ای از روش‌های شاخص استخراج کلیدواژه بر روی دادگان مورد نظر آزمون شد و دادگان مورد ارزیابی قرار گرفت. از نتایج آزمایش‌ها مشخص شد که در مجموع، روش رتبه‌بندی موقعیت دارای عملکرد بهتری نسبت به مابقی روش‌هاست؛ همچنین، به این خاطر که کلیدواژه‌های موجود در دادگان نوروژه، بعضی به صورت لغوی و بعضی دیگر به صورت معنایی استخراج شده‌اند، این دادگان قابلیت ارزیابی روش‌هایی را که با هر یک از این دو رویکرد (لغوی و معنایی) به استخراج کلیدواژه‌ها می‌پردازند دارد.

[1] Precision
[2] Recall
[3] Micro
[4] Macro

[DOI: 10.61186/jsdp.21.4.113]

برای بهبود کیفیت دادگان می‌توان اقداماتی انجام داد؛ از جمله اینکه کلیدواژه‌های استخراج‌شده بازبینی و اصلاح دوباره شوند و کلیدواژه‌های متنوع‌تر از نظر انطباق با متن چکیده و دقت آن به دادگان اضافه شود؛ همچنین حجم دادگان را افزایش قابل توجهی داد تا از آن در روش‌های مبتنی بر یادگیری ماشین به ویژه یادگیری عمیق استفاده شود.

۷-مراجع 7-References

[1] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "PerKey: A Persian News Corpus for Keyphrase Extraction and Generation," in *9th International Symposium on Telecommunication: With Emphasis on Information and Communication Technology, IST 2018*, 2019, pp. 460–465.

[2] Y. MATSUO and M. ISHIZUKA, "Keyword Extraction From a Single Document Using Word Co-Occurrence Statistical Information," *Int. J. Artif. Intell. Tools*, vol. 13, no. 01, pp. 157–169, 2004.

[3] M. S. Paukkeri and T. Honkela, "Likey: Unsupervised language-independent keyphrase extraction," *ACL 2010 - SemEval 2010 - 5th Int. Work. Semant. Eval. Proc.*, pp. 162–165, 2010.

[4] A. J. P. Tixier, F. D. Malliaros, and M. Vazirgiannis, "A graph degeneracy-based approach to keyword extraction," *EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc.*, pp. 1860–1870, 2016.

[5] P. Meladianos, A. J. P. Tixier, G. Nikolentzos, and M. Vazirgiannis, "Real-time keyword extraction from conversations," *15th Conf. Eur. Chapter Assoc. Comput. Linguist. EACL 2017 - Proc. Conf.*, vol. 2, pp. 462–467, 2017.

[6] B. Škrlj, A. Repar, and S. Pollak, "RaKUn: Rank-based Keyword Extraction via Unsupervised Learning and Meta Vertex Aggregation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019, vol. 11816 LNAI, pp. 311–323.

[7] K. Bennani-Smires, C. Musat, A. Hossmann, M. Baeriswyl, and M. Jaggi, "Simple unsupervised keyphrase extraction using sentence embeddings," in *CoNLL 2018 - 22nd Conference on Computational Natural Language Learning, Proceedings*, 2018, pp. 221–229.

[8] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features," *Inf. Sci. (Ny).*, vol. 509, pp. 257–289, 2020.

[9] M. Azarafza and M. Feizi-Derakhshi, "Textrank-based microblogs keyword extraction method for Persian language," 2020.

[10] E. Mehrabi, A. Mohebi, and A. Ahmadi, "Improved keyword extraction for persian academic texts using RAKE algorithm; case study: Persian theses and dissertations," *Iran. J. Inf. Process. Manag.*, vol. 37, no. 1, pp. 197–228, 2021.

[11] S. Duari and V. Bhatnagar, "Complex Network based Supervised Keyword Extractor," *Expert Syst. Appl.*, vol. 140, 2020.

[12] S. Lazemi, H. Ebrahimpour-Komleh, and N. Noroozi, "PAKE: a supervised approach for Persian automatic keyword extraction using statistical features," *SN Appl. Sci.*, vol. 1, no. 12, 2019.

[13] A. Sharifi and M. A. Mahdavi, "Supervised approach for keyword extraction from Persian documents using lexical chains," *Signal Data Process.*, vol. 15, no. 4, pp. 95–110, 2019.

[14] H. Veisi, N. Aflaki, and P. Parsafard, "Variance-based features for keyword extraction in Persian and English text documents," *Sci. Iran.*, vol. 27, no. 3 D, pp. 1301–1315, 2020.

[15] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "Persian Keyphrase Generation Using Sequence-to-Sequence Models," *ICEE 2019 - 27th Iran. Conf. Electr. Eng.*, pp. 2010–2015, 2019.

[16] F. Liu, X. Huang, W. Huang, and S. X. Duan, "Performance evaluation of keyword extraction methods and visualization for student online comments," *Symmetry (Basel).*, vol. 12, no. 11, pp. 1–20, 2020.

[17] Y. Wang and J. Zhang, "Keyword extraction from online product reviews based on bi-directional LSTM recurrent neural network," *IEEE Int. Conf. Ind. Eng. Eng. Manag.*, vol. 2017-Decem, pp. 2241–2245, 2018.

[18] M. Tang, P. Gandhi, M. A. Kabir, C. Zou, J. Blakey, and X. Luo, "Progress Notes Classification and Keyword Extraction using Attention-based Deep Learning Models with BERT," 2019.

[19] S. N. Kim, O. Medelyan, M. Y. Kan, and T. Baldwin, "SemEval-2010 Task 5: Automatic keyphrase extraction from scientific articles," in *ACL 2010 - SemEval 2010 - 5th International Workshop on Semantic Evaluation, Proceedings*, 2010, pp. 21–26.

[20] A. Hulth, "Improved automatic keyword extraction given more linguistic knowledge," pp. 216–223, 2003.

[21] X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," *Proc. Natl. Conf. Artif. Intell.*, vol. 2, pp. 855–860, 2008.

[22] T. D. Nguyen and M.-Y. Kan, "Keyphrase Extraction in Scientific Publications," *Asian Digit. Libr. Look. Back 10 Years Forg. New Front.*, pp. 317–326, 2008.

[23] N. Giarelis and N. Karacapilidis, "Deep learning and embeddings-based approaches for keyphrase extraction: a literature review," *Knowl. Inf. Syst.*, 2024.

[24] G. Ashqar and A. Mutlu, "A Comparative Assessment of Various Embeddings for Keyword Extraction," *HORA 2023 - 2023 5th Int. Congr. Human-Computer Interact. Optim. Robot. Appl. Proc.*, 2023.

[DOI: 10.61186/jsdp.21.4.113]

[25] L. C. Chen and K. H. Chang, “An entropy-based corpus method for improving keyword extraction: An example of sustainability corpus,” *Eng. Appl. Artif. Intell.*, vol. 133, 2024.
[26] Z. H. Amur, Y. K. Hooi, G. M. Soomro, H. Bhanbhro, S. Karyem, and N. Sohu, “Unlocking the Potential of Keyword Extraction: The Need for Access to High-Quality Datasets,” *Appl. Sci.*, vol. 13, no. 12, 2023.
[27] P. He, J. Huang, and M. Li, “Text Keyword Extraction Based on GPT,” *Proc. 2024 27th Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2024*, pp. 1394–1398, 2024.
[28] S. Mohtaj, B. Roshanfekr, A. Zafarian, and H. Asghari, “Parsivar: A language processing toolkit for Persian,” in *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pp. 1112–1118, 2019.
[29] Najaf Project, “Synonyms & Antonyms Set in Persian,” 2021. http://najafproj.ir/datasets/syn-ant-set-in-persian/.
[30] R. Mihalcea and P. Tarau, “TextRank: Bringing Order into Texts,” *2004 Empir. Methods Nat. Lang. Process.*, 2004.
[31] C. D. Manning, P. Raghavan, and H. Schutze, “Introduction to Information Retrieval,” in *Introduction to information retrieval*, Stanford: Cambridge University Press, 2010.
[32] C. Florescu and C. Caragea, “PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents,” *Proc. ACL*, pp. 1105–1115, 2017.
[33] F. Boudin, “Pke: An open source python-based keyphrase extraction toolkit,” *COLING 2016 - 26th Int. Conf. Comput. Linguist. Proc. COLING 2016 Syst. Demonstr.*, pp. 69–73, 2016.
[34] A. Hosseini, “PERKE,” 2021.

محمدامین طاهری در حال حاضر دانشجوی کارشناسی‌ارشد مهندسی کامپیوتر در دانشگاه علم‌وصنعت ایران است. ایشان مدرک کارشناسی خود را در رشتهٔ مهندسی کامپیوتر در سال ۱۴۰۰ از همان دانشگاه دریافت کرد. زمینه‌های فعالیت و پژوهش وی شامل یادگیری ماشین، یادگیری عمیق، علم داده، بینایی کامپیوتر و پردازش زبان طبیعی است.
نشانی رایانامهٔ ایشان عبارت‌است از:
taheri_m96@comp.iust.ac.ir

محمدابراهیم شناسا عضو هیأت علمی دانشگاه آزاد واحد تهران شمال است. وی کارشناسی‌ارشد خود را در سال ۱۳۸۸ از دانشگاه آزاد واحد علوم و تحقیقات دریافت کرد. زمینه‌های پژوهشی مورد علاقهٔ ایشان متن‌کاوی، پردازش زبان طبیعی، یادگیری عمیق و داده‌کاوی است.
نشانی رایانامهٔ ایشان عبارت‌است از:
eshenassa@gmail.com

بهروز مینایی بیدگلی دانش‌آموختهٔ دانشگاه ایالتی میشیگان آمریکا در رشتهٔ علوم و مهندسی کامپیوتر با تخصص هوش مصنوعی و داده‌کاوی است. ایشان در حال حاضر دانشیار دانشکدهٔ کامپیوتر دانشگاه علم‌وصنعت ایران است و یک گروه پژوهشی در زمینهٔ داده‌کاوی و بازی‌های رایانه‌ای در این دانشگاه راه‌اندازی کرده‌است. زمینه‌های پژوهشی مورد علاقهٔ ایشان متن‌کاوی، پردازش زبان طبیعی و یادگیری ماشین هستند.
نشانی رایانامهٔ ایشان عبارت‌است از:
b_minaei@iust.ac.ir

سید علی حسینی مدرک کارشناسی و کارشناسی‌ارشد خود را در رشتهٔ علوم کامپیوتر به ترتیب از دانشگاه صنعتی شریف و دانشگاه شهید بهشتی و مدرک دکترای خود را در هوش مصنوعی از دانشگاه ژیرونا دریافت کرد و در حال حاضر، پژوهش‌گر پسادکترا و سرتیم پژوهشی آزمایشگاه پردازش دانش در دانشگاه علم‌وصنعت ایران است.
نشانی رایانامهٔ ایشان عبارت‌است از:
mrhossayni@gmail.com

[DOI: 10.61186/jsdp.21.4.113]