Keyword and Phrase Extraction from Persian Texts: A Review of the literature

Atefeh Kalantari*
PhD Candidate; Knowledge and Information Science Department; School of Education & Psychology; Shiraz University; Shiraz, Iran; Academic Librarian; Nursing & Midwifery Faculty; Qazvin University of Medical Sciences; Qazvin, Iran;
Email: atefeh_kalantari@shirazu.ac.ir

Abdolrasool Jowkar
PhD in Knowledge & Information Science; Professor; School of Education & Psychology; Shiraz University; Shiraz, Iran;
Email: ajowkar2003@yahoo.com

Seyed Mostafa Fakhrahmad
PhD in Computer Engineering; Associate Professor; Department of Computer Science and Engineering & IT; Shiraz University; Shiraz, Iran Email: fakhrahmad@shirazu.ac.ir

Javad Abbaspour
PhD in Knowledge & Information Science; Assistant Professor; School of Education & Psychology; Shiraz University; Shiraz, Iran;
Email: javad.abbaspour@gmail.com

Hajar Sotudeh
PhD in Knowledge & Information Science; Associate Professor; School of Education & Psychology; Shiraz University; Shiraz, Iran;
Email: sotudeh@shirazu.ac.ir

Massoud Mortazavi
PhD Candidate; Knowledge & Information Science Department; School of Education & Psychology; Ferdowsi University of Mashhad; Mashhad, Iran; Research & Development Specialist; Pars Azarakhsh Co.; Tehran, Iran;
Email: mortazavi.massoud@mail.um.ac.ir

Amir Javadi
PhD in Medical Informatics; Assistant Professor; Social Medicine Department; Qazvin University of Medical Sciences; Qazvin, Iran;
Email: javadi_a@yahoo.com

Zahra Pourbahman
PhD Candidate in Computer Engineering; AmirKabir University of Technology; Tehran, Iran Email: pourbahman@aut.ac.ir

* Corresponding Author

Iranian Journal of Information Processing and Management

Received: 10, Oct. 2019 Accepted: 14, Jun. 2020

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 36 | No. 2 | pp. 563-592
Winter 2021
https://doi.org/10.35050/JIPM010.2020.011

Abstract: Keyword and phrase extraction is a prerequisite of many natural language processing tasks. However, a review on the related Persian and English literature showed that a few studies have already been done on how to extract keywords and phrases from Persian texts. Thus, aiming to shed light on the research status of Keyword and phrase extraction from Persian texts, the present study reviews the Persian and English publications which have assessed their research ideas over Persian texts. We also focus on each of the studies to challenge their methodologies, implementations and evaluation methods and measures.

To our knowledge, a total number of 14 Persian and 6 English papers exist which have worked on the extraction of Persian keywords and phrases. Investigating on the papers revealed that they were mostly based on statistical and linguistic information. A majority of the papers suffered from the lack of either appropriate methodologies or lucid explanation of their research ideas. They generally used non-standard datasets and vague or problematic metrics to evaluate the experimental systems. Generally speaking, except 3 papers that appropriately reported their proposed methods, the other papers lacked reproducibility and generalizability. Hence, their results cannot be confidently used as a benchmark in evaluating future works, and their proposed ideas cannot be employed in developing applications for extraction of key words and phrases from Persian texts.

Keywords: Extraction, Key Words, Key Phrases, Natural Language Processing, Persian Language, Review

استخراج کلمات و عبارات کلیدی از متون فارسی

(مروری بر پژوهش‌های صورت‌گرفته)

عاطفه کلانتری
دانشجوی دکتری علم اطلاعات و دانش‌شناسی (بازیابی اطلاعات)؛ دانشگاه شیراز؛ شیراز، ایران؛ مسئول کتابخانهٔ دانشکدهٔ پرستاری و مامایی؛ دانشگاه علوم پزشکی قزوین؛ قزوین، ایران؛
پدیدآور رابط atefeh_kalantari@shirazu.ac.ir

عبدالرسول جوکار
دکتری علم اطلاعات و دانش‌شناسی؛ استاد؛ بخش علم اطلاعات و دانش‌شناسی؛ دانشکدهٔ روان‌شناسی و علوم تربیتی؛ دانشگاه شیراز؛ شیراز، ایران؛
ajowkar2003@yahoo.com

سیدمصطفی فخراحمد
دکتری مهندسی کامپیوتر؛ دانشیار؛ بخش مهندسی و علوم کامپیوتر و فناوری اطلاعات؛ دانشگاه شیراز؛ شیراز، ایران fakhrahmad@shirazu.ac.ir

جواد عباس‌پور
دکتری علم اطلاعات و دانش‌شناسی؛ استادیار؛ بخش علم اطلاعات و دانش‌شناسی؛ دانشکدهٔ روان‌شناسی و علوم تربیتی؛ دانشگاه شیراز؛ شیراز، ایران؛
javad.abbaspour@gmail.com

هاجر ستوده
دکتری علم اطلاعات و دانش‌شناسی؛ دانشیار؛ بخش علم اطلاعات و دانش‌شناسی؛ دانشکدهٔ روان‌شناسی و علوم تربیتی؛ دانشگاه شیراز؛ شیراز، ایران؛
sotudeh@shirazu.ac.ir

مسعود مرتضوی نصرآباد
دانشجوی دکتری علم اطلاعات و دانش‌شناسی (بازیابی اطلاعات)؛ دانشگاه فردوسی مشهد؛ مشهد، ایران؛ کارشناس بخش تست و توسعه؛ شرکت پارس‌آذرخش؛ تهران، ایران mortazavi.massoud@mail.um.ac.ir

امیر جوادی
دکتری انفورماتیک پزشکی؛ استادیار گروه پزشکی اجتماعی؛ دانشکدهٔ پزشکی؛ دانشگاه علوم پزشکی قزوین؛ قزوین، ایران javadi_a@yahoo.com

زهرا پوربهمن
دانشجوی دکتری مهندسی کامپیوتر (نرم‌افزار)؛ دانشکدهٔ مهندسی کامپیوتر؛ دانشگاه صنعتی امیرکبیر؛ تهران، ایران؛
pourbahman@aut.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۳۹۸/۰۷/۱۸ | **پذیرش:** ۱۳۹۹/۰۳/۲۵ | مقاله برای اصلاح به مدت ۳۷ روز نزد پدیدآوران بوده است.

چکیده: استخراج کلمات/ عبارات کلیدی متن، پیش‌نیاز بسیاری دیگر از وظایف حوزهٔ پردازش زبان طبیعی است. اما بررسی متون فارسی و انگلیسی این حوزه نشان می‌دهد که تلاش‌های انگشت‌شماری برای استخراج کلمات/ عبارات کلیدی از متون فارسی صورت‌گرفته است. ازاین‌رو، این مقاله با هدف تعیین موقعیت کنونی پردازش زبان طبیعی فارسی، و به‌طور خاص، استخراج کلمات/ عبارات کلیدی از متون فارسی به مرور خلاصه‌ای از مقالات فارسی و انگلیسی منتشرشده در این حوزه که

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران
(ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۶ | شمارهٔ ۲ | ص‌ص ۵۶۳-۵۹۲
زمستان ۱۳۹۹
https://doi.org/10.35050/JIPM010.2020.011

از متـون فارسـی بـرای آزمـودن ایده‌هـا اسـتفاده کرده‌انـد، می‌پـردازد. سـپس، هـر مقالـه را از نظـر روش‌شناسـی، نحـوهٔ اجـرا و پیاده‌سـازی، روش ارزیابـی، و معیارهـای آن مـورد تعمـق قـرار داده و بـه چالـش می‌کشـد.
در مجمـوع، ۱۴ مقالـهٔ فارسـی و ۶ مقالـهٔ انگلیسـی بـه اسـتخراج کلمـات و عبـارات کلیـدی از متـون فارسـی پرداخته‌انـد. روش بیشـتر ایـن مقـالات اسـتفاده از اطلاعـات آمـاری و زبان‌شـناختی بـوده اسـت. اکثـر ایـن مقـالات یـا در روش‌شناسـی انتخاب‌شـده ایـراد دارنـد و یـا نویسـندگان نتوانسـته‌اند ایـدهٔ پیشـنهادی خـود را به‌وضـوح بـرای خواننـده تبییـن نماینـد. در بسـیاری از مقـالات از مجموعـهٔ دادهٔ اسـتانداردی بـرای ارزیابـی سیسـتم اسـتفاده نشـده و نحـوهٔ محاسـبهٔ معیارهـای ارزیابـی مبهـم یـا دارای اشـکال اسـت.
در مجمـوع، به‌جـز ۳ مقالـه کـه روش اجراشـده را به‌نحـو نسـبتاً قابل‌قبولـی گـزارش کرده‌انـد، سـایر مقـالات قابلیـت تکرارپذیـری و تعمیـم ندارنـد. ایـن اسـت کـه نمی‌تـوان از آن‌هـا به‌عنـوان معیـار پایـه‌ای بـرای ارزیابـی سیسـتم‌های آینـده اسـتفاده کـرد یـا از ایـدهٔ مطرح‌شـده در آن‌هـا بـا اطمینـان در سـاخت و توسـعهٔ نرم‌افزارهـای کاربـردی و عملـی در حـوزهٔ اسـتخراج کلمـات کلیـدی اسـتفاده نمـود.

کلیدواژه‌هـا: اسـتخراج کلمـات کلیـدی، اسـتخراج عبـارات کلیـدی، پـردازش زبـان طبیعـی، زبـان فارسـی، بررسـی مـروری

۱. مقدمه

منظـور از پـردازش زبـان طبیعـی کلیـهٔ تلاش‌هایـی اسـت کـه متخصصـان ایـن حـوزه انجـام می‌دهنـد تـا ادراک، فهـم، تجربـه، دانـش و به‌نوعـی هـوش انسـان را به‌گونـه‌ای در قالـب کدهـای برنامه‌نویسـی بـه سیسـتم‌های کامپیوتـری آمـوزش دهنـد کـه بـا حداقـل نیـاز بـه حضـور و دخالـت انسـانِ متخصـص موضوعـی و اطلاعاتـی (مثـل کتابـداران موضوعـی سـابق)، بهتریـن و مرتبط‌تریـن اطلاعـات در کوتاه‌تریـن زمـان ممکـن بـه کاربـر ارائـه شـود. بدین‌منظـور، وظیفـهٔ اصلـی و دشـوار پـردازش زبـان طبیعـی بـه ده‌هـا وظیفـهٔ کوچک‌تـر تفکیـک شـده کـه انجـام و یـا بهبـود هـر یـک از آن‌هـا گامـی اسـت در راسـتای تحقـق هـدف غائـی پـردازش زبـان طبیعـی. اسـتخراج کلمـات/ عبـارات کلیـدی یکـی از آن ده‌هـا وظیفـهٔ کوچک‌تـر اسـت کـه خـود زیربنـا و گام نخسـت بسـیاری دیگـر از وظایـف ایـن حـوزه اسـت. شـاید بتـوان گفـت کـه نخسـتین تلاش‌هـا در زمینـهٔ اسـتخراج خـودکار کلمـات/ عبـارات کلیـدی همزمـان بـا تـلاش بـرای نمایه‌سـازی خـودکار/ ماشـینی در حـوزهٔ کتابـداری و علـم اطلاعـات صـورت گرفتـه اسـت.

طـی سـالیانی کـه پـردازش زبـان طبیعـی و به‌تبـع آن اسـتخراج کلمـات/ عبـارات کلیـدی مـورد توجـه جامعـهٔ پژوهشـی حـوزهٔ علـوم کامپیوتـر قـرار گرفتـه، مطالعـات و پژوهش‌هـای

متعددی در زمینهٔ پردازش متون انگلیسی و استخراج کلمات/ عبارات کلیدی از آن‌ها صورت گرفته است که کثرت جمعیت انگلیسی‌زبان و نیز پذیرش زبان انگلیسی به‌عنوان زبان علمی از سوی جوامع علمی را می‌توان از مهم‌ترین علل این امر دانست. گذشته از این، طرح ایده‌ها و الگوریتم‌های مبتنی بر یادگیری ماشینی، شبکه‌های عصبی، و غیره نیز منجر به بهبود قابل‌توجهی در نتایج حاصل از پردازش زبان انگلیسی و استخراج عبارات کلیدی از آن شده است[1].

اما بررسی متون فارسی و انگلیسی این حوزه نشان می‌دهد که تلاش‌های انگشت‌شماری برای استخراج کلمات/ عبارات کلیدی از متون فارسی صورت گرفته است. همان‌طور که پیش‌تر اشاره شد، استخراج کلمات/ عبارات کلیدی پیش‌نیاز بسیاری دیگر از وظایف پردازش زبان طبیعی است. یکی از علل این امر آن است که پردازش زبان فارسی به اذعان پژوهشگران متعدد، به‌دلیل ماهیت عناصر زبان‌شناختی و تحریری آن با چالش‌های متعددی روبه‌روست. بنابراین، وظایف مختلف پردازش زبان فارسی در نخستین مراحل آن‌چنان با چالش‌ها مواجه می‌شوند که کمتر به نتایج مطلوب رسیده و در همان مراحل اولیه متوقف می‌گردند. ازاین‌رو، پژوهشگران فارسی‌زبانِ حوزهٔ پردازش زبان طبیعی پیاده‌سازی و آزمون ایده‌هایشان را بر پایهٔ پیکره‌ها و متون انگلیسی طرح‌ریزی می‌کنند. بدین‌ترتیب، پردازش زبان فارسی کمترین پیشرفت عملی را در عرصهٔ وظایف مختلف تجربه می‌کند و در نتیجه، شاهد هستیم که با وجود پیشرفت قابل توجه در حوزهٔ ترجمهٔ ماشینی از زبان انگلیسی به بسیاری از زبان‌ها و بالعکس، ترجمهٔ ماشینی از زبان فارسی به انگلیسی و بالعکس کاستی‌ها و نقاط ضعف متعدد دارد. در حوزه‌های دیگر استفاده‌کننده از نتایج پردازش زبان طبیعی نیز چنین وضعیتی حاکم است. برای نمونه، یکی از این حوزه‌ها، حوزهٔ سیستم‌های توصیه‌گر محتوامحور است که در آن با پردازش متون و استخراج ویژگی از آن‌ها (که کلمات/ عبارات کلیدی از مهم‌ترین آن‌هاست)، گزینه‌های مرتبط با آن‌ها به کاربران متقاضی پیشنهاد می‌شود.

بنابراین، به‌منظور تعیین موقعیت کنونی پردازش زبان طبیعی فارسی و به‌طور خاص

---

۱. در صورت تمایل به کسب اطلاعات تکمیلی در خصوص روش‌های استخراج کلمات/ عبارات کلیدی می‌توانید به مقالهٔ زیر مراجعه فرمایید:

S. Siddiqi, & A. Sharan. 2015. Keyword and Keyphrase Extraction Techniques: A Literature Review. Int J Comput Appl. 109 (2):18–23.

استخراج کلمات/ عبارات کلیدی از متون فارسی، این مقاله به مرور مقالات فارسی و انگلیسی منتشرشده در این حوزه که از متون فارسی برای آزمودن ایده‌های خود استفاده کرده‌اند، پرداخته و سپس، هر مقاله را از نظر روش‌شناسی، نحوهٔ اجرا و پیاده‌سازی، روش ارزیابی و معیارهای آن مورد تعمق قرار داده و به چالش می‌کشد.

۲. روش پژوهش

مرور نظام‌مند متون که بیشتر در حوزهٔ علوم پزشکی رایج است، از سال ۱۹۹۳ توسط شرکت «کاکرین»[1] و با هدف بررسی اثر مداخلات بهداشتی[2] که از طریق کارآزمایی‌های بالینی تصادفی[3] مورد آزمایش قرار گرفته‌اند، شروع شده است. مرور نظام‌مند متون عبارت است از ارزیابی انتقادی تمام مطالعاتی که به یک پرسش پژوهشی خاص پرداخته‌اند. پژوهشگری که مرور نظام‌مند را انجام می‌دهد، از یک روش سازمان‌یافته برای شناسایی، گردآوری، و ارزیابی پیکرهٔ متون حول موضوع مورد نظر استفاده می‌کند و بدین‌منظور، از مجموعه‌ای از معیارهای ازپیش‌تعیین‌شده بهره می‌گیرد (Onwuegbuzi 2016).

بر این اساس، پژوهش حاضر یک مرور نظام‌مند متون[4] است که با هدف پاسخ‌گویی به این پرسش که «چگونه می‌توان کلمات/ عبارات کلیدی را از متون فارسی استخراج نمود؟» انجام شده است. به‌منظور یافتن کلیهٔ مطالعات صورت‌گرفته در این حوزه، از ترکیبات مختلف عبارات جدول ۱، برای جست‌وجو و بازیابی مقالات و پایان‌نامه‌های فارسی احتمالی در پایگاه‌های اطلاعاتی SID، IranDoc، ISC، Civilica، Google Scholar، magiran و غیره استفاده شده است.

جدول ۱. ترکیب‌های مختلف عبارات برای جست‌وجو و بازیابی مقالات و پایان‌نامه‌های فارسی

استخراج	خودکار	کلمات/ کلمه‌های کلیدی	متون	فارسی
تخصیص	اتوماتیک	عبارات/ عبارت‌های	متن	
نمایه‌سازی		واژگان/ واژه‌های	زبان	

همچنین، ترکیب‌های مختلف عبارات جدول ۲، برای جست‌وجو و بازیابی مقالات و

1. Cochrane
2. healthcare interventions
3. randomized clinical trials
4. systematic literature review

پایان‌نامه‌های انگلیسی احتمالی در پایگاه‌های اطلاعاتی Scopus، Google Scholar و Pubmed (به‌دلیل وجود رشتهٔ انفورماتیک پزشکی) مورد استفاده قرار گرفته‌اند.

جدول ۲. ترکیب‌های مختلف عبارات برای جست‌وجو و بازیابی مقالات و پایان‌نامه‌های انگلیسی

Persian	Text	Important	Keywords	Automatic	Extraction
Farsi	Texts		Keyphrases		Assignment
	Language		Words		Indexing

افزون بر نتایج جست‌وجوهای صورت‌گرفته و با استفاده از جست‌وجوی استنادی عمیق، فهرست منابع و مآخذ مقالات دارای شرط ورود به مطالعه نیز مورد کاوش قرار گرفته و از میان آن‌ها مدارک مرتبط احتمالی گزینش و به مطالعه وارد شده‌اند.

همچنین، مقالات و پایان‌نامه‌های متعددی بررسی شده‌اند که از میان آن‌ها ۱۴ مقالهٔ فارسی و ۶ مقالهٔ انگلیسی دارای شرط ورود به مطالعه بوده‌اند، و بنابراین، بقیهٔ مقالات از چرخهٔ مطالعه حذف شده‌اند.

- شرط ورود به مطالعه:

مقاله یا پایان‌نامه به‌زبان فارسی یا انگلیسی باشد، اما برای پیاده‌سازی و آزمودن آن حتماً از متون و پیکره‌های فارسی استفاده شده باشد.

۳. مروری بر پژوهش‌های صورت‌گرفته

اگر بپذیریم که رویکرد و چگونگی استخراج کلمات یا عبارات کلیدی نسبت بسیار نزدیکی با نمایه‌سازی خودکار دارد، آنگاه می‌توان گفت که تاریخچهٔ شروع آن در دنیا به اوایل دههٔ ۱۹۵۰ میلادی برمی‌گردد. اما نخستین تلاش‌ها در این حوزه در کشور ما از اوایل دههٔ ۱۳۸۰ شمسی یعنی حدود ۵۰ سال بعد از آن تاریخ آغاز شده است.

«تشکری» که در سال ۱۳۸۰ در پایان‌نامهٔ کارشناسی ارشد خود روی روش‌های شاخص‌گذاری خودکار برای متون فارسی کار کرده بود، در سال ۱۳۸۲ به‌همراه «میبدی»، نخستین نمایه‌ساز خودکار متون فارسی را طراحی کرد. آن‌ها برای ارزیابی سیستم پیشنهادی خود یک مجموعهٔ دادهٔ فارسی در حوزهٔ کامپیوتر ایجاد کردند که شامل ۴۵۰

چکیدهٔ مقاله و پایان‌نامه، همراه با ۳۲ عبارت پرس‌وجو[1] و قضاوت ربط[2] مربوطه بود (تشکری و میبدی ۱۳۸۲). با توجه به محدودیت منابع و متون فارسی دیجیتال در آن زمان، این کار را می‌توان نقطهٔ قوت کار آن‌ها دانست. سیستم طراحی‌شده توسط آن‌ها به نام «نمایه‌ساز»، با استفاده از قانون «زیف»[3] تمام واژه‌های متون مجموعهٔ داده را رتبه‌بندی می‌کرد. آن‌ها دریافتند که ۲۰ درصد از واژگان متمایز، بیش از ۸۵ درصد از کل واژگان متون را تشکیل داده‌اند. بنابراین، فهرستی از ۱۵۰ واژهٔ فاقد خاصیت تعیین‌کنندگی برای مجموعهٔ داده تهیه کرده و آن‌ها را حذف نمودند. اما دربارهٔ ادامهٔ فرایند نمایه‌سازی، در متن مقاله‌شان توضیحی نداده‌اند. از سایر توضیحات و استناد به برخی منابع برای توجیه انتخاب نمایه‌سازی کنترل‌نشدهٔ پس‌همارا این‌گونه برمی‌آید که آن‌ها تمام واژه‌های باقی‌مانده را واژه‌های نمایه‌ای در نظر گرفته‌اند. آن‌ها مقادیر دو معیار دقت و بازیافت را با انجام عملیات ریشه‌یابی و بدون آن محاسبه کردند و دریافتند که ریشه‌یابی، بازیافت را افزایش می‌دهد، ولی باعث کاهش دقت می‌گردد (همان).

پژوهشگران دیگری هم دست به طراحی سامانه‌های نمایه‌سازی خودکار زده‌اند. از آن جمله «بشیری، کربلائی و موسوی» یک سیستم نمایه‌سازی به‌نام «سینا» طراحی کردند که پس از حذف واژگان ایستا، کلمات موجود در متن ورودی را ریشه‌یابی می‌کرد. اساس عملکرد الگوریتم ریشه‌یابی «سینا»، همچون ریشه‌یاب انگلیسی «پورتر»[4] حذف پسوندها و پیشوندها بر اساس قوانین ازپیش‌تعریف‌شده و استثناهای تعیین‌شده بود. پس از اعمال عملیات ریشه‌یابی روی متن ورودی و حذف مجدد واژگان ایستا، نمایه‌ساز «سینا» به کلمات باقی‌ماندهٔ متن بر اساس الگوهای TF-IDF[5]، Lnu، ltn و ntc وزن می‌داد تا بهترین کلمات به‌عنوان نمایهٔ متن تعیین گردند. آن‌ها در نهایت، با مقایسهٔ مقادیر دقت و بازیافت به‌دست‌آمده از طریق هر الگوی وزن‌دهی دریافتند که TF-IDF و Lnu بهتر از سایر الگوها می‌توانند به تعیین کلمات نمایه‌ای متون فارسی کمک کنند (۱۳۸۴).

آن‌ها همچنین، برای ارزیابی سیستم نمایه‌ساز «سینا»، از مجموعه داده‌ای که «تشکری و میبدی» (۱۳۸۲) ایجاد کرده بودند، استفاده کرده و با محاسبهٔ دقت و بازیافت تعیین کلمات نمایه‌ای با انجام ریشه‌یابی و بدون آن دریافتند که استفاده از ریشه‌یاب، علاوه بر

1. query
2. relevance judgement
3. Zipf
4. Porter
5. Term Frequency-Inverse Document Frequency

این‌که میانگین دقت را افزایش می‌دهد، باعث کاهش بیش از ۵۰ درصدی حجم نمایهٔ ساخته‌شده از متون هم می‌شود (همان).

نکتهٔ مبهم در مورد مقالهٔ «بشیری، کربلائی و موسوی» (۱۳۸۴) این است که چگونگی استفاده از مجموعهٔ دادهٔ «تشکری و میبدی» (۱۳۸۲) برای ارزیابی نمایه‌ساز «سینا» را توضیح نداده‌اند. بنابراین، این سئوال قابل طرح است که آیا اساساً این مجموعهٔ داده برای ارزیابی سیستم طراحی‌شده توسط آن‌ها مناسب بوده یا خیر.

«گزنی» برای تعیین عبارات کلیدی متن از سه معیار تعداد رخداد کلمه (تعیین حد بالا و پایین برای تعداد تکرار کلمات)، مجاورت مکانی کلمات (فاصلهٔ کمتر برخی واژه‌ها، احتمالاً بر جنبهٔ خاصی از یک موضوع دلالت می‌کند)، و موقعیت مکانی آن‌ها نسبت به‌هم (واژه‌هایی که به یک موضوع خاص مرتبط هستند، احتمالاً به‌هم نزدیک‌تر قرار می‌گیرند) استفاده کرد. وی استفاده از این ویژگی‌ها را ساده و مقرون به‌صرفه دانسته و معتقد بود که با توجه به ماهیت مقاله‌های فنی، احتمال وجود کلماتی با بیش از یک معنا و یا چند کلمهٔ هم‌معنا در چنین مقالاتی اندک است (۱۳۸۵). او پس از حذف سیاههٔ واژگان ایستا، برای ترکیب واژه‌هایی که حروف آغازین آن‌ها مشترک است (ظاهراً هم‌ریشه هستند)، به مقایسهٔ هر جفت واژه به‌صورت حرف‌به‌حرف در فهرست الفبایی واژه‌ها پرداخت و کلماتی را که ۴ حرف آغازین مشابه داشتند با هم ترکیب کرد. با این حال، از نظر وی، با توجه به ماهیت زبان فارسی این رقم بین ۳ تا ۵ قابل تغییر بود. وی امکان خطای این روش تطبیق را کمتر از ۰/۵ درصد می‌دانست. «گزنی» ایدهٔ پیشنهادی‌اش را روی ۵۰ مقاله آزمایش کرد و عبارات کلیدی به‌دست‌آمده را برای ارزیابی به ۸ متخصص نمایه‌سازی داد. او بدون آن‌که به نتایج ارزیابی متخصصان اشاره نماید، نتیجه گرفت که انتخاب خودکار عبارت‌های کلیدی به‌نحوی که بیانگر موضوع کلی مقاله باشند، عملی است و این عبارت‌ها تا حدود زیادی شبیه عبارت‌هایی هستند که توسط انسان انتخاب می‌شوند (همان).

کار «گزنی» اگرچه از نخستین تلاش‌ها برای استخراج عبارات کلیدی از متون فارسی است، اما بیشتر جنبهٔ نظری دارد تا انجام کار پژوهشی؛ چراکه به نظر می‌رسد، بیشتر پیش‌فرض‌های مطرح‌شده به‌جای این‌که حاصل نتیجه‌گیری از طریق گردآوری داده‌ها و تحلیل آن‌ها یا دست کم استناد به اسناد و مدارکی باشد که درستی این پیش‌فرض‌ها را قبلاً تائید کرده‌اند، مبتنی بر نظریه‌هایی بر اساس احتمال و حدس و گمان است.

در همین راستا «عربی نرئی، وحیدی اصل و مینایی بیدگلی» پس از حذف واژگان ایستا به ریشه‌یابی کلمات پرداختند و پس از آن مجدداً اقدام به حذف واژگان ایستا نمودند. آن‌ها با در نظر گرفتن یک حد فرکانس بالا و یک حد فرکانس پایین، کلمات بین این دو مقدار را به‌عنوان کلمات کاندیدا تعیین کردند. سپس، با استفاده از روش Luhn، به هر جمله یک فاکتور اهمیت نسبت دادند. آنگاه، به‌منظور کاهش تعداد کلمات باقی‌مانده، با استفاده از یک لغت‌نامهٔ الکترونیکی فارسی، کلمات هم‌معنا را با یکی از کلماتی که همان معنا را داشت، جایگزین کردند. سپس، کلمات هر مدرک را با استفاده از رابطهٔ اعم[1] در قالب یک درخت سازماندهی کرده و هر دو کلمه را با نزدیک‌ترین ریشهٔ مشترک‌شان در درخت جایگزین کردند. آن‌ها ادعا کردند که این دو کار باعث همگون شدن مستندات می‌شود و در نتیجه، مدارکی که دارای مضمون مشابهی هستند، کلمات کلیدی آن‌ها نیز تقریباً یکسان می‌شود (۱۳۸۶).

آن‌ها در پژوهشی دیگر، ۱۰۰ متن فارسی را از ۵ طبقهٔ مختلف در نظر گرفته و کلمات کلیدی آن‌ها را به‌صورت دستی استخراج نمودند. آنگاه به‌منظور ارزیابی ایدهٔ پیشنهادی‌شان برای استخراج کلمات کلیدی، میزان شباهت کلمات کلیدی استخراج‌شده توسط سیستم با کلمات کلیدی تخصیص‌داده‌شده به‌صورت دستی را محاسبه کردند. نتایج نشان‌دهندهٔ ۶۱/۳۲ درصد شباهت بود. به‌علاوه، نتایج آزمایشات آن‌ها نشان داد که کلمات کلیدی با اولویت بالا (۵ کلمهٔ کلیدی نخست) در هر طبقهٔ موضوعی شباهت زیادی (بین ۸۵-۷۵ درصد) به‌هم دارند (۱۳۸۶).

اگرچه توضیحات کلی «عربی نرئی، وحیدی اصل و مینایی بیدگلی» (۱۳۸۶) درک ایدهٔ کلی آن‌ها را تسهیل نموده، اما حذف جزئیات مهمی همچون نحوهٔ جایگزینی کلمات هم‌معنا و چگونگی برخورد با مسئلهٔ رفع ابهام کلمات و نحوهٔ ایجاد درخت بر اساس رابطه اعم و ابزار احتمالی استفاده‌شده، صحت و دقت نتایج گزارش‌شده توسط آن‌ها را زیر سئوال می‌برد؛ ضمن آن‌که نه‌تنها دربارهٔ ملاک انتخاب متون و چگونگی تخصیص کلمات کلیدی به آن‌ها توضیحی ارائه نشده، بلکه حتی نحوه و معیار ارزیابی آن‌ها از عملکرد سیستم پیشنهادی‌شان نیز بسیار مبهم است.

«فرهودی و یاری» با هدف طبقه‌بندی متون فارسی، اقدام به استخراج عبارات کلیدی

1. hypernym

کردند. آن‌ها از ۸۰۰ مدرک متعلق به ۸ دستهٔ موضوعی مجموعهٔ دادهٔ «روزنامهٔ همشهری» استفاده کرده و پس از حذف واژگان ایستا، برای متون هر دستهٔ موضوعی واژه‌نامه‌ای ایجاد کردند که کلمات رایج آن دستهٔ موضوعی را به‌همراه فراوانی آن‌ها در متون همان دسته دربرمی‌گرفت. به‌علاوه، واژه‌نامهٔ دیگری شامل تمام کلمات موجود در تمام متون نیز تولید و از TF-IDF به‌عنوان معیار انتخاب ویژگی برای تعیین کلمات رایج در هر واژه‌نامه استفاده کردند. سپس، برای هر مدرک بُرداری ایجاد کرده و از TF-CRF[1] برای وزن‌دهی به کلمات هر مدرک استفاده نمودند. سپس، ماتریسی ایجاد کردند که ردیف‌های آن نمایانگر مدارک موجود در نمونه و ستون‌های آن نمایانگر تمامی واژه‌های موجود در واژه‌نامهٔ کلی بود و الگوریتم‌های یادگیری ماشینی KNN[2] و SVM[3] را روی این ماتریس پیاده نمودند تا مشخص شود کدام‌یک در طبقه‌بندی متون فارسی موفق‌تر عمل می‌کند (Farhoodi & Yari 2010).

کاستی پژوهش Farhoodi & Yari (2010) آن بود که آن‌ها ارزیابی خاصی از میزان موفقیت سیستم در انتخاب صحیح کلمات کلیدی ارائه نکردند، بلکه تنها موفقیت نهایی سیستم در تعیین طبقهٔ درست برای مدارک آزمایشی را گزارش کردند؛ درحالی‌که به نظر می‌رسد، نتیجهٔ نهایی به‌شدت تحت تأثیر میزان دقت و بازیافت مرحلهٔ استخراج کلمات کلیدی است. با وجود این، با توجه به این‌که ارزیابی داخلی یکی از روش‌های ارزیابی سیستم‌های استخراج کلمات کلیدی است، موفقیت سیستم در تعیین طبقهٔ درست متون را نیز می‌توان به‌منزلهٔ موفقیت روش استفاده‌شده برای استخراج کلمات کلیدی تعبیر کرد.

«کیان و زاهدی» از استخراج کلمات کلیدی از متون فارسی استفاده کردند تا دسترس‌پذیری محتوای وب از طریق موتورهای جست‌وجو را بهبود بخشند. آن‌ها پس از پیش‌پردازش و حذف واژگان ایستا از متون مورد آزمایش، مقادیر مجموعه‌ای از ویژگی‌ها را برای هر یک از n-gramهای مستخرج از متن محاسبه کرده و از آن‌ها به‌عنوان پارامترهای تابع نمره‌دهی استفاده کردند. سپس، با استفاده از الگوریتم ژنتیک، ویژگی‌های مورد استفاده‌شان را بهینه ساختند. آن‌ها مجموعهٔ داده‌ای شامل محتوای متنی ۴۵۰۰ صفحه وب از سایت‌های رسمی[4] و غیررسمی[5] فارسی ایجاد کرده و از میان آن‌ها محتوای ۶۰۰ صفحه را

1. Frequency-Category Relevancy Factor
2. K-Nearest Neighbor
3. Support Vector Machine
4. official
5. non-official

انتخاب نمودند و از خوانندگان متخصص خواستند تا به هر صفحه کلیدواژه‌هایی تخصیص دهند. سپس، ۱۰۰ مدرک را به‌عنوان مجموعهٔ آزمون و ده صفحهٔ اول نتایج موتورهای جست‌وجوی «گوگل»، «یاهو» و «ام‌اس‌ان»[1] را به‌عنوان پارامترهای سازگاری تعیین کردند و به‌منظور ارزیابی اثربخشی روش پیشنهادی خود مقادیر معیارهای دقت، بازیافت، و F1 را به‌ازای اندازه‌های مختلف مجموعهٔ آموزشی گزارش کردند. آن‌ها TF-IDF نرمال‌سازی‌شده را مهم‌ترین ویژگی رتبه‌بندی برای الگوریتم رتبه‌بندی موتورهای جست‌وجو دانستند. آن‌ها همچنین، ادعا کردند که با استفاده از دو تابع نمره‌دهی پیشنهادی‌شان، بازیافتی بهتر و دقتی در حد بهترین پژوهش‌های صورت‌گرفته روی مدارک فارسی به دست می‌آید و افزون بر این، به‌دلیل استفاده از نتایج موتورهای جست‌وجو برای بهینه‌سازی استخراج کلمات کلیدی، دسترس‌پذیری از طریق موتورهای جست‌وجوی عمومی بهبود می‌یابد (Kian & Zahedi 2011).

آنها توضیحات مبسوط، شفاف و روانی در خصوص روش پیشنهادی خود ارائه کرده‌اند، اما نحوهٔ محاسبهٔ مقادیر معیارهای دقت، بازیافت و F1 در مقالهٔ آن‌ها نامعلوم است و مشخص نیست کدام صفحات وب از میان نتایج ۱۰ صفحهٔ نخست صحیح تلقی شده‌اند. حتی پیش‌تر از آن، مشخص نیست که از کلمات کلیدی استخراج‌شده توسط سیستم پیشنهادی‌شان چگونه برای جست‌وجو استفاده شده است. تعداد و مشخصات کلمات کلیدی استخراج‌شده توسط سیستم نیز در هیچ کجای مقاله گزارش نشده است.

«حسینی خوزانی و بیات» برای استخراج کلمات و عبارات کلیدی از متون فارسی از الگوریتم n-gram برای شناسایی کلماتی با ریشهٔ مشترک استفاده کردند و با استفاده از TF-IDF به هر یک از کلمات وزنی دادند و ماتریسی از این اوزان ایجاد نمودند که از آن برای استخراج کلیدواژه‌های یک یا چندکلمه‌ای استفاده کردند. آنگاه با کمک عبارات کلیدی استخراج‌شده تمام جملات کلیدی متون مورد آزمایش را استخراج نمودند. آن‌ها ایدهٔ خود را روی مجموعهٔ داده‌ای شامل ۱۲۰۰ مدرک از ۷ حوزهٔ موضوعی مجموعهٔ دادهٔ «همشهری» آزمایش کردند و نتایج آن را با الگوریتم هم‌رخدادی کلمات که توسط «ماتسوئو» و ایشیزوکا»[2] آزمایش شده بود، مقایسه کردند. آن‌ها تعداد کلمات غیرتکراری، تعداد کلمات کلیدی به‌دست‌آمده از الگوریتم هم‌رخدادی برای هر مدرک، و تعداد

1. MSN
2. Matsuo and Ishizuka

عبارات کلیدی یک و دوکلمه‌ای به‌دست‌آمده را گزارش کردند (Hosseini Khozani, S. M., H., & Bayat, H. 2011).

آن‌ها نه‌تنها از معیارهای رایج همچون دقت و بازیافت برای ارزیابی ایده‌شان استفاده نکردند، بلکه در بخش بحث و نتیجه‌گیری مقاله نیز هیچ تفسیری در زمینهٔ مقادیر گزارش‌شده در قالب جدول ارائه ننمودند تا خواننده دریابد مفهوم این مقادیر چیست و آیا نتیجهٔ به‌دست‌آمده بهتر از الگوریتم هم‌رخدادی بوده یا خیر. اما استفاده از مجموعهٔ دادهٔ استانداردی همچون پیکرهٔ «روزنامهٔ همشهری» را می‌توان به‌عنوان نکته‌ای مثبت در کار آن‌ها دانست.

در مطالعه‌ای دیگر، «پروین» و همکاران از اصطلاح‌نامه استفاده کردند تا تمام واژه‌های مترادف را با عضوی از خانوادهٔ ترادف‌شان که زودتر از سایرین در متن رخ داده بود، جایگزین کنند. آنگاه، روابط والد و فرزندی کلمات، تضاد/ ترادف و شمول را تعیین کردند. سپس، به کلمات مترادف/ متضاد، وزن یک و به کلماتی که واژهٔ رأس‌شان نیز در همان متن موجود بود، وزن یک به‌علاوهٔ آلفا دادند. آن‌ها ایدهٔ خود را در دو بخش مورد آزمایش قراردادند. در بخش اول، از یک طبقه‌بند ساده استفاده کردند تا کارایی روش پیشنهادی‌شان را نشان دهند و معیارهای صحت، آنتروپی[1] و خلوص[2] را گزارش کردند. در بخش دوم، از الگوریتم خوشه‌بندی «کامیانگین»[3] و اطلاعات متقابل نرمال‌سازی‌شده به‌عنوان معیار ارزیابی استفاده کردند (Parvin et al. 2012).

آن‌ها به‌منظور ارزیابی ایدهٔ پیشنهادی‌شان، ۴۰۰ متن از ۵ دستهٔ موضوعی مختلف از «روزنامهٔ همشهری» را به‌عنوان مجموعهٔ داده جمع‌آوری کردند و با ایجاد یک ماتریس n×m، یک فضای ویژگی تشکیل دادند که در آن، n تعداد کلماتی بود که حداقل در یکی از مقالات مجموعهٔ داده به‌عنوان کلمهٔ رأس شناسایی شده بود. موجودیت ستون j ام و ردیف iام در این ماتریس، فراوانی کلمهٔ رأس jدر مقالهٔ iام بود و m، تعداد مقالات مجموعهٔ داده بود. آن‌ها این ماتریس را یک بار با استفاده و بار دیگر بدون استفاده از اصطلاح‌نامه‌ای که بر اساس دستنامهٔ حری[4] تولید کرده بودند، تکمیل کردند. نتایج نشان

1. entropy 2. purity 3. K-Means

۴. به این رفرنس استناد شده است:

A. Hori. 2003. A manual to make and develop a multilingual thesaurus. Tehran: Scientific Documentation Center (in Persian)

داد که در صورت استفاده از اصطلاح‌نامه، سیستم دچار سردرگمی کمتری می‌شود و طبقهٔ موضوعی مدارک بیشتری را به‌درستی تعیین می‌کند (Parvin et al. 2012).

نکتهٔ تأمل برانگیز این است که افرادی که با فرایند تولید و به‌روزرسانی اصطلاح‌نامه‌ها آشنایی دارند، می‌دانند که این فرایند تا چه حد پیچیده و زمان‌بر است و حتی اصطلاح‌نامه‌های معتبر تخصصی که توسط تیم‌های آکادمیک هدایت و به‌روزرسانی می‌شوند، هرگز به مرحلهٔ تکامل و توقف تولید نمی‌رسند. بنابراین، کفایت اصطلاح‌نامهٔ تولیدشده توسط «پروین» و همکارانش مورد سئوال است. از سوی دیگر، توضیحات ارائه‌شده توسط آن‌ها دربارهٔ روش اجرا و نحوهٔ وزن‌دهی به کلمات کلیدی کاندیدا آنقدر مبهم و پیچیده است که قابلیت تکرار توسط خواننده را ندارد.

«کیان و زاهدی» به‌منظور افزایش دقت، بدون کاهش قابل توجه در مقدار بازیافت، فاز پس‌پردازشی طراحی نمودند که در آن رشته‌هایی از کلمات جالب توجه[1] را با استفاده از فرایند مهندسی معکوس از مجموعهٔ دادهٔ آموزشی استخراج کردند. سپس، منطقهٔ همسایگی هر رشته را مشخص کردند تا تعیین کنند که واژه‌های کاندیدای استخراج‌شده با یک روش، چقدر به کلمات انتخاب‌شده توسط انسان نزدیک است. آن‌ها برای هر رشتهٔ جالب توجه دو مفهوم همسایهٔ راست و چپ تعریف کردند و هر واژه‌ای را که با یک کلیدواژه در مجموعهٔ آموزشی رابطهٔ همسایهٔ چپ یا همسایهٔ راست داشت، به‌عنوان یک AAS برچسب زدند. سپس، برای انتخاب کلمات کلیدی نهایی روش پس‌پردازشی را بر مبنای قانون احتمال کل[2] تعریف کردند و یک تابع نمره‌دهی ایجاد کردند که هرچه مقدار آن بیشتر باشد، به‌معنای احتمال بالاتر رخداد یک واژهٔ کلیدی است و بدین‌ترتیب، کلمات کلیدی نهایی را از میان کلمات کاندیدا با بالاترین نمره انتخاب کردند (Kian & Zahedi 2013).

آن‌ها ایدهٔ خود را روی ۸۰۰ مدرک از موضوعات مختلف مجموعهٔ دادهٔ «روزنامهٔ همشهری» آزمایش کردند که در آن از چند کاربر خواسته بودند به هر یک، کلمات و عبارات کلیدی دلخواهشان را تخصیص دهند. آن‌ها به‌منظور ارزیابی ایدهٔ خود، چندین آزمایش ترتیب دادند و هر بار نتایج را با استفاده از AAS و بدون آن برای ارزیابی کلمات کلیدی انتخاب‌شده گزارش کردند. نتایج نشان داد که در تمام آزمایش‌ها استفاده از AAS

1. attention attractive strings (AAS) 2. total probability theorem

منجر به افزایش دقت و کاهش بازیافت می‌شود.

آنچه درک ایدهٔ اجراشده توسط «کیان و زاهدی» (۲۰۱۳) را دشوار می‌سازد، این است که اولاً تعریف رشته‌های جالب توجه در مقاله و ثانیاً نحوهٔ اجرای مهندسی معکوس برای استخراج چنین رشته‌هایی مبهم است. به‌علاوه، آن‌ها برای ارزیابی ایدهٔ خود از روش‌های ارائه‌شده در مقالاتی استفاده کرده‌اند که برای زبان فارسی طراحی نشده بودند. این است که به نظر می‌رسد چگونگی اجرای این روش‌ها روی متن فارسی جای توضیح بیشتری دارد که در مقالهٔ آن‌ها خالی به نظر می‌رسد.

«احمدی و حسینی‌خواه» به‌منظور استخراج کلمات کلیدی از متن، از دو نوع شبکهٔ عصبی (LVQ و MLP) استفاده کردند و پس از یکدست‌سازی متنِ ۱۰۰ خبر فارسی و حذف کلمات کمتر از سه حرف و نیز واژگان ایستا، کلمات باقی‌مانده را به‌عنوان کلمات کلیدی کاندیدا در نظر گرفتند. سپس، به‌ازای هر کلمهٔ کلیدی کاندیدا فاکتورهای کمّی از قبیل TF، IDF، کلمات عنوان (تخصیص ۰ و ۱) و میانگین موقعیت کلمه در متن را محاسبه کرده و شبکهٔ عصبی‌ای ایجاد کردند که تعداد نورون‌های لایهٔ ورودی آن برابر ۴ (تعداد فاکتورهای کمّی‌شده) و تعداد نورون‌های لایهٔ خروجی آن برابر ۲ (کلیدی بودن یا نبودن هر کلمه) بود. آن‌ها از ۸۰ خبر، برای آموزش دو نوع شبکهٔ عصبی (پرسپترون چندلایه و LVQ) استفاده نمودند و هر مدل را با در نظر گرفتن دسته‌بندی موضوعی و بدون آن‌ها اجرا کردند و هر بار مقادیر دقت و بازیافت را گزارش نمودند. نتایج نشان داد که با در نظر گرفتن دسته‌بندی موضوعی متون و به‌شرط استفاده از اوزان مخصوص هر دستهٔ موضوعی نتایج بهبود می‌یابد و روش MLP هم عملکرد بهتری نسبت به LVQ دارد (۱۳۹۲). اگر از ابهام موجود در نگارش نحوهٔ پیاده‌سازی آزمایش‌ها توسط آن‌ها بگذریم، مقالهٔ آن‌ها یکی از معدود مقالات قابل قبول استخراج عبارات کلیدی از متون فارسی است.

«حسن‌پور و مدنی» از استخراج عبارات کلیدی به‌منظور دسته‌بندی متون فارسی استفاده کردند. آن‌ها لغت‌نامه‌ای از متون مجموعهٔ آموزشی ایجاد نموده و به‌کمک «فارس‌نت» برای هر یک از کلمات آن، بُرداری از کلمات هم‌معنا ایجاد کردند. سپس، تعداد حضور هر یک از کلمات لغت‌نامه و مفاهیم موجود در بُردار هر یک را تعیین کردند و برای رفع ابهام و تعیین مفهوم برنده، هر کلمه را به دسته‌ای از متون تخصیص دادند که بیشترین تکرار را در آن دسته داشته است. آن‌ها با هدف کاهش فضای مشخصه از روش آماری $\chi 2$ استفاده کرده و فرض کردند عبارتی که تعداد تکرار بیشتری در یک

دسته دارد، می‌تواند مشخصهٔ متمایزکنندهٔ خوبی برای آن دسته باشد. بدین‌منظور، ابتدا، ماتریس «عبارت-دسته» را ایجاد و سپس، میزان اثربخشی هر مشخصه را در هر دسته محاسبه کردند. آن‌ها در نهایت، گزارش کردند که استفاده از «فارس‌نت» منجر به بهبود ۲ درصدی در دقت نتایج به‌دست‌آمده از دسته‌بندی می‌گردد (۱۳۹۳).

نکته قابل تأمل دربارهٔ مقالهٔ «حسن‌پور و مدنی» (۱۳۹۳)، این است که هیچ راهنمایی دربارهٔ این‌که هر یک از اجزای فرمول بر چه چیز دلالت دارد و چگونه چنین ترکیبی به نتیجهٔ مورد نظر می‌رسد، ارائه نشده و بنابراین، نحوهٔ استفاده از این رویکرد برای کاهش فضای مشخصه چندان قابل درک نیست. این مقاله نیز همچون اغلب مقالات این حوزه فاقد ویژگی تکرارپذیری است؛ چراکه مراحل انجام کار به‌نحوی قابل درک برای خواننده تشریح نشده است.

«نوریان و یدالله‌زادهٔ طبری» پس از حذف واژگان ایستای اسناد متنی مجموعهٔ دادهٔ «روزنامهٔ همشهری» و ریشه‌یابی کلمات باقی‌مانده، فراوانی ریشهٔ کلمات و پرتکرارترین کلمات پیکره را تعیین کردند. آنگاه کلمات پرتکرار هر یک از دسته‌های موضوعی پیکره را تعیین کردند و با استفاده از آزمون آماری مجذور کای، کلماتی را که در هر سند دارای ارزش اطلاعاتی بالایی بودند، شناسایی کردند. آن‌ها این کار را با هدف دسته‌بندی اسناد فارسی انجام داده و از این کلمات به‌عنوان ورودی دو دسته‌بند مبتنی بر شبکه‌های عصبی (الگوریتم پس‌انتشار و شبکه‌های باور عمیق) استفاده کردند. در نهایت، پس از آزمون و خطاهای متعدد برای دستیابی به مقادیر مطلوب پارامترهای مختلف، مقادیر دقت، بازیافت، معیار F و زمان محاسبهٔ این دو دسته‌بند را گزارش کردند و نتیجه گرفتند که دسته‌بندی با شبکه‌های باور عمیق به‌مراتب بهتر و سریع‌تر از شبکه‌های عصبی عمل می‌کند (۱۳۹۴).

با وجود این‌که «نوریان و یدالله‌زادهٔ طبری» (۱۳۹۴) توضیحات مناسبی دربارهٔ الگوریتم پس‌انتشار و شبکه‌های باور عمیق ارائه کرده‌اند، اما توضیحی دربارهٔ نحوهٔ اعمال آن‌ها نداده‌اند. به‌علاوه، فرمول استفاده‌شده برای محاسبهٔ مجذور کای مملو از پارامترهایی است که چیستی آن‌ها برای خواننده نامشخص است. در مجموع، به نظر می‌رسد که کار آن‌ها توسط محققان دیگر قابل تکرار مجدد نیست.

«ویسی و افلاکی» پس از حذف واژگان ایستای ۵۰۰ متن از سایت‌های خبری فارسی، فراوانی کلمات هر متن و نیز تعداد کلمات هر متن را محاسبه کردند و کلماتی با بیشترین تکرار را به‌عنوان کلمات کلیدی کاندیدا در نظر گرفتند و آن‌ها را به‌صورت دوتایی با

هم جفت کردند. آن‌ها از یک واژه‌نامه[1] و الگوریتم لسک[2] برای اندازه‌گیری شباهت جفت‌ها استفاده کردند و در نهایت، بهترین کلمات را به‌عنوان کلمات کلیدی نهایی انتخاب کردند (۱۳۹۴).

آن‌ها برای ارزیابی روش پیشنهادی خود مقادیر معیارهای دقت و بازیافت به‌دست‌آمده از روش فرکانس عبارت و روش پیشنهادی را برای ۲۰۰ سند خبری از ۵ موضوع محاسبه کردند. مبنای محاسبهٔ این مقادیر، کلمات کلیدی پیش‌فرضی بود که توسط سایت‌های خبری به‌همراه هر خبر منتشر شده بود. نتایج نشان‌دهندهٔ عملکرد بهتر و قابل توجه روش آن‌ها در استخراج ۵ عبارت کلیدی برای هر متن خبری بود. آن‌ها در پایان، ارزیابی انسجام کلمات کلیدی استخراج‌شده با استفاده از روش آنتروپی را نیز پیشنهاد کردند.

نکتهٔ قابل ذکر در مورد مقالهٔ «ویسی و افلاکی» (۱۳۹۴) این است که در مقایسه با سایر مقالات این حوزه وضوح و شفافیت بیشتری در ارائهٔ نحوهٔ انجام کار و نیز مبنا و معیار ارزیابی ایدهٔ پیشنهادی وجود دارد و همین امر تکرار آن را امکان‌پذیر می‌سازد. اما در مورد فارسی بودن زبان اسناد خبری که ایدهٔ پیشنهادی روی آن‌ها پیاده‌سازی شده، جز در چکیدهٔ مقاله اشارهٔ دیگری نشده است.

«باسره» و همکاران کوشیدند تجربهٔ پژوهشگرانی به نام‌های «حدود[3] و عبدالدایم»[4] (۲۰۱۴) را روی زبان فارسی پیاده‌سازی نمایند و هجده ویژگی آماری را از طریق هفت شیوهٔ تصمیم‌گیری چندمعیاره مورد آزمایش قرار دهند. آن‌ها مجموعهٔ داده‌ای شامل ۲۴۴ متن خبری ایجاد کرده و آن‌ها را به‌روشی نامعلوم به‌صورت دستی برچسب‌زنی کردند. سپس، عملیات ریشه‌یابی، حذف واژگان ایستا و اعداد را روی این مجموعهٔ داده پیاده نمودند و مقادیر همهٔ هجده ویژگی آماری مورد نظر را استخراج کرده و عبارات کلیدی کاندیدای ۲۴۴ متن خبری را استخراج نمودند. آن‌ها توزیع و نسبت واژگان کلیدی به واژگان غیرکلیدی را ۵ درصد در برابر ۹۵ درصد گزارش کردند؛ اما دربارهٔ نحوهٔ دستیابی به چنین نسبتی توضیحی ارائه ندادند. آن‌ها مجموعهٔ داده‌ها را برای یادگیری به هفت الگوریتم تصمیم‌گیری داده و مدل حاصل را با معیار F، روی داده‌های آزمون ارزیابی کردند و برای کلمات کلیدی عنوان، امتیاز دو برابر در نظر گرفتند. نتایج نشان داد که دو معیار

1. www.parsi.wiki
2. Lesk
3. Haddoud
4. Abdeddaim

MLE[1] و KLD[2] هنگام استفاده از الگوریتم‌های تصمیم‌گیری، کارایی بالایی دارند. اما اگر از الگوریتم‌های تصمیم‌گیری چندمنظوره استفاده نشود، بالاترین کارایی مربوط به دو ویژگی IDF و DPM-Index است (باسره و همکاران ۱۳۹۴).

آن‌ها با مقایسهٔ نتایج حاصل از اجرای هفت الگوریتم تصمیم‌گیری چندمعیاره و نتایج حاصل از اجرای شیوهٔ آماری IDF دریافتند که الگوریتم‌های تصمیم‌گیری روی متون فارسی کارایی لازم را ندارند. بنابراین، استفاده از IDF به‌تنهایی با همهٔ معیارهای دیگر برابری می‌کند. از نظر آن‌ها برای افزایش دقت باید از ترکیبی از شیوه‌های تصمیم‌گیری استفاده کرد. آن‌ها برای ارزیابی و محاسبهٔ دقت و بازیافت به‌جای استفاده از انطباق دقیق، از انطباق نسبی استفاده کردند، اما دربارهٔ چگونگی این کار توضیحی ندادند. از نظر آن‌ها پیچیدگی پیش‌پردازش و استخراج ویژگی از زبان فارسی از عوامل مؤثر بر دقت سیستم است و به‌کارگیری ابزار معنایی در نظر گرفتن هم‌رخدادی کلمات، به‌کارگیری ترکیبی از الگوریتم‌های تصمیم‌گیری چندمعیاره و در نظر گرفتن معیارهای مناسب‌تر ارزیابی برخی از راهکارهای بهبود دقت سیستم هستند (همان).

«معادی و فولادی» از ترکیبی از ویژگی‌های آماری، قوانین سادهٔ زبان، و بردار تکرار موقعیت برای استخراج کلمات کلیدی از متون فارسی استفاده کردند و پس از پیش‌پردازش متن، از ویژگی‌هایی همچون فراوانی واژه‌ها[3] و طول عمر[4] کلمات برای وزن‌دهی به واژه‌های متن و رتبه‌بندی آن‌ها استفاده کردند. آن‌ها متن پیش‌پردازش‌شده را با استفاده از یک واژه‌نامه شامل واژگان کاربردی[5] غربال کرد و با محاسبهٔ تعداد تکرار و وزن هر کلمه در متن، جدولی از کلیدواژه‌ها و وزن و موقعیت هر یک ایجاد کردند و بر اساس آن، وزن رتبه‌ای اولیهٔ[6] هر کلمه و هم‌رخدادی ۳۰ کلمه با بیشترین وزن مرحلهٔ قبل را محاسبه کردند. روش محاسبهٔ وزن، واریانس تکرار موقعیت برای هر کلمه، هم‌رخدادی و شباهت کلمات به‌گونه‌ای مختصر و مبهم در متن مقاله توضیح داده شده به‌نحوی که تکرار محاسبهٔ آن‌ها را برای خواننده(گان) ناممکن می‌سازد (Maadi & Fouladi 2015).

در هر حال، «معادی و فولادی» برای ارزیابی سیستم پیشنهادی خود پیکره‌ای از متن اصلی ۱۰۰ مقالهٔ فارسی با موضوعات مختلف تولید کردند و کلیدواژه‌هایی را که نویسندهٔ

---

1. Maximum Likelihood Estimation
2. Kullback-Leibler Divergence
3. TF
4. life time (نسبت تعداد کلمات بین اولین و آخرین رخداد یک کلمه به تعداد کل کلمات متن)
5. functional words
6. initial ranking weight

هر مقاله به آن تخصیص داده بود، به‌عنوان نتایج مطلوبِ ارزیابی جدا و ذخیره کردند. آن‌ها مقادیر معیارهای دقت، بازیافت و F-measure را برای ۵ و ۱۰ کلمهٔ کلیدی شناسایی‌شده توسط سیستم گزارش کرده و مطابق انتظار دریافتند که افزایش تعداد کلمات، بازیافت را افزایش و دقت را کاهش می‌دهد (همان).

یکی از نکات سئوال‌برانگیز دربارهٔ کار آن‌ها این است که چطور نتیجه گرفته‌اند که روش پیشنهادی آن‌ها میانگین هارمونیک را بهبود بخشیده است؛ حال‌آن‌که، حتی در جدولی که خودشان گزارش کرده‌اند، شاهد کاهش این مقدار در ۱۰ واژه در مقایسه با ۵ واژه هستیم؟! آن‌ها انتخاب کلیدواژه‌های غیراستاندارد توسط نویسندگان مقالات را مشکل اصلی بسیاری از مقالات می‌دانند. اگرچه میزان بازیافت کلیدواژه‌های نویسنده در مقایسه با کلیدواژه‌های پیشنهادی سیستم قابل قبول نیست، اما به‌نظر ایشان مقدار معیار دقت افزایش یافته است. دربارهٔ نحوهٔ محاسبهٔ دقت و بازیافت توسط آن‌ها توضیحی در مقاله‌شان ارائه نشده، اما به نظر می‌رسد برای این کار تمام عبارات، به کلمه‌های سازنده‌شان شکسته شده‌اند. آن‌ها پس از برشمردن برخی از محدودیت‌های پژوهش، نتایج را امیدبخش و قابل قبول توصیف کردند.

«احمدی و حبیبی» برای استخراج کلمات کلیدی مجموعه‌ای از اسناد فارسی در حوزهٔ مهندسی کامپیوتر به جداسازی کلمات و حذف واژگان ایستا پرداختند. آنگاه مجموعهٔ داده‌ای شامل مقادیر ویژگی‌های TF، IDF، کلمات عنوان و میانگین موقعیت کلمه در مدرک ایجاد کردند و از آن برای کشف دانش با استفاده از شبکه‌های عصبی استفاده کردند. این پژوهشگران به‌منظور تعیین مقدار بهینهٔ تعداد نورون‌ها و لایه‌ها آزمایشات متعددی انجام دادند که در نهایت، بهترین نتیجه را از شبکه‌ای با یک لایهٔ پنهان و ۱۰ نورون به دست آوردند. آن‌ها این شبکه را روی مجموعهٔ داده‌ها اعمال نموده و از الگوریتم‌های مختلفی برای ایجاد مدل استفاده کردند و به‌منظور ارزیابی عملکرد این الگوریتم‌ها از طریق ماتریس آشفتگی، صحت[1]، نرخ خطا، دقت، بازیافت و ویژگی[2] را محاسبه کردند. آن‌ها در پایان به مقایسهٔ نتایج به‌دست‌آمده از اجرای روش پیشنهادی خود با نتایج گزارش‌شده در چهار مقالهٔ دیگر پرداخته و ادعا کردند که ایدهٔ پیشنهادی آن‌ها در مقایسه با سایر ایده‌ها (به‌جز یکی) عملکرد بهتری داشته است. آن‌ها در نهایت، نتیجه گرفتند که با استفاده از

1. accuracy 2. specificity

شبکهٔ عصبی می‌توان کلمات کلیدی را با دقت قابل قبولی از متون فارسی استخراج کرد (۱۳۹۵).

نکتهٔ حائز اهمیت دربارهٔ مقالهٔ «احمدی و حبیبی» (۱۳۹۵) این است که هیچ اطلاعات کاملی دربارهٔ حجم مجموعهٔ داده و ملاک محاسبهٔ مقادیر معیارهای گزارش‌شده در مقاله ارائه نشده است.

«راد» و همکاران با استفاده از یک اصطلاح‌نامه کلمات هم‌معنا، اجداد و وابسته‌ها را شناسایی کردند تا از آن‌ها برای وزن‌دهی استفاده کنند. آن‌ها عدم توجه به زنجیرهٔ واژگانی را اِشکال اصلی روش‌های موجود در بازیابی متن در زبان فارسی دانستند و از میان کلمات هم‌معنای موجود در متن یکی را به‌عنوان نماینده انتخاب کردند و به‌ازای هر بار تکرار هر یک از کلمات هم‌معنای آن، یک امتیاز به امتیازات کلمهٔ نماینده اضافه نمودند. سپس، برای کلماتی که با کلمات نماینده رابطهٔ اعم، اخص، و هم‌بسته داشتند، وزن یک چهارم در نظر گرفتند. در نهایت، کلمات دارای بیشترین وزن و کلمات پدر را به‌عنوان کلمهٔ کلیدی معرفی نمودند (۱۳۹۵).

آن‌ها به‌منظور پیاده‌سازی روش پیشنهادی خود از مدل فضای برداری استفاده کردند و با انجام آزمایش‌هایی، حد آستانهٔ ۳ و ۴ امتیاز را برای انتخاب یک کلمه به‌عنوان کلمهٔ کلیدی تعیین نمودند. سپس، از ۶۹۰ مقاله از ۵ دستهٔ موضوعی سایت «روزنامهٔ همشهری» برای ارزیابی روش خود استفاده نمودند. آن‌ها به‌جای استفاده از معیارهای استاندارد ارزیابی سیستم‌های استخراج کلمات کلیدی، درصد کلمات اضافی اشتباه و درصد کلمات اصلی به‌اشتباه حذف‌شده را در موضوعات مختلف گزارش کردند. به‌علاوه، در مقالهٔ آن‌ها مشخص نیست که اصولاً کلمات کلیدی درست برای هر متن چه بوده که بر اساس آن‌ها این مقادیر محاسبه شده است (راد و همکاران ۱۳۹۵).

«باسره» و همکاران در مقاله‌ای که ظاهراً گزارش همان پروژهٔ سال ۱۳۹۴ از زاویه‌ای دیگر است، برای هر عبارت، ۱۸ ویژگی آماری را محاسبه نموده و عبارات عنوان و خلاصهٔ اخبار را استخراج کردند و به عبارات عنوان دو برابر سایر عبارات امتیاز دادند. آنگاه، به مقایسهٔ مقادیر ویژگی‌های عبارات پرداخته و با فرض مسئلهٔ استخراج عبارات کلیدی به‌عنوان یک دسته‌بندی دو کلاسه، شش الگوریتم دسته‌بندی را روی مجموعهٔ داده پیاده‌سازی نمودند و مقادیر معیار F1 به‌دست‌آمده را با هم مقایسه کردند و به این

نتیجه رسیدند که الگوریتم دسته‌بند جنگل تصادفی[1] بهترین نتایج را تولید می‌کند (۱۳۹۶). آن‌ها از امتیاز تعلق‌گرفته به هر عبارت توسط دسته‌بند برای ایجاد لیست رتبه‌بندی‌شده و انتخاب n عبارت برتر به‌عنوان عبارات کلیدی استفاده کردند؛ زیرا از نظر آن‌ها عدم توازن داده‌گان در پژوهش‌های استخراج عبارات کلیدی باعث دقت بالا در شناسایی عبارات غیرکلیدی و دقت پایین در شناسایی عبارات کلیدی می‌گردد. اما، یافته‌ها نشان داد که اگر از امتیاز تعلق‌گرفته به هر عبارت توسط دسته‌بند جنگل تصادفی برای ایجاد یک لیست رتبه‌بندی‌شده و انتخاب n عبارت برتر به‌عنوان عبارات کلیدی استفاده شود، هیچ‌یک از روش‌های نمونه‌برداری نمی‌تواند نتایج را بهبود بخشد؛ حال آن‌که، اگر از خروجی دسته‌بند برای تصمیم‌گیری نهایی دربارهٔ کلیدی بودن یا نبودن عبارات کاندیدا استفاده شود، شیوهٔ نمونه‌برداری SYN-OS با توزیع ۵۰-۵۰ می‌تواند F1 سیستم را تا سه برابر افزایش دهد و میزان درستی نتایج را به‌اندازهٔ زمانی برساند که از امتیاز تعلق‌گرفته به هر عبارت برای تعیین کلیدی بودن یا نبودن آن استفاده می‌شود.

نکتهٔ قابل ذکر در مورد مقالهٔ «باسره» و همکاران (۱۳۹۶) این است که اگرچه آن‌ها در این مقاله توضیحات بیشتری در خصوص نحوهٔ ایجاد مجموعهٔ دادهٔ پژوهش خود ارائه کرده‌اند، اما همچنان نحوهٔ برچسب‌زنی اخبار و این‌که چه کسی این کار را انجام داده، برای خواننده مبهم است. به‌علاوه، توضیحات ارائه‌شده، به‌ویژه در بخش «نحوهٔ دسته‌بندی» به‌قدری مبهم است که درک مقاله و نتایج حاصل از آن را برای خواننده دشوار می‌سازد و محتوای جدول‌ها نیز به‌خوبی دسته‌بندی نشده و مقایسه را دشوار می‌کند. به‌علاوه، آن‌ها در مقالهٔ پیشین خود طول مناسب عبارات کاندیدا را حداکثر سه کلمه تعیین کرده بودند؛ با این حال، در این مقاله از عباراتی با طول حداکثر ۴ کلمه برای تعیین عبارات کلیدی استفاده کرده‌اند و توضیحی در این مورد نداده‌اند. به‌هرشکل، مقالهٔ آن‌ها از قوی‌ترین پژوهش‌های حوزهٔ استخراج عبارات کلیدی از متون فارسی است.

«رضایی» و همکاران نیز با هدف بهبود طبقه‌بندی متون فارسی از اصطلاح‌نامهٔ «اصفا» برای یافتن کلمات هم‌خانواده، مترادف، متضاد، اعم و اخص و وابسته تک‌تک کلمات باقی‌مانده از مرحلهٔ پیش‌پردازش و وزن‌دهی به آن‌ها استفاده کردند و هر جا در متن، یک کلمه یا یکی از مترادف‌های آن دیده می‌شد، یک واحد به وزن آن اضافه می‌کردند.

1. random forest

به‌همین‌ترتیب، حضور کلمات هم‌خانواده، اعم و اخص یک کلمه در متن منجر به افزایش وزن آن به‌ترتیب، به‌اندازه ۰/۴ و ۰/۲ می‌شد. آن‌ها به‌منظور ارزیابی ایدهٔ خود روش طبقه‌بندی متون و مدل فضای برداری را روی ۶۹۰ مقاله از ۵ دستهٔ موضوعی «روزنامهٔ همشهری»، شبیه مجموعهٔ دادهٔ استفاده‌شده توسط «راد» و همکاران (۱۳۹۵)، پیاده کردند و این کار را به‌کمک الگوریتم طبقه‌بندی KNN انجام دادند و نتایج حاصل از عدم استفاده از اصطلاح‌نامه، استفاده از اصطلاح‌نامه تنها در سطح مترادف و استفاده از اصطلاح‌نامه با روش پیشنهادی خود را برای چهار روش 1-KNN، 3-KNN، 5-KNN و MLP ارائه نمودند. بررسی نتایج نشان‌دهندهٔ برتری روش 1-KNN در قیاس با سایر روش‌ها بود (۱۳۹۶).

با این‌که «رضایی» و همکاران (۱۳۹۶) به‌طور قابل قبولی نحوهٔ برآورد مقادیر بهینهٔ ضرایب استفاده‌شده برای کلمات هم‌خانواده و اعم و اخص را تشریح کرده‌اند، اما محتوای مقالهٔ آن‌ها پراکندگی زیادی دارد و علی‌رغم این‌که هدف‌شان از وزن‌دهی به کلمات کلیدی افزایش جامعیت جست‌وجو بود، در مقالهٔ خود رابطه‌ای بین وزن‌دهی به کلمات و اثر آن بر جست‌وجو برقرار نکرده‌اند.

«شریفی و مهدوی» ایدهٔ ایجاد زنجیره‌های لغوی برای متون و ارائهٔ آن‌ها به یک سیستم یادگیری ماشینی با ناظر را روی عنوان، چکیده و واژگان کلیدی ۱۰۲ مقاله از سه مجلهٔ علمی-پژوهشی فارسی در زمینهٔ علوم رایانه اجرا و ارزیابی نمودند. آن‌ها پس از برچسب‌گذاری نحوی، تعیین عبارات اسمی و ریشه‌یابی واژه‌ها، با استفاده از اطلاعات نحوی و کتابخانهٔ «هضم»[1]، کلمات کلیدی کاندیدای یک و دوکلمه‌ای بخشی از متون مجموعهٔ داده را استخراج نمودند. آنگاه، پس از رفع ابهام کلمات کاندیدا از نسخهٔ دوم هستان‌شناسی «فارس‌نت» برای ایجاد زنجیرهٔ لغوی بر اساس روش «گالی و مک‌کیوان»[2] استفاده کردند. آن‌ها برای هر واژهٔ کاندیدا ده ویژگی شامل شش ویژگی مربوط به زنجیرهٔ لغوی و چهار ویژگی آماری را تعیین کردند. آنگاه با ایجاد بردار ویژگی برای هر واژه، دادهٔ آموزشی تولید کرده و از مجموعهٔ دسته‌بندهای «کتابخانهٔ وکا»[3]، برای تعیین مقدار بردارهای ویژگی متن ورودی استفاده کردند تا کلیدی بودن یا نبودن هر واژهٔ کاندیدا را تعیین کنند (۱۳۹۷).

آن‌ها از نرم‌افزار «وکا» برای ارزیابی ایدهٔ خود استفاده کرده و داده‌ها را با چند

1. HAZM 2. Galley and McKeown 3. Weka

دسته‌بند مختلف ارزیابی کردند و هر بار مقادیر معیارهای دقت، بازیافت و F را گزارش نمودند. نتایج نشان داد که ماشین بردار پشتیبان با بالاترین دقت نسبت به سایر دسته‌بندها واژگان کلیدی متن را تعیین می‌کند، اما در عین حال، کمترین بازیافت و در نتیجه، کمترین مقدار F را نیز به دست می‌دهد. به‌همین‌دلیل، آن‌ها آزمایش را با حذف چهار مورد از ویژگی‌های مربوط به زنجیره‌های لغوی تکرار کردند و این بار با استفاده از دسته‌بند «بیز»[1] ساده به بالاترین مقدار معیارهای بازیافت و F دست یافتند و نتیجه گرفتند که این چهار ویژگی تأثیر زیادی در نتایج به‌دست‌آمده ندارند. آن‌ها در پایان نتیجه گرفتند که استفاده از ویژگی‌های معنایی در کنار ویژگی‌های آماری باعث بهبود نتایج می‌شود و استفاده از دسته‌بند «بیز» ساده بهترین نتایج را در تعیین واژگان کلیدی و غیرکلیدی حاصل می‌نماید. آن‌ها کم‌کردن تعداد واژگان کاندیدای اولیه را در افزایش دقت تعیین واژگان کلیدی مؤثر دانسته و استفاده از اصطلاح‌نامه‌ها یا پیکره‌های تخصصی را برای پردازش متون تخصصی و تعیین ارتباط میان واژگان پیشنهاد نمودند (همان).

شریفی و مهدوی (۱۳۹۷) برای دستیابی به نتایج قابل قبول ناچار شدند ۴ ویژگی از ۶ ویژگی‌ای را که از طریق زنجیره‌های لغوی به دست آورده بودند، حذف کنند. اما گزارشی که نشان دهد آیا اساساً حذف تمامی این ویژگی‌ها لطمه‌ای به نتایج حاصل از ویژگی‌های صرفاً آماری می‌زد یا خیر، ارائه نکرده‌اند. سؤالی که پیش می‌آید این است که آیا تأثیر حاصل از دو ویژگی به‌دست‌آمده از زنجیره‌های لغوی، ارزش کاری را که برای استخراج چنین زنجیره‌هایی باید انجام شود، دارد یا خیر؟

٤. بحث و نتیجه‌گیری

جست‌وجوها منجر به شناسایی ۱۴ مقالهٔ فارسی و ۶ مقالهٔ انگلیسی شد که به موضوع استخراج کلمات و عبارات کلیدی از متون فارسی پرداخته‌اند. خلاصه‌ای از این مقالات به‌ترتیبِ زمان انتشار در جدول ۳، آمده است.

روش‌ها و اطلاعات استفاده‌شده در این مقالات را می‌توان به چند دسته تقسیم کرد:

استفاده از اطلاعات آماری استخراج‌شده از متون (برای نمونه: بشیری، کربلائی و موسوی ۱۳۸۴؛ عربی نرئی، وحیدی اصل و مینایی بیدگلی ۱۳۸۶؛ Farhoodi and Yari

1. Bayes

2010؛ نوریان و یدالهزادهٔ طبری ۱۳۹۴؛ Maadi and Fouladi 2015؛ احمدی و حبیبی ۱۳۹۵)، استفاده از اطلاعات زبانشناختی (برای نمونه: بشیری، کربلائی و موسوی ۱۳۸۴؛ عربی نرئی، وحیدی اصل و مینایی بیدگلی ۱۳۸۶؛ Kian and Zahedi 2011؛ راد و همکاران ۱۳۹۵)، روشهای یادگیری ماشینی (برای نمونه: حسنپور و مدنی ۱۳۹۳؛ باسره و همکاران ۱۳۹۴؛ شریفی و مهدوی ۱۳۹۷) و سایر روشها و اطلاعات (برای نمونه، قانون Zipf: تشکری و میبدی ۱۳۸۲؛ شبکهٔ عصبی: احمدی و حسینیخواه ۱۳۹۲؛ الگوریتم خوشهبندی کامیانگین: Parvin et al., 2012؛ الگوریتم لسک: ویسی و افلاکی ۱۳۹۴؛ الگوریتمهای دستهبندی: باسره و همکاران ۱۳۹۶؛ اصطلاحنامه و هستانشناسی: راد و همکاران ۱۳۹۵، شریفی و مهدوی ۱۳۹۷ و Parvin et al. 2012). بررسیها نشان داد که اکثر این مقالات یا در روششناسی انتخابشده ایراد دارند و یا نویسندگان نتوانستهاند ایدهٔ پیشنهادی خود را بهوضوح برای خواننده تبیین نمایند (برای نمونه: نوریان و یدالهزادهٔ طبری ۱۳۹۴؛ حسنپور و مدنی ۱۳۹۳؛ احمدی و حسینیخواه ۱۳۹۲؛ Kian, Zahedi 2013؛ Parvin et al. 2012؛ عربی نرئی، وحیدی اصل و مینایی بیدگلی ۱۳۸۶؛ تشکری و میبدی ۱۳۸۲). بهعلاوه، نحوهٔ ارزیابی ایدهها اغلب مبهم یا دارای اشکال است. در بسیاری از مقالات از مجموعهٔ دادهٔ استانداردی برای ارزیابی سیستم استفاده نشده (برای نمونه: احمدی و حبیبی ۱۳۹۵؛ باسره و همکاران ۱۳۹۴) و نحوهٔ محاسبهٔ معیارهای ارزیابی مبهم یا دارای اشکال است (برای نمونه: احمدی و حبیبی ۱۳۹۵؛ Maadi and Fouladi 2015؛ Kian and Zahedi 2011؛ راد و همکاران ۱۳۹۵). برخی از مقالات از معیارها و ملاکهای استاندارد برای ارزیابی سیستم استفاده نکردهاند (برای نمونه: بشیری و همکاران ۱۳۸۴؛ HosseiniKhozani and Bayat 2011؛ عربی نرئی، وحیدی اصل و مینایی بیدگلی ۱۳۸۶؛ Farhoodi and Yari 2010) و برخی دیگر نیز نتوانستهاند خواننده را متقاعد کنند که ایدهٔ بنیادین آنها ارزشمند و مؤثر است (برای نمونه: شریفی و مهدوی ۱۳۹۷؛ باسره و همکاران ۱۳۹۴). در برخی مقالات، جداول و دادههایی آمده که در متن بهخوبی تشریح نشدهاند و درک آنها برای خواننده دشوار یا ناممکن است (برای نمونه باسره و همکاران ۱۳۹۶؛ Khozani and Bayat 2011؛ Maadi and Fouladi 2015) و برخی نیز از ابزار ناکارآمدی برای پیادهسازی ایدهٔ خود استفاده کردهاند و همین امر نتایج را تحت تأثیر قرار داده است (برای نمونه: Parvin et al. 2012؛ احمدی و حسینیخواه ۱۳۹۲).

در مجموع، میتوان گفت که بهجز ۳ مقالهٔ «ویسی و افلاکی» (۱۳۹۴)، «احمدی و حسینیخواه» (۱۳۹۲)، و «باسره» و همکاران (۱۳۹۶) که روش اجراشده را بهنحو نسبتاً

قابل قبولی گزارش کرده و معیارها و نحوهٔ ارزیابی‌شان واضح و قابل درک است و نتایج واقع‌بینانه‌ای را گزارش کرده‌اند، سایر مقالات قابلیت تکرارپذیری و تعمیم ندارند. بنابراین، نمی‌توان از آن‌ها به‌عنوان معیار پایه‌ای برای ارزیابی سیستم‌های آینده استفاده کرد یا از ایدهٔ مطرح‌شده در آن‌ها با اطمینان در ساخت و توسعهٔ نرم‌افزارهای کاربردی و عملی در حوزهٔ استخراج کلمات کلیدی استفاده نمود.

جدول ۳. خلاصهٔ روش‌ها، ابزارها، معیارها، مجموعهٔ داده‌ها و زبان مقالات مرورشده

نویسندگان	ابزار کمکی	روش	مجموعهٔ داده	معیارهای ارزیابی گزارش‌شده	زبان مجموعهٔ داده	زبان انتشار مقاله
میبدی و تشکری (۱۳۸۲)	–	حذف واژگان ایستا، قانون zipf	۴۵۰ چکیدهٔ مقاله و پایان‌نامهٔ فارسی و ۳۲ پرس‌وجو	Precision, Recall	فارسی	فارسی
بشیری، کربلائی و موسوی (۱۳۸۴)	–	حذف واژگان ایستا، ریشه‌یابی بر اساس ریشه‌یاب انگلیسی Porter	۴۵۰ چکیدهٔ مقاله و پایان‌نامهٔ فارسی و ۳۲ پرس‌وجو	Precision, Recall	فارسی	فارسی
گزنی (۱۳۸۵)	سیاههٔ واژگان بازدارندهٔ موضوعی	ویژگی‌های آماری	۵۰ مقالهٔ فنی	گزارش‌نشده	فارسی	فارسی
عربی نرئی، وحیدی اصل و مینایی بیدگلی (۱۳۸۶)	لغت‌نامه الکترونیکی	روش Luhn	۱۰۰ متن فارسی از ۵ طبقهٔ موضوعی مختلف	درصد شباهت	فارسی	فارسی
احمدی و حسینی‌خواه (۱۳۹۲)	فارس‌نت	شبکهٔ عصبی، یادگیری با ناظر	۶۲۳ سند از ۵ دستهٔ موضوعی از مجموعهٔ دادهٔ «همشهری»	Precision, Recall, F1	فارسی	فارسی
حسن‌پور و مدنی (۱۳۹۳)	آنتولوژی لغوی فارس‌نت	یادگیری ماشینی، انتخاب ویژگی با الگوریتم X^2، TF-IDF نرمال‌شده	۶۲۳ سند از ۵ دستهٔ موضوعی از مجموعهٔ دادهٔ «همشهری»	Precision, Recall F1	فارسی	فارسی

نویسندگان	ابزار کمکی	روش	مجموعهٔ داده	معیارهای ارزیابی گزارش‌شده	زبان مجموعهٔ داده	زبان انتشار مقاله
نوریان و یداله‌زادهٔ طبری (۱۳۹۴): دسته‌بندی/ طبقه‌بندی	کتابخانهٔ هضم و کتابخانه NLTK زبان پایتون	شبکهٔ عصبی	مجموعهٔ دادهٔ «همشهری»	Precision, Recall, F-measure کارایی	فارسی	فارسی
ویسی و افلاکی (۱۳۹۴)	لغت‌نامه	آنالیز آماری	۲۰۰ متن خبری از سایت‌های خبری فارسی زبان	Precision, Recall	فارسی	فارسی
باسره و همکاران (۱۳۹۴)	–	۱۸ ویژگی آماری، هفت شیوهٔ تصمیم‌گیری چندمعیاره	۲۴۴ متن خبری فارسی	Precision, Recall, F-measure	فارسی	فارسی
حسینی‌خوزانی و بیات (۲۰۱۱)		روش آماری، الگوریتم N-Gram، TF*IDF	۱۲۰۰ مقاله از ۷ دسته پیکرهٔ «همشهری»	تعداد کلمات غیرتکراری هر مدرک، تعداد کلمات کلیدی به‌دست‌آمده از الگوریتم هم‌رخدادی، تعداد عبارات کلیدی یک و دوکلمه‌ای	فارسی	انگلیسی
کیان و زاهدی (۲۰۱۱)		الگوریتم ژنتیک	۴۵۰۰ صفحهٔ وب از سایت‌های رسمی و غیررسمی فارسی	Precision, Recall, F-measure	فارسی	انگلیسی
پروین و همکاران (۲۰۱۲)	اصطلاح‌نامه محقق‌ساخته	طبقه‌بند ساده، الگوریتم خوشه‌بندی کامیانگین	۶۹۰ مقاله از مجموعهٔ دادهٔ «همشهری»	NMI, Accuracy, Purity, Entropy	فارسی	انگلیسی
کیان و زاهدی (۲۰۱۳)		رشته‌های جالب توجه، قانون احتمال کل، منطقهٔ اطمینان	۸۰۰ مدرک از موضوعات مختلف مجموعهٔ دادهٔ «همشهری»	Precision, Recall, F-measure	فارسی	انگلیسی

نویسندگان	ابزار کمکی	روش	مجموعهٔ داده	معیارهای ارزیابی گزارش‌شده	زبان مجموعهٔ داده	زبان انتشار مقاله
معادی و فولادی (۲۰۱۵)		ترکیبی از ویژگی‌های آماری، قوانین سادهٔ زبان، بردار تکرار- موقعیت کلمات	۱۰۰ مقالهٔ فارسی منتخب از مجموعهٔ مقالات و مجلات معتبر در موضوعات مختلف	Precision, Recall, F-measure	فارسی	انگلیسی
احمدی و حبیبی (۱۳۹۵)	نرم افزار متلب	شبکهٔ عصبی	مجموعه‌ای از اسناد حوزهٔ مهندسی کامپیوتر	Precision, Recall, Accuracy, Error Rate	فارسی	فارسی
راد و همکاران (۱۳۹۵)	ترکیبی از اصطلاح‌نامهٔ اسلامی و اصفا	اطلاعات زبان‌شناختی، خوشه‌بندی	۶۹۰ مقاله در ۵ دستهٔ موضوعی از سایت «روزنامهٔ همشهری»	Precision, F-measure, NMI, Purity, Entropy	فارسی	فارسی
باسره و همکاران (۱۳۹۶)		ترکیبی از شیوه‌های زبان‌شناختی، یادگیری ماشینی با ناظر، ویژگی‌های آماری، دسته‌بند جنگل تصادفی	۲۴۴ متن خبری از اخبار ۳۲ خبرگزاری	Precision, Recall, F-measure	فارسی	فارسی
رضایی و همکاران (۱۳۹۶)	اصطلاح‌نامهٔ اصفا	یادگیری ماشینی با ناظر، مدل فضای برداری، الگوریتم طبقه‌بندی KNN	۶۹۰ مقاله در ۵ دستهٔ موضوعی از سایت «روزنامهٔ همشهری»	صحت، درصد کلماتی که به‌اشتباه جزء کلمات اصلی شناسایی شده بودند، درصد کلماتی که به‌اشتباه جزء کلمات اضافی در نظر گرفته شده بودند.	فارسی	فارسی
شریفی و مهدوی (۱۳۹۷)	فارس‌نت، کتابخانهٔ هضم، نرم‌افزار وکا	یادگیری ماشینی با ناظر	۱۰۲ مقاله از ۳ مجلهٔ علمی- پژوهشی فارسی در زمینهٔ علوم رایانه	Precision, Recall F-measure	فارسی	فارسی

فهرست منابع

احمدی، عباس، و طیبه حسینی‌خواه. ۱۳۹۲. استخراج کلمات کلیدی یک متن با استفاده از شبکه‌های عصبی. در *دهمین کنفرانس بین‌المللی مهندسی صنایع*. تهران: دانشگاه تهران.

احمدی، علی‌اصغر، و مریم حبیبی. ۱۳۹۵. استخراج کلمات کلیدی فارسی با استفاده از تکنیک‌های داده‌کاوی. در *چهارمین کنفرانس بین‌المللی در مهندسی برق و کامپیوتر*. تهران: مؤسسهٔ آموزش عالی صالحان، دانشکده مدیریت دانشگاه تهران.

باسره، مریم، ولی درهمی، و سجاد ظریف‌زاده. ۱۳۹۶. ارائهٔ روشی برای استخراج خودکار عبارات کلیدی از اخبار وب پارسی. *مجله مهندسی برق دانشگاه تبریز*. ۴۷ (۸۱): ۸۵۷-۸۶۶.

_____، و صادق طاهرزاده. ۱۳۹۴. به‌کارگیری شیوه‌های آماری و تصمیم‌گیری چندمعیاره در استخراج عبارات کلیدی از صفحات خبری وب پارسی. در *هفتمین کنفرانس بین‌المللی فناوری اطلاعات و دانش*. ارومیه: دانشگاه ارومیه.

بشیری، حسن، فاطمه کربلائی، و شیرین موسوی. ۱۳۸۴. طراحی و ارزیابی نمایه‌ساز خودکار متون فارسی. در *یازدهمین کنفرانس بین‌المللی کامپیوتر*. تهران: انجمن کامپیوتر ایران، پژوهشگاه دانش‌های بنیادی، پژوهشکده علوم کامپیوتر.

تشکری، مسعود، و محمدرضا میبدی. ۱۳۸۲. ساخت یک نمایه‌ساز خودکار برای متون فارسی. در *یازدهمین کنفرانس مهندسی برق*. شیراز: دانشگاه شیراز.

حسن‌پور، حمید، و صبا مدنی. ۱۳۹۳. بهبود دقت سیستم دسته‌بندی خودکار اسناد فارسی به‌کمک هستان‌شناسی فارس‌نت. *مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات* ۳ (۱): ۴۵-۵۵.

راد، فرهاد، حمید پروین، آتوسا دهباشی، و بهروز مینایی. ۱۳۹۵. ارائه روشی جدید برای شاخص‌گذاری خودکار و استخراج کلمات کلیدی برای بازیابی اطلاعات و خوشه‌بندی متون. *فصلنامه پردازش علائم و داده‌ها* ۱۳ (۱): ۸۷–۱۰۰.

رضایی، وحیده، مجید محمدپور، حمید پروین، و صمد نجاتیان. ۱۳۹۶. ارائه روشی برای استخراج کلمات کلیدی و وزن‌دهی کلمات برای بهبود طبقه‌بندی متون فارسی. *فصلنامه پردازش علائم و داده‌ها* ۱۴ (۴): ۵۵–۷۸.

شریفی، عطیه، و محمدامین مهدوی. ۱۳۹۷. رویکردی باناظر در استخراج واژگان کلیدی اسناد فارسی با استفاده از زنجیره‌های لغوی. *پردازش علائم و داده‌ها* ۴: ۹۵-۱۰۹.

عربی نرئی، سمیه، مجتبی وحیدی اصل، و بهروز مینایی بیدگلی. ۱۳۸۶. استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی. در *اولین کنفرانس داده‌کاوی ایران*، ۱–۹. تهران: دانشگاه صنعتی امیر کبیر، مؤسسهٔ پژوهشی داده‌پردازان گیتا IDMC۰۱-https://www.civilica.com/Paper. (دسترسی در ۱۳۹۸/۷/۴-).

گزنی، علی. ۱۳۸۵. استخراج خودکار عبارت‌های کلیدی از متون مقاله‌های فارسی. *کتابداری و اطلاع رسانی*

۹ (۳): ۹۷–۱۰۸.

نوریان، زهرا، و میثم یداله‌زاده‌طبری. ۱۳۹۴. دسته‌بندی اسناد فارسی با استفاده از شبکه‌های عصبی. در *سومین کنفرانس بین‌المللی دستاوردهای نوین در علوم مهندسی و پایه*. اوکراین.

ویسی، هادی، و نیلوفر افلاکی. ۱۳۹۴. استخراج کلمات کلیدی متن فارسی با استفاده از آنالیز آماری. در *کنفرانس بین‌المللی مهندسی و علوم کاربردی*. دوبی.

References

Farhoodi, M., & A. Yari. 2010. Applying Machine Learning Algorithms for Automatic Persian text Classification. Proceedings of 6th International Conference on Advanced Information Management and Service, IMS2010, with ICMIA2010 - 2nd International Conference on Data Mining and Intelligent Information Technology Applications (pp. 318–323). Venetian, Macao.

Hosseine Khozani, S. M. H., & Bayat, H. 2011. Specialization of Keyword Extraction Approach to Persian Texts. Proceedings of the 2011 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2011, 112–116. https://doi.org/10.1109/SoCPaR.2011.6089124 (accessed June 23, 2019)

Kian, H. H., & M. Zahedi. 2011. An Efficient Approach for Keyword Selection: Improving Accessibility of Web Contents by General Search Engines. *International Journal of Web & Semantic Technology (IJWesT)*, 2(4). https://doi.org/10.5121/ijwest.2011.2406 (accessed Aug. 28, 2019)

_____. 2013. Improving Precision in Automatic Keyword Extraction Using Attention Attractive Strings. *Arab J Sci Eng*, 38, 2063–2068. https://doi.org/10.1007/s13369-013-0573-6 (accessed Aug. 28, 2019)

Maadi, M., & K. Fouladi. 2015. Providing a Method for Extracting Keywords in the Persian Language. *International Academic Journal of Innovative Research*, 2(11), 34–42. Retrieved from www.iaiest.com (accessed June 23, 2019)

Onwuegbuzie, A. J., & R. Frels. 2016. Seven Steps to a Comprehensive Literature Review: A Multimodal and Cultural Approach. Los Angeles: SAGE Publications.

Parvin, H., A. Dahbashi, S. Parvin, & B. Minaei-Bidgoli. 2012. Improving Persian Text Classification and Clustering Using Persian Thesaurus. *Advances in Intelligent and Soft Computing*, 151 AISC, 493–500. https://doi.org/10.1007/978-3-642-28765-7_59 (accessed Feb. 24, 2018)

عاطفه کلانتری

متولد سال ۱۳۶۱، دانشجوی دکتری علم اطلاعات و دانش‌شناسی (بازیابی اطلاعات) دانشگاه شیراز است. ایشان هم‌اکنون مسئول کتابخانهٔ دانشکدهٔ پرستاری و مامایی دانشگاه علوم پزشکی قزوین است. پردازش زبان طبیعی، سامانه‌های توصیه‌گر، خدمات مرجع و بازیابی اطلاعات شخصی‌سازی‌شده از جمله علائق پژوهشی وی است.

عبدالرسول جوکار

متولد سال ۱۳۲۷، دارای مدرک دکتری از «دانشگاه لافبورو» Loughborough University of Tecnology در انگلستان است. ایشان هم‌اکنون استاد گروه علم اطلاعات و دانش‌شناسی دانشگاه شیراز است.

روش‌شناسی تحقیق، رفتارهای اطلاع‌یابی و ... از جمله علائق پژوهشی وی است.

سیدمصطفی فخراحمد

متولد سال ۱۳۵۹، دارای مدرک تحصیلی دکتری در رشتهٔ مهندسی کامپیوتر از دانشگاه شیراز است. ایشان هم‌اکنون دانشیار بخش مهندسی و علوم کامپیوتر دانشگاه شیراز است.

داده‌کاوی، متن‌کاوی، پردازش زبان طبیعی و طراحی سیستم‌های خبره از جمله علائق پژوهشی وی است.

جواد عباس‌پور

متولد سال ۱۳۵۶، دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی از دانشگاه تهران است. ایشان هم‌اکنون استادیار گروه علم اطلاعات و دانش‌شناسی دانشگاه شیراز است.

مسائل و چالش‌های بازیابی زبان فارسی، طراحی و ارزیابی نظام‌های بازیابی اطلاعات و سامانه‌های توصیه‌گر از جمله علائق پژوهشی وی است.

هاجر ستوده

متولد سال ۱۳۵۰، دارای مدرک تحصیلی دکتری در رشتهٔ علم اطلاعات و دانش‌شناسی است. ایشان هم‌اکنون دانشیار دانشگاه شیراز است.

علم‌سنجی، دسترسی آزاد به اطلاعات علمی، بازیابی استنادی و پردازش زبان طبیعی از جمله علائق پژوهشی وی است.

مسعود مرتضوی نصرآباد

متولد سال ۱۳۶۵، دانشجوی دکتری علم اطلاعات و دانش‌شناسی دانشگاه فردوسی مشهد است. ایشان هم‌اکنون کارشناس تست و توسعهٔ شرکت پارس‌آذرخش است.

کتابخانهٔ دیجیتال، استانداردهای فایل دیجیتال متنی، کاربرد هوش مصنوعی در جست‌وجو و بازیابی اطلاعات و کاربرد ابزارهای موبایلی در کتابخانهٔ دیجیتال از جمله علائق پژوهشی وی است.

امیر جوادی

متولد سال ۱۳۴۴، دارای مدرک تحصیلی دکتری در رشتهٔ انفورماتیک پزشکی از دانشگاه علوم پزشکی تهران است. ایشان هم‌اکنون استادیار دانشگاه علوم پزشکی قزوین است.

یادگیری ماشینی، هوش مصنوعی و تحلیل‌های آماری از علائق پژوهشی وی است.

زهرا پوربهمن

متولد سال ۱۳۶۵، دانشجوی دکتری مهندسی کامپیوتر، گرایش نرم‌افزار دانشگاه صنعتی امیرکبیر تهران است.

رویکردهای مختلف هوش مصنوعی نظیر بازیابی اطلاعات متنی، پردازش زبان طبیعی، یادگیری عمیق، پردازش تکاملی و سیستم‌های خبره از جمله علائق پژوهشی ایشان است.