Language Related Research
E-ISSN: 2383-0816
https://lrr.modares.ac.ir
https://doi.org/10.52547/LRR.12.6.16

Tarbiat Modares University

Vol. 12, No. 6, Tome 66
pp. 499-531
January & February 2022

An Overview of Text Mining in Language Studies: The Computational Approach to Text Analytics

Hadi Masjedy[1], *Seyyed Mohammad Reza Adel*[2*], *Seyyed Mohammad Reza Amirian*[3], & *Gholamreza Zareian*[4]

Received: 6 October 2019
Received in revised form: 4 January 2020
Accepted: 25 February 2020

Abstract

Text mining' refers to the computational process of unstructured text analytics for extracting latent linguistic layers and themes. It is especially significant as content or thematic analysis in descriptive and interpretive studies. This process begins with structuring simple texts and proceeds with summarizing, classifiing, modelling, evaluating and interpreting the inherent textual concepts and patterns. Given that this method counts as an interdisciplinary innovation especially in discoursal studies, it is to be pursued more intensively in academic studies. Despite the multitude of English studies in this area, there has been little interest to date in text mining amongst Iranian researchers as evidenced by the critically limited number of local Persian and English studies. Thus looking into the theory and practice of text mining and its major analytic tools and methods in Persian and English, this paper aims to prepare the ground for utilizing this methodology in language studies.

***Keywords*:** Text mining, unstructured texts, content analysis, thematic analysis, natural language processing

1. PhD in TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0003-3610-6651
2. Corresponding author, Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar,Iran; *Email: sm.adel@hsu.ac.ir*, https://orcid.org/0000-0002-1136-8973
3. Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0003-3719-3902
4. Associate Professor of TEFL, Hakim Sabzevari University, Sabzevar, Iran; https://orcid.org/0000-0002-9084-608X

The last two decades faced a major increase in the rate and accuracy of knowledge generation in language studies due to advances in interdisciplinary studies of applied linguistics and computer sciences. At the heart of methodological innovations especially in discourse studies lies 'text mining' whose merits have only recently been appreciated by researchers. 'Text mining', 'text data mining' or 'Text Analysis' is the use of different data mining algorithms and methods like natural language processing and linguistic as well as statistical techniques to derive linguistic features, significant patterns and valuable themes from the unstructured texts through collecting unstructured data, pre-processing and cleansing them to detect and remove anomalies and processing and controlling operations (Zhou et al, 2012). These processes are further broken down into feature extraction, structural analysis, text summary, text classification, text clustering, and association analysis. Text mining is actually a complicated procedure of extracting valuable, significant patterns and trends from a large number of textual data used for such functions as product suggestion analysis, social media opinion mining, and sentiment or trend analysis (He, 2013).

Dating back to Feldman and Dagan (1995), text mining is an innovative methodology with a relatively short history which is often integrated with corpus analysis to computationally analyze a large body of unstructured texts as potential informative sources of insight. As a subfield of data mining in computer sciences and an interdisciplinary method, text mining borrows from corpus and computational linguistics, whose main purpose is to extract the meta-characters representing textual features (Pons-Porrata et al, 2007). Zhou et al (2017) believe that despite its short history, text mining has been remarkably evolved into the mainstream research methodology in many interdisciplinary areas in the wake of increasingly rapid developments in data mining.

Hashimi et al (2015) explained the steps involved in text mining as a semi-automated process of collecting, structuring and then analyzing textual data as follows: (a) collecting unstructured data from a variety of sources like

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]

textual documents, social media, web pages, mails, blogs, etc. using specialized corpora for organization, (b) pre-processing and cleansing the data for removing the anomalies to unveil latent valuable information using text mining tools, (c) unstructured data conversion into relevant structured formats, (d) discovering the underlying data patterns using word structures, sequences and frequency, and (e) extracting useful knowledge and storing them in a secure database for evaluation, later retrieval, trend analysis and possible decision-making. Text mining aslso makes use of lexicometrics dealing with frequency and co-occurrence analysis of vocabulary to derive structures from texts; sentiment analysis is an application of lexicometrics looking for positive or negative emotions in documents and has been used in social media analysis for evaluating public opinion (Shangzhen & Lemen, 2016).

Text mining is an *area of inquiry* that in itself deserves to be pursued more intensively in future studies and this paper, thus, is an attempt to review its basic principles, procedures and top analytic tools and to raise researchers' awareness of the virtues of text mining.

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]

[DOR: 20.1001.1.23223081.1400.12.6.27.3]

[DOI: 10.52547/LRR.12.6.16]

دوماهنامهٔ بین‌المللی
د۱۲، ش٦ (پیاپی ٦٦) بهمن و اسفند ۱٤۰۰، صص٤۹۹- ۵۳۱
مقاله پژوهشی
http://dorl.net/dor/20.1001.1.23223081.1400.12.6.27.3

نگرشی به «متن‌کاوی» در پژوهش‌های زبانی: رویکرد رایانشی در تحلیل متون

هادی مسجدی[۱]، سید محمدرضا عادل[۲*]، سید محمدرضا امیریان[۳]، غلامرضا زارعیان[۴]

۱.دکتری زبان انگلیسی دانشگاه حکیم سبزواری، سبزوار، ایران.
۲. دانشیار زبان انگلیسی دانشگاه حکیم سبزواری، سبزوار، ایران.
۳. دانشیار زبان انگلیسی دانشگاه حکیم سبزواری، سبزوار، ایران.
٤. دانشیار زبان انگلیسی دانشگاه حکیم سبزواری، سبزوار، ایران.

تاریخ دریافت: ۱۳۹۸/۰۷/۱٤ تاریخ پذیرش: ۱۳۹۸/۱۲/۰٦

چکیده

«متن‌کاوی» به فرایند رایانشی تحلیل متون بدون ساختار و استخراج لایه‌های زبانی پنهان و مضامین موجود در آن‌ها گفته می‌شود. این روش، اهمیت ویژه‌ای در تحلیل محتوا یا مضمون پژوهش‌های توصیفی و تفسیری دارد. در این فرایند، نخست متون ساده ساختارمند شده وسپس مفاهیم و انگاره‌های نهفتهٔ آن خلاصه‌سازی، طبقه‌بندی، مدل‌سازی، ارزیابی و تفسیر می‌شوند. نظر به اینکه این روش به‌ویژه در مطالعات گفتمان به‌منزلهٔ یک نوآوری میان‌رشته‌ای به‌شمار می‌آید، سزاوار است استفاده از آن در مطالعات دانشگاهی کشور با جدیت بیشتری دنبال شود. مع‌الوصف، به‌رغم گستردگی کمی و کیفی پژوهش‌های بین‌المللی در این حوزه، جای خالی این پژوهش‌ها در مقالات فارسی و انگلیسی داخل کشور بسیار احساس می‌شود. از این‌رو، این مقاله در نظر دارد از رهگذر کنکاش نظری و عملی روش‌های متن-کاوی و ارزیابی ابزارها و روش‌های اصلی آن در زبان فارسی و انگلیسی، بستری مناسب برای بهره-مندی از ظرفیت‌های این روش‌شناسی در مطالعات زبانی فراهم سازد.

واژه‌های کلیدی: متن‌کاوی، متون بدون ساختار، تحلیل محتوا، تحلیل مضمون، پردازش طبیعی زبان.

* نویسندهٔ مسئول مقاله: E-mail: sm.adel@hsu.ac.ir

۱. مقدمه

در این بخش، نخست به تبیین ماهیت متن‌کاوی می‌پردازیم و سپس تفاوت آن را با فرایندهای مشابه دیگر بررسی خواهیم کرد. متن‌کاوی[۱]، داده‌کاوی متنی[۲] یا تجزیه و تحلیل متن[۳] فرایندی است که با استفاده از الگوریتم‌های مختلف داده‌کاوی و روش‌هایی مانند پردازش زبان طبیعی[۴] و تکنیک‌های زبان‌شناختی وآماری، داده‌های کمی متون بدون‌ساختار را استخراج و پردازش کرده و از آن‌ها در جهت تولید دانش و اطلاعات با کیفیت و تحلیل‌پذیر استفاده می‌کند. هِرست[۵] (2009) متن‌کاوی را کشف و استخراج رایانشی اطلاعات جدید و مجهول از منابع نوشتاری مختلف و ایجاد ارتباط میان آن‌ها در جهت تولید فرضیه‌های جدید می‌داند و از دیدگاه ویتن[۶] (2005) این یک زمینهٔ جدید برای جمع‌آوری اطلاعات معنی‌دار از متن زبان طبیعی و فرایندی مفید برای استخراج اطلاعات برای اهداف و نیازهای ویژه است. درواقع، متن‌کاوی به‌مثابهٔ فرایند تحلیلی هوشمندی است که با استخراج مفاهیم صریح وغیرصریح از متن‌ها، ارزش افزوده‌ای به محتوای آن‌ها می‌دهد.

مفهوم متن‌کاوی برای اولین‌بار در اواخر قرن بیستم پدیدار شد و بیش از دو دهه است که به‌منزلهٔ حوزه‌ای میان‌رشته‌ای با تلفیق فناوری‌های رایانه‌ای، مفاهیم پردازش زبان و داده‌کاوی به حیات خود ادامه می‌دهد. باوجود اینکه متن‌کاوی در آغاز روندی کند و ناامیدکننده داشت، با رشد سریع فناوری اطلاعات، پیشرفت‌های چشمگیری در این حوزه پدید آمد و دیری نگذشت که به سنتی مألوف در کاوش‌های زبانی مبتنی‌بر متن بدل شد. در سال‌های اخیر نیز، با ظهور عصر داده‌های بزرگ، متن‌کاوی بویژه در تجزیه و تحلیل اطلاعات و زمینه‌های مرتبط با آن به‌تدریج به‌منزلهٔ رشته‌ای نوظهور در کانون پژوهش‌های زبانی قرار گرفته است (Shangzhen & Lemen, 2016). برآوردها حاکی از آن است که بیش از ۸۰ درصد از داده‌های رایانه‌ای بدون ساختار بوده و بسیاری از نهادهای آموزشی، پژوهشی و اقتصادی به‌رغم برخورداری از حجم انبوهی از داده‌های خام مرتبط با سوابق فعالیت‌هایشان، در پردازش و ساختارمند کردن این اطلاعات ناتوان بوده و برخی ار آن‌ها به همین دلیل از گردونهٔ رقابت خارج می‌شوند (Dolgun et al., 2009).

در اینجا لازم است به دو اصطلاح داده‌کاوی[۷] و وب‌کاوی[۸] و رابطهٔ آن با متن‌کاوی هم اشاره‌ای گذرا داشته باشیم. باید گفت در همهٔ این فرایندها ویژگی‌های مشترکی وجود دارد

که عبارت‌اند از نیمه‌خودکار بودن، کاوش و کشف انگاره‌های[۹] جدید و مفید، استفاده از تکنیک‌های مشابهی چون طبقه‌بندی یا خوشه‌بندی و برخورداری از خروجی‌های یکسان. اما وجه تمایز آن‌ها در این است که داده‌کاوی به تجزیه و تحلیل متون دسته‌بندی‌شده و ساختارمند پایگاه‌دادگان[۱۰] می‌پردازد درحالی‌که منابع متن‌کاوی صرفاً پیکره‌های[۱۱] تشکیل‌شده از متون بدون‌ساختار زبان طبیعی هستند که ابتدا باید ساختارمند و سپس داده‌کاوی شوند. به‌علاوه در وب‌کاوی داده‌های ورودی به‌شکل نیمه‌ساختارمند (ایمیل‌ها، وبلاگ‌ها و...) یا ساختارمند (پایگاه‌های داده) بوده و این داده‌ها برخلاف متن‌کاوی صرفاً زبانی نبوده و می‌توانند به‌صورت صدا و تصویر باشند و البته وب‌کاوی بیشتر مستلزم جست‌وجوی داده‌های معین و آشناست، درحالی که ورودی داده‌کاوی متنی را اطلاعات ناشناخته و مجهول تشکیل می‌دهند.(Achtert et al., 2006)

باید گفت تازمانی که داده‌های خام موجود در متون بدون‌ساختار در فرایند متن‌کاوی پردازش نشود، امکان تحلیل و تفسیر و داده‌کاوی رایانشی آن وجود ندارد که این امر در متن‌کاوی با استفاده از فناوری یادگیری ماشینی و ازطریق طبقه‌بندی کلی متن‌ها، خلاصه‌سازی آن‌ها، آشکار کردن مفاهیم واژگان، مقایسهٔ متون مشابه و بررسی روابط میان آن‌ها انجام می‌پذیرد (Çalis et al., 2013). این بدان معناست که پردازش و تجزیه و تحلیل متن می‌تواند کاری سهل و ممتنع باشد، چراکه از یک سو متن‌ها متداول‌ترین شیوهٔ رسمی برای تبادل اطلاعات بوده و امکان استفاده از این فناوری حتی برای کاربران معمولی در محیط خانه میسر است و از سویی دیگر، متون زبان طبیعی، بدون ساختار و مبهم بوده و استخراج اطلاعات قابل‌فهم، با کیفیت و مفید و ساماندهی آن‌ها در حوزه‌های مختلف از میان متون متفرقه کاری نسبتاً دشوار است. به‌علاوه، معنا و مفهوم کلمات و متن‌ها در موقعیت‌های گوناگون متفاوت بوده و به‌سادگی می‌تواند به برداشت‌های اشتباه منجر شود و همین امر آن را به فرایندی دشوار بدل می‌سازد (Cohen & Hunter, 2008). آنچه اهمیت و دشواری این فرایند را دوچندان می‌سازد ضرورت پردازش هوشمند و سریع و بازتولید محتوایی ساختارمند از میان صدها فایل شخصی یا رسمی و ایمیل‌ها یا رسانه‌های اجتماعی است. برای نمونه، تحلیل محتوای نظرسنجی‌های برخط مستلزم فرایند پیچیدهٔ بررسی بازخورد شرکت‌کنندگان برحسب مضامین مشترک پاسخ‌های آن‌هاست که خود نیازمند

واکاوی اصطلاحات کلیدی در محتوای متن نظرسنجی و استخراج انگاره‌های نهفتهٔ معنایی از آن‌هاست.

به‌نظر می‌رسد استفاده از روش متن‌کاوی در دههٔ آینده باتوجه به حجم روزافزون تولید محتوا در رشته‌های مختلف و عدم امکان تجزیه و تحلیل ذخیره‌سازی‌های انبوه به‌صورت سنتی، امری اجتناب‌ناپذیر باشد. در این میان، به‌رغم وجود پیکره‌های جمع‌آوری‌شده در زبان فارسی، این روش چندان مورد اقبال پژوهشگران ایرانی قرار نگرفته است. لذا در این مقاله درنظر داریم ضمن آشنایی مخاطبان فارسی با پژوهش‌های بین‌المللی و معدود مطالعات بومی انجام‌گرفته در این حوزه و بررسی کاربردهای متنوع آن، با رویکردی عملی به واکاوی ابزارها و نرم‌افزارهای رایج متن‌کاوی و ارزیابی نقاط قوت و ضعف برخی از آن‌ها بپردازیم. در گام بعدی برآنیم تا با ارائه و بررسی نمونه‌هایی از تکنیک‌های اصی آن، پژوهشگران زبان فارسی را به‌سمت بهره‌گیری هرچه بیشتر از قابلیت‌های این روش‌شناسی سودمند رهنمون باشیم.

۲. پیشینهٔ تحقیق

مفهوم متن‌کاوی روی داده‌های بدون‌ساختار اولین‌بار در اواخر قرن بیستم توسط فلدمن و داگان[۱۲] مطرح شد. از آن پس پژوهش‌ها در این زمینه به‌سرعت ادامه پیدا کرد. به‌اعتقاد سو و ژاو[۱۳](2017) پیشرفت‌های بی‌سابقهٔ داده‌کاوی متنی به‌رغم تاریخچهٔ کوتاه آن، این حوزه را به جریان اصلی روش‌شناختی در پژوهش بدل ساخته است. پانزپوراتا[۱۴] و همکاران (2007) متن‌کاوی را استخراج اطلاعات از مجموعه‌های متون بدون‌ساختار می‌دانند که وظیفهٔ اصلی آن به‌منزلهٔ حوزه‌ای میان‌رشته‌ای، کشف قوانین بالقوه و روندها در متون طولانی است. درحقیقت، متن‌کاوی را می‌توان یک فناوری تولید دانش دانست که موضوع اصلی آن زبان طبیعی بوده و به واکاوی مفهومی عناصر قابل‌تشخیص، گره‌های گرامری در ساختار متن، اصطلاحات، حقایق علمی و سایر اَشکال زبانی می‌پردازد و تمرکز آن عمدتاً بر مدل‌سازی، استخراج ویژگی‌های خاص و کشف انگاره‌های متنی از طریق طبقه‌بندی، خوشه‌بندی وسایر جنبه‌ها قرار دارد (Cohen, 1999).

در این بخش ابتدا به پاره‌ای از پژوهش‌های بین‌المللی که در زمینه‌های مختلف متن کاوی

صورت پذیرفته است، اشاره خواهیم کرد و سپس چند نمونه از این مطالعات در متون فارسی را برمی‌شمریم. یاسینی و حج[۱۵] (2010) با رویکردی نو به بررسی محتوای عاطفی متون، روابط دوستی و بیان احساسات در شبکه‌های اجتماعی برخط پرداختند تا ماهیت خاص این سایت‌ها و زبان مورد استفادهٔ آن‌ها را کشف کنند و توانستند چارچوب جدیدی برای توصیف مؤلفه‌های تعاملات عاطفی در شبکه‌های اجتماعی پیشنهاد کنند که مدل مناسبی برای جمع‌آوری و پردازش داده‌ها محسوب می‌شود. در پژوهش کریگل[۱٦] و همکاران (2009) خوشه‌بندی‌های متنی منسجم مبتنی‌بر اسناد وب مورد بررسی قرار گرفت که در آن روش‌های نوین متن‌کاوی اسناد وب با فرایندهای فعلی خوشه‌بندی مقایسه شدند؛ عملکرد روش پیشنهادی به‌وسیلهٔ مدل‌های متن‌کاوی مبتنی‌بر مفهوم مورد تجزیه و تحلیل قرار گرفت و انگاره‌های مفیدی به‌منظور بهبود کارایی خوشه‌بندی پیشنهاد شد. گاپتا[۱۷](2009) نیز مطالعاتی درمورد ارتقای جنبه‌هایی ازبرنامه‌های کاربردی متن‌کاوی مانند بازیابی اطلاعات، ردیابی موضوع، خلاصه‌سازی، طبقه‌بندی، خوشه‌بندی، پیوستگی مفاهیم، تصویرسازی اطلاعات و پاسخ‌گویی به سؤالات انجام داد. قره چپوغ و عباسی خلیفه‌لو[۱۸] (2011) دریافتند که با بهره‌گیری از متن‌کاوی و تکنیک‌های پردازش زبان طبیعی می‌توان به درک قابل‌قبولی از معناشناسی داده‌های بدون‌ساختار متون وب دست پیدا کرد. وو[۱۹] و همکاران (2006) نیز با بررسی عملکرد الگوریتم‌های متن‌کاوی، به‌کارگیری تکنیک‌های جدید در افزایش بهره‌وری روش‌های انگاره‌محور را مفید ارزیابی کردند. همچنین در دو پژوهش متفاوت (Liptha et al., 2010; Wang et al., 2015) مدل‌های باکیفیت‌تری از خوشه‌بندی از طریق استخراج ساختار معنایی جملات در اسناد به‌دست آمد.

یکی از زمینه‌های بدیع پژوهشی در متن‌کاوی حوزهٔ داده‌کاوی آموزشی است. یک نظام آموزشی دارای تعداد زیادی از داده‌های آموزشی است که شامل اطلاعات دانش‌آموزان، معلمان، فارغ‌التحصیلان، داده‌های منابع، و غیره است (Swamy & Hanumanthappa, 2012). امروزه از رهگذر پیشرفت در روش‌شناسی متن‌کاوی آموزشی و بهبود کیفیت مدل‌های آن و پشتیبانی این فناوری‌ها از دور دوم تحول در همهٔ حوزه‌های یادگیری (Baker & Inventado, 2014) امکان کمک به مدارس و دانشگاه‌ها برای تمرکز بر روی اطلاعات و داده‌های کلیدی مربوط به فراگیران و کسب نتایج بهتر آموزشی فراهم شده است (Sin &

(Muthu, 2015 ازاین‌رو، با توجه به اینکه اطلاعات متنی و کلامی دانش‌آموزان در محیط-های یادگیری می‌تواند منعکس‌کنندهٔ وضعیت روان‌شناختی افراد و کنش‌های گفتاری[۲۰] آنان در اجرای یک رفتار خاص باشد، واکاوی این اطلاعات به اولیای مدرسه این امکان را می‌دهد تا وضعیت ذهنی فراگیران را به‌موقع درک کرده و بحران‌های رفتاری را پیش‌بینی کنند و مشاوره و پشتیبانی‌های مربوط به آن را ارائه دهند (Zhou, 2012). از نگاه لی[۲۱] و همکارانش (2014)، اطلاعات و دانش استخراج‌شده از داده‌کاوی متنی می‌تواند در پیش‌بینی و جهت‌گیری رفتار و عملکرد دانش‌آموزان و بهبود مدیریت سازمان‌ها در نظام آموزشی بسیار مؤثر باشد. پژوهش‌های مارکز و مویا[۲۲] (2011) در دانشگاه شیلی براساس مدل‌های خودکار متن‌کاوی و ویژگی‌های نهفتهٔ معنایی متون انجام شد که هدف آن ارزیابی و تخمین سطح دانش فراگیران و مهارت‌های تف1ر پیشرفته با استفاده از سؤالات عینی و داده‌های متنی مفهومی در جهت تولید مواد آموزشی بود. پژوهش‌های کورتز و سیلوا [۲۳](2008) درمورد نظام آموزشی پرتغال هم نشان داد که براساس داده‌های آموزشی می‌توان پیش‌بینی‌های درست و دقیقی برای بهبود نظام یاددهی و یادگیری انجام داد. به‌علاوه، از متن‌کاوی در زمینهٔ تکنیک‌های آموزش به‌منظور تحلیل برنامهٔ درسی و مباحث پژوهشی و عملکرد دانش‌آموزان هم استفاده شده است (Alfiani & Wulandari, 2015). مثلاً در مطالعهٔ موردی لیان[۲۴] و همکاران (2012)، بازخورد زبان‌آموزان درمورد کیفیت دورهٔ آموزشی از طریق واکاوی محتوای مثبت و منفی پیام کوتاه آنان ارزیابی شد.

حال بد نیست به گزیده‌ای از معدود کارهای پژوهشگران داخلی هم اشاره‌ای داشته باشیم. تیمورپور و همکاران (۱۳۸۸) با بهره‌گیری از یک روش نوین متن‌کاوی موسوم به SUTC اقدام به دسته‌بندی هوشمند متون مقالات معتبر پژوهشگران ایرانی در حوزهٔ فناوری نانو کردند تا بتوان از این اطلاعات برای سیاست‌گذاری و پایش علمی دستاوردهای پژوهشی این حوزه بهره جست. لاکتراشی و همکاران (۱۳۹٤) در مقاله‌ای با استفاده از پردازش زبان طبیعی به تبیین مزایا و معایب الگوریتم‌های دسته‌بندی موضوع‌محور خبری پرداخت. امامی و قائمی (۱۳۹٤) هم در پژوهشی از یک مدل پیشنهادی در دسته‌بندی موضوعی متون فارسی برگرفته از پیکرهٔ چندهزار متنی «همشهری» به‌منزلهٔ معتبرترین منبع زبان طبیعی در فارسی بهره جستند. آن‌ها این دسته‌بندی پیشنهادی را براساس روش‌های یادگیری انتقالی انجام

دادند که در آن امکان تعمیم و به‌کارگیری دانش حل مسئله در سایر مسائل مرتبط بررسی می‌شود. نتایج این پژوهش نشان داد که این روش دسته‌بندی موضوعی با استفاده از کلمات کلیدی از دقت و صحت بالایی برخوردار است. هاشمی (۱۳۹۳) نیز با استفاده از مدل تلفیقی «گزینش ویژگی» و «الگوریتم‌های IG و CFS یادگیری ماشینی» برای دسته‌بندی چندگانهٔ متون فارسی، بهینه بودن این مدل پیشنهادی را اثبات کرد.

متن‌کاوی با تمرکز بر متون زبان انگلیسی و آن هم در حوزهٔ فناوری اطلاعات و درجهت بهبود الگوریتم‌های آن صورت پذیرفته است. در این میان، سهم متون فارسی به‌رغم حجم روزافزون متون زبان فارسی بسیار اندک و تنها محدود به تکنیک‌های فنی متن‌کاوی بوده است. البته کارهای ارزشمند فراوانی در زمینهٔ تحلیل متن و گفتمان بدون رویکرد رایانشی مانند پژوهش مزینانی و همکاران (۱۳۹۷) موسوم به text analysis صورت پذیرفته است که درخور توجه است ولی هنوز جای پژوهش‌های صرفاً زبانی مبتنی‌بر آنالیز عددی با رویکرد رایانشی موسوم به text analytics در راستای تولید دانش و فرآورده‌های زبان‌شناسی فارسی بسیار خالی است. لذا در ادامه درنظر داریم با ارائهٔ قابلیت‌های عملی این فناوری در بخش کاربردها و ابزارها، بستری مناسب برای پژوهشگران ایرانی در جهت شناخت توانمندی‌های این روش‌شناسی در حوزهٔ زبان فراهم آوریم.

۳. کاربردها

متن‌کاوی در ابتدا برای تحلیل بازخورد مشتریان در شرکت‌های عمدهٔ فروشندگی و بازاریابی استفاده می‌شد، ولی به‌تدریج از این روش‌شناسی نوظهور در زمینه‌های مختلف پژوهشی ازجمله مطالعات زبانی استفاده شد. در این بخش به شرح مختصری از اهم کاربردهای مرتبط با مسائل زبانی خواهیم پرداخت.

۳ـ۱. تحلیل محتوا و مضمون

یکی از کاربردهای متن‌کاوی، تحلیل محتوا یا تحلیل مضمون داده‌های کمی و کیفی در حوزه‌های زبان‌شناسی و مطالعات رسانه است که شامل زیرمجموعه‌هایی همچون تحلیل گفتمان و تحلیل موضوعی و واژگانی می‌شود و از تکنیک‌هایی مانند بسامد واژگانی،

هم‌بستگی واژگانی و مفهومی، گروه‌بندی واژگانی و غیره استفاده می‌کند. به‌طور کلی، پیش از تحلیل محتوا یا مضمون باید فرایند رمزگذاری به یکی از سه روش زیر اعمال شود: روش سنتی که در آن رمزگذاری مبتنی‌بر متن انجام می‌شود، روش هدایت‌شده[۲۵] که رمزگذاری را متناسب با نظریه یا فرضیهٔ خاصی انجام می‌دهد و روش تلخیصی[۲۶] که متن را براساس واژگان کلیدی جست‌وجو، مقایسه و تفسیر می‌کند (Hsieh & Shannon, 2005).

از تحلیل محتوا می‌توان در پژوهش‌های توصیفی و تفسیری بطور توأمان استفاده کرد. البته در پژوهش‌های تفسیری برخلاف طرح‌های توصیفی، داده‌ها آزادانه‌تر تفسیر می‌شوند وبه همین دلیل، روش تحلیل مضمون برای این طرح‌ها مناسب‌تر و شایع‌تر است و از طرفی در پژوهش‌های توصیفی هم به‌دلیل وابستگی بیشتر به متن، استفاده از تحلیل محتوا بسیار بیشتر به چشم می‌خورد. در روش تحلیل مضمون، باید نخست مانند تحلیل محتوا فرایند رمزگذاری انجام شده و سپس براساس آن، تحلیل موضوعی و تفسیر مفاهیم متون صورت پذیرد. برای تحلیل محتوای مصاحبه‌های ساختارمند و تفسیر صحیح نظرسنجی‌ها، می‌توان از توزیع فراوانی، جدول‌بندی‌های متقاطع[۲۷] و آزمون خی‌دو[۲۸] استفاده کرد و داده‌ها را به‌شکل گرافیکی نیز ارائه داد. در روش‌های تحلیل موضوعی به‌جای دسته‌بندی متن‌ها، می‌توان با استفاده از تحلیل استقرایی محتوای بازخوردها، مضامین کلی آن‌ها را استخراج کرد. این بازخوردها عبارات معناداری هستند که می‌توانند هنگام تجزیه و تحلیل مجموعهٔ داده‌ها ایجاد روشنگری کرده و برای سازمند کردن مضامین مشابه به‌کار روند. از تحلیل مضمون این سلسله از اطلاعات برای تصمیم‌سازی در عرصه‌های مختلف استفاده می‌شود.

۳ـ۲. پژوهش‌سنجشی ادبیات دانشگاهی

کاربرد دیگر متن‌کاوی در زمینهٔ کتاب‌سنجی[۲۹] و زیرمجموعهٔ آن علم‌سنجی[۳۰] است که هدف آن تحلیل مقالات علمی و سنجش ادبیات نوشتاری دانشگاهی از طریق آنالیز عددی میزان استنادها، تحلیل محتوای مقالات و اندازه‌گیری ضریب تأثیر مجلات علمی توسط الگوریتم‌های خودکار است. این ابزارها میزان محبوبیت نویسندگان، مقالات، مجلات و میزان تأثیر آن‌ها را اندازه‌گیری می‌کنند(De Bellis, 2009). به‌علاوه در این حوزه ویژگی‌های دیگری نظیر بسامد عبارات و الگوهای نحوی در متون نیز بررسی می‌شوند. کتاب‌سنجی بجز تجزیه و

تحلیل کتاب‌ها و مقالات علمی زیرمجموعه‌های دیگری هم دارد. برای نمونه دگرسنجی[۳۱]، تعداد بارگیری مقالات، کتاب‌ها و ویدئوهای تحت‌پوشش رسانه‌های نوشتاری و اجتماعی را اندازه‌گیری می‌کند. تولید، توزیع و استفاده از اخبار و تعیین میزان محبوبیت وبگاه‌ها هم به ترتیب در اطلاع‌سنجی[۳۲] و وب‌سنجی[۳۳] بررسی می‌شوند.

یکی دیگر از تکنیک‌های کتاب‌شناسی، نقشه‌برداری علمی[۳٤] است که ساختارهای مفهومی، ذهنی و اجتماعی حوزه‌های پژوهشی را واکاوی می‌کند. ازجمله این موارد می‌توان به واکاوی ساختار مفهومی اسناد براساس اصطلاحات اصلی و کلیدی مشترک و تحلیل ارتباط بین نویسندگان مشترک و مقالات و شبکه‌های ارتباطی‌شان اشاره کرد(Cobo et al., 2011). نقشه‌برداری علمی امکان تولید نقشه‌های ارتباطی میان گروه‌های علمی پژوهشی را نیز فراهم می‌سازد. اسناد علمی را می‌توان با استفاده از تکنیک‌های «تحلیل هم‌واژگانی، هم‌رخدادی و هم‌نویسندگی[۳۵]» بررسی کرده و الگوهای مفهومی آن را مشخص کرد (He, 1999). مجموعهٔ مفاهیم با استفاده از روابط میان واژگان کلیدی و روابط بین استنادها در حوزه‌ای خاص واکاوی شده است و نقشهٔ آن ترسیم می‌شود. به چنین پژوهش‌هایی که از طریق متن‌کاوی بر روی روابط مفهومی در مقالات انجام می‌شوند مفهوم‌کاوی[۳٦] گفته می‌شود که جملات را نه‌تنها در سطح کلمات و وجوه مختلف شکلی و محتوایی آن‌ها بلکه در سطح معنایی آن واکاوی می‌کند.

۳ـ۳. متن‌کاوی مفهومی نظرسنجی‌ها و وب‌نوشت‌ها

در متن‌کاوی مفهومی نظرسنجی‌های باز[۳۷]، روند زمان‌بر و پرهزینهٔ کدگذاری و خواندن هزاران پاسخ در زمان کوتاهی تبدیل به خروجی‌های ساختارمند و اطلاعات مفید و قابل‌تفسیر می‌شود. در تحلیل نظرسنجی‌ها، معمولاً سؤالاتی که پاسخ‌های محدود و گزینه‌ای دارند به‌راحتی قابل تفسیر و کمی شدن هستند، ولی درمورد سؤالات تشریحی و پاسخ‌باز این‌گونه نیست؛ این نوع پرسش‌نامه‌ها اطلاعات غنی‌تر و باارزش‌تری ارائه داده و با تولید اطلاعات غیرقابل‌پیش‌بینی، منبع مهم دانش‌افزایی به‌شمار می‌آیند و لذا نیازمند مفهوم‌کاوی ویژه‌ای از حیث واژگان و پاسخ‌های بدون‌ساختار هستند که از عهدهٔ یک فرد ساخته نیست و متن‌کاوی در این زمینه بسیار مددرسان خواهد بود (Agrawal & Batra, 2013).

پیوند مفهومی[۳۸] با برقراری ارتباط میان مفاهیم مشترک اسناد وب می‌تواند در ایجاد شبکهٔ درهم‌تنیده‌ای از داده‌ها برای سامان‌دهی دانش‌ها راهگشا باشد. در حال حاضر، در پژوهش‌های متن‌کاوی مفهومی در وب بیشتر از پایگاه دادهٔ WordNet استفاده می‌شود که هم‌بستگی واژگانی را بررسی می‌کند و از نمونه‌های انگلیسی و فارسی آن می‌توان به Princeton Wordnet و EuroWordnet و فردوس نت وابسته به آزمایشگاه فناوری وب دانشگاه فردوسی مشهد اشاره کرد. در سایر پژوهش‌ها نیز، مجموعه دادهٔ دیگری به‌نام ConceptNet به‌کار رفته که در آن هم‌بستگی‌های فرمی(شکلی)، اجتماعی و زمانی در میان واژگان مورد جست‌وجو قرار می‌گیرند. برای نمونه، در پژوهش آیدین و همکاران (2013) از الگوریتم خوشه‌بندی برای مفهوم‌کاوی استفاده شده که در آن واژه‌های پیش‌فرض، دسته‌بندی شده و از آن‌ها برای شناسایی مفاهیم نهفته در اسناد وب استفاده شده است. یک مورد دیگر در مفهوم‌کاوی، کاربرد یکی از الگوریتم‌های هوش مصنوعی موسوم به Latent Dirichlet است که به جست‌وجوی واژه‌هایی می‌پردازد که امکان همایندی با واژگان کلیدواژه‌ها را داراست و از این طریق، انگاره‌های معنایی نهفته در جملات یا متون را کشف می‌کند(Blei et al., 2003). در آینده می‌توان با بهره‌گیری از استانداردهای ایجادشده در فناوری وب معنایی۳.۰، مراحل متن‌کاوی واژگانی و مفهومی را قبل از قرار دادن متون در اینترنت به انجام رسانید تا خواندن و درک متون برای موتورهای جست‌وجو بهینه‌تر شود و مثل قبل تنها محدود به جست‌وجوی کلیدواژه‌ها نشده و امکان تفسیر سایت‌ها و دسترسی صحیح‌تر برای کاربران فراهم شود(Arslan, 2011).

۳ـ۴. متن‌کاوی آموزشی در نظام‌های یاددهی و یادگیری

ازجمله کاربردهای دیرظهور اما در حال گسترش متن‌کاوی، داده‌کاوی آموزشی در سیستم‌های مدیریت یادگیری است که فرایندی برای تبدیل داده‌های خام سیستم‌های آموزشی به اطلاعات مفید برای برنامه‌ریزی درسی است. از تکنیک‌های آن می‌توان به پیش‌بینی، خوشه‌بندی و استخراج روابط[۳۹] داده‌های آموزشی اشاره کرد که برای مطالعهٔ رفتار و عملکرد دانش‌آموزان به‌کار می‌روند. در اسلوب یادگیری تعاملی[۴۰]، داده‌های بدون-ساختار مبتنی‌بر متن تولید می‌شوند و می‌توان از متن‌کاوی به‌منزلهٔ روش جدید یادگیری در

ارزیابی دانش فراگیران، غربالگری رفتار آنان، ارزشیابی برنامهٔ درسی، گروه‌های یادگیری جامعه‌بنیاد[٤١]، یادگیری زودهنگام، هشدارهای مربوط به بحران‌های رفتاری، پیش‌بینی اثر یادگیری، تصویرسازی وضعیت یادگیری و بهبود نتایج آموزشی و عملکردی دانش‌آموزان استفاده کرد(Kumar & Vijayalakshmi, 2011) . به‌علاوه، روش‌های داده‌کاوی آموزشی هر ساله با دقت بیشتری پدیده‌های مرتبط به یادگیری دانش‌آموزان در سامانه‌های هوشمند برخط را مدسازی و اعتبارسنجی می‌کنند تا بتوان با گذشت زمان نتایج آن‌ها را قابل تعمیم‌تر کرد (Baker & Inventado, 2014).

۳ـ۵. کاربردهای عمومی

به‌جز موارد یاد شده، متن‌کاوی کاربردهای متعدد دیگری هم در حوزه‌های غیرزبانی دارد که برخی از آن‌ها عبارت‌اند از: تحلیل شبکه‌های اجتماعی فیس‌بوک، توییتر و... برای اهداف امنیتی، پردازش ایمیل‌ها برای فیلتر کردن خودکار هرزنامه‌ها[٤٢]، تحلیل داده‌های کتابخانه‌های دیجیتالی برای تسهیل پژوهشگری، اعتبارسنجی محتوایی[٤٣] داده‌های کلان، ایجاد نمودار دسته‌بندی مطالب، متن‌کاوی برنامه‌های کاربردی در شبکه‌های جهانی برای ایجاد سامانه‌های برخط مشاوره، پیش‌بینی مسائل آینده براساس شواهد آماری، بهبود جست‌وجوی پیشینه‌های تحقیق با توسعهٔ مطالعات بسامد واژگانی و حوزه‌های دیگری همچون نشر، رسانه، ارتباطات، بازاریابی، بهداشت و درمان و... (Gupta & Lehal, 2009). درادامه به شرح و بررسی تفصیلی برنامه‌های رایج متن‌کاوی به‌ویژه در زبان فارسی خواهیم پرداخت.

۴. ابزارهای متن‌کاوی

در این بخش درصدد هستیم نخست به بررسی ابزارهای رایج متن‌کاوی و برخی از ویژگی‌های آن‌ها بپردازیم و سپس نگاهی اجمالی به برنامه‌های کاربردی متن‌کاوی در زبان فارسی برای پژوهشگران ایرانی بیندازیم.

۴ـ۱. نرم‌افزارهای برتر پردازش متن

نهادهای پژوهشی با تکیه بر بهترین نرم‌افزارهای تجزیه و تحلیل متن و الگوریتم‌های

پردازش زبان طبیعی به منظور شناسایی انگاره‌ها، مضامین و موضوعات خاص، اقدام به کشف و تفسیر اطلاعات و تصمیم‌سازی بر پایهٔ داده‌های منابع بزرگ متنی بدون‌ساختار مانند نتایج نظرسنجی‌های برخط، توییت‌ها، مکالمات ضبط‌شده، ایمیل‌ها و اسناد دیگر می‌کنند. با توجه به اینکه شیوه‌های سنتی تحلیل متون به‌صورت دستی قادر به بررسی و درک زبان اسناد متنی در حجم زیاد نیستند، ازاین‌رو اخیراً نهادهای پژوهش‌محور به نرم‌افزارهای خودکارسازی انبوه متون روی آورده‌اند. از جمله مزایای این نرم‌افزارها می‌توان به واکاوی سریع متون بدون‌ساختار و ساختارمند در حجم بالا از منابع مختلف، کسب رهیافت جدید از داده‌های متنی برای اقدامات مقتضی و استفاده از داده‌های نیازسنجی در پاسخ‌گویی به مخاطبان اشاره کرد(Pena-Ayala, 2014).

برخی از ویژگی‌های اصلی نرم‌افزارهای داده‌کاوی متنی عبارت‌اند از: توانایی وارد کردن و بازیابی متون از منابع مختلف با قالب‌های چندگانه، استفاده از الگوریتم‌های پردازش زبان طبیعی برای تشخیص زبان، پردازش و ارزیابی میزان خوانایی متون، طبقه‌بندی موضوعی، واکاوی و دسته‌بندی واحدسازی صرفی[44]، برچسب‌گذاری بخشی از گفتار، تصویرسازی‌های متنوع داده‌های متون پردازش‌شده برای تفسیر آسان متن، نمایش وضعیت و کشف ارتباط میان وضعیت‌ها، نمایش نتایج نمودارهای تعاملی، استفاده از یک رابط کاربرپسند[45] و انعطاف‌پذیر برای ادغام و نمایش موضوعات، مدیریت نمودارهای جریانی[46]، مدیریت جدول‌ها و تغییر زبان و پشتیبانی از زبان‌های زندهٔ دنیا (Hashimi et al., 2015). در بخش پیش‌رو درنظر داریم با ارائهٔ گلچینی از ابزارهای برتر متن‌کاوی و برخی از ویژگی‌های آن‌ها، یک مرجع کاربردی و جامع در زمینهٔ گزینش دقیق‌تر نرم‌افزار در اختیار پژوهشگران زبان فارسی قرار دهیم. در این راستا، مراجعه به گزارش پژوهشی جِمِرت[47] (2000) در دانشگاه آمستردام و کائِر و چاپرا[48] (2016) بسیار رهگشا خواهد بود.

جدول ۱: ابزارهای برتر متن‌کاوی

Table 1: Premier text mining tools

الف) ابزارهای عمومی (Ergün, 2017):
IBM SPSS Text Analytics, Intellexer, Inxight, LanguageWare, Language Computer Corporation, Lexalytics, LexisNexis, Luminoso, Mathematica, Medallia, Megaputer Intelligence, NetOwl, RapidMiner, SAS Text Miner and Teragram, Semantria, Smartlogic –

Semaphore, STATISTICA Text Miner, Sysomos, Textalytics, WordStat, Xpresso, Angoss, Attensity, AUTINDEX, Autonomy, Averbis, Basis Technology (which can perform analysis in more than twenty languages), Clarabridge, Complete Discovery Source, Endeca Technologies, Expert System S.p.A., FICO Score, General Sentiment
ب) ابزارهای متن‌باز[۴۹] (Kaur & Chopra, 2016):
LIBSVM, Open NLP Apache, ROST CM, Weka, GATE, Orange, Bow, UIMA, Ling Pipe,
ج) ابزارهای بصری (Wang et al, 2015):
TagCrowd, Yippy, WordMosaic, AbcYa, VocabGrabber, Microsoft WordSift, WordItOut, Excel, Wordle, Tagul, Tagxedo, Jason Davies' Word Cloud Generator
د) ابزارهای تجاری(Gemert, 2000):
IBM, SAS, IBM SPSS, IDOL Server Autonomy, Darwin Oracle, SQL Server Microsoft, Clear forest Thomoson Reuters, Themescpease Cartia, Founder of Peking University, TRS text mining software Beijing Tuols
ه) ابزارهای تجزیه و تحلیل یادگیری.(Xu & Zhao, 2017) :
LIWC, Cohere, Sobek, Rapid Miner, Dissertation Browse, EduMiner, GCS, Toreador
و) ابزارهای تحلیل محتوا (Ergün, 2017):
TABARI, TACT (Text Analysis Computing Tools), Tapor Tools, Text Analysis Tool, Textalyser, Textanz, TextArc, TEXTPACK, TextQuest, VBPro, Wordcruncher, WORDij, Wordle, WordStat, Yoshikoder, Concordance, Crawdad Technologies LLC, Diction, General Inquirer, Hamlet II, INTEXT, Leximancer, Minnesota Contextual Content Analysis (MCCA), Profiler Plus, PROTAN, SALT Software, PASW Text Analytics for Surveys, T-LAB Tools for Text Analysis
ز) ابزارهای علم‌سنجی (Cobo et al., 2017):
Network Workbench Tool, Science of Science (Sci2) Tool, VantagePoint, VOSViewer, Pajek, UCINET, Cytoscape, Bibexcel, CiteSpace II, CoPalRed, IN-SPIRE, Leydesdorff's Software

در حال حاضر، متن‌کاوی هنوز در مرحلهٔ تحقیق و توسعه است، ولی باید اذعان کرد که در این زمینه، ابزارهای تجاری به‌دلیل ویژگی‌های خاص خود گوی سبقت را از نرم‌افزارهای متن‌باز ربوده‌اند. ازجمله این ویژگی‌ها می‌توان به امکان دسترسی عمده به داده‌های جهانی، کاربرد متنوع، پشتیبانی از چندین زبان و قالب و قابلیت پردازش تحلیلی داده‌های ساختارمند، نیمه‌ساختارمند و کاملاً بدون‌ساختار اشاره کرد. با این حال، بیشتر نرم‌افزارهای تجاری به دلیل کیفیت بالا و عدم دسترسی آسان، بسیار گران‌قیمت بوده و برای نهادهای کوچک پژوهشی مناسب نیستند(Feng, 2008). از طرفی، برخی از ابزارهای متن‌باز به‌دلیل استفاده از الگوریتم‌های ثابت، دامنهٔ کاربرد محدود و الزامات قالبی خاص و عدم پشتیبانی از

زبان‌های مختلف به‌ویژه فارسی، فعلاً از توانایی تحلیل انواع متون بدون‌ساختار برخوردار نیستند. با این وجود، کاربرد نرم‌افزارهای متن‌باز به‌دلیل سهولت دسترسی آن‌ها در بین کاربران شایع‌تر است، ضمن اینکه قابلیت رفع و پشتیبانی نیازهای ویژه را نیز تاحد زیادی دارا هستند. در ادامه با گزینشی دقیق‌تر از میان ابزارهای مزبور، چند برنامهٔ کاربردی‌تر برای متن‌کاوی زبان فارسی و مختصات آن‌ها در جدول زیر معرفی کنیم.

جدول ۲: ابزارهای متن‌باز متن‌کاوی

Table 2: Open source text mining tools

ابزار	نوع	تکنیک	ویژگی	وبگاه	توضیحات بیشتر
Textalyser	برخط	تجزیه وتحلیل متن + کلید واژگان	آنالیز متن	Http://textalyser.net/	پشتیبانی ورودی‌های متنی وURL
IBM SPSS Predictive Analytics	تجاری	مدل‌سازی/آنالیز پیشگو+الگوریتم‌های هوش مصنوعی	داده‌کاوی+متن-کاوی+آنالیز آماری	Http://www.ibm.com/analytics/us/en/technology/spss/	
Intellexer	تجاری	پردازش زبان طبیعی	آنالیز متن و مدیریت اطلاعات+مقایسه و مقوله‌بندی اسناد	Http://www.intellexer.com/knowledge_management.html	
Lexalytics	تجاری	تکنیک‌های یادگیری ماشینی+ قواعد خبره صنعتی+ پردازش زبان طبیعی+ پیشرفته	تحلیل احساسات+مقوله-سازی+استخراج مؤلفه‌ی مذکور	Https://www.lexalytics.com/	
SAS Text Miner	تجاری	مدل‌های پیش-گو+یادگیری ماشینی+ زبان طبیعی+تکنیک‌های داده‌کاوی پردازش	آنالیز و پردازش متن+کشف مضمون سند	Http://www.sas.com/en_us/software/analytics/text-miner.html	سیستم عامل مورد نیاز: HP/UX on Itanium, - IBM 64-Bit Enabled AIX, Linux (x86-64),- Microsoft Windows - (x86-64), 64-Bit Enabled Solaris - on SPARC, Solaris on - x64
GATE	متن‌باز	پردازش زبان طبیعی	پردازش متن	Https://gate.ac.uk/	
Rapidminer	متن‌باز	یادگیری ماشینی	داده‌کاوی+تجزیه و تحلیل متن+پردازش متن	Http://docs.rapidminer.com/	

۴ـ۲. متن‌کاوی زبان فارسی

تجزیه و تحلیل متون فارسی به‌رغم سابقهٔ دیرینهٔ حوزهٔ هوش مصنوعی در آزمایشگاه پردازش متن و فناوری وب دانشگاه تهران و فردوسی مشهد، همچنان جنبهٔ عملیاتی و کاربردی در ابعاد گسترده پیدا نکرده است (بنائی، ۱۳۹٤). برای انجام پژوهش‌های عملی در این حوزه در گام نخست لازم است بدانیم که هرگونه عملیات تجزیه و تحلیل نیازمند فرایند آماده‌سازی و پیش‌پردازش متن‌هاست. ازاین‌رو، در این قسمت برخی از برنامه‌های پردازش متون فارسی را تشریح می‌کنیم و سپس نگاهی اجمالی به ابزارهای برنامه‌های پردازش متون فارسی با اقتباسی اندک از آزمایشگاه فناوری وب دانشگاه فردوسی (http://wtlab.um.ac.ir/fa/) خواهیم داشت.

۴ـ۲ـ۱. عادی‌ساز[۵۰] متن فارسی

برای یکدست کردن متون فارسی لازم است نخست واژگان عادی سازی شوند. در این فرایند باید برای یکپارچه کردن متن، تمام نویسه‌های[۵۱] آن به‌شکل استاندارد درآیند، چراکه فارسی جزو زبان‌های بدون‌ساختار بوده و نمی‌توان از قالب‌های پیش‌فرض موجود برای آن استفاده کرد. برخی عادی‌سازهای متون زبان فارسی برای پژوهشگران زبان فارسی عبارت‌اند از: الف) Persian Pre-processor: PrePer ب) HAZM ج) Virastyar د) Parsivar. برای بهینه‌سازی و رفع معضلات پردازش متن، ابتدا باید ایست‌واژه‌ها[۵۲] مانند حروف ربط، اضافه و مثل آن را از متن حذف کرد، زیرا این واژگان علی‌رغم پرتکرار بودن، دارای ارزش محتوایی و معنایی کم‌تری هستند و حذف آن‌ها سبب یکسان‌سازی مؤلفه‌های متنی خواهد شد. در زبان فارسی چندین برنامهٔ کاربردی شناخت ایست‌واژه طراحی شده که مهم‌ترین آن‌ها عبارت‌اند از: الف) Persian stopwords collection ب) Hazm stop words ج) mhbashari stopword list. در این میان، لیست مجموعهٔ «هضم» از شهرت بیشتری برخوردار است.

۴ـ۲ـ۲. برچسب‌گذار انواع دستوری واژگان[۵۳] فارسی

به‌کمک این برنامه می‌توان نوع و نقش فعل، اسم، صفت، قید و... را در متن طی فرایند برچسب‌گذاری واژگان تشخیص داد. ضرورت این عمل از این جهت است که نوع و نقش بسیاری از واژگان موجود در متن بالقوه مبهم بوده و با توجه به جایگاه آن‌ها در جمله تعیین

می‌شود. لذا فرایند برچسب‌گذاری کمک زیادی به ابهام‌زدایی خواهد کرد. ازجمله مهم‌ترین این ابزارها در فارسی عبارت‌اند از: الف) farsiNLPTools ب)HAZM : کتابخانهٔ برنامه قدرت‌مند پایتون[٥٤] برای متون فارسی. ج) Maryam Tavafi POS Tagger د) Persian Toolbox: ابزاری چندمنظوره برای پردازش زبان طبیعی و) Parsivar.

۴ـ۲ـ۳. تحلیل‌گر واحدهای صرفی (توکن‌ساز) فارسی

از این ابزار برای بخش‌بندی متن طبق واحدهایی همچون کلمه، پاراگراف، نمادهای معنای و غیره به‌منزلهٔ واحدهای مستقل معنایی و بخش‌بندی مجدد متن براساس تنها یکی از این واحدها استفاده می‌شود. از نمونه‌های فارسی آن می‌توان به موارد زیر اشاره کرد: الف) HAZM ب) polyglot: یک ابزار چندزبانه و چندمنظوره برای بسیاری ار کاربردهای پردازش زبان طبیعی مانند تحلیل احساسات، برچسب زدن دستوری واژگان و غیره ج) tok-tok د) Parsivar.

۴ـ۲ـ۴. ریشه‌یاب[٥٥] فارسی

از این ابزار برای حذف پیشوند، پسوند و میانوند به منظور استخراج ریشهٔ واژگان بر-طبق قواعد ساخت واژه و سپس تشخیص انواع دستوری واژگان براساس ریشه به‌دست آمده استفاده می‌شود. تاکنون روش مؤثری برای حذف پیشوندها ارائه نشده است. در تلاشی که در آزمایشگاه فناوری وب انجام شده است، سعی شده تا براساس آنالیزهای آماری و داده‌کاوی، پسوندها حذف شوند که این روش هم می‌تواند راهی برای تشخیص ریشه باشد. اگر این فرایند برحسب تک‌واژه‌ها و بدون توجه به بافت واژه باشد به آن ریشه-یاب لغوی[٥٦] گفته می‌شود، ولی اگر براساس کلمات هم‌نشین و بستر کاربرد آن‌ها پردازش انجام شود به آن ریشه‌یاب نحوی[٥٧] می‌گوییم که فرایندی به‌مراتب پیچیده‌تر و البته دقیق‌تر به شمار می‌آید. در زبان انگلیسی بیشتر از ابزار porter استفاده می‌شود، ولی بهترین ابزارهای فارسی عبارت‌اند از: الف) PersianStemmer: در این ابزار از الگوریتم مبتنی‌بر تطابق انگاره[٥٨] استفاده می‌شود که در ارزیابی‌ها، گوی سبقت را از سایر ریشه‌یاب‌های فارسی ربوده است چراکه بسیاری از معضلات ریخت‌شناسی[٥٩] فارسی را برطرف می‌سازد. ب) Perstem: دارای قابلیت افزوده ریخت‌شناسی و تحلیل صرفی ج) Parsivar.

۴ـ۲ـ۴. پیکره‌های موجود در زبان فارسی

ازجمله پیکره‌های دیگر فارسی می‌توان به موارد زیر اشاره کرد: مجموعه اشعار فارسی و متون کلاسیک Farsi Poem Corpus که حاوی متون ٤٨ شاعر فارسی و دارای بیش از ٨ میلیون واژه و بالغ بر ۱۵۰ هزار مصراع شعری است. مجموعه‌های مفید دیگری که برای تحلیل احساسات به‌کار می‌روند پیکرهٔ NRC-Persian-Lexicon و polyglot هستند. به علاوه، پژوهشگران حوزهٔ زبان می‌توانند از دو پیکرهٔ تک‌زبانهٔ TMC:Tehran Monolingual Corpus برگرفته از دو مجموعهٔ همشهری و وبگاه خبرگزاری ایسنا و VOA Persian Corpus با حدود ٨ میلیون واژه نیز بهره جویند.

۵. تکنیک‌ها و روش‌های متن‌کاوی

فرایند متن‌کاوی شامل مراحل انتخاب متن، پیش پردازش متن، تبدیل متن و تولید ویژگی‌ها، گزینش ویژگی‌ها، داده‌کاوی و کشف الگو، و تفسیر و ارزیابی است، اما به‌طور اجمالی می‌توان فرایند متن‌کاوی را در سه مرحله خلاصه کرد: الف) آماده‌سازی یا پیش‌پردازش متن، ب) پردازش متن و استفاده از الگوریتم‌های داده‌کاوی به‌منظور کشف انگاره‌های پنهان و روابط مفهومی اجزای متن، ج) تحلیل متن برای ارزیابی برون‌داد و کشف دانش (Saura & Bennett, 2019). این فرایند با به‌کارگیری روشمند تکنیک‌های بازتولید و بازیابی اطلاعات مفید از منابع بزرگ داده‌های متنی شکل می‌گیرد. این قابلیت در روش‌های سنتی، محدود به بازیابی واژگان کلیدی پیش‌فرض بود، ولی روش‌های جدید، فراتر از سطح واژه و با استفاده از الگوریتم‌های محاسباتی پردازش زبان طبیعی و جستارهای[٦٠] مبتنی‌بر آن به بررسی، تفسیر، شناسایی، استخراج، تلفیق و واکاوی خودکار اطلاعات متنی می‌پردازند (Shi et al.,2014). در بخش پایانی، به شرح تکنیک‌های متن‌کاوی می‌پردازیم و چگونگی تبدیل آن‌ها به اطلاعات کاربردی در حوزه‌های زبانی را بررسی می‌کنیم.

۵ـ۱. خلاصه‌سازی

پردازش و خلاصه‌سازی سریع اسناد کلان و فشرده‌سازی اطلاعات انبوه برای برآورده کردن نیازها و اهداف ویژهٔ کاربران اهمیت بسزایی دارد. ازاین‌رو، اسناد ورودی در

خلاصه‌سازهای متنی از طریق حذف جزئیات اضافی و با محوریت مفهوم اصلی متن به صورت یک خروجی فشرده و قابل استفاده درمی‌آیند. باوجوداین، آموزش تفسیر مفهومی متون به نرم‌افزار کاری دشوار است (Gupta & Lehal, 2009). بدیهی است که خروجی اولیهٔ الگوریتم‌های فشرده‌ساز به زبان ماشینی بوده و تنها برای بازتولید خودکار متن اولیه بکار می‌رود. لذا این خروجی در بدو امر برای کاربران قابل استفاده نیست و برون‌داد باز-سازی‌شدهٔ نهایی است که حائز اهمیت است. در زمینهٔ خلاصه‌سازی افرادی هم هستند که به طور سنتی و البته حرفه‌ای، متون را بدون استفاده از تکنیک‌های متن‌کاوی، چکیده‌نویسی و نمایه‌سازی می‌کنند (Agrawal & Batra, 2013). پژوهشگران زبان فارسی می‌توانند از میان خلاصه‌سازهای متون به برنامهٔ کاربردی Persian-Summarization مراجعه کنند.

۵ـ۲. بازیابی سند[۶۱]

یکی دیگر از تکنیک‌های اصلی متن‌کاوی استفاده از روش بازیابی سند و در مرحلهٔ تکمیلی آن بازیابی اطلاعات است. در این روش برای شناسایی و بازیابی یک سند متنی به ابَردادهها[۶۲] نیاز است. ابَردادهها به هر نوع اطلاعات توصیفی ساختاریافته مانند موضوع، زبان، نویسنده و واژگان کلیدی گفته می‌شود که برای شناسایی، توصیف، مکان‌یابی و مدیریت منابع شبکه‌ای به‌کار می‌روند (Weiss et al., 2010). به‌عبارت کلی‌تر، ابَردادهها داده‌هایی دربارهٔ داده‌ها هستند. در کتابخانه‌ها، از کاتالوگ‌ها برای جست‌وجو و شناسایی اسناد از طریق منابع حاوی ابَرداده استفاده می‌شود. امروزه با وجود شناسایی خودکار ابَردادهها، هنوز هم روش‌های دستی استخراج و شناسایی آن‌ها در رشتهٔ کتابداری آموزش داده می‌شود. در متن‌کاوی خودکار ابَردادهها باید هر واژه به‌طور جداگانه در مجموعه سندهای متنی مربوط به آن نمایه شده که این امر نیازمند استفادهٔ هم‌زمان از چندین تکنیک بازیابی است. در مرحلهٔ بعد متن‌های برگشتی باید پردازش شوند تا اطلاعات موردنظر کاربران با توجه به جستارهای قبلی خودشان خلاصه‌سازی و استخراج شود که به آن بازیابی اطلاعات گفته می‌شود و به‌جای تلفیق همهٔ اسناد متنی، بخش‌های مختلف آن را به طور مجزا و مستقل بررسی شوند تا از تداخل محتوایی پرهیز گردد (Agrawal & Batra, 2013).

۵ـ۳. ارزیابی شباهت بین اسناد[۶۳]

ازجمله تکنیک‌های متن‌کاوی، ارزیابی شباهت بین اسناد مختلف متنی است که به‌منزلهٔ یک زمینهٔ پژوهشی و چالش مهم مطرح است. برای مثال، می‌توان به مسئلهٔ تخصیص اسناد و خوشه‌بندی آن‌ها در دسته‌های پیش‌فرض اشاره کرد که تکنیک‌های مختلف آن را تنها با بکارگیری معیارها و شاخص‌های علمی پژوهشی می‌توان سنجش و ارزیابی نمود. در اینجا مسئلهٔ طبقه‌بندی محتوایی متون اسناد بدون‌ساختار در دسته‌بندی‌های ازپیش‌تعریف‌شده مطرح می‌شود که با روش موسوم به «واژگان کنترل‌شده» انجام پذیرفته و صرفاً از اطلاعات محتوای اسناد برای برچسب‌گذاری‌های نمادین استفاده می‌کند. این‌کار تکنیکی دیرینه برای بازیابی اطلاعات در کتابخانه‌های سنتی است که دسترسی به محتوا و مطالب آن را از طریق جست‌وجوی موضوع، نویسنده یا عنوان ممکن می‌سازد (Vashishta & Jain, 2011).

۵ـ۴. خوشه‌بندی[۶۴] و طبقه‌بندی[۶۵] اسناد متنی

خوشه‌بندی/ طبقه‌بندی خودکار برای نمایه‌سازی[٦٦]، استخراج خودکار ابَرداده‌ها، ابهام‌زدایی واژگانی[٦٧]، ساماندهی کاتالوگ‌های انبوه اینترنتی و ساختارمند کردن آن‌ها اهمیت ویژه‌ای دارند. خوشه‌بندی اسناد به یک یا چندین دسته به نوع متن و کارکرد آن بستگی دارد و به شکلی انجام می‌شود که داده‌های موجود در هر زیرمجموعه برخی از ویژگی‌های مشترک طبقات مشابه را هم دارا باشند (Patel & Sharma, 2014). برخلاف خوشه‌بندی که مبنای معینی نداشته و هر متن را هم‌زمان در چندین زیرعنوان قرار می‌دهد، در طبقه‌بندی، اسناد متنی در زیرگروه‌های پیش‌فرض دسته‌بندی می‌شوند تا درون‌مایهٔ اصلی متن مشخص شود. تشخیص درون‌مایهٔ اصلی سند با قرار دادن آن در مجموعه عنوان‌های ازپیش تعیین‌شده است که از آن می‌توان در مرتبط کردن متون و سامان‌دهی اطلاعات علمی بهره جست (Bhushan et al., 2014). پژوهشگران حوزهٔ زبان می‌توانند برای خوشه‌بندی/طبقه‌بندی متون فارسی از پیکره‌های بزرگ Hamshahri و Bijankhan Corpus استفاده کنند. مجموعهٔ بیژن‌خان حاوی حدود دو و نیم میلیون کلمه برچسب دستی و چهل برچسب نحوی فارسی در موضوعات متنوع سیاسی، فرهنگی، علمی و غیره است.

۵ـ۵. تصویرسازی[۶۸] یا متن‌کاوی بصری

تصویرسازی یکی ازروش‌هایی است که هم به‌منزلهٔ خروجی نهایی و هم به‌منزلهٔ تکنیک متن‌کاوی مورد استفاده قرار می‌گیرد. این روش برای نمایش گرافیکی روندها، انگاره‌ها وروابط اجزای متن در قالب‌ها و طرح‌های مختلفی مانند انواع نمودارها بویژه نمودارهای جریانی استفاده می‌شود. تبدیل متون بدون‌ساختار به اطلاعات بصری، به مخاطب در پردازش ذهنی بهتر و سریع‌تر نتایج کمک می‌کند. تصویرسازی داده‌ها اغلب براساس فراوانی واژگان متون انجام می‌پذیرد (Burley, 2010). به‌دلیل تلاش‌های جیم فِلانِگان[۶۹] در اواخر قرن بیستم امکان استفاده از فناوری ابرساز[۷۰] برای درک مضمون متون و اختلاف گفتمان‌های سیاستمداران در بازه‌های زمانی متفاوت به‌کمک ابربرچسب‌ها[۷۱] و براساس معنی و بسامد واژگان فراهم شد. به‌علاوه، امکان مقایسهٔ چندبعدی متون با استفاده از این تکنیک‌ها نیز به‌وجود آمد (Bilgin & Çamurcu, 2008). به اعتقاد هِرست و رازنر[۷۲] (2008) ابر-برچسب‌ها از مؤثرترین تکنیک‌ها در تصویرسازی داده‌ها به‌شمار می‌آیند.

در اینجا باید به این نکته اشاره کرد که هر کدام از این تکنیک‌ها تنها برای واکاوی بخشی از اطلاعات پنهان در اسناد متنی به‌کار می‌روند و بدیهی است که پژوهشگران باید با توجه به ملاک‌های موردنظر خود و ارزیابی اثربخشی هر روش اقدام به انتخاب هوشمندانهٔ تکنیک مناسب کنند و شناسایی و جداسازی هر گونه اطلاعات خاص مستلزم نیازسنجی و هدف‌گذاری ویژه خواهد بود. دراین‌صورت نتیجهٔ نهایی کار، یک قطعهٔ سفارشی از متن یا نوعی ارائهٔ بصری پیش‌بینی‌شده است که برای کشف دانش در زمینهٔ موردنظر مفید خواهد بود.

با توجه به نوپا بودن این دانش و اقبال روزافزون محافل پژوهشی به آن، به‌نظر می‌رسد متن کاوی سهم بسزایی در پژوهش‌ها و طرح‌های دانش‌بنیان علوم انسانی در دههٔ آینده داشته باشد. با این وجود، هنوز این دانش در مقایسه با مفهوم‌کاوی متن در مرحلهٔ تحقیق و توسعه بوده و کمبود پژوهش‌ها و ابزارهای کاربردی در این زمینه مشهود است چراکه تکنیک‌های آن از عمق کافی برخوردار نبوده و قابلیت استخراج و درک ارزش واقعی محتوای نهفته همهٔ متن‌ها را دارا نیستند. به‌علاوه، در دسته‌بندی مضامین نوظهور نیاز به بازخورد مداوم برای رصد، ادغام و ساماندهی موضوعات و مفاهیم آن‌ها به شکلی معنادار و پرهیز از اشتباهات معمول متن‌کاوی در دسته‌بندی‌ها احساس می‌شود. ژرف‌اندیشی و پیچیدگی ویژهٔ

متون پژوهش‌های کیفی نیز سبب دشواری در تفسیر معنای این مباحث و درنتیجه عدم شفافیت متن‌کاوی می‌شود و این کار مستلزم آموزش مجدد الگوریتمی برای هر مجموعه دادهٔ جدید خواهد بود که نیازمند تلاش بیشتر در این حوزه خواهد بود. همکاری‌های میان‌رشته‌ای زبان‌شناسان رایانشی و پژوهشگران علوم کامپیوتر در کشور می‌تواند به بومی‌سازی روش‌های متن کاوی محتوای متون فارسی هم کمک کند و افزایش حجم و تعداد پیکره‌ها در این زمینه را درپی داشته باشد.

۶. پی‌نوشت‌ها

1. text mining
2. text data mining
3. text analytics
4. natural language processing (NLP)
5. Hearst
6. Witten
7. data mining
8. web mining
9. patterns
10. database
11. corpora
12. Feldman & Dagan
13. Xu & Zhao
14. Pons-Porrata et al
15. Yassine & Hajj
16. Kriegel et al
17. Gupta
18. Gharehchopogh & Abbasi Khalifehlou
19. Wu et al
20. speech acts
21. Lee et al
22. Marquez & Moya
23. Cortez & Silva
24. *Leong* et al
25. directed content analysis
26. summative content analysis
27. cross tabulations
28. chi-square

29. bibliometrics
30. scientometrics
31. altmetrix
32. webometrix
33. informetrics
34. science mapping
35. co-word, co-occurrence & co-author analyses
36. concept mining
37. open-ended surveys
38. concept linkage
39. relational extraction
40. interactive learning mode
41. community-based learning groups
42. spams
43. content validation
44. tokenizations
45. user friendly interface
46. flowcharts
47. Gemert
48. Kaur & Chopra
49. open-source (text mining) tools
50. normalizer
51. characters
52. stopwords
53. Part of speech (POS) tagger
54. Python
55. stemmer
56. stemming
57. lemmatization
58. pattern matching
59. morphology
60. queries
61. document retrieval
62. metadata
63. document similarity assessment
64. clustering
65. classification
66. indexing
67. lexical disambiguation
68. visualization
69. Jim Flanagan
70. cloud technology

71. tag clouds
72. Hearst & Rosner

۷. منابع

- امامی، ا.، و قائمی، ر. (۱۳۹٤). دسته بندی موضوعی متون فارسی با استفاده از تکنیک‌های یادگیری انتقالی. *کنفرانس بین‌المللی پژوهش‌های کاربردی در فناوری اطلاعات، کامپیوتر و مخابرات*، تربت حیدریه، شرکت مخابرات خراسان رضوی. https://www.civilica.com/Paper-ITCC01-ITCC01_445.html.
- بنائی، م. (۱۳۹٤). مقدمه‌ای بر پردازش متون فارسی با پایتون. http://www.bigdata.ir/1394/08/ .
- تیمورپور، ب.، و سپهری، م.، و پزشک، ل. (۱۳۸۸). روشی نوین برای دسته بندی هوشمند متون علمی (مطالعهٔ موردی مقالات فناوری نانو متخصصان ایران). *سیاست علم و فناوری*، ۲(۲)، ۱-۱٤. https://www.sid.ir/fa/journal/ViewPaper.aspx?id=105127 .
- لاکتراشی، ط.، و بهشتی، ه. و بهرام پور، ا. (۱۳۹٤). چالش های استفاده از پردازش زبان طبیعی (NLP)در زبان فارسی. *دومین کنفرانس ملی توسعه علوم مهندسی*،تنکابن. https://civilica.com/doc/386108.
- مزینانی ا.، و علیزاده ع.، وآزاد ع.(۱۳۹۷). تحلیل گفتمان موضوعی: تلفیقی از رویکرد گفتمانی ـ تاریخی و تفسیر موضوعی قرآن کریم. *جستارهای زبانی*، ۹ (۳) ، ۱-۳۳.
- هاشمی، س.م.، و درفشان، ز.، و جوادی، ص.، و مولازاده دزفولی، م. (۱۳۹۳). استفاده از تکنیک‌های متن‌کاوی برای دسته‌بندی متون فارسی با مجموعه دادهٔ همشهری، *کنفرانس بین‌المللی مهندسی، هنر و محیط زیست*. https://civilica.com/doc/372406.

References

- Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., Gorman, M., & Zimek, I. (2006). Finding hierarchies of subspace clusters. *LNCS: Knowledge discovery in databases: PKDD. Lecture Notes in Computer Science* (Vol. 4213, 446–453).
- Agrawal, R., & Batra, M. (2013). A detailed study on text mining techniques. *International Journal of Soft Computing and Engineering (IJSCE)*. ISSN: 2231-

2307, Vol. 2, Issue-6.

- Alfiani, P.A., & Wulandari, A. F. (2015). Mapping student's performance based on data mining approach (a case study). *Ital. Oral Surg.* **3**, 173–177.
- Arslan, A.A. (2011). *Türkçe Metinlerden Anlamsal Bilgi Çikarimi Için Bir Veri Madenciligi Uygulamasi.* Baskent Üniversitesi Fen Bilimleri Enstitüsü, Yüksek Lisans Tezi.
- Aydin, C.R., Erkan, A., Güngör, T., & Takçi, H. (2013). Sözlük Tabanli Kavram Madenciligi. *Türkçeiçin bir Uygulama*, 30. Ulusal Bilisim Kurultayi, November 2013, Ankara.
- Banai, M. (2015). Introduction to the processing of Persian texts with Python. http://www.bigdata.ir/1394/08/. [In Persian]
- Baker, R.S., & Inventado, P.S. (2014). Educational data mining and learning analytics. *Learning Analytics: From Research to Practice*, 61–75. Springer, New York.
- Bhushan, J., Pushkar, W., Shivaji, K., & Nikhil, K. (2014). Searching research papers using clustering and text mining. *International Journal of Emerging Technology and Advanced Engineering*, 4(4).
- Bilgin, T.T., & Çamurcu, A.Y. (2008). *Çok Boyutlu Veri Görsellestirme Teknikleri*, Akademik.
- Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Burley, D. (2010). Information visualization as a knowledge integration tool. *International Journal of Knowledge Management Practice*, 11(4).
- Çalis, K., Gazdagi, O., & Yildiz, O. (2013), Reklam Içerikli Epostalarin Metin Madenciligi Yöntemleriile Otomatik Tespiti. *Bilisim Teknolojileri Dergisi*, 6(1), 1-7.
- Cobo, M.J., López-Herrera, A.G., Herrera-Viedma, E. & Herrera, F. (2011). Science Mapping Software Tools: Review, Analysis, and Cooperative Study among Tools.

Journal of the American Society for Information Science and Technology, 62(7). 1382–1402.

- Cohen, W.W. (1999). What can we learn from the web? *In Proceedings of the Sixteenth International Conference on Machine Learning* (ICML'99), 515-521.
- Cohen, K.B. & Hunter, L. (2008). Getting started in text mining. *PLoS Comput Biol.* 4(1): e20.
- Cortez, P., & Silva, A. (2008). Using data mining to predict secondary school student performance. *5th Annual Future Business Technology Conference*, vol. 2003, No. 2000,.5–12.
- De Bellis, N. (2009). *Bibliometrics and citation analysis: from the Science citation index to cybermetrics***.** Scarecrow Press.
- Dolgun, M.Ö., Özdemir, T.G., & Oguz, D. (2009). Veri madenciliginde yapisal olmayan verininanalizi: Metin ve web madenciligi. *Istatistikçiler Dergisi*, 2, 48-58.
- Emami, A., & Ghaemi, R. (2015). Thematic classification of Persian texts using transitional learning techniques. *International Conference on Applied Research in Information Technology, Computer and Telecommunication*. Torbat Heydariyeh, Khorasan Razavi Telecommunication Company. https://www.civilica.com/Paper-ITCC01-ITCC01_445.html. [In Persian]
- Ergün, M. (2016). Using the techniques of data mining and text mining in educational research. *Participatory Educational Research (PER)* Special Issue 2016-III, pp., 140-151. http://www.partedres.com ISSN: 2148-6123.
- Feldman, R., & Dagan, I. (1995). KDT knowledge discovery in texts. In Proceedings of the First International Conference on Knowledge Discovery *(KDD95)*.
- Feng, Q. (2008). A comprehensive review of data excavation tools in China [J]. (10): 11-13.
- Gemert, J.V. (2000). Text Mining Tools on the Internet: an overview.*ISIS technical report series*, Vol. 25. Department of Computer Science, University of

Amsterdam,The Netherlands. http://www.science.uva.nl/research/isis.

- Gharehchopogh, F. S., & Abbasi Khalifehlou, Z. (2011). Analyzing information
- extraction methods from text mining and natural language processing perspectives. *AWER Procedia Information Technology & Computer Science, 2nd World Conference on Information Technology.*
- Gupta, V. (2009). A survey of text mining techniques and applications. *Journal of*
- *Emerging Technologies in Web Intelligence*, 1(1).
- Gupta, V. & Lehal, S. (2009). A survey of text mining technique and applications, *Journal of Emerging Technologies in Web Intelligence*, vol.1, No. 1. doi: 10.4304/jetwi.1.1.60-76
- Hashemi, S.M., Darfshan, Z., Javadi, S., & Molazadeh Dezfuli, M. (2014). Using text mining techniques to classify Persian texts with Hamshahri data set. *International Conference on Engineering, Art and Environment*. https: // civilica .com / doc / 372406. [In Persian]
- Hashimi, H., Hafez A., & Mathkour, H. (2015). Selection criteria for text mining approaches. *Computers in Human Behavior* 51 (B): 729–733.
- He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends*, 48(1). 133-159.
- Hearst, M. (2009). What is text mining? *http://www.sims.berkeley.edu/~hearst/.*
- Hearst, M.A. & Rosner, D., (2008). Tag clouds: Data analysis tool or social signaller?. *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual,* 7-10 Ocak, 160-160. DOI: 10.1109/HICSS.2008.422.
- Hsieh, H. F., & Shannon, S.E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277-1288. doi: 10.1177/10497323 05276687.
- Kaur, A. & Chopra, D. (2016). Comparison of text mining tools. *5th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO*). pp. 186-192. doi:10.1109/icrito.2016.7784950.

- Kriegel, H. P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *Transactions on Knowledge Discovery from Data* (New York, NY: ACM), 3(1), 1–58.
- Kumar, S.A., & Vijayalakshmi, M. (2011). A novel approach in data mining techniques for educational data. In: *3rd International Conference on Machine Learning Computing* (ICMLC 2011) A, no. Icmlc,. 152–154.
- Laktrashi, T., Beheshti, H., & Bahrampour, A. (2015). Challenges of using natural language processing (NLP) in Persian. *Second National Conference on Engineering Science Development*. Tonekabon. https://civilica.com/doc/386108. [In Persian]
- Lee, S.J., Liu, Y., & Popović, Z. (2014). Learning individual behavior in an educational game: a data-driven approach. In: *Proceedings of 7th International Conference on Educational Data Mining*, 114–121.
- Leong, C. K., Ee, H., & Ak, K. (2012). Mining sentiments in SMS texts for teaching evaluation [J]. *Expert Systems with Applications*, 39(3): 2584~2589.
- Liptha, L. R., Raja, K., & Tholkappia A. G. (2010). Text clustering using concept-based mining model. *International Journal of Electronics and Computer Science Engineering* (ISSN: 2277-1956).
- Mazinani, A., Alizadeh, A., & Azad, A.R. (2018). Topic-based discourse analysis: a blend of discourse-historical: approach and the quran's thematic exegesis. *LRR, 9* (3):1-33. RL. http://lrr.modares.ac.ir/ article-14-169-fa.html. [In Persian]
- Marquez, B., & Moya, L. (2011). An automatic text comprehension classifier based on mental models and latent semantic features [A]. *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies* [C]. New York: ACM Press, 2011:158-162.
- Patel, R., & Sharma, G. (2014). A survey on text mining techniques. *InternationalJournal of Engineering and Computer Science*, 3(5).
- Pena-Ayala, A. (Ed.) (2014). *Educational data mining: Applications and trends* (Vol.

524). Springer.

- Pons-Porrata. A., Berlanga-Llavori, R., & Ruiz-Shulcloper, J. (2007). Topic discovery based on text mining techniques [J]. *Information Processing & Manmanagement*, 43(3): 752-768.
- Saura, J.R., & Bennett, D.R.A. (2019). A three-stage method for data text mining: using ugc in business intelligence analysis. *Symmetry*, *11*(4), 519. doi:10.3390/sym11040519. [In Persian]
- Shangzhen, L., & Lemen, C. (2016). Research on the application of text mining in Chinese information analysis comment [J]. *Information Science*, (08): 153-159.
- Shi, C., Verhagen, M. & Pustejovsky, J. (2014). A conceptual framework of online natural language processing pipeline application. *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, Dublin, Ireland. 53–59.
- Sin, K., & Muthu, L. (2015). Application of big data in education data mining and learning analytics—a literature review. *ICTACT J. Soft Comput. : Special Issue Soft Comput. Models Big Data* **5**(4), 1035–1049.
- Swamy, M., & Hanumanthappa, M. (2012). Predicting academic success from student enrolment data using decision tree technique. *Int. J. Appl. Inf. Syst.* 4(3), 1–6.
- Teymourpour, B., Sepehri, M., & physician, L. (2010). A new method for intelligent classification of scientific texts (Case study of nanotechnology articles by Iranian experts). *Science and Technology Policy*, *2*(2), 1-14. https://www.sid.ir/fa/journal/ViewPaper .aspx? id = 105127. [In Persian]
- Vashishta, S., & Jain, Y. K. (2011). Efficient retrieval of text for biomedical domain using data mining algorithm. *International Journal of Advanced Computer Science and Applications*, 2(4).
- Wang, L., Wang, G. & Alexander, C.A. (2015). Big data and visualization: Methods, challenges and technology progress. *Digital Technologies*, 1(1), 33-38.
- Weiss S.M., Indurkhya N., & Zhang T. (2010). Information retrieval and text mining. *Fundamentals of Predictive Text Mining. Texts in Computer Science.* Springer,

London.

- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques,* 2nd Edition. San Francisco: Morgan Kaufmann.
- Wu, S. T., Li, Y., & Xu, Y. (2006). Deploying approaches for pattern refinement in text mining. In *Proceedings of the Sixth International Conference on Data Mining*.
- Xu, Y., & Zhao, R. (2017). The literature review of text data mining. *Science Discovery*. 5(6), 2017,438-443. doi: 10.11648/j.sd.20170506.18
- Yassine, M., & Hajj, H. (2010). A framework for emotion mining from text in online social networks. In *IEEE international conference on data mining workshops*(pp. 1136–1142). Sydney, NSW: IEEE publications.
- Zhou, Y., Zhang, Y., Vonortas, N., & Williams, J. (2012). *A text mining model for strategic alliance discovery.* 45th Hawaii International Conference on System Sciences.