Language Related Research
E-ISSN: 2383-0816
https://lrr.modares.ac.ir
https://doi.org/10.52547/LRR.12.6.11

Tarbiat Modares University

Vol. 12, No. 6, Tome 66
pp. 339-367
January & February 2022

Keywords Extraction from Persian Thesis Using Statistical Features and Bayesian Classification

Behzad Hejazi[1] & *Jalal A. Nasiri*[2*]

Received: 23 October 2019
Received in revised form: 22 December 2019
Accepted: 13 January 2020

Abstract

Keyword extraction aims to extract words that are able to represent the corpus meaning. Keyword extraction has a crucial role in information retrieval, recommendation systems and corpora classification. In Persian language, keyword extraction is known as hard task due to Persian's inherent complication. In this research work, we aim to address keyword extraction with a combination of statistical and Machine Learning as a novel approach to this problem. First the required preprocessing is applied to the corpora. Then three statistical methods and Bayesian classifier was utilized to the corpora to extract the keywords pattern. Also, a post processing methods was used to decrease the number of True Positive outputs. It should be pointed out that the built model can extract up to 20 keywords and they will be compared with keywords in the corresponding corpus. The evaluation results indicate that the proposed method, could extract keywords from scientific corpora (Specifically Thesis and Dissertations) with a good accuracy.

Keywords: Extraction, Bayesian Classification, statistical features, preprocessing, post-processing

1. M.Sc Student of Artificial Intelligence, Islamic Azad University North Tehran Branch, Tehran, Iran; ORCID: https://orcid.org/0000-0001-6772-8715
2. Corresponding author: Department of Computational Linguistics, Information Science Research Department, Iranian Research Institute for Information Science and Technology (IRANDOC), Tehran, Iran; *Email: j.nasiri@irandoc.ac.ir,*
ORCID: https://orcid.org/ 0000-0003-1821-5037

1. Introduction

Automated keyword extraction is the process of identifying document terms and phrases that can appropriately represent the subject of our writing. With the proliferation of digital documents today, extracting keywords manually can be impractical. Many applications such as auto-indexing, summarization, auto-classification, and text filtering can benefit from this process since the keywords provide a compact display of the text. Automated keyword generation can be broadly classified into two categories: keyword allocation and keyword extraction.

In keyword allocation, a set of potential keywords is selected from a set of controlled vocabularies, while keyword extraction examines the words in the text. Keyword extraction methods can be broadly classified into four groups: statistical approaches, linguistic approaches, machine learning approaches, and hybrid approaches.

2. Literature Review

working on Persian words is a big challenge for the paucity of sufficient research. The inadequacy of text pre-processing programs has made it more complex than the Latin language. Also, the presence of large dimensions of input data is one of the challenges that has always arisen in such researches and this problem becomes more apparent due to the variety of Persian written forms (Gandomkar, 2017, p. 233:256). In Moin Maedi's article (2015, p. 34:42) A method for extracting keywords in Persian language is presented. This article extracts keywords from each text separately and without seeing another text as training data.

In the article by Mohammad Razaghnouri (2017, P. 16:27) using the Word2Vec method and the TIF-IDIF frequency, they created a question and answer system in Persian, which is a new work due to the use of Word2Vec in Persian. However, with size reduction techniques and Word2Vec, this 72% success rate can be enhanced in the future.

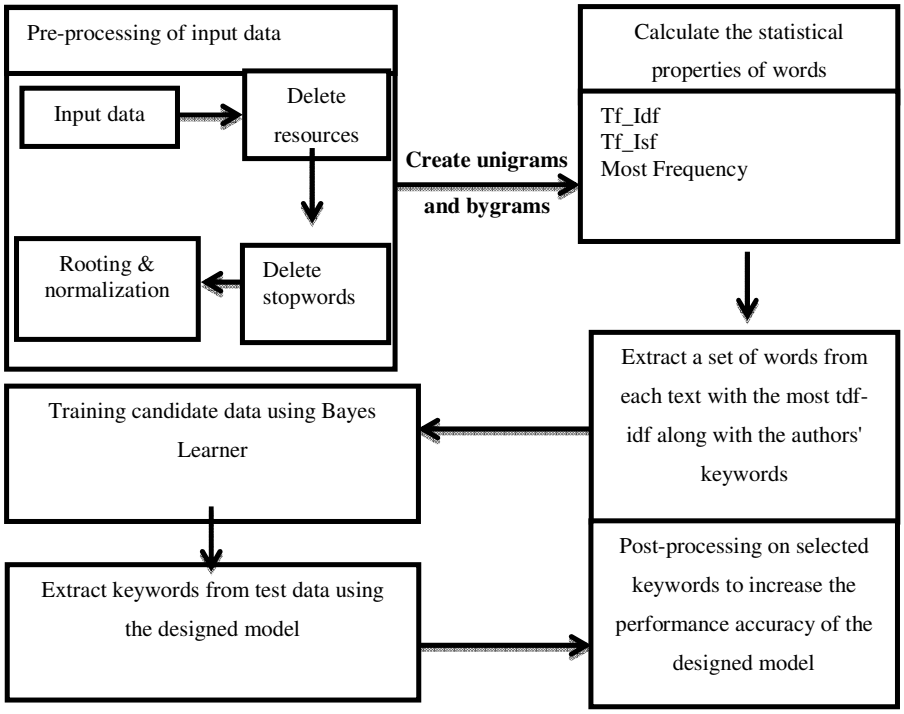
3. Methodology

Accordingly, the current paper examines the integration of statistical keyword extraction methods with the Naive Bayes Classifier. Initially, we integrated input texts which are dissertations in Persian by using preprocessing (deletion of stop words, etymology, etc.) methods. Then, using the available statistical features, each word has been given a certain weight. Then, the valuable words of each text were selected and the proposed model was taught using the selected category, then the selected words were processed by the trained model, and at the end, the words extracted from the final model were evaluated using the keywords suggested by the authors themselves. Figure 1 depicts all the steps performed.

4. Results

Literature review shows that this is the first time that these combinations are used to extract Persian keywords, so that unlike other studies, each text is as a sample for category input and words as its properties, however, in this paper the words of each text input are categorized and words are extracted using statistical methods that are considered as features. The choice of keywords by the authors has always been a personal decision and people may not make a single decision to choose a set of words for a single text.

Figure 1
Proposed research framework for keyword extraction



The current paper attempts to create a model and program with a new approach, due to the small number of input documents, which to extract keywords without dependence on the orientation of dissertations and the meaning of their words and only by using statistical features of words in each text. According to Tables 1 and 2, the developed model is able to extract a maximum of 20 keywords from each dissertation with an overall accuracy of 98.1%, in best condition which that is the use of a maximum frequency feature. The keywords written in each dissertation with 84% and 98% accuracy, correspond to one-word and two-word expressions, respectively.

Table 1
Evaluation criteria for Bayesian outputs in different states of statistical Features

Statistical Features	Accuracy	Recall	F1-Score	Precision
Tf_Idf, Most Frequent, Tf_Isf	97.2%	0.98	0.98	0.98
Most Frequent	98.1%	0.982	0.99	0.99
Tf_Idf, Tf_Isf	99.8%	0.91	0.94	0.99

Table 2
Evaluation of post-processing test data for outputs that have been categorized by keyword.

Step	Statistical Features	Recall	F1-Score	Precision	Number of words	Number of keywords that selected by writers
Uni-Grams	Most Frequent	0.84	0.323	0.2	210	42
By-Grams	Most Frequent	0.98	0.888	0.8	158	34

.

دوماهنامهٔ بین‌المللی
د۱۲، ش٦ (پیاپی ٦٦) بهمن و اسفند ۱٤۰۰، صص۳۳۹- ۳٦۷
مقاله پژوهشی
http://dorl.net/dor/20.1001.1.23223081.1400.12.6.29.5

استخراج کلیدواژگان پایان‌نامهٔ فارسی با استفاده از ویژگی آماری و دسته‌بند بیز

بهزاد حجازی[۱]، جلال‌الدین نصیری[*۲]

۱. دانشجوی کارشناسی ارشد هوش مصنوعی، دانشگاه آزاد اسلامی واحد تهران شمال، تهران، ایران.
۲. استادیار دانشکدهٔ علوم ریاضی، دانشگاه فردوسی، مشهد، ایران.

تاریخ دریافت: ۱۳۹۸/۰۸/۱ تاریخ پذیرش: ۱۳۹۸/۱۰/۲۳

چکیده

هدف اصلی استخراج کلمات کلیدی انتخاب مجموعه‌ای از لغات در متن است که می‌تواند موضوع اصلی متن را بازگو کند. استخراج کلیدواژگان در بازیابی اطلاعات، سیستم‌های پیشنهاددهندهٔ متنی و دسته‌بندی متون، نقش مهم را ایفا می‌کند. در زبان فارسی باتوجه به پیچیدگی ذاتی زبان فارسی استخراج کلیدواژگان به‌مراتب دشوارتر شده است. در این پژوهش سعی شده است با رویکرد نوین ترکیبی آماری و یادگیری ماشین به استخراج کلیدواژگان پرداخته شود. ابتدا باتوجه به ساختار زبان فارسی پیش‌پردازهای لازم برای حذف کلمات و علائم نگارشی صورت می‌گیرد. سپس با استفاده از سه نوع ویژگی آماری و دسته‌بند بیز سیستم به‌صورت خودکار الگوی کلمات کلیدی با کلمات عادی را آموزش می‌بیند. همچنین پس‌پردازش کارا برای کم کردن کلمات مثبت کاذب در چارچوب پیشنهادی طراحی شده است. گفتنی است که مدل ساخته‌شده قادر به شناسایی تعداد حداکثر ۲۰ کلیدواژه در هر پایان‌نامه است و این کلمات با کلیدواژگان نوشته‌شده در هر متن مقایسه و ارزیابی می‌شوند. نتایج ارزیابی‌های متنوع نشان می‌دهد روش پیشنهادی با دقت مناسبی توانسته است کلمات کلیدی نوشتارهای فارسی علمی (پایان‌نامه و رساله) را استخراج کند.

واژه‌های کلیدی: استخراج کلیدواژگان، دسته‌بند بیز، ویژگی‌های آماری، پیش‌پردازش، پس‌پردازش.

---

* نویسندهٔ مسئول مقاله: E-mail: j.nasiri@irandoc.ac.ir

۱. مقدمه

استخراج خودکار کلمات کلیدی فرایند شناسایی اصطلاحات و عبارات سندی است که به‌طور مناسب می‌تواند نمایندهٔ موضوع نوشتار ما باشد. امروزه با گسترش تعداد اسناد دیجیتال، استخراج کلمات کلیدی به‌صورت دستی می‌تواند یک موضوع غیرعملی باشد. از زمانی که کلیدواژگان نمایشی فشرده از متن را فراهم کرده‌اند، بسیاری از کاربردها مانند نمایه‌سازی خودکار، خلاصه‌سازی، طبقه‌بندی خودکار و فیلترینگ متن می‌توانند از این فرایند سود ببرند.

تولید کلمات کلیدی خودکار را می‌توان به‌طور گسترده به دو دسته تقسیم کرد: تخصیص کلیدواژه و استخراج کلمات کلیدی. در تخصیص کلمهٔ کلیدی، مجموعه‌ای از کلمات کلیدی محتمل از مجموعه‌ای واژگان کنترل‌شده انتخاب می‌شوند، درحالی که در استخراج کلیدواژه، کلمات موجود در نوشتار موردبررسی قرار می‌گیرند. روش‌های استخراج کلیدواژه می‌توانند به‌طور عمده به چهار گروه تقسیم شوند: رویکردهای آماری، رویکردهای زبانی، رویکردهای یادگیری ماشین و رویکردهای ترکیبی.

دسته‌بندی متن بخش مهمی در داده‌کاوی است که متنی را به یک یا چند کلاس یا دسته‌های ازپیش‌تعیین‌شده اختصاص می‌دهد. چندین شکل از مجموعه‌های متنی مانند مقالات خبری، کتابخانه‌های دیجیتالی و صفحات وب منابع مهم اطلاعات هستند. ازاین‌رو، طبقه‌بندی متن یک جهت تحقیقاتی مهم در علوم کتابداری، اطلاع‌رسانی و علوم کامپیوتر است. بسیاری از کاربردهای داده‌کاوی را می‌توان به‌منزلهٔ یک مشکل طبقه‌بندی متن مدل‌سازی کرد. این کاربردها فیلترینگ اخبار، سازماندهی اسناد، بازیابی، تحلیل احساسات، فیلترینگ هرزنامه‌ها و استخراج کلیدواژگان است.

به‌دلیل نبود پژوهش‌های کافی، کار بر روی کلمات فارسی چالشی بزرگ است. کافی نبودن برنامه‌های مرتبط با پیش‌پردازش متن این امر را نسبت‌به زبان لاتین پیچیده‌تر کرده است. همچنین، وجود ابعاد بزرگ داده‌های ورودی یکی از چالش‌هایی است که همواره در این گونه پژوهش‌ها به‌وجود آمده است و به‌دلیل تنوع در صورت‌های نگارشی زبان فارسی (گندمکار، ۱۳۹۶، صص. ۲۳۳ـ ۲۵۶) این مشکل، بیشتر نمایان می‌شود.

باتوجه به مسائل یادشده، این مقاله به بررسی ادغام روش‌های استخراج آماری کلمات کلیدی با دسته‌بند بیز ساده می‌پردازد. در ابتدا با استفاده از پیش‌پردازش (حذف

ایست‌واژه‌ها، ریشه‌یابی و ...) به یکپارچه‌سازی متن‌های ورودی که پایان‌نامه‌هایی به زبان فارسی است، پرداخته شده است. سپس با استفاده از ویژگی‌های آماری موجود، به هر کلمه وزن مشخصی اعطا شده است. در مرحلهٔ بعد کلمات باارزش هر متن انتخاب و با استفاده از دسته‌بند انتخابی به آموزش مدل مطرح‌شده پرداخته شده است. سپس پس‌پردازش کلمات انتخاب‌شده توسط مدل آموزش‌دیده صورت می‌پذیرد. در انتهای کار با بهره‌گیری از کلیدواژگان مطرح‌شده توسط خود نویسندگان به ارزیابی کلمات استخراج‌شده از مدل نهایی پرداخته شده است.

باتوجه به مطالعات انجام‌شده، این اولین‌بار است که برای استخراج کلیدواژگان فارسی از این ترکیب‌ها استفاده می‌شود، به‌طوری که برخلاف آنکه همانند سایر پژوهش‌های صورت‌گرفته، هر متن به‌منزلهٔ یک نمونه برای ورودی دسته‌بند و کلمات به‌منزلهٔ ویژگی‌های آن باشند، در این پژوهش کلمات هر متن ورودی‌های دسته‌بند هستند و با بهره‌گیری از روش‌های آماری که به‌منزلهٔ ویژگی‌ها درنظر گرفته شده‌اند، به استخراج کلمات پرداخته می‌شود. در این پژوهش فرض بر این است که با بهره‌گیری از اطلاعات آماری داده‌های ورودی، مدلی هوشمند برای استخراج کلیدواژگان با دقت بالا ساخته شده است.

ادامهٔ مقاله بدین شکل سازمان‌دهی شده است که در بخش دوم به بررسی پژوهش‌های انجام‌شده در زبان لاتین و فارسی پرداخته شده است. در بخش سوم به شرح چارچوب پیشنهادی پرداخته و قسمت‌های مختلف برنامه تعریف شده است. در بخش چهارم نتایج عملی موردبررسی قرار می‌گیرد. بدین شکل که نحوهٔ جمع‌آوری داده‌ها و عملیاتی که به روی آن‌ها صورت می‌پذیرد به‌طور کامل شرح داده می‌شود و نتایج آن موردتحلیل قرار می‌گیرد. در بخش پنجم نتیجه‌گیری مقاله موردبحث قرار گرفته و در بخش آخر به ایده‌های آینده برای انجام پژوهش‌های مختلف در این مبحث پرداخته شده است.

۲. مرور ادبیات

۲ـ۱. پژوهش‌های انجام‌شده در حوزهٔ زبان لاتین

در تازه‌ترین پژوهش‌هایی که در این حوزه انجام شده می‌توان به آتانو دی[1] (2018, p.92, 105) اشاره کرد که در آن باتوجه به تجزیه و تحلیل احساسات به ارزیابی عملکرد

محصولات و خدمات، از محتویات تولیدشده توسط کاربر می‌پردازد. رویکردهای تجزیه و تحلیل احساسات مبتنی‌بر زنجیره‌های نحوی، زمانی که اطلاعات آموزشی کافی نیست برمبنای یادگیری ترجیح داده می‌شوند. واژگان موجود فقط دارای یگانگی و نمرهٔ احساسات خود هستند. مشاهده می‌شود که در استخراج عبارات چندکلمه‌ای، کلمه‌های احساس‌شده به‌وسیلهٔ ترکیبی از تک‌کلمه‌ای‌ها[۲] با تقویت‌کننده‌ها نتایج بهبودیافته را نشان می‌دهد. چنین واژگان عاطفی در دسترس عموم نیست. این مقاله روشی را برای ایجاد عباراتی به‌نام جملات چندبخشی[۳] ارائه می‌دهد. رویکرد مبتنی‌بر قاعدهٔ پیشنهادی، نمرات احساسات چندبخشی را از یک مجموعهٔ تصادفی شامل بررسی محصول و رتبه‌بندی عددی مربوطه در مقیاس پنج‌گانه استخراج می‌کند. به‌علاوهٔ روش‌هایی که گفته شد، قوانین اضافی دیگری نیز وجود دارند که می‌توانند به همراه روش‌های گفته‌شده به‌کار گرفته شوند و کیفیت عبارات کاندید را بهبود بخشند. چنین ایده‌هایی در مقالهٔ رافائل[۴] (2010, p.190, 193) به‌کار گرفته شده است. در روش آن‌ها اگر یک عبارت در ٤٠٠ کلمهٔ اول متن ظاهر نشود آنگاه به‌منزلهٔ عبارت کاندید جهت پیش‌بینی کلیدواژگان درنظر گرفته نمی‌شود. ایدهٔ پشت شرط آن‌ها این است که اگر یک عبارت در ابتدای متن به‌کار نرود پس احتمال کمی وجود دارد که عبارت کلیدی باشد. یک رویکرد مشابه توسط نیومن[۵] (2010, p.150, 153) به‌کار گرفته شده که محدودهٔ استخراج عبارات کاندید را به ۲۰۰۰ کلمهٔ اول هر متنی محدود می‌کند. در مقاله لانگ[۶] (2010, p.21, 26) نیز رویکردی به‌کار گرفته شده که در آن، انتخاب عبارات کاندید را به بخش‌های چکیده، مقدمه، کارهای مرتبط، عناوین موجود در متن و اولین خط هر پاراگراف برای ادامهٔ متن محدود می‌کنند. این روش کیفیت نتایجشان را بهبود بخشیده است.

لیتواک[۷] (2008, p.17, 24) به بررسی کارایی روش‌های استخراج کلمات کلیدی متمرکز و بدون نظارت بر استخراج خلاصهٔ اسناد پرداختند. برای نشان دادن اسناد متنی، نمایندگی نحو براساس گراف استفاده شده است.

در مطالعهٔ دیگری هوان[۸] (2006, p.275, 284) یک الگوریتم استخراج کلمات کلیدی خودکار ارائه داد که می‌توان آن را در هر دو روش یادگیری بدون نظارت و تحت‌نظارت استفاده کرده است. الگوریتم ارائه‌شده هر سند متن را به‌منزلهٔ شبکه‌ای معنایی مدل می‌کند. عبارات کلیدی براساس داینامیک ساختاری شبکهٔ معنایی استخراج می‌شوند.

ترنی[۹] (2002) روش استخراج کلمات کلیدی را به‌منزلهٔ یک مسئلهٔ یادگیری ماشین نظارت‌شده ارائه کرد. در این روش از ۹ ویژگی برای امتیاز دادن به عبارات کاندید استفاده شده است. سپس عبارات کلیدی براساس ویژگی‌ها از میان کاندیده‌ها استخراج می‌شوند. ویتن[۱۰] (1999) یک الگوریتم استخراج کلمات کلیدی ساده و کارآمد (KEA) را ارائه داد که از الگوریتم بیز ساده برای استخراج کلمات کلیدی مبتنی‌بر دامنه استفاده می‌کند. در این روش، عبارات کلیدی ممکن باید با روش‌های لغوی تعیین شود و عبارات کلیدی خوب توسط الگوریتم یادگیری ماشین به‌دست آید. کمر[۱۱] (2005, p.657, 669) اثربخشی روش‌های خودکار استخراج کلیدواژه‌ها و روش‌های یادگیری را در مقالات علمی به زبان انگلیسی موردبررسی قرار دادند. روش‌های استخراج کلمات کلیدی با روش‌های مختلف یادگیری ماشین ارزیابی می‌شوند و نتایج تجربی نشان می‌دهد که الگوریتم C4.5 دارای بالاترین نرخ پیش‌بینی برای دامنه است. در تحقیقات عملی صورت‌گرفته در اُنان[۱۲] (2016, p.232, 247) که برای استخراج کلمات کلیدی بر روی کتابخانهٔ دیجیتال ACM صورت گرفته است از ترکیب روش‌های مختلف و مقایسهٔ آن‌ها با هم پرداخته است که بهترین نتایج مربوط به ترکیب الگوریتم‌های دسته‌بندی‌کننده[۱۳] و جنگل تصادفی[۱٤] است.

۲ـ۲. پژوهش‌های انجام‌شده در حوزهٔ زبان فارسی

همچنین مقالاتی در این حوزه در زبان فارسی به چاپ رسیده است که درادامه به اختصار توضیحاتی درموردشان داده می‌شود. در مقاله معین ماعدی[۱۵] (2015, p.34, 42) روشی برای استخراج کلمات کلیدی در زبان فارسی ارائه شده است که این مقاله به استخراج کلیدواژگان از هر متن به صورت جداگانه و بدون دیدن متن دیگری به‌عنوان داده‌های آموزش، می‌پردازد. روش پیشنهادی مقاله با ترکیب ویژگی‌های آماری، بردار موقعیت تکرار را برای هر کلمه و قوانین سادهٔ زبان ارائه داده است. روش پیشنهادی در سه مرحله ارائه شده است که پیش‌پردازش متن، وزن‌دهی به کلمات، ایجاد بردار موقعیت تکرار کلمه در متن و محاسبهٔ شباهت و استفاده از اسناد تشکیل‌شده برای این مطالعه در زبان فارسی موردارزیابی قرار گرفته است. از این روش نتیجه می‌شود که زبان فارسی را می‌توان به‌منزلهٔ دستاورد جدید و امیدوارکننده برای زبان‌های دیگر درنظر گرفت. باوجوداین، در

بررسی نتایج حاصل از نتایج به‌دست آمده می‌توان مشکلات زیر را در بررسی نتایج به‌دست آورد: ۱. انتخاب کلمات توصیفی به‌جای کلمات کلیدی توسط نویسنده، ۲. استفاده از اختصار انگلیسی به‌جای کلمهٔ انتخاب‌شدهٔ فارسی. مانند کلمهٔ «بردارهای ماشین پشتیبان» در کلمات کلیدی، اما استفاده از کلمهٔ «SVM» در متن، ۳. استفادهٔ کوچک در متن یا به‌منزلهٔ اشارهٔ مختصر به‌ویژه کلمات مهم در مقالات، معمولاً در ابتدای متن مانند عنوان، انتزاعی یا مقدمهٔ بیشتر برای حضور وجود دارد و اغلب نویسندگان به ضمایر یا مرجع آن‌ها اشاره دارند، ٤. وجود کلمات و عبارات مربوط به سبک نوشتاری نویسنده، ٥. ایجاد ابهام در مرحلهٔ تبدیل فایل PDF به متن، ٦. احترام نگذاشتن به زبان فارسی.

در مقالهٔ محمد رزاغنوری[۱۶] (2017, p.16, 27) با استفاده از روش ورد تو وک[۱۷] و بسامد تی‌اف ـ ای‌دی‌اف سیستم پرسش و پاسخ را در زبان فارسی به‌وجود آوردند که باتوجه به استفاده از ورد تو وک در زبان فارسی کاری تازه بوده است، گرچه با تکنیک‌های کاهش ابعاد و ورد تو وک می‌توان این ۷۲ درصد میزان موفقیت را در آینده ارتقا داد.

در مقاله بهنام ثابتی[۱۸] (2018) به شرح میراث‌تکس[۱۹] پرداخته‌اند که می‌شود گفت بزرگ‌ترین مجموعه اسناد در حوزهٔ زبان فارسی است که دارای بیش از ۲/٤ میلیون سند و دارای بیش از ۱/٤ میلیارد کلمه است، اگرچه هنوز به نسبت زبان انگلیسی دارای مقدار کم‌تری است. در مقالهٔ مرتضی اخوت[۲۰] (2010, p.94,101) برای برچسب‌گذاری متون فارسی نحوی از روش مارکوف[۲۱] استفاده کردند. به‌طوری که برای برچسب‌گذاری کلمات فارسی ابتدا متن فارسی را به انگلیسی ترجمه می‌کردند، سپس متن انگلیسی را برچسب‌گذاری می‌کردند. ایراد این روش این است که در حین ترجمه یک سری کلمات نامشخص باقی می‌ماندند که دقت برای این کلمات پایین است. اما برای کلماتی که درست ترجمه شده‌اند که بخش عمده‌ای از متن هستند دارای دقت ۹۸ درصد است و به‌طور میانگین با احتساب کلمات درست و غلط تشخیص داده شده دارای دقت ۹۷درصد است.

محمد[۲۲] (2011, p.112,116) با استفاده از متن‌های روزنامهٔ همشهری به استخراج کلمات کلیدی یک‌کلمه‌ای و دوکلمه‌ای پرداخته است. اما جواب‌هایی که استخراج شده مورد ارزیابی واقع نشده است. حال که بررسی پژوهش‌های پیشین انجام شده است، نوآوری‌های انجام شده در این پژوهش موردتحلیل قرار گرفته است. این پژوهش سه نوآوری بارز دارد که

بدین شرح است:

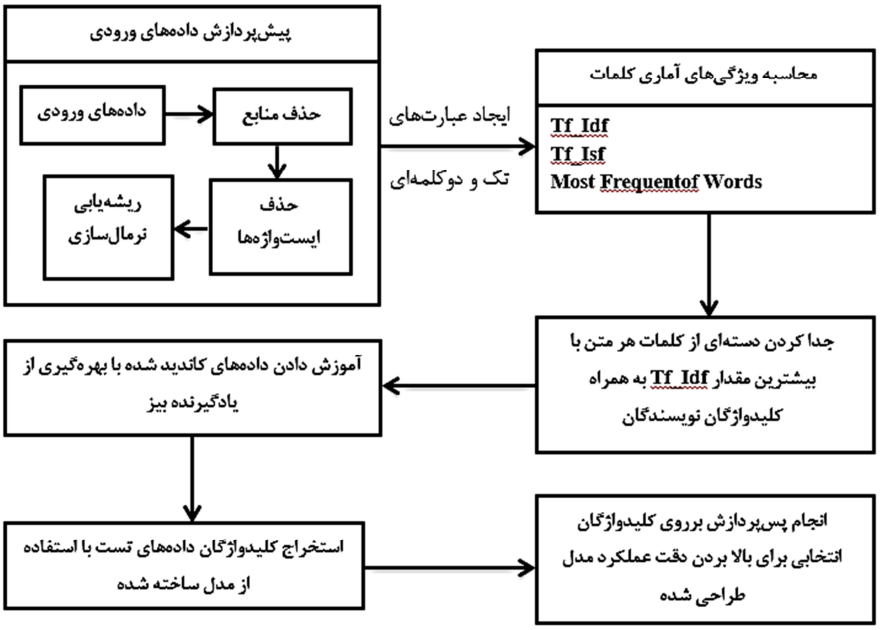
۱) استخراج کلیدواژه‌ها برمبنای ویژگی‌های آماری هر کلمه به جای ویژگی‌های هر سند،

۲) کاربردی‌سازی الگوریتم‌های استخراج کلیدواژهٔ لاتین برای متون علمی فارسی باتوجه به پیچیدگی‌های زبان علمی فارسی،

۳) طراحی پس‌پردازش مناسب جهت کارا کردن چارچوب پیشنهادی باتوجه به ویژگی‌های زبان فارسی.

۳. چارچوب پیشنهادی

در این قسمت راهکار پیشنهادی برای استخراج کلیدواژگان از متون علمی توضیح داده شده است. نمودار مراحل پیشنهادی در شکل ۱ به‌تصویر کشیده شده است. همان‌گونه که از شکل مشخص است، ابتدا داده‌های ورودی دریافت شده‌اند، سپس پیش‌پردازش‌های لازم بر روی داده‌ها انجام شد. در گام بعدی عبارت‌های تک‌کلمه‌ای و دوکلمه‌ای ایجاد می‌شوند و ویژگی‌های آماری آن‌ها به‌دست می‌آید. سپس انتخاب کلمات برای داده‌های آموزش انجام شده است. در مرحلهٔ بعد به‌وسیلهٔ کلمات انتخابی و یادگیرندهٔ بیز به استخراج کلمات از داده‌های تست پرداخته شده است. سپس پردازش‌های لازم بر روی کلمات انتخابی مدل یادگیرنده پرداخته می‌شود تا بدین وسیله دقت عملکرد برنامه بهبود یابد و کلمات کلیدی برنامه استخراج شوند. در گام آخر، با بهره‌گیری از معیارهای ارزیابی استاندارد تعریف‌شده که در ادامهٔ مقاله به شرح آن‌ها نیز پرداخته شد، سعی می‌شود دقت عملکرد مدل ساخته‌شده نمایش داده شود.

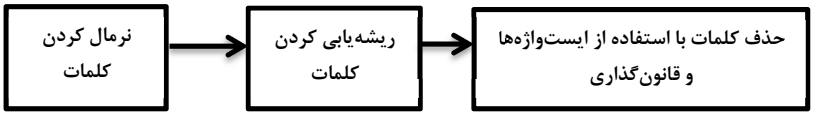


شکل ۱: چارچوب پیشنهادی پژوهش برای استخراج کلیدواژگان

Figure 1: Proposed research framework for keyword extraction

۳ـ۱. پیش‌پردازش متن

پیش‌پردازش متن به مفهوم یکدست کردن متن، شناسایی و ازبین بردن کلماتی در متن است که به‌منزلهٔ واژه‌ای مهم در متن تلقی نمی‌شوند و عموماً این کلمات در تمامی اسناد موجود است و باعث تمایز نوشته‌ها با یکدیگر نمی‌شوند.



شکل ۲: نمودار مراحل پیش‌پردازش داده‌ها

Figure 2: Chart of data preprocessing steps

برای پیش‌پردازش متن می‌شود از ریشه‌یابی و نرمال کردن کلمات استفاده کرد. این امر باعث می‌شود تا کلمات تبدیل به ریشه‌شان شوند و تمامی فاصله‌ها و نیم‌فاصله‌ها استاندارد شوند. برای انجام این کار از کتابخانهٔ هضم در زبان برنامه‌نویسی پایتون استفاده شده است. این کتابخانه توسط گروه سبحه ساخته شده و به‌صورت رایگان در اختیار کاربران قرار گرفته است. نمونه‌ای از تغییرات کلمات، قبل و بعد از استفاده از این برنامه در جدول ۱ آمده است.

جدول ۱: مقایسهٔ کلمات قبل و بعد از ریشه‌یابی توسط برنامهٔ هضم

Table 1: Compare words before and after rooting by Hazm program

نمونهٔ اولیهٔ کلمات	کلمات نرمال و ریشه‌یابی‌شده
خصوصی‌سازی	خصوص‌ساز
سازمانی	سازمان
تکنیک‌های داده‌کاوی	تکنیک داده‌کاو
دانش‌آموزان	دانش‌آموز

یکی از ایرادات این قسمت آن است که این برنامه کلماتی مانند «استخوان» را نیز تبدیل به «استخو» می‌کند، اما به‌دلیل زیاد بودن واژگان و حجم زیاد ابعاد به‌اجبار باید از این برنامه استفاده کرد.

ایست‌واژه‌ها کلماتی مانند حروف اضافه و افعال هستند که در پیش‌پردازش از متن ما حذف می‌شوند. نمونه‌ای از این لغات که در این پژوهش مورداستفاده قرار گرفته است در جدول ۲ آمده است.

جدول ۲: نمونه‌ای از کلماتی که به‌منزلهٔ ایست‌واژه به‌کار گرفته شده‌اند

Table 2: An example of words used as a StopWord

از	اگه	به	و	رفت	آمد	آنکه	شکل	انجام	همچون	جدول	=
این	ولی	آن	همچنین	را	شد	تا	می	سپس	نمودار	+	تا

در این قسمت بعضی مواقع برای یکپارچگی متن قوانینی توسط خود پژوهشگران علاوه‌بر حذف ایست‌واژه‌ها نیز وضع می‌شود که در این پژوهش نیز شرط‌هایی برای واژگان

گذاشته شده، که شامل این موارد است: ۱) حذف اعداد، ۲) حذف علائم نگارشی، ۳) حذف منابع و صفحات ابتدایی متون که شامل تقدیر و تشکر و تعهدنامه‌ها هستند.

یکی از موارد مشکل‌ساز برای انجام این کار در زبان فارسی به‌طور نمونه آن است که زبان‌های برنامه‌نویسی به‌طور مستقیم قادر به شناسایی و حذف اعداد به زبان فارسی نیستند و باید ابتدا آن‌ها را به زبان انگلیسی برگرداند و سپس به حذف آن‌ها اقدام کرد.

۳ـ۲. استخراج عبارات تک‌کلمه‌ای و دوکلمه‌ای

اولین اقدامی که در شناسایی عبارات کاندید درنظر گرفته می‌شود آن است که آیا کلمات تک استخراج شوند یا دنباله‌ای از کلمات. استخراج عبارات چندکلمه‌ای یا همان N-gram یکی از روش‌های رایج در استخراج عبارات کلیدی است. مفهوم و هدف N-gram استخراج گروه‌هایی از n کلمه‌ای پشت‌سر هم در یک جمله است. متغیر n تعداد کلمات را تعیین می‌کند. در این پژوهش باتوجه به آنکه اکثریت کلیدواژگان مطرح‌شده توسط نویسندگان یک‌کلمه‌ای و دوکلمه‌ای هستند، عبارات تک‌کلمه‌ای و دوکلمه‌ای استخراج شده‌اند.

۳ـ۳. وزن‌دهی به کلمات

در این بخش با استفاده از ویژگی‌های آماری به کلمات مربوط به هر سند وزن مشخصی تخصیص داده می‌شود. در این روش برای تعیین کلمات کلیدی از روش‌های آماری فرکانس کلمه و معیار TF-IDF و معیار TF-ISF استفاده شده است. روش استخراج کلیدی آماری می‌تواند مستقل از دامنه باشد و نیازی به اطلاعات آموزشی نیز نداشته باشد (Beliga, 2015, vol.1/p.20).

۳ـ۳ـ۱. بسامد کلمه ـ معکوس بسامد سند [۲۳]

این روش فرکانس کلمه در یک سند را ارائه می‌دهد. معیار بهتر این است که فرکانس کلمه در سند با مقایسهٔ ارتباط آن با اسناد دیگر به‌دست آورده شود. این معیار بسامد کلمه ـ معکوس بسامد سند یا به اختصار تی.اف ـ آی.دی.اف نام دارد که میزان خاص بودن کلمه در متن موردنظر را نشان می‌دهد. برای محاسبهٔ آن از فرمول زیر استفاده می‌شود (Lott,

.(2012

$$TFIDF(t,d,n,N)=TF(t,d)\times IDF(n,N) \quad (۱)$$

تی ـ اف از رابطهٔ زیر و از تقسیم فرکانس کلمه بر تعداد کل کلمات سند به‌دست می‌آید. ای ـ دی ـ اف نیز از رابطهٔ زیر محاسبه می‌شود که در آن، N تعداد اسنادی از پیکره است که دارای کلمهٔ موردنظر است و n تعداد کل اسناد پیکره است.

$$TF(t,d)=\frac{m_{td}}{\sum_{k} m_{kd}} \quad (۲)$$

$$IDF(n,N)=\log\frac{(N+1)}{(n+1)} \quad (۳)$$

با استفاده از این معیار کلماتی از یک سند که در اسناد دیگر کم‌تر استفاده شده‌اند امتیاز بالاتری دارند.

۳ـ۳ـ۲. استخراج کلمات کلیدی براساس تکرار[۲۴]

شایع‌ترین روش استخراج کلمهٔ کلیدی (MF) است، مؤلفه‌های متداول در متن را به‌منزلهٔ کلمات کلیدی مشخص می‌کند. برای نشان دادن اسناد متن، ماتریس اصطلاحی استفاده می‌شود. در این بازنمایی، مقدار رخداد اصطلاح برای یک جملهٔ خاص ارزش یک یا صفر را به ترتیب مربوط به وقوع یا عدم رخداد اصطلاح در آن جمله می‌گیرد. براساس این نمایندگی، نمرهٔ فرکانس برای هر اصطلاح با محاسبهٔ تعداد وقایع اصطلاح در ماتریس محاسبه می‌شود. برای هر اصطلاح (t_k)، نمرهٔ فرکانس به‌صورت زیر تعریف می‌شود (Rossi, 2014, p.17, 37).

$$MF(t_k)=\sum_{s_l\in S} occurence_{t_k,s_l} \quad (٤)$$

۳ـ۳ـ۳. بسامد کلمه ـ معکوس بسامد جمله[۲۵]

بسامد کلمه ـ معکوس بسامد جمله اندازه‌ای آماری است که سازگاری با «TF-IDF» را به جملات متن می‌دهد. در اصطلاح TF-ISF، استخراج کلمات کلیدی براساس فرکانس هر جمله از متن سند به‌منزلهٔ یک بردار از وزن درنظر گرفته شده است (Neto, 2000, p.41, 55). در اندازه‌گیری TF-ISF، نرخ t_k از طریق فراوانی معکوس جملهٔ s به‌وسیلهٔ فراونی محاسبه

می‌شود (Fiori, 2014):

$$\text{TF-ISF(S)} = \sum_{t \in S} freq\,(t_k) \times isf\ (t_k) \qquad (۵)$$

$$isf\ (t_k) = 1 - \frac{\log(n(t_k))}{\log(n)} \qquad (۶)$$

جایی که n نشان‌دهندهٔ تعداد جملات در مجموعه اسناد است. در این پژوهش هر کلمه یک جمله درنظر گرفته شده است.

۳ـ۴. استفاده از یادگیرندهٔ بیز ساده

استخراج عبارات کلیدی می‌تواند به‌منزلهٔ فرایند یادگیری ماشین درنظر گرفته شود. این‌گونه روش‌ها در کل به‌صورت زیر عمل می‌کنند: در ابتدا یک مجموعه از سندهای آموزشی برای سیستم فراهم می‌شود که کلمات کلیدی به‌صورت دستی برای هر کدام از سندهای متنی این مجموعه انتخاب شده است. سپس سیستم با استفاده از این مجموعه آموزش داده می‌شود و بعد از آن سیستم این توانایی را دارد که برای اسناد جدید، کلمات کلیدی انتخاب کند. در روش‌های یادگیری ماشین یک الگوریتم یادگیری، مانند ماشین‌های بردار پشتیبانی، بیز ساده، درخت تصمیم‌گیری، برای ساخت یک مدل طبقه‌بندی استفاده می‌شود. در ساخت مدل، یک مجموعه آموزش اسناد با برچسب استفاده می‌شود و مدل دارای یک مجموعه آزمون از اسناد معتبر است.

بیز ساده[۲۶]

الگوریتم بیز یک الگوریتم یادگیری آماری است که براساس قضیهٔ بیز است. این الگوریتم در زمینهٔ معرفی، استفاده و یادگیری دانش احتمالی واضح است. براساس طبقه‌بندی فرض استقلال شرطی آن است که هزینه‌های محاسباتی موردنیاز را ساده می‌کند. ازاین‌رو، الگوریتم می‌تواند به‌خوبی قیاس کند و به‌راحتی می‌تواند در تعدادی از حوزه‌ها، ازجمله مشتق متا استفاده شود. این الگوریتم دقت بالا و سرعت را در مجموعه‌داده‌های بزرگ به ارمغان می‌آورد و نتایج مشابهی را با الگوریتم‌های طبقه‌بندی دیگر مانند تصمیم‌گیری درختان و شبکه‌های عصبی ارائه می‌دهد.

همسایگی کلاس احتمال پیشین

$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)} \quad (۷)$$

پیش‌بینی احتمال پیشین

$$P(c \mid x) = P(x_1 \mid c) \times P(x_2 \mid c) \times ... \times P(x_n \mid c) \times P(c) \quad (۸)$$

۳ـ۵. معیارهای ارزیابی

برای ارزیابی عملکرد پیش‌بینی‌شده در روش‌های استخراج کلمات کلیدی آماری، الگوریتم‌های طبقه‌بندی و روش‌های ترکیبی از دقت طبقه‌بندی، اندازه‌گیری F و منطقهٔ تحت‌منحنی (AUC) به‌منزلهٔ معیارهای ارزیابی استفاده می‌شود. دقت طبقه‌بندی (ACC) یکی از معیارهای به‌کاررفته در بررسی طبقه‌بندی‌هاست. این نسبت، نسبت مثبت واقعی و منفی واقعی در کل تعداد موارد است که توسط معادله نشان داده شده است:

$$ACC = \frac{TN + TP}{TP + FP + FN + TN} \quad (۹)$$

جایی که TN، TP، FP و FN نشان‌دهندهٔ تعداد منفی واقعی، تعداد مثبت واقعی، تعداد مثبت کاذب و تعداد منفی‌های کاذب است. دقت (PRE) درصد مثبت واقعی نسبت‌به مجموع مثبت واقعی و مثبت کاذب است که توسط معادلهٔ زیر ارائه شده است:

$$PRE = \frac{TP}{TP + FP} \quad (۱۰)$$

در (REC)، نسبت مثبت‌های واقعی نسبت‌به مثبت‌های واقعی و منفی‌های کاذب است که توسط معادلهٔ زیر داده می‌شود:

$$REC = \frac{TP}{TP + FN} \quad (۱۱)$$

مقدار F مقداری بین (۰ و ۱) است. این میانگین هارمونیک دقیق و فراخوانی است که توسط معادلهٔ زیر تعیین می‌شود:

$$F\_measure = \frac{2 \times PRE \times REC}{PRE + REC} \quad (۱۲)$$

محدودهٔ تحت‌منحنی (AUC) یکی دیگر از شاخص‌های معمول برای ارزیابی طبقه‌بندی‌ها است. این برابر با احتمالی است که یک طبقه‌بندی نمونهٔ مثبت تصادفی انتخاب‌شده را بالاتر از یک انتخاب منفی انتخاب می‌کند. مقدار آن از ۰ به ۱ طول می‌کشد. مقادیر بالاتر AUC نشان‌دهندهٔ عملکرد بهتر الگوریتم‌های طبقه‌بندی است.

۴. نتایج تجربی

۴ـ۱. عملکرد و تحلیل تجربی

تمامی داده‌ها که شامل ۱٤۱ فایل پایان‌نامه با فرمت ورد هستند، از مجموعهٔ ایران‌داک گردآوری شده است. این پایان‌نامه‌ها مربوط به رشته‌های علوم انسانی، فنی و مهندسی، هنر، علوم پایه و پزشکی هستند که به‌ترتیب هرکدام از آن‌ها دارای ۱٤، ٤۸، ۲۳، ۳۱ و ۲۵ پایان‌نامه از مجموع داده‌های ما را سهیم‌اند. گفتنی است که تعداد کلیدواژگان مطرح‌شده در جدول ۳ شامل کلیدواژگان بیش از دوکلمه‌ای نیز است، ولی در این پژوهش فقط آن دسته از عباراتی که تک‌کلمه‌ای و دوکلمه‌ای هستند درنظر گرفته شده‌اند.

جدول ۳: اطلاعات مربوط به داده‌های ورودی

Table 3: Input data information

گرایش	تعداد اسناد	تعداد کلمات	تعداد کلیدواژگان
انسانی	۱٤	۳۰۷۱٦۲	۷۰
فنی	٤۸	۷۲٤۹۸۱	۲۳۹
هنر	۲۳	٥۳٤۹۰۳	۱۱٦
پایه	۳۱	٤۲۲۰۰۳	۱٥۰
پزشکی	۲٥	۲۹٤۹۰۹	۱۰۹
جمع کل	۱٤۱	۲۲۸۳۹۸٥	٦۸٤

ابتدا داده‌های ورودی مورد پیش‌پردازش قرار گرفته‌اند. در گام بعدی هر متن به‌صورت

لیستی از عبارات تک‌کلمه‌ای و دوکلمه‌ای درنظر گرفته شده‌اند. سپس ویژگی‌های آماری گفته‌شده برای هر کدام از این عبارات محاسبه می‌شود و ماتریسی شامل ٥ ستون ایجاد می‌شود، که سه ستون آن ویژگی‌های آماری گفته‌شده، یک ستون مربوط به عبارت‌های ورودی و ستون آخر که نشان هر کلمه است. اگر این کلمه توسط نویسندهٔ پایان‌نامه کلیدواژه باشد، نشان ۱ و درغیر این صورت نشان صفر را می‌گیرد. نمونه‌ای از این ماتریس برای عبارت‌های تک‌کلمه‌ای در جدول ٤ آمده است.

جدول ۴: نمونه‌ای از ماتریس داده‌های ورودی تک‌عبارتی

Table 4: An example of a single-phrase (Unigrams) input data matrix

کلمات	Tf-idf	Tf-isf	Mf	Label
فرهنگ	۰.۰۲٤٥۷۷	۰.۰۰۸٦۱٤	٥۱	۰
استقرار	۰.۰۰٥۲٦۸	۰.۰۰۲۰۲۳	۱	۰
پرسش‌نامه	۰.۰۰۳٦٦۸	۰.۰۰۱۰۸۲	۳	۰
نوآور	۰.۰۰۱۹۸۷	۰.۰۰۰٥۱۳	۱	۱

در این مرحله برای کم کردن اختلاف زیاد تعداد کلیدواژه‌ها و غیرکلیدواژه‌ها، باتوجه به بیشترین مقدار ویژگی TF-IDF، تعداد ٥۰ عبارات تک‌کلمه‌ای و ٥۰ عبارات دوکلمه‌ای از هر متن به‌صورت جداگانه انتخاب می‌شوند و به همراه تمام کلیدواژگان مطرح‌شده توسط نویسندگان جدا می‌شوند. این کلمات انتخاب‌شده همراه دو ویژگی دیگر آن‌ها در ماتریسی قرار داده می‌شوند. پس در این حالت از هر متن حدود ۱۰۰ الی ۱۱۰ کلمه در ماتریس ساخته‌شده موجود است.

در گام بعدی به‌طور تصادفی کلمات مربوط به یک متن را از ماتریس ورودی جدا کردیم و این ماتریس را به‌منزلهٔ داده‌های آموزش به یادگیرنده بیز داده می‌شود و تمامی کلمات متنی که جدا شده بودند به‌منزلهٔ داده‌های تست درنظر گرفته می‌شوند. این مرحله را ۲۱ بار انجام می‌پذیرد و هر بار کلمات یک متن را به‌طور کامل به‌منزلهٔ داده‌های تست درنظر گرفته می‌شود. در انتها باتوجه به معیارهای ارزیابی گفته‌شده به تحلیل نتایج پرداخته می‌شود. همان‌طور که در جدول ٥ دیده می‌شود، این آزمایشات با ترکیب‌های مختلف ویژگی‌های آماری مطرح‌شده نیز مقایسه شده است.

جدول ۵: میزان معیارهای ارزیابی برای خروجی دسته‌بند بیز در حالت‌های مختلف ویژگی‌های آماری
Table 5: Evaluation criteria for Bayesian outputs in different states of statistical Features

ویژگی‌های آماری	Accuracy	Recall	F1-Score	Precision
Tf_Idf, Most Frequent, Tf_Isf	۹۷.۲٪	۰.۹۸	۰.۹۸	۰.۹۸
Most Frequent	۹۸.۱٪	۰.۹۸۲	۰.۹۹	۰.۹۹
Tf_Idf, Tf_Isf	۹۹.۸٪	۰.۹۱	۰.۹٤	۰.۹۹

باتوجه به نتایج به‌دست آمده، هنگامی که ویژگی بیشترین فرکانس را از ماتریس داده‌ها حذف می‌شود یادگیرنده بیز قادر به تشخیص داده‌ها با نشان ۱ نیست. بیشترین دقت نیز برای جدا کردن کلیدواژگان از متن نیز مربوط به وقتی است که بیشترین فرکانس به‌تنهایی موردبررسی قرار می‌گیرد.

جدول ۶: میزان ارزیابی داده‌های تست برای خروجی‌هایی که دسته‌بند آن‌ها را کلیدواژه تشخیص داده است.
Table 6: Evaluation of test data for outputs whose keywords have been categorized

ترکیب ویژگی‌های آماری	Recall	F1-Score	Precision	تعداد کلمات	تعداد کلیدواگان انتخابی نویسندگان
Tf_Idf, Most Frequent, Tf_Isf	۰.۹۲	۰.۰۲	۰.۰۱	۷٤۳۳	۷٦
Most Frequent	۰.۹٥	۰.۰۳	۰.۰۲	٤٦۱۸	۷٦

اما آنچه بیشتر از همه در این مبحث حائز اهمیت است، نتایج مربوط به داده‌هایی از تست است که یادگیرنده به‌منزلهٔ کلیدواژه ارائه داده است. باتوجه به آزمایشات انجام‌شده هنگامی که بیشترین فرکانس از ویژگی‌ها حذف می‌شود، یادگیرنده قادر به تشخیص هیچ‌کدام از کلیدواژگان نیست. باتوجه به جدول ٦، در سطر دوم که بیشترین فرکانس فقط استفاده شده است نتایج بهتری از سطر اول که تمامی ویژگی‌ها به بیز انتقال پیدا کرده است، نشان می‌دهد. در سطر دوم هم تعداد کلیدواژگان واقعی بیشتری تشخیص داده شده و Recall بیشتری دارد

و از طرفی دیگر تعداد کلماتی که یک واقعی نیستند و فقط صرفاً شبیه به کلمات کلیدی نویسنده هستند حدود ۳۰۰۰ عبارت کم‌تر از حالت سطر اول است.

حال برای اینکه تعداد کلیدواژگان انتخابی دارای محدویت شوند و میزان دقت مدل طراحی‌شده را بالاتر برود، ابتدا افعال و کلمات کم‌تر از سه حرف حذف می‌شوند. سپس اقدام به جست‌وجوی کلمات انتخابی در بخش‌هایی از متن شده است که دارای اهمیت روان‌شناختی بیشتری هستند. باتوجه به اینکه بخش‌های چکیده و مقدمه در هر متن باید نشان‌دهندهٔ محتوای آن متن باشند، این راهکار درنظر گرفته شده است. نتایج پژوهش لانگ (2010, p.21, 26) نشان می‌دهد که عموماً کلمات مهم و اثرگذار هر سند حداقل یک‌بار در بخش‌های چکیده و مقدمه آمده‌اند. از آنجایی که انتخاب این بخش‌ها به‌صورت اتوماتیک و هوشمند دشوار است، ۱۰ درصد ابتدای پایان‌نامه برای اعمال پس‌پردازش انتخاب شده است. نتیجهٔ به‌دست‌آمده این است که اگر کلمات در بخش چکیده و فصل اول پایان‌نامه‌ها وجود نداشته باشند از مجموع کلمات انتخابی حذف می‌شوند. در مرحلهٔ بعد هرکدام از عبارت‌های تک‌کلمه‌ای و دو کلمه‌ای مربوط به داده‌های تست را به‌صورت جداگانه باتوجه به بیشترین مقدار فرکانس آن‌ها لیست می‌شوند و فقط ۱۰تای اول آن‌ها درنظر گرفته می‌شوند. البته این تعداد برای قسمت عبارت‌های دوکلمه‌ای ممکن است کم‌تر نیز باشد. پس در این صورت خروجی انتهایی مدل طراحی‌شده حدود حداکثر ۲۰ عبارت تک‌کلمه‌ای و دوکلمه‌ای است که برای هریک از ۲۱ متن به‌منزلهٔ دادهٔ تست درنظر گرفته شده است. همان‌طور که در جدول ۷ مشاهده می‌شود، پس از انجام عملیات گفته‌شده تعداد کلیدواژگان انتخابی بسیار محدودتر می‌شوند.

جدول ۷: میزان ارزیابی داده‌های تست بعد از پس‌پردازش برای خروجی‌هایی که دسته‌بند آن‌ها را کلیدواژه تشخیص داده است.

Table 7: Evaluation of post-processing test data for outputs that have been categorized by keyword.

مرحله	ویژگی آماری	Recall	F1-Score	Precision	تعداد کلمات	تعداد کلیدواژگان انتخابی نویسندگان
عبارت‌های تک‌کلمه‌ای	Most Frequent	۰.۸٤	۰.۳۲۳	۰.۲۰	۲۱۰	٤۲
عبارت‌های دوکلمه‌ای	Most Frequent	۰.۹۸	۰.۸۸۸	۰.۸۰	۱٥۸	۳٤

همان‌طور که از جدول ۷ مشخص است مجموع کلمات انتخابی بعد از پردازش گفته‌شده برابر با ۳٦۸ کلمه شده است، یعنی تعداد ٤۲٥ کلمه نسبت‌به بهترین حالت قبل از پردازش که در جدول ٦ آمده، کم شده است. درصد Precision داده‌های تست که در جدول ۷ آمده است مربوط به ۱۰ کلمهٔ اول مجموع کلمات انتخابی است که باتوجه به مقدار بیشترین فرکانسشان لیست شده‌اند. برای مثال داده‌های خروجی مربوط به یک پایان‌نامه که مربوط به گرایش انسانی است در جدول ۸ و ۹ آورده شده است. همان‌طور که مشاهده می‌کنید، در انتها تعداد ۱۰ عبارت تک‌کلمه‌ای و تعداد ٥ عبارت دوکلمه‌ای به‌منزلهٔ خروجی مدل طراحی‌شده موجود است. نویسندهٔ این پایان‌نامه عبارت‌های {ساختار سازمانی، نوآوری، بروکراتیک، آموزش عالی} را به‌منزلهٔ کلیدواژگان درنظر گرفته است. همان‌گونه که مشاهده می‌شود، تمامی کلیدواژگان انتخابی نویسنده در خروجی مدل ساخته‌شده، آمده است. علاوه‌بر آن‌ها، ۱۱ کلمهٔ دیگر نیز به‌منزلهٔ کلیدواژه تشخیص داده شده‌اند که همگی آن‌ها به غیر از دو کلمه {توجه، هم‌بستگی} نشان‌دهندهٔ مفهوم متن موردنظر است.

جدول ۸: نمونه‌ای از خروجی برنامه برای کلمات تک عبارتی

Table 8: An example of a program output for unigram words

نشان نویسنده	نشان تشخیص داده شده توسط مدل	بیشترین فرکانس	کلمات
۱	۱	۷۸	نوآور
۰	۱	۷٦	سازمان
۰	۱	٥۷	توجه
۰	۱	٤۸	کارکن
۰	۱	٤۸	سیستم
۰	۱	٤۳	دانشگاه
۰	۱	٤۰	آموزش
۱	۱	۳۷	بروکراتیک
۰	۱	۳٥	همبستگی
۰	۱	۳٤	انسان

جدول ۹: نمونه‌ای از خروجی برنامه برای کلمات دو عبارت

Table 9: An example of a program output for bygram words

نشان نویسنده	نشان تشخیص داده شده توسط مدل	بیشترین فرکانس	کلمات
۰	۱	٦۱	خلاق نوآور
۱	۱	٤۳	ساختار سازمان
۰	۱	۳۹	قوانین مقررات
۰	۱	۳۳	سلسله‌مراتب
۱	۱	۲٦	آموزش عالی

۴ـ۲. ابزار مورداستفاده

در این پژوهش از زبان برنامه‌نویسی رایگان پایتون[۲۷] استفاده شده است. برای پیش‌پردازش متن‌های ورودی نیز همان‌طور که بیان شد از برنامهٔ رایگان هضم[۲۸] استفاده شده است.

۵. نتیجه

انتخاب کلیدواژگان توسط نویسندگان همواره تصمیمی وابسته به شخص بوده است و ممکن است افراد مختلف برای انتخاب مجموعه‌ای از لغات برای یک متن تصمیمی واحد نگیرند. در این پژوهش سعی بر آن بوده است که با رویکری نوین، باتوجه به تعداد کم اسناد ورودی، مدل و برنامه‌ای ساخته شود که بدون وابستگی به گرایش پایان‌نامه‌ها و معنای کلمات آن‌ها و صرفاً با استفاده از ویژگی‌های آماری کلمات در هر متن، به استخراج کلیدواژگان پرداخته شود. همان‌طور که در جدول ۵ و ۷ نیز نشان داده شده است، مدل طراحی‌شده در بهترین شرایط که مربوط به استفاده از یک ویژگی بیشترین فرکانس است، قادر است با دقت کلی ۹۸.۱ درصد، از هر پایان‌نامه تعداد حداکثر ۲۰ کلیدواژه را استخراج کند. کلیدواژگان نوشته‌شده در هر پایان‌نامه نیز با دقت‌های ۸٤ درصد و ۹۸ درصد که به ترتیب مربوط به عبارت‌های تک‌کلمه‌ای و دوکلمه‌ای هستند، دربر می‌گیرد. برای مثال نیز یکی از نتایج به‌دست آمده بر روی یکی از پایان‌نامه‌ها در جداول ۸ و ۹ نیز آورده و نشان داده شده است که مدل طراحی‌شده از این پایان‌نامه تعداد ۵ عبارت دوکلمه‌ای و ۱۰ عبارت تک‌کلمه‌ای را به‌منزلهٔ کلیدواژه تشخیص داده است که ۱۰۰ درصد کلیدواژگان نوشته‌شده در پایان‌نامه در آن آمده است.

۶. کارهای آینده

از این مدل طراحی‌شده سازمان‌هایی نظیر ایران‌داک، پایگاه مجلات تخصصی نور و خبررسانی‌های فارسی نظیر همشهری برای نمایه‌سازی، استخراج کلیدواژگان و فیلترینگ می‌توانند بهره ببرند. یکی از مشکلات بزرگی که در این فرایند با آن مواجه هستیم، ریشه‌یابی و یکپارچه کردن متن است که این امر باعث می‌شود تا پیش‌پردازش مناسبی صورت نگیرد و کلمات زائد به‌شکل مناسبی حذف نشوند. این عوامل پیچیدگی ابعاد ماتریس ورودی برای دسته‌بند را بالا می‌برد و باعث پایین آمدن دقت می‌شود. در آینده سعی بر آن می‌شود به‌جای استفاده از برنامهٔ هضم، از برنامه فارسی‌یار[۲۹] استفاده شود و نتایج به‌دست آمده موردبررسی قرار داده شوند. همچنین، می‌توان به‌جای استفاده از دسته‌بند بیز از دیگر دسته‌بندها استفاده کرد و نتایج به‌دست آمده را با این دسته‌بند مقایسه کرد.

۷. پی‌نوشت‌ها

1. Atanu Dey
2. unigrams
3. senti-n-gram
4. Rafeal
5. Neumann
6. Loung
7. Litvak
8. Huan
9. Turney
10. Witten
11. Hachoen Kerner
12. Onan
13. bagging
14. random forest
15. Moien Maadi
16. Mohammad Razzaghnoori
17. word2vec
18. Behnam Sabeti
19. miras text
20. Morteza Okhovvat
21. Hidden Markov
22. Moammad

23. Term Frequency Inverse Document Frequency (TF-IDF)
24. most frequent of words
25. term frequency inverse sentence frequency (TF-ISF)
26. naïve bayes
27. python
28. sobhe.ir
29. text-mining.ir

۸. منابع

- گندمکار، ر. (۱۳۹٦). توسیع معنایی در زبان فارسی؛ مطالعهٔ موردی افعال. *جستارهای زبانی*، ۵ (۵۳)، ۲۳۳ـ ۲۵٦.

References

- Gandomkr, R. (2017). Semantic expansion in Persian; A case study of verbs. *Language Related Research*. *5*(53), 233-256.[In Persian]
- Aytug Onan, Serdar Korukoglu, Hasan Bulut. (2016). *Ensemble of keyword extraction methods and classifiers in text classification. Expert Systems with Applications*. 232-247.
- Atanu Dey, Mamata Jenamani, Jitesh J.Thakkar. (2018).Senti-N-Gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications.* 92–105.
- Rafea, S. El-Beltagy & A. (2010). Kp-miner: Participation in semeval-2" Proceedings of the 5th International Workshop. *ACL2010.* 190–193
- Neumann, K. E. and G. (2010).Dfki keywe: Ranking keyphrases extracted from scientific articles. In *Proceedings of the 5th International Workshop on Semantic Evaluation.* ACL 2010*,* 150–153.
- Luong, T. Nguyen and M. (2010).WINGNUS: Keyphrase extraction utilizing document logical structure. *Proceedings of the 5th International Workshop.* 21–26.
- Turney, P. D, (2002).Mining the Web for Lexical Knowledge to Improve Keyphrase Extraction: Learning from Labeled and Unlabeled Data P.D. *ArXiv*.
- Witten, I. H. , Paynter, G. W. , Frank, E. , Gutwin, C. (1999). KEA: Practical

automatic keyphrase extraction. In *Proceedings of the Fourth ACM Conference on Digital Libraries.*

- HaCohen-Kerner. (2005).Automatic extraction and learning of keyphrases from scientific articles. *Lecture Notes in Computer Science.* 657-669.
- Moien Maadi, Kazim Fouladi. (2015).Providing a method for extracting keywords in the Persian language. *International Academic Journal of Innovative Research.* 34-42.
- Mohammad Razzaghnoori, Hedieh Sajedi , Iman Khani Jazani. (2017).Question classification in Persian using word vectors and frequencies. *Journal Cognitive Systems Reaserch.* 16-27.
- Behnam Sabeti, Hossein Abedi Firouzjaei, Ali Janalizadeh Choobbasti. (2018).MirasText: An Automatically Generated Text Corpus for Persian. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*
- Morteza Okhovvat, Behrouz Minaei Bidgoli. (2010). *A hidden Markov model for Persian part-of-speech Tagging.* 94–101.
- Sayyid Mohammad Hoseini Khozani, Hosein Bayat. (2011).Specialization of keyword extraction approach to Persian texts. *International Conference of Soft Computing and Pattern Recognition (SoCPaR).*. 112-116.
- Beliga, S. , Mestrovic, A. , & Martincic-Ipsic, S. (2015).An overview of graph-based keyword extraction methods and approaches. *Journal of Information and Organizational Sciences. 39* (1), 1–20.
- Lott, B. 2012. *Survey of keyword extraction techniques. UNM Education.*
- Rossi, R. G. , Maracini, R. M. , & Rezende, S. O. (2014). Analysis of domain independent statistical keyword extraction methods for incremental clustering. *Learning and Nonlinear Models.* 17-37.
- Neto, L. J. , Santos, A . D. , Kaestner, C. A. (2000). Document cluster- ing and text summarization. *In Proceedings of the 4th International Conference on Practical*

Applications of Knowledge Discovery and Data Mining. 41-55.

- Fiori, A. (2014). Innovative document summarization techniques: Revolutionizing knowledge understanding. *Advances in Data Mining and Database Management.*
- Litvak, M. , & Last, M. (2008).Graph-based keyword extraction for single-document summarization. *In Proceedings of the Workshop on Multi-source Multilingual Information Extraction and Summarization.* 17-24.
- Huan, C. , Tian, Y. (2006).Keyphrase extraction us- ing semantic network structure analysis. *In Proceedings of the Sixth International Conference on Data Mining.* 275-284.