Tabriz Journal of Electrical Engineering (TJEE), vol. 52, no. 2, Summer 2022 Serial no. 100
DOI: 10.22034/tjee.2022.15426

# Persian Keyphrase Generation Using Transfer Learning

Marziea Rahimi*, Erfan Jalili Jalal, Hossain Alirezayi

Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran.
marziea.rahimi@shahroodut.ac.ir; up202111337@edu.fe.up.pt; hoseinalirezaee@mail.um.ac.ir

**Short Abstract**
Automatic keyphrase generation plays an important role in many text analysis and natural language processing tasks. Many existing methods are bound to select keyphrases from the terms and phrases that are present in the target text. This handicap could be overcome using sequence-to-sequence methods. However, many such methods need huge datasets for training which pose a challenge for low-resource languages such as Persian. Transfer learning where a pre-trained model is adapted to a new task specified with a smaller dataset is very useful in such circumstances. In this paper, we present a sequence-to-sequence method utilizing a transformer model for Persian keyphrase generation. Accordingly, a corpus of 70K Persian scientific abstracts and their corresponding keyphrases have been gathered. A pretrianed MT5 mdel is fine-tuned on this corpus for the task of Persian keyword generation. The resulted model is compared to several other keyphrase generation methods. The results indicate that the proposed method can outperform existing methods at least by 2.71 percent.

**Keywords**
Keyphrase generation, keyphrase extraction, transformer models, Persian corpus, abstractive summarization, sequence-to-sequence learning.

## 1- Short Introduction (4-5 lines)

Persian keyphrase generation methods are extractive where the keyphrases are selected from the present phrases in the target text. While human authors almost equally use absent keyphrases which are not mentioned in the target text itself [1]. Sequence-to-sequence models especially the ones based on transfer learning such as MT5 [2] can overcome this handicap. We have gathered a sufficiently big corpus of scientific abstract and keyphrase string pairs to implement such a model for Persian texts.

## 2- Proposed Work and Methodology (including comparison, simulation/experimental results and discussion)

We have scraped Persian open access journal websites to gather a sufficiently big corpus of scientific abstract and keyphrase string pairs. After preprocessing the original texts, a corpus of 56422 entries is provided. The corpus is comprised of abstracts on 13 diverse topics. Diversity is an important feature for general-purpose text analysis datasets.

Table 1- Statistics of the tokens in *keyphrase* strings

| Feature | Value |
|---|---|
| Total number of tokens | 422624 |
| Mean | ۷.۵ |
| Standard deviation | 3.75 |
| Maximum number of tokens per string | 164 |
| Minimum number of tokens per string | 1 |
| Absent tokens rate | 28.86 |

Table 2-Statistics of the tokens in abstracts

| Feature | Value |
|---|---|
| Total number of tokens | 6961233 |
| Mean | 123.53 |
| Standard deviation | 41.98 |
| Maximum number of tokens per abstract | 584 |
| Minimum number of tokens per abstract | 25 |

This corpus is used to fine-tune a multilingual version of a pre-trained transformer [3] model which is called MT5 [2]. This model has a text-to-text encoder-decoder structure. We also provide a comparison between our fine-tuned mT5 model and other existing models in Table 3; As shown, The mT5 obtained the best results among other available keyphrase generation models.

Table 3-Comparison of Mt5 with other existing models

| Model | P@10 | R@10 | F1@10 | P@5 | R@5 | F1@5 |
|---|---|---|---|---|---|---|
| TFIDF | 23.18 | 27.19 | 23.18 | 30.43 | 20.68 | 24.62 |
| Yake | 12.5۱ | 16.01 | 14.04 | 16.9۹ | 10.89 | 13.27 |
| SR | 17.62 | 23.68 | 20.21 | 18.11 | 12.16 | 14.55 |
| MR | 12.78 | 16.81 | 14.51 | 14.11 | 9.47 | 11.34 |
| TR | 13.79 | 17.95 | 15.60 | 16.29 | 10.91 | 13.07 |
| MT5 | 38.09 | 28.84 | 32.82 | 38.14 | 28.81 | 32.82 |

## 3- Conclusion (4-5 lines)

We proposed a dataset for the Farsi keypharas generation task and Also fined-tuned an mT5 pre-trained model using this dataset. Finally, results show that our fine-tuned transformer model outperforms other available keyphrase generation models.

## 4- References (2-3 references)

[1] Çano E, Bojar O (2020) Two Huge Title and Keyword Generation Corpora of Research Articles. In: LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings. pp 6663–6671.
[2] Xue L, Constant N, Roberts A, et al (2021) mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 483–498
[3] Vaswani A, Shazeer N, Parmar N, et al (2017) Attention Is All You Need. Advances in neural information processing systems 8:8–15

# تولید کلمات کلیدی متون فارسی با استفاده از یادگیری انتقالی

**مرضیه رحیمی**
استادیار، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران

**عرفان جلیلی جلال**
دانشجوی دکتری، دانشکده مهندسی انفورماتیک، دانشگاه پورتو، پورتو، پرتغال

**حسین علیرضایی**
فارغ التحصیل کارشناسی دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران

**چکیده**
تولید خودکار کلمات کلیدی، نقش مهمی در بسیاری از کاربردهای تحلیلی متن و زبان‌های طبیعی، به‌ویژه در دسته‌بندی و بازیابی سریع متون دارد. بسیاری از روش‌های کنونی محدود به انتخاب کلماتی هستند که صریحاً در متن ذکر شده‌اند. استفاده از روش‌های دنباله‌به‌دنباله قادر است این نقصان را برطرف کند. البته استفاده از این روش‌ها معمولاً مستلزم وجود پیکره‌های عظیم است که برای زبان‌های کم‌منبع مثل فارسی یک چالش محسوب می‌شود. در چنین موقعیت‌هایی، یادگیری انتقالی که در آن یک مدل پیش‌آموخته بر روی یک وظیفه جدید با مجموعه کوچکتری از داده‌ها تطبیق داده می‌شود، می‌تواند راه‌گشا باشد. در این مقاله، برآنیم تا با استفاده از یک روش دنباله‌به‌دنباله مبتنی بر شبکه‌های عمیق انتقالی، به تولید کلمات کلیدی برای متون علمی فارسی بپردازیم. در همین راستا، پیکره متنوعی از ۷۰هزار مقاله تخصصی به زبان فارسی و کلمات کلیدی متناظرشان جمع‌آوری شده است. سپس شبکه انتقالی پیش‌آموخته MT5 با استفاده از این پیکره، برای وظیفه تولید کلمات کلیدی، تنظیم و بازآموزی شده است. مدل حاصل، با چندین روش دیگر مقایسه شده است. نتایج این مقایسه حاکی از برتری حداقل ۲٫۷۱ درصدی آن بر روش‌های موجود است.

**کلمات کلیدی**
تولید عبارات کلیدی، استخراج عبارات کلیدی، روش‌های دنباله‌به‌دنباله، شبکه‌های عمیق انتقالی، پیکره فارسی، خلاصه‌سازی چکیده‌ای.

نام نویسنده مسئول: مرضیه رحیمی
ایمیل نویسنده مسئول: marziea.rahimi@shahroodut.ac.ir

تاریخ ارسال مقاله: ۱۴۰۱/۰۱/۳۰
تاریخ(های) اصلاح مقاله: ۱۴۰۱/۰۴/۲۲
تاریخ پذیرش مقاله: ۱۴۰۱/۰۶/۱۲

## ۱- مقدمه

امروزه، همه ما در کار و زندگی روزمره با حجم بزرگی از اطلاعات دیجیتال و به‌ویژه متون روبرو هستیم. متون خبری، ایمیل‌ها، کتاب‌ها، گزارش‌های تکنیکی و علمی، گزارشات پزشکی، ارتشی یا قضایی نمونه‌هایی از این متون هستند. اگر هریک از این متون به شکل مناسبی و با مجموعه کوتاهی از کلمات و عبارات کلیدی توصیف شوند خوانندگان با سرعت بیشتری می‌توانند از محتوای آن‌ها آگاه شده و تصمیمات کاراتری برای ادامه کار بگیرند. این، یکی از دلایلی است که کلمات یا عبارات کلیدی نقش مهمی را در توصیف عناصر دیجیتال و به‌ویژه متون بازی می‌کنند. کلمات یا عبارات کلیدی یک متن، مجموعه‌ای کوچک از چند عبارت و کلمه را گویند که نماینده‌ای از موضوعات و مفاهیم تشکیل‌دهنده آن متن هستند. این مجموعه معمولاً کوچک از کلمات و عبارات، همچنین نقش مهمی در دسته‌بندی و بازیابی سریع متون در کتابخانه‌های دیجیتال بازی می‌کنند و نیز می‌توانند در بسیاری از کاربردهای دیگر از جمله خلاصه‌سازی [۱, ۲]، طراحی سامانه‌های پرسش‌وپاسخ [۳, ۴]، طراحی سیستم‌های توصیه‌گر [۵] و ایده‌کاوی [۶, ۷] موثر باشند.

در بسیاری از موارد، به‌ویژه در مورد متون خبری یا گزارش‌های علمی، نویسنده متن، کلمات کلیدی مناسب و باکیفیتی را از دیدگاه تخصصی خود برای متن ارائه می‌کند. ولی در بیشتر موارد چنین کلماتی وجود ندارند یا دارای کیفیت مناسبی نیستند. به همین دلیل است که تولید یا استخراج خودکار مجموعه کلمات کلیدی از اهمیت بسیار بالایی برخوردار است.

روش‌های خودکار تولید کلمات کلیدی [۸–۱۰] را می‌توان به دو دسته کلی روش‌های استخراجی و روش‌های مبتنی بر داده تقسیم کرد. روش‌های دسته اول، عموماً روش‌های بی‌ناظر هستند که خود به دو زیردسته روش‌های گرافی و روش‌های آماری قابل تقسیم‌اند. دسته دوم عموماً روش‌های باناظر هستند. این دسته را نیز می‌توان به دو زیردسته روش‌های دسته‌بندی و روش‌های چکیده‌ای[۱] شکست. در چند سال اخیر، روش‌های چکیده‌ای، که عموماً به صورت دنباله‌به‌دنباله[۲] انجام می‌شوند و دارای ارتباط تنگاتنگ با یادگیری عمیق هستند، مورد توجه بسیار قرار گرفته و نتایج برجسته‌ای را نیز

[۱] Abstractive

[۲] Sequence-to-sequence

در مقایسه با سایر روش‌های موجود، تولید نموده‌اند. این روش‌ها در صورتی قابل اجرا هستند که داده برچسب‌دار موجود باشد. این نکته، توسعه آن‌ها را به‌ویژه برای زبان‌هایی چون زبان فارسی که منابع‌شان محدود است، با چالش مواجه می‌کند. تا جایی که ما می‌دانیم، در زبان فارسی، تنها دو مجموعه داده بزرگ مناسب برای تولید کلمات کلیدی ارائه شده‌اند که یکی [۱۱] حاوی ۵۵۳هزار متن خبری و کلمات کلیدی آن‌هاست و دیگری [۱۲] حاوی ۲۶۰هزار چکیده مربوط به رساله‌ها و پایان‌نامه‌های علمی است. از بین این دو، فقط مجموعه اول در دسترس عموم است. از طرف دیگر، استفاده از روش‌های انتقالی و بهره‌گیری از روش‌های پیش‌آموخته می‌تواند، نیاز ما را به مجموعه‌های عظیم تا حدی تقلیل دهد.

در این مقاله، هدف ما تولید عبارات کلیدی با استفاده از یک روش انتقالی است. برای انجام این کار، یک مجموعه داده فارسی برچسب‌دار که برای کاربرد تولید و استخراج کلمات کلیدی مناسب است، را تولید و معرفی کرده‌ایم. همچنین از یک روش انتقالی[۳] برای تولید کلمات کلیدی در زبان فارسی استفاده خواهیم کرد که تا پیش از این انجام نشده است. نتایج این روش را با مجموعه‌ای از روش‌های پایه و همچنین جدید برای تولید و استخراج کلمات کلیدی، بر روی مجموعه داده مذکور، مقایسه خواهیم نمود.

## ۲- مرور کارهای پیشین

در این بخش، به مرور روش‌ها و مجموعه داده‌های موجود برای استخراج و تولید کلمات کلیدی خواهیم پرداخت و البته تمرکز ما بر معرفی کارهای مربوط به زبان فارسی خواهد بود. همانطور که پیش از این بیان شد، روش‌های خودکار تولید کلمات کلیدی، در دو دسته کلی روش‌های استخراجی و روش‌های مبتنی بر داده قرار دارند. روش‌های دسته اول، عمدتا روش‌های بی‌ناظر هستند که خود به دو زیردسته روش‌های گرافی و روش‌های آماری تقسیم می‌شوند. بیشتر روش‌های دسته دوم، باناظر هستند. این دسته نیز خود شامل دو زیردسته روش‌های دسته‌بندی و روش‌های چکیده‌ای است.

تا قبل از معرفی روش‌های دنباله‌به‌دنباله، یک روال کلی رایج در روش‌های خودکار استخراج کلمات کلیدی این بوده است که برای عبارات متن که معمولاً هم به صورت n-gramهای متن استخراج می‌شوند وزن یا رتبه‌ای تعیین می گردد. سپس برمبنای این وزن یا رتبه، برای انتخاب یا عدم انتخاب آن عبارت تصمیم گرفته می‌شود. حال بسته به اینکه در فرایند تعیین وزن از کلمات کلیدی یا همان برچسب‌های مجموعه داده استفاده شود یا خیر، دو دسته کلی باناظر و بی‌ناظر شکل گرفته‌اند که در بالا به آن اشاره کردیم. با در نظر داشتن این روال کلی، در ادامه، به معرفی چند نمونه اخیر از هریک از این دسته‌ها خواهیم پرداخت. در این مقاله، از روش‌های چکیده‌ای با عنوان "تولید کلمات کلیدی" و سایر روش‌ها را با عنوان "استخراج کلمات کلیدی" نیز یاد خواهیم نمود چرا که روش‌های چکیده‌ای محدود به انتخاب کلمات کلیدی از بین کلمات متن نیستند.

روش‌های خودکار استخراج و تولید کلمات کلیدی از دهه ۹۰ میلادی به گستردگی مورد توجه بوده‌اند. در این دهه، بیشتر روش‌ها، از ویژگی‌های ساده آماری کلمات یا n-gramها در متن استفاده نموده و روش‌های باناظر بسیار مورد توجه بوده‌اند. یک نمونه پرکابرد از این روش‌ها، روش KEA [13] است که از دو ویژگی TF_IDF[۴] و اولین رخداد استفاده می‌کند. ویژگی اولین رخداد، نماینده این است که چه کسری از متن سند، قبل از اولین رخداد کلمه یا عبارت مورد نظر، قرار گرفته است. در واقع، هم موقعیت و هم فراوانی رخداد یک کلمه در تصمیم‌گیری برای انتخاب آن موثرند. نمونه دیگر تلاش‌های این دهه، تلاش در جهت بهبود روش‌های موجود با استفاده از ابزارهایی مثل الگوریتم‌های تکاملی است. به عنوان مثال GenEx [14] از الگوریتم ژنتیک برای تنظیم پارامترهای روشی به نام Extractor [15] برای بهبود عملکرد آن استفاده کرده است.

در دهه‌های بعد، روش‌های بدون ناظر با اقبال بیشتری مواجه شدند. در این روش‌ها، علاوه بر استفاده از ویژگی‌های ساده آماری، ویژگی‌ها و پردازش‌های مبتنی بر گراف هم مورد استفاده قرار گرفتند. یک نمونه پرکابرد از این روش‌ها، روش TFIDF است. این روش، یکی از روش‌های پایه رایج برای مقایسه در تولید و استخراج کلمات کلیدی است. در این روش، دنباله n-gramها به عنوان عبارات متن، استخراج شده و بر اساس رابطه مشهور TF-IDF وزن‌دهی می‌شوند به این صورت که فراوانی‌ها برای هر دنباله عبارت محاسبه می‌گردد. به عنوان نمونه دیگر، در روش KPMiner [16]، ابتدا تعدادی از کلمات متن، به عنوان کلمات کاندید انتخاب شده و سپس این کلمات بر اساس ویژگی‌های آماری و موقعیتی، وزن‌دهی می‌شوند. انتخاب کلمات کاندید براساس دو قانون ابتکاری فراوانی رخداد بیشتر از یک حد آستانه و عدم رخداد در بخش‌های انتهایی سند، صورت می‌گیرد. نمونه دیگر روش‌ها که در سالهای اخیر معرفی شده است، روش YAKE [17] است که پنج ویژگی مختلف مرتبط با فراوانی و موقعیت یک کلمه را برای تعیین وزن آن عبارت یا کلمه در نظر می‌گیرد. ویژگی اول مرتبط با بزرگی و کوچکی حروف یک کلمه است. بر این مبنا که کلماتی که با حروف بزرگ نوشته می‌شوند از اهمیت ویژه‌ای برخوردارند. ویژگی دوم، موقعیت کلمه در سند است بدین معنی که کلمات ابتدای سند، اهمیت بالاتری دارند. ویژگی سوم فراوانی کلمات درمتن مورد نظر است. ویژگی چهارم نماینده میزان معنی داری یک کلمه یا در واقع اهمیت آن به عنوان نماینده یک مفهوم است. اساس محاسبه این ویژگی بر تنوع کلماتی است که قبل از کلمه مورد نظر، در یک پنجره آمده‌اند. هرچه این کلمات متنوع‌تر باشند، کلمه مورد نظر، کم‌اهمیت‌تر است. در نهایت، ویژگی پنجم، تعداد جملاتی است که یک کلمه در آن‌ها ظاهر شده است. هرچه تعداد این جملات بیشتر باشد، کلمه از اهمیت بالاتری برخوردار است. در نهایت همه این ویژگی‌ها در یک تابع ابتکاری با هم ترکیب می‌شوند تا وزن کلمات محاسبه گردد. یک نمونه دیگر از این روش‌ها به نام Attention Rank (AR) [18] که در سال‌های اخیر معرفی شده است، از مکانیزم توجه در شبکه‌های عمیق انتقالی استفاده می‌کند. به این ترتیب که با استفاده از میزان توجه به کلمات که از شبکه انتقالی استخراج می‌شود، درجه اهمیت هر کلمه کاندید از سند را به عنوان یک کلمه کلیدی مشخص می‌نماید.

از زیردسته روش‌های مبتنی بر گراف، می‌توان روش Single Rank (SR) [19] را نام برد که با توجه به نقش نحوی کلمات متن، فقط نام‌ها و صفات را استخراج کرده و سپس برای هر سند گرافی را می‌سازد که راس‌های آن دنباله‌هایی از همین نام‌ها و صفات بوده و یال‌هایش نماینده میزان ارتباط همرخدادی بین کلمات هستند. سپس بر اساس الگوریتم TextRank [20] کلمات کلیدی را استخراج می‌نماید. الگوریتم دیگری از این دسته TopicRank [21] (TR) است که اطلاعات موضوعی را نیز به الگوریتم Single Rank اضافه می‌کند. به این صورت که رئوس گراف بجای عبارات اسمی، موضوعات یا در واقع خوشه‌هایی از کلمات مرتبط با یک موضوع هستند. یعنی، بعد از تعیین نقش کلمات و استخراج دنباله‌های اسم‌ها و صفات، این دنباله‌ها خوشه‌بندی شده و راس‌های گراف را تشکیل می‌دهند. الگوریتم دیگری به نام Multipartite Rank (MR) [22] نیز سعی در به‌کارگیری اطلاعات موضوعی در تعیین عبارات کلیدی دارد. در این روش از یک گراف چندبخشی برای نمایش موضوعی عبارات کاندید استفاده می‌شود. یعنی هرچند همچنان خوشه‌هایی از عبارات به عنوان

[۳] Transfer learning

[۴] Term Frequency_ Inverse Document Frequency

موضوع ساخته‌می‌شوند ولی رئوس گراف همچنان تک‌عبارت هستند و ارتباط موضوعی آن‌ها با هم در هم‌بخش بودنشان در گراف نهفته است.

در چند سال اخیر، با اوج‌گیری توسعه روش‌های دنباله‌به‌دنباله در حوزه شبکه‌های عصبی عمیق و اعمال این روش‌ها بر تولید خودکار کلمات کلیدی، روش‌های چکیده‌ای [۲۳] که زیردسته‌ای از روش‌های باناظر هستند، به شدت مورد توجه قرار گرفته‌اند. یکی از تفاوت‌های اصلی روش‌های این دسته با سایر روش‌های بررسی‌شده تاکنون، امکان تولید عباراتی است که صریحا در سند مورد نظر ذکر نشده‌اند. بررسی دادگان انگلیسی موجود، نشان داده است که کاربر متخصص انسانی، در تخصیص کلمات کلیدی به یک متن، خود را محدود به کلمات حاضر در متن نمی‌کند و کلماتی را که در متن غایبند و به صورت ضمنی در متن مورد اشاره قرار گرفته‌اند هم به همان میزان به کار می‌برد. با توجه به این امر، محدود کردن روش‌های خودکار به انتخاب کلمات حاضر در متن، یک نقصان محسوب می‌شود و روش‌های چکیده‌ای قادرند این نقصان را رفع نمایند [۲۳].

بسیاری از روش‌های فوق، هرچند برای زبان انگلیسی ارائه شده‌اند، محدود به زبان انگلیسی نبوده و قابل اعمال بر زبان‌های دیگر، از جمله فارسی نیز هستند، اما کارهایی که در زمینه تولید و استخراج کلمات و عبارات کلیدی در زبان فارسی منتشر شده‌اند بسیار اندک است. مجموعه داده‌های کلمات کلیدی منتشرشده در زبان انگلیسی [۲۳]، چه بزرگ و چه کوچک، بسیارند. ولی در مورد زبان فارسی، تا جایی که ما می‌دانیم تنها یک نمونه [۱۱] منتشرشده از چنین دادگانی وجود دارد و سایر کارهای انجام‌شده در فارسی، دادگان خود را منتشر نکرده‌اند. در ادامه به معرفی کارهای فارسی خواهیم پرداخت.

یکی از روش‌های ارائه‌شده به زبان فارسی [۲۴] که بر روی مجموعه‌ای از چکیده و کلمات کلیدی ۱۰۲ مقاله فارسی انجام شده است، از زنجیره‌های لغوی کلمات برای تعیین عبارات کلیدی استفاده نموده است. به این ترتیب که ابتدا زنجیره‌های لغوی کلمات تشکیل می‌شود. سپس بر مبنای این زنجیره‌ها، ۱۰ ویژگی برای هر کلمه استخراج می‌گردد و این ویژگی‌ها در اختیار یک دسته‌بند قرار می‌گیرند تا کلمات کلیدی تعیین شوند. نمونه دیگر [۲۵]، برای محاسبه وزن کلمات از بردار تعبیه آن کلمه و محاسبه فاصله بردار هر کلمه با سایر کلمات استفاده می‌شود. این روش بر روی مجموعه‌ای از ۲۰۰۰ متن خبر استخراج‌شده از پایگاه خبری YJC و کلمات کلیدی مربوط به آن‌ها اجرا و ارزیابی شده است. محرابی و همکاران، در پژوهش خود [۲۶]، سعی کرده‌اند الگوریتم RAKE [27] را با ایجاد تغییراتی در شیوه وزن‌دهی آن، بهبود بخشند. این پژوهش بر روی پایگاه داده‌ای شامل ۵۰۰ چکیده و کلمات کلیدی مربوطه از پایان‌نامه‌های ثبت‌شده در پایگاه «گنج»، اجرا و ارزیابی شده است. نمونه دیگر از کارهای فارسی مقاله‌ای [۱۱] است که در آن پیکره Perkey شامل ۵۵۳ هزار متن خبر و کلمات کلیدی مربوط به آن‌ها، معرفی شده است. هفت روش مختلف استخراج کلمات کلیدی بر روی این دادگان اجرا و نتایج آن منتشر شده است. به نظر می‌رسد، پیکره Perkey، بزرگترین و تنها پیکره منتشرشده فارسی در زمینه استخراج و تولید کلمات کلیدی است. در روش PAKE [28]، از شش ویژگی آماری که پنج تا از آن‌ها بر مبنای فراوانی محاسبه شده‌اند و یکی بر اساس موقعیت کلمه، وزنی به عبارات مختلف متن، تخصیص داده می‌شود و عبارات کلیدی برمبنای این وزن انتخاب می‌شوند. این روش بر روی مجموعه‌ای از ۱۵۷۰ متن از مقالات و رساله‌های علمی فارسی و کلمات کلیدی آن‌ها که از پایگاه‌های مختلف فارسی جمع‌آوری شده‌اند، پیاده‌سازی شده است. ویسی و همکاران [29] نیز چهار روش جدید وزن‌دهی را بر مبنای ویژگی‌هایی عموما آماری ارائه کرده‌اند که بر روی مجموعه‌ای از ۵۰۰ سند خبری و کلمات کلیدی آن‌ها اجرا شده است. حجازی و نصیری نمونه دیگری از روش‌های آماری [۳۰] را برای استخراج کلمات کلیدی ارائه کرده‌اند که ترکیبات مختلفی از ویژگی های آماری را بررسی کرده و کلمات کلیدی را با استفاده از یک دسته‌بند انتخاب می‌نماید. مجموعه داده مورد استفاده شامل ۱۴۱ پایان‌نامه فارسی و کلمات کلیدی آن‌هاست. به طور مشابه باسره و همکاران [۳۱] مجموعه‌ای از ۱۸ ویژگی آماری را برای دسته‌بندی کلمات با استفاده از جنگل تصادفی پیشنهاد کرده‌اند. این روش بر روی مجموعه‌ای از ۲۴۴ سند خبری فارسی پیاده‌سازی و ارزیابی شده است. محسنی و فیلی [۱۲] از شبکه LSTM به صورت رمزنگار-رمزگشا[۵] استفاده می‌کنند تا عبارات کلیدی را برای متون به صورت باناظر تولید نمایند و در این تحقیق چندین شیوه بازنمایی مختلف برای کلمات مورد بررسی قرار گرفته است. پیکره ارائه‌شده برای آموزش و ارزیابی مدل‌ها متشکل از ۲۶۰هزار متن از پایان‌نامه‌های ارشد و دکتری فارسی است که در دسترس عموم قرار نگرفته است.

یکی از روش‌های موثری که امروزه برای تولید کلمات کلید مورد استفاده قرار می‌گیرد، یادگیری انتقالی، به‌ویژه با استفاده از شبکه‌های انتقالی است. سایر انواع یادگیری انتقالی در برخی کاربردهای زبان فارسی [۳۲] به کار برده شده‌اند ولی تاکنون برای تولید کلمات کلیدی فارسی مدل یا روشی از این دست، ارائه نشده است. در مقاله کنونی، از شبکه انتقالی MT5 [33] استفاده خواهیم کرد. هدف ماتغییر ساختار این شبکه نیست بلکه شبکه را برای تطبیق با هدف مقاله، تنظیم و بازآموزی خواهیم کرد. فرایند این کار در ادامه توصیف خواهد شد.

## ۳- روش انتقالی پیشنهادی برای تولید کلمات کلیدی فارسی

در این روش، از MT5 [33]، نسخه چندزبانه شبکه انتقالی T5 [34] برای تولید کلمات کلیدی استفاده شده است که یک شبکه متن‌به‌متن است. یعنی در یک ساختار رمزنگار-رمزگشا، یک متن را به عنوان ورودی دریافت کرده و در خروجی نیز یک متن تولید خواهد کرد. شبکه با کمک تولید بردارهای تعبیه کلمات که به صورت مبتنی بر بافت[۶] تولید می‌شوند، نگاشت بین دنباله ورودی و خروجی را یاد می‌گیرد. این روال هم در مرحله پیش‌آموزش و هم در مرحله تنظیم و بازآموزی مورد استفاده قرار می‌گیرد. ساختار اصلی آن بسیار مشابه معماری پایه شبکه‌های انتقالی [۳۵] است.

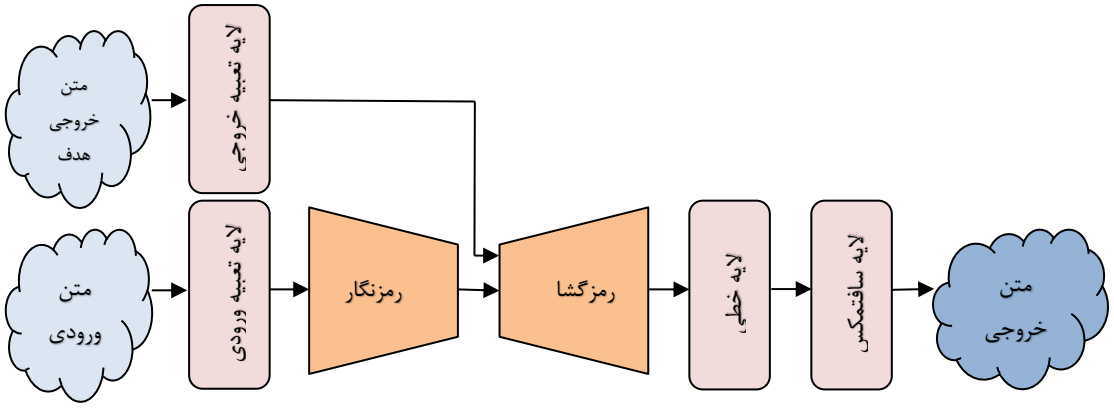
این شبکه دارای ساختار رمزنگار-رمزگشا است که هردو از n لایه تشکیل شده‌اند و هر لایه مبتنی بر مکانیزم خود-توجه است. در سمت رمزنگار، چنانکه در شکل ۱ نمایش داده شده است، یک متن یا به عبارت دیگر، دنباله‌ای از کلمات به عنوان ورودی در اختیار شبکه قرار می‌گیرد. شبکه قرار است نگاشت بین این متن و متن هدف را یاد بگیرد. ابتدا بردارهای تعبیه برای تک‌تک کلمات دنباله ورودی، تولید می‌شوند. این کار توسط یک مدل تعبیه آموزش‌دیده انجام می‌شود. سپس تمامی این بردارها به هم پیوسته و بردار تعبیه ورودی را می‌سازند. این بردار تعبیه در اختیار رمزنگار قرار می‌گیرد. این مرحله از کار در ساختار شبکه انتقالی، توسط لایه تعبیه انجام می‌شود. سپس رمزنگار با تکیه بر مکانیزم خودتوجه[۷] ارتباط بین کلمات مختلف را رمز کرده و بردارهای جدیدی را تولید می‌کند که متاثر از کلمات اطراف کلمه جاری بوده و ارتباط بین این کلمات را در خود نهفته‌اند [۳۵]. در واقع می‌توان گفت مکانیزم خودتوجه هر کلمه را به صورت میانگین وزن‌داری از بردارهای کلمات اطرافش بازنمایی می‌کند [۳۴]. شبکه های انتقالی از ساختارهای بازگشتی یا پیچشی استفاده نمی‌کنند.

حال وظیفه رمزگشا، یادگیری نگاشت بین این بردارها با بردارهای تعبیه کلمات دنباله خروجی هدف است. بردارهای تعبیه خروجی نیز به روشی مشابه

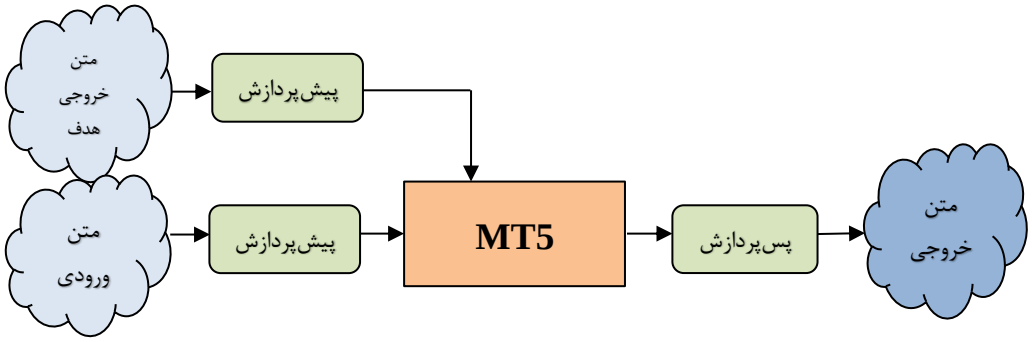
[۴] Encoder-Decoder
[۵] Contextual
[۶] Self-attention

ورودی، قابل تولید هستند. رمزگشا ساختاری مشابه رمزنگار داشته و متکی بر مکانیزم خودتوجه است. پس از تکمیل فرایند آموزش، رمزگشا قادر خواهد بود بردارهای تولیدشده توسط رمزنگار، برای هر متن ورودی را به یک بردار بنگارد که بعد توسط لایه خطی انتهایی تبدیل به یک بردار خیلی بزرگتر و درواقع به اندازه مجموعه لغات متن آموزشی می‌شود. مقادیر این بردار، توسط لایه سافتمکس به مقادیر بین صفر و یک (احتمال) تبدیل می‌گردند. حال می‌توان کلمه متناظر با درایه حاوی بیشترین احتمال را به عنوان کلمه خروجی انتخاب کرد. دنباله این کلمات انتخابی، متن خروجی را می‌سازد.

در روش پیشنهادی از همین ساختار به صورت پیش‌آموخته استفاده شده و تنظیم و بازآموزی آن برروی مجموعه داده ساخته‌شده انجام می‌گردد. در مرحله تنظیم و بازآموزی، متون ورودی و خروجی در اختیار شبکه پیش‌آموخته قرار گرفته و نگاشت‌های یادگرفته‌شده توسط شبکه، بر اساس این داده‌های جدید به‌روز می‌شوند. روش پیشنهادی شامل بخش تنظیم و بازآموزی است.




**شکل ۱- ساختار کلی شبکه انتقالی**




**شکل ۲- ساختار کلی روش انتقالی مورد استفاده در این مقاله برای تولید کلمات کلیدی فارسی**

**جدول ۱- موضوعات مقالات گردآوری‌شده و نسبت هرکدام**

| موضوع | بزرگی (٪) |
|---|---|
| حقوق و سیاست | ۱۳/۳۶ |
| محیط زیست و جغرافیا | ۱۱/۹۵ |
| کشاورزی و دامپروری | ۱۱/۹۳ |
| علوم پایه و مهندسی | ۱۰/۹۶ |
| علوم اجتماعی و روانشناسی | ۸/۲۷ |
| ادبیات و زبان‌شناسی | ۶/۹۰ |
| دین و فلسفه | ۶/۵۴ |
| علوم اقتصادی | ۶/۰۶ |
| مدیریت | ۵/۵۷ |
| پزشکی و دامپزشکی | ۵/۵۱ |
| فرهنگ و هنر | ۴/۱۰ |
| تاریخ | ۳/۵۶ |
| ورزش | ۳/۲۲ |

روند کلی مورد استفاده ما برای تولید کلمات کلیدی، در این روش، در شکل ۱ نمایش داده شده است. همانطور که در شکل پیداست، ورودی و خروجی، هر دو متن‌های بدون ساختار هستند و اندازه دنباله متنی ورودی لزوماً یکی نیست.

البته، برای ارائه این متن به شبکه، یک مقدار حداکثر برای طول دنباله متنی ورودی در نظر گرفته می‌شود که در صورت کوچکتر بودن دنباله از این طول، مابقی به صورت خودکار با یک کاراکتر خاص پر می‌شود. قبل از ارائه متن به شبکه، یک مرحله پیش‌پردازش وجود دارد که شامل تمیز کردن و آماده‌سازی متن برای کاربرد تولید کلمات کلیدی است و در ادامه مفصلاً معرفی خواهد شد. ورودی در اختیار شبکه پیش‌آموخته قرار می‌گیرد و طی فرایند تنظیم و بازآموزی، وظیفه جدید تولید کلمات کلیدی را که به واسطه دنباله‌ها ورودی و خروجی مشخص شده است یاد می‌گیرد. سپس در مرحله پس‌پردازش، دنباله تولیدشده به شکل رایج دنباله‌های کلمات کلیدی که با کاما از هم جدا شده‌اند بازگردانده می‌شود.

بنابراین، همانطور که توصیف شد، برای تولید کلمات کلیدی با استفاده از این شبکه، مجموعه کلمات کلیدی هر متن به عنوان یک رشته متنی در نظر گرفته می‌شود. به این ترتیب، دنباله ورودی، خود متن و دنباله خروجی، رشته کلمات کلیدی آن متن است. روشن است که برای این منظور به مجموعه بزرگی از متون فارسی و رشته کلمات و عبارات کلیدی متناظرشان نیازمندیم. به این منظور مجموعه‌ای از ۷۰هزار چکیده مقالات علمی فارسی و کلمات کلیدی تخصیص‌یافته به آن‌ها توسط نویسندگانشان، جمع‌آوری شده است که در ادامه به معرفی این مجموعه و شیوه آماده‌سازی و پیش‌پردازش آن خواهیم پرداخت.

### ۳-۱- مجموعه داده

برای پیاده‌سازی روش مذکور در بخش قبل، نیاز به مجموعه داده‌ای بزرگ و برچسب‌دار است. در ایجاد یک پیکره برای آموزش مدل‌های زبانی بزرگ، تنوع از اهمیت بالایی برخوردار است. تحقیقات اخیر نشان می‌دهند که تنوع متون پیکره‌ی مورد استفاده برای آموزش این مدل‌ها می‌تواند تعمیم‌پذیری مدل را در کاربردهای مختلف به شکل چشمگیری افزایش دهد [۳۶, ۳۷]. همچنین، تنوع پیکره‌ی آموزشی می‌تواند قابلیت مدل را برای تطبیق خود با یک دامنه جدید، تنها با دیدن مجموعه کوچکی از داده‌های آن دامنه، افزایش دهد [۳۸]. از طرفی، بیشتر مجموعه داده‌های مورد استفاده برای کاربرد تولید کلمات کلیدی، چه بزرگ و چه کوچک، از متون علمی جمع‌آوری شده‌اند [۸]. این را می‌توان در مقابل کاربردهای مشابه مانند خلاصه‌سازی دید که در آن‌ها، وزنه متون خبری سنگین‌تر است. یکی از دلایل این امر می‌تواند کیفیت کلمات کلیدی مشخص شه توسط نویسنده، در متون علمی باشد. همچنین وجود فراداده‌های متنوع می‌تواند یکی دیگر از جذابیت‌های مقالات علمی باشد.

با تکیه بر نکات فوق، و با توجه به عدم در دسترس بودن مجموعه‌ای به اندازه کافی بزرگ و متنوع از متون مقالات علمی فارسی، مجموعه‌ای از ۷۰هزار چکیده فارسی از مقالات منتشرشده در حدود ۱۰۰ پایگاه مختلف مجلات علمی فارسی با دسترسی باز، به همراه کلمات کلیدی و تعدادی فراداده دیگر گردآوری شده است. برای این کار ابتدا لیستی از مجلات علمی فارسی در موضوعات متنوع که دسترسی به مقالات آن‌ها باز و بوده است به عنوان هسته اولیه تهیه شد و سپس با خزش هر آدرس، چکیده مقالات، کلمات کلیدی اختصاص‌یافته به هرکدام و سایر اطلاعات مورد نیاز، از این صفحات جمع‌آوری شد. از آنجا که کلمات کلیدی مذکور، معمولاً توسط نویسندگان مقالات که متخصصین حوزه علمی مربوطه هستند، به آن مقاله تخصیص یافته است، می‌توان کیفیت این مجموعه را مناسب ارزیابی کرد. مجموعه داده در یک فایل با پسوند Jsonline ذخیره شده است که هر خط آن نماینده یک داده و شامل کلیدهای متناظر با اطلاعات مقابل است: عنوان فارسی، عنوان انگلیسی، چکیده فارسی، چکیده انگلیسی، کلمات کلیدی فارسی و کلمات کلیدی انگلیسی و آدرس. مجموعه گردآوری‌شده حاوی کلمات بسیار متنوعی از حوزه‌های موضوعی مختلف است. جدول ۱ آمار این تنوع را نشان می‌دهد.

برای کاربرد مورد نظر در این مقاله، فقط به اطلاعات چکیده و کلمات کلیدی فارسی نیازمندیم. بنابراین زیرمجموعه‌ای شامل این دو کلید، آماده شده است که در ادامه به معرفی ویژگی‌های آماری مربوط به مجموعه خواهیم پرداخت.

### ۳-۲- پیش‌پردازش و آماده‌سازی داده‌ها برای کاربرد تولید عبارات کلیدی

همانطور که در بخش قبل نیز اشاره شد، مجموعه آماده‌شده برای این مقاله شامل چکیده‌های فارسی و کلمات کلید متناظرشان است. برای تنظیم و بازآموزی شبکه انتقالی MT5 و تولید عبارات کلیدی، نیاز است چکیده، به عنوان متن ورودی، و مجموعه عبارات کلیدی، به عنوان متن خروجی، به شکل مناسبی در اختیار شبکه قرار بگیرند. به این منظور، باید به گونه‌ای به شبکه نشان داد که در متن هدف، کدام توکن‌ها با هم تشکیل یک عبارت را داده‌اند. معمولا در مجلات علمی، کلمات در هر عبارت کلیدی با فاصله از هم جدا می‌شوند و وجود ویرگول به معنی پایان یک عبارت است. اما با توجه به اینکه فاصله در تمام متن، برای جداسازی کلمات یا همان توکن‌ها مورد استفاده قرار می‌گیرد، استفاده از آن برای نشان دادن ارتباط کلمات در یک عبارت مناسب به نظر نمی‌رسد. به همین منظور، تصمیم گرفتیم ارتباط توکنهای هر عبارت را با "_" یا همان زیرخط نشان دهیم و عبارت مختلف با یک فاصله از هم جدا شوند. این، یک مرحله از پیش‌پردازش را در روش پیشنهادی تشکیل می‌دهد که در ادامه شرح داده شده است.

قبل از انجام مرحله فوق، برای آماده‌سازی مجموع داده، چکیده‌هایی با طول کمتر از ۵۰ و بیشتر از ۱۰۰۰ در صورت وجود و همچنین چکیده‌هایی که تعداد کلمات کلیدی متناظرشان کمتر از دو کلمه بوده است حذف شده‌اند. به این ترتیب، تعداد متون مجموعه به ۵۶۴۲۲ کاهش یافت.

پس از حذف چکیده‌های نامطلوب در مرحله قبل، اعداد و ایست‌واژه‌ها، هم از چکیده‌ها و هم از مجموعه کلمات کلیدی حذف می‌شوند. سپس علائم هم از چکیده‌ها حذف می‌گردند. همچنین علامت "_" یا همان زیرخط در صورت وجود از کلمات کلیدی حذف می‌گردد. سپس فاصله میان کلمات منفرد در هر عبارت کلیدی با زیرخط جایگزین شده و پس از آن سایر علائم به غیر از زیرخط از کلمات کلیدی نیز حذف می‌گردد. به این ترتیب، اگر مجموعه کلمات کلیدی مربوط به یک چکیده به این صورت باشد:

مدیریت سازمانی، مدیریت رسانه، فرهنگ ارتباطات رسانه‌ای، روندپژوهی

پس از اعمال تغییرات مذکور، رشته کلمات کلیدی به صورت زیر در خواهد آمد. بر این اساس، وجود زیرخط در خروجی مدل هم به عنوان بخشی از یک عبارت کلیدی تفسیر خواهد شد و کاراکتر فاصله، جداکننده آن‌ها محسوب می گردد:

مدیریت_سازمانی مدیریت_رسانه فرهنگ_ارتباطات_رسانه‌ای روندپژوهی

تبدیل خروجی خام به شکل یک دنباله از عبارات کلیدی که با ویرگول از هم جدا شده‌اند، مرحله پس‌پردازش را در روش ما تشکیل می‌دهد. پس از انجام

پیش‌پردازش‌های اشاره‌شده در بالا، ویژگی‌های شمارشی و آماری توکن‌ها و عبارات (دنباله‌هایی از یک یا چند کلمه) در کلمات کلیدی و توکن‌ها (کلمات منفرد) در مجموعه چکیده‌ها محاسبه شده‌اند. نتایج این محاسبات در جداول ۲ تا ۴ خلاصه شده است. منظور از توکن‌های غایب در جدول ۲، توکن‌هایی است که در چکیده متناظر هر مجموعه از کلمات کلیدی به طور صریح آورده نشده‌اند. نرخ کلمات غایب $p(A, K)$ به این صورت محاسبه می‌شود که اشتراک بین توکن‌های هر چکیده و کلمات کلیدی متناظر آن محاسبه شده و بر تعداد توکن‌های چکیده تقسیم می‌گردد. رابطه (۱) این نحوه محاسبه را نشان می دهد.

$$p(A,K)=100\times\frac{|A\cap K|}{|A|} \qquad (۱)$$

در این رابطه، $A$ نماینده مجموعه توکن‌های چکیده و $K$ نماینده مجموعه توکن‌های کلمات کلیدی است. همین مقدار برای عبارات هم محاسبه و در جدول ۳ گزارش شده است.

**جدول ۲- ویژگی‌های آماری توکن‌ها در مجموعه کلمات کلیدی**

| عنوان ویژگی | مقدار |
|---|---|
| تعداد کل | ۴۲۲۶۲۴ |
| میانگین | ۷/۵ |
| انحراف معیار | ۳/۷۵ |
| بیشترین تعداد توکن | ۱۶۴ |
| کمترین تعداد توکن | ۱ |
| نرخ غایب | ۲۸/۸۶ |

**جدول ۳- ویژگی‌های آماری عبارات در مجموعه کلمات کلیدی**

| عنوان ویژگی | مقدار |
|---|---|
| تعداد کل | ۲۱۴۶۱۴ |
| میانگین | ۳/۸۱ |
| انحراف معیار | ۱/۳۱ |
| بیشترین تعداد عبارات | ۲۴ |
| کمترین تعداد عبارات | ۱ |
| نرخ غایب | ۴۳/۳۹ |

**جدول ۴- ویژگی‌های آماری توکن‌ها در مجموعه چکیده‌ها**

| عنوان ویژگی | مقدار |
|---|---|
| تعداد کل | ۶۹۶۱۲۳۳ |
| میانگین | ۱۲۳/۵۳ |
| انحراف معیار | ۴۱/۹۸ |
| بیشترین تعداد توکن‌ها | ۵۸۴ |
| کمترین تعداد توکن‌ها | ۲۵ |

## ۴- آزمایشات و نتایج

همانطور که پیش از این بیان شد، هدف ما جمع‌آوری یک پیکره مناسب برای تولید کلمات کلیدی فارسی و همچنین پیاده‌سازی یک روش تولید کلمات کلیدی چکیده‌ای با استفاده از یادگیری انتقالی است. پس از گردآوری و آماده‌سازی پیکره، چنانکه در بخش قبل توصیف شد، مجموعه داده به سه قسمت آموزش، اعتبارسنجی و آزمون تقسیم شد. مجموعه آموزش دارای ۵۲۴۲۲ سند است و مجموعه‌های آزمون و اعتبارسنجی، هریک دارای ۲هزار سند هستند. سپس از شبکه پیش‌آموخته MT5 استفاده کردیم و تنظیم و بازآموزی شبکه با استفاده از مجموعه آموزش انجام شد تا شبکه برای وظیفه تولید کلمات کلیدی آماده شود. برای تنظیم فراپارامترها در این مرحله، از مجموعه اعتبارسنجی استفاده نمودیم. در نتیجه، نرخ یادگیری در این آزمایشات 2e-4 در نظر گرفته شد. همچنین اندازه دسته ۲ در نظر گرفته شده است. پس از ساخت مدل، مجموعه آزمون به مدل حاصل ارائه شد و کلمات کلیدی برای هریک از اسناد (چکیده‌ها) موجود در این مجموعه تولید شدند.

برای انجام مقایسه، چندین روش استخراج کلمات کلیدی از بین روش‌های معرفی‌شده در بخش ۲، پیاده‌سازی شده و توسط آن‌ها، کلمات کلیدی برای همان مجموعه تولید شده‌اند. هریک از این روش‌ها با معیارهای دقت، بازیابی و معیار F۱ برای n کلمه اول هر مجموعه از کلمات کلیدی ارزیابی شده‌اند. لیست این روش‌ها به همراه نتایج ارزیابی آن‌ها در جدول ۵ و ۶ و ۷ آمده است. در جدول ۵، معیارهای دقت، بازیابی و F1 بر مبنای دو کلمه اول، در جدول ۶ بر مبنای پنج کلمه اول و در جدول ۷ برمبنای ۱۰ کلمه اول محاسبه شده‌اند.

**جدول ۵- نتایج تولید کلمات کلیدی بر روی مجموعه داده پیشنهادی با احتساب دو کلمه اول**

| مدل | P@2 | R@2 | F1@2 |
|---|---|---|---|
| TFIDF | ۴۵/۸۲ | ۱۲/۴۶ | ۱۹/۵۹ |
| Yake [17] | ۲۱/۴۲ | ۵/۴۸ | ۸/۷۳ |
| SR [19] | ۱۳/۲۵ | ۳/۶۷ | ۵/۷۵ |
| MR [22] | ۱۳/۹۵ | ۳/۸۱ | ۵/۹۹ |
| TR [21] | ۱۶/۹۲ | ۴/۴۹ | ۷/۱۰ |
| RNN2RNN[39] | ۴۷/۱۰ | ۱۳/۳۲ | ۲۰/۴۶ |
| AR[18] | ۲۹/۹۴ | ۸/۶۳ | ۱۳/۴۱ |
| MT5 [33] | **۵۹/۹۲** | **۱۷/۵۷** | **۲۷/۱۷** |

**جدول ۶- نتایج تولید کلمات کلیدی بر روی مجموعه داده پیشنهادی با احتساب پنج کلمه اول**

| مدل | P@5 | R@5 | F1@5 |
|---|---|---|---|
| TFIDF | ۳۰/۴۳ | ۲۰/۶۸ | ۲۴/۶۲ |
| Yake [17] | ۱۶/۹۹ | ۱۰/۸۹ | ۱۳/۲۷ |
| SR [19] | ۱۸/۱۱ | ۱۲/۱۶ | ۱۴/۵۵ |
| MR [22] | ۱۴/۱۱ | ۹/۴۷ | ۱۱/۳۴ |
| TR [21] | ۱۶/۲۹ | ۱۰/۹۱ | ۱۳/۰۷ |
| RNN2RNN[39] | ۳۷/۳۴ | ۲۵/۰۷ | ۳۰/۰۰ |
| AR[18] | ۲۴/۷۱ | ۱۵/۲۸ | ۱۸/۸۹ |
| MT5 [33] | **۳۸/۱۴** | **۲۸/۸۱** | **۳۲/۸۲** |

**جدول ۷- نتایج تولید کلمات کلیدی بر روی مجموعه داده پیشنهادی با احتساب ۱۰ کلمه اول**

| مدل | P@10 | R@10 | F1@10 |
|---|---|---|---|
| TFIDF | ۲۳/۱۸ | ۲۷/۱۹ | ۲۳/۱۸ |
| Yake [17] | ۱۲/۵۱ | ۱۶/۰۱ | ۱۴/۰۴ |
| SR [19] | ۱۷/۶۲ | ۲۳/۶۸ | ۲۰/۲۱ |
| MR [22] | ۱۲/۷۸ | ۱۶/۸۱ | ۱۴/۵۱ |
| TR [21] | ۱۳/۷۹ | ۱۷/۹۵ | ۱۵/۶۰ |
| RNN2RNN[39] | ۳۷/۰۳ | ۲۵/۳۷ | ۳۰/۱۱ |
| AR[18] | ۲۲/۸۹ | ۱۹/۵۳ | ۲۱/۰۸ |
| MT5 [33] | **۳۸/۰۹** | **۲۸/۸۴** | **۳۲/۸۲** |

همچنین یک روش دنباله‌به‌دنباله دیگر [۳۹] که در این مقاله از آن به عنوان مدل RNN2RNN یاد خواهد شد، برای مقایسه با روش انتقالی پیاده‌سازی شده است. این مدل، دارای ساختار رمزنگار-رمزگشا است. در این شبکه، برای رمزنگار و رمزگشا هردو از RNN[۸] استفاده شده است. این مدل، متن چکیده را به عنوان متن ورودی و دنباله کلمات کلیدی منتسب به هر چکیده را به عنوان متن هدف دریافت می‌کند. این شبکه، انتقالی نیست و به روش انتقالی آموزش نمی‌بینند. در واقع، این شبکه بدون هیچ دانش قبلی، فقط با مجموعه داده معرفی‌شده، آموزش می‌بینند.

همانطور که در جدول ۵ و جدول ۶ و جدول ۷ می‌بینید، بهترین نتایج برای تمامی معیارها توسط روش انتقالی به دست آمده است. این نکته قابلیت اینگونه روش‌ها را برای تولید کلمات کلیدی نشان می‌دهد.

در جداول فوق می‌بینید که معیار F1 با افزایش n از دو به پنج، افزایش داشته است، درحالی که از پنج به ده کاهش داشته است. این در حالی است که در هردو حالت میزان دقت روند کاهشی و میزان بازیابی روند افزایشی داشته است. برای درک علت این موضوع باید به این نکته توجه شود که تعداد کلمات کلیدی هدف در بسیاری از متون مجموعه داده ما کوچکتر از ۱۰ است و به همین دلیل گاهی با افزایش n در محاسبه دقت، صورت کسر که اشتراک بین کلمات کلیدی هدف و کلمات کلیدی تولیدشده است، با افزایش n، تغییری نمی‌کند ولی مخرج بزرگتر می‌شود که باعث کوچک شدن بیشتر مقدار می‌شود. در چنین مواردی بازیابی معمولا تغییر چندانی نمی‌کند. همین نکات باعث می‌شود که میزان کاهش دقت بیشتر از میزان افزایش بازیابی باشد و در نتیجه، مقادیر F1 در نتایج برای ۱۰ کمتر از ۵ باشند.

بعد از مدل پیشنهادی، بهترین نتیجه توسط مدل RNN2RNN تولید شده است. در واقع، این نتیجه نشان می‌دهد که مدل‌های دنباله‌به‌دنباله بهتر از روش‌های استخراجی عمل می‌کنند و مدل انتقالی موفق‌تر از مدل دنباله‌به دنباله دیگر عمل کرده است. این رخداد در واقع به این علت است که شبکه انتقالی، ابتدا روی یک مجموعه عظیم از زبان فارسی آموزش دیده است و روابط بین کلمات فارسی را خوب آموخته است یا به عبارت ساده‌تر، قادر است فارسی را بهتر بفهمد و بعد ما از این توانایی استفاده کرده و یک کاربرد جدید را به آن آموخته‌ایم. به همین دلیل است که شبکه انتقالی بهتر از شبکه RNN2RNN که آموخته‌های آن فقط برگرفته از مجموعه داده ما است، یا از مدل Attention Rank که فقط از مقادیر توجه تولیدشده توسط شبکه انتقالی در یک روال بی‌ناظر استفاده می‌کند، بهتر عمل کرده است.

بعد از نتایج عمیق بهترین نتایج در جداول فوق توسط روش TFIDF به دست آمده است و بعد از آن، بهترین نتیجه را روش Single-Rank تولید کرده است. جالب اینجاست که هم در مورد روش‌های آماری و هم در مورد روش‌های گرافی، هرچه که مدل ساده‌تر است نتایج بهتری تولید شده‌اند. در واقع به نظر می‌رسد اطلاعات فراوانی و همرخدادی کل کلمات متن بدون توجه به نقش دستوریشان، نتیجه بهتری را تولید کرده است. شبکه‌های عمیق انتقالی هم متکی بر همین اطلاعات هستند و یکی از دلایل عملکرد بهتر آن‌ها می‌تواند همین نکته باشد. از طرفی با توجه عملکرد بهتر TR از MR، به نظر می‌رسد اطلاعات موضوعی، یعنی در نظر گرفتن ارتباط بین خوشه‌های کلمات بجای اتکا به تک‌کلمات بهتر عملکرده است این نکته می‌تواند یک سرنخ خوب برای ادامه کار ما بر روی روش‌های دنباله‌به‌دنباله باشد. هرچند روش‌های دنباله‌به‌دنباله، به صورت سربه‌سر عمل می‌کنند و مرحله تعیین ویژگی‌ها به صورت جداگانه اجرا نمی‌شود ولی این نکته می‌تواند سرنخ موثری در استفاده مناسب از مکانیزم توجه و تعریف یک ساختار قوی برای شبکه باشد.

## ۵- نتیجه‌گیری

در این مقاله، یک پیکره‌ی جدید از مقالات علمی فارسی معرفی شده است. این پیکره، شامل ۷۰هزار چکیده از مقالات منتشرشده در حدود ۱۰۰ پایگاه مختلف مجلات علمی فارسی با دسترسی باز، به همراه کلمات کلیدی و تعدادی ویژگی دیگر از جمله عنوان فارسی، معادل‌های انگلیسی این ویژگی‌ها است. این پیکره شامل بیش از ۱۳ موضوع متنوع است که یک ویژگی مهم برای مجموعه داده‌هایی از این دست محسوب می‌گردد. دو ویژگی چکیده و کلمات کلیدی فارسی از این مجموعه برداشته شده و پس از انجام پیش‌پرداش‌هایی که شرح آن در مقاله رفته است، برای تولید کلمات کلیدی فارسی به روش چکیده‌ای آماده شده است. این زیرمجموعه از داده‌ها شامل ۵۶۴۲۲ چکیده است که ۵۴۴۲۲ نمونه از آن برای تنظیم و بازآموزی یک شبکه پیش‌آموخته انتقالی با عنوان MT5 استفاده شده است. مدل حاصل از این مرحله برای تولید کلمات کلیدی مجموعه آزمون مشتمل بر ۲هزار چکیده، استفاده شده است. این مدل همچنین با چند نمونه از مدل‌های موجود تولید کلمات کلیدی، بر روی مجموعه مذکور مقایسه شده است که نتایج این مقایسه برتری چشمگیر مدل MT5 را نشان می‌دهد.

## مراجع


[1] R. C. Belwal, S. Rai, and A. Gupta, "A new graph-based extractive text summarization using keywords or topic modeling," Journal of Ambient Intelligence and Humanized Computing, vol. 12, no. 10, pp. 8975–8990, 2021, doi: 10.1007/s12652-020-02591-x.

[2] H. Li, J. Zhu, J. Zhang, C. Zong, and X. He, "Keywords-Guided Abstractive Sentence Summarization," Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 8196–8203, Apr. 2020, doi: 10.1609/aaai.v34i05.6333.

[3] W. Wong, J. Thangarajah, and L. Padgham, "Contextual question answering for the health domain," Journal of the American Society for Information Science and Technology, vol. 63, no. 11, pp. 2313–2327, 2012, doi: 10.1002/asi.22733.

[4] A. Willis, G. Davis, S. Ruan, L. Manoharan, J. Landay, and E. Brunskill, "Key Phrase Extraction for Generating Educational Question-Answer Pairs," in Proceedings of the Sixth (2019) ACM Conference on Learning @ Scale, 2019, pp. 1–10, doi: 10.1145/3330430.3333636.

[5] A. Chaudhuri, N. Sinhababu, M. Sarma, and D. Samanta, "Hidden features identification for designing an efficient research article recommendation system," International Journal on Digital Libraries, vol. 22, no. 2, pp. 233–249, 2021, doi: 10.1007/s00799-021-00301-2.

[6] S. Riaz, M. Fatima, M. Kamran, and M. W. Nisar, "Opinion mining on large scale data using sentiment analysis and k-means clustering," Cluster Computing, vol. 22, no. S3, pp. 7149–7164, May 2019, doi: 10.1007/s10586-017-1077-z.


[۸] Recurrent Neural Network

[7] U. Rahardja, T. Hariguna, and W. M. Baihaqi, "Opinion Mining on E-Commerce Data Using Sentiment Analysis and K-Medoid Clustering," in 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media), 2019, pp. 168–170, doi: 10.1109/Ubi-Media.2019.00040.

[8] E. Cano and O. Bojar, "Keyphrase Generation: A Multi-Aspect Survey," in 2019 25th Conference of Open Innovations Association (FRUCT), 2019, vol. 5, pp. 85–94, doi: 10.23919/FRUCT48121.2019.8981519.

[9] S. Siddiqi, "Keyword and Keyphrase Extraction Techniques : A Literature Review," International Journal of Computer Applications, vol. 109, no. 2, pp. 18–23, 2015, doi: 10.5120/19161-0607.

[10] E. Papagiannopoulou and G. Tsoumakas, "A review of keyphrase extraction," WIREs Data Mining and Knowledge Discovery, vol. 10, no. 2, pp. 1–59, Mar. 2020, doi: 10.1002/widm.1339.

[11] E. Doostmohammadi, M. H. Bokaei, and H. Sameti, "PerKey: A Persian News Corpus for Keyphrase Extraction and Generation," in 2018 9th International Symposium on Telecommunications (IST), 2018, pp. 460–465, doi: 10.1109/ISTEL.2018.8661095.

[12] M. Mohseni and H. Faili, "Title Generation and Keyphrase Extraction from Persian Scientific Texts," in 2020 25th International Computer Conference, Computer Society of Iran (CSICC), 2020, pp. 1–6, doi: 10.1109/CSICC49403.2020.9050113.

[13] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning, "KEA: practical automatic keyphrase extraction" in Proceedings of the fourth ACM conference on Digital libraries - DL '99, 1999, pp. 254–255, doi: 10.1145/313238.313437.

[14] P. D. Turney, "Learning algorithms for keyphrase extraction," Information Retrieval, vol. 2, no. 4, pp. 303–336, 2000, doi: 10.1023/A:1009976227802.

[15] P. D. Turney, "Learning to Extract Keyphrases from Text," Dec. 1999.

[16] S. R. El-Beltagy and A. Rafea, "KP-Miner: A keyphrase extraction system for English and Arabic documents," Information Systems, vol. 34, no. 1, pp. 132–144, 2009, doi: 10.1016/j.is.2008.05.002.

[17] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes, and A. Jatowt, "YAKE! Collection-Independent Automatic Keyword Extractor," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10772 LNCS, pp. 806–810, 2018.

[18] H. Ding, and X. Luo, "AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions." In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1919-1928, 2021.

[19] X. Wan and J. Xiao, "CollabRank: Towards a collaborative approach to single-document keyphrase extraction," Coling 2008 - 22nd International Conference on Computational Linguistics, Proceedings of the Conference, vol. 1, no. August, pp. 969–976, 2008.

[20] R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004 - A meeting of SIGDAT, a Special Interest Group of the ACL held in conjunction with ACL 2004, 2004, vol. 85, pp. 404–411.

[21] A. Bougouin, F. Boudin, and B. Daille, "TopicRank: Graph-Based Topic Ranking for Keyphrase Extraction," in 6th International Joint Conference on Natural Language Processing, IJCNLP 2013 - Proceedings of the Main Conference, 2013, pp. 543–551.

[22] F. Boudin, "Unsupervised Keyphrase Extraction with Multipartite Graphs," in Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 2018, vol. 2, pp. 667–672, doi: 10.18653/v1/N18-2105.

[23] E. Çano and O. Bojar, "Two Huge Title and Keyword Generation Corpora of Research Articles," in LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, 2020, pp. 6663–6671.

[24] A. Sharifi and M. A. Mahdavi, "Supervised approach for keyword extraction from Persian documents using lexical chains," Signal and Data Processing, vol. 15, no. 4, pp. 95–110, Mar. 2019, doi: 10.29252/jsdp.15.4.95.

[۲۵] امید حاجی پور، سعیده سادات سدیدپور، «استخراج خودکار کلمات کلیدی متون کوتاه فارسی با استفاده از word2vec. »، پدافند الکترونیکی و سایبری، جلد ۸، شماره ۲، صفحات ۱۰۵-۱۱۴.

[26] E. Mehrabi, A. Mohebi, and A. Ahmadi, "Improved keyword extraction for persian academic texts using RAKE algorithm; case study: Persian theses and dissertations," Iranian Journal of Information Processing and Management, vol. 37, no. 1, pp. 197–228, 2021, doi: 10.52547/jipm.37.1.197.

[27] S. Rose, D. Engel, N. Cramer, and W. Cowley, "Automatic Keyword Extraction from Individual Documents," in Text Mining, Chichester, UK: John Wiley & Sons, Ltd, 2010, pp. 1–20.

[28] S. Lazemi, H. Ebrahimpour-Komleh, and N. Noroozi, "PAKE: a supervised approach for Persian automatic keyword extraction using statistical features," SN Applied Sciences, vol. 1, no. 12, pp. 1–4, 2019, doi: 10.1007/s42452-019-1627-5.

[29] H. Veisi, N. Aflaki, and P. Parsafard, "Variance-based features for keyword extraction in Persian and English text documents," Scientia Iranica, vol. 27, no. 3 D, pp. 1301–1315, 2020, doi: 10.24200/SCI.2019.50426.1685.

[30] B. Hejazi and J. A. Nasiri, "Keywords Extraction from Persian Thesis Using Statistical Features and Bayesian Classification," Language Related Research, vol. 12, no. 6, pp. 339–367, 2022, doi: 10.52547/LRR.12.6.11.

[۳۱] مریم باسره، ولی درهمی، سجاد ظریف‌زاده. «ارائه روشی برای استخراج خودکار عبارات کلیدی از اخبار وب پارسی»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۳، صفحات ۸۵۷-۸۶۶، ۱۳۹۶.

[۳۲] سعید دهقانی اشکذری، ولی درهمی، علی‌محمد زارع بیدکی، محمداحسان بصیری. «عقیده‌کاوی در زبان فارسی مبتنی بر یادگیری انتقالی»، مجله مهندسی برق دانشگاه تبریز، جلد ۵۰، شماره ۳، صفحات ۱۲۱۵-۱۲۲۴، ۱۳۹۹.

[33] Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer," in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 483–498, doi: 10.18653/v1/2021.naacl-main.41.

[34] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, pp. 1–67, Oct. 2019.

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., "Attention Is All You Need," Advances in neural information processing systems, vol. 8, no. 1, pp. 8–15, Jun. 2017.

[36] C. Rosset, "Turing-NLG: A 17-billion-parameter language model by Microsoft," Microsoft Blog, 2020. [Online]. Available: https://www.microsoft.com/en-us/research/blog/turing-nlg-a-17-billion-parameter-language-model-by-microsoft/.

[37] Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N. and Presser, S., "The Pile: An 800GB Dataset of Diverse Text for Language Modeling," Dec. 2020.

[38] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S.,, "Language models are few-shot learners," Advances in Neural Information Processing Systems, vol. 2020-Decem, no. NeurIPS, 2020.

[39] Meng, R., Zhao, S., Han, S., He, D., Brusilovsky, P. and Chi, Y., 2017. Deep keyphrase generation. arXiv preprint arXiv:1704.06879.