مقاله ترویجی
Doi: 10.22034/popsci.2022.306089.1130

متن‌کاوی :مفاهیم و روش‌ها

فاطمه جلالی شیجانی[1]، مجید شیرزاد[2]

تاریخ دریافت:۹ شهریور ۱۴۰۰ تاریخ پذیرش: ۱۶ آذر ۱۴۰۰

چکیده:

در عصر حاضر، حجم عظیمی از اطلاعات موجود در محیط وب، اسناد و مقالات متنی هستند. در سالیان اخیر، توجه بسیار زیادی در حوزه بین‌المللی و ملی به متن‌کاوی شده است. افزایش چشم‌گیر داده‌های متنی، پژوهشگران را بر آن داشته است که به دنبال روش‌هایی جهت کاوش در این داده‌ها باشند. طبیعی است که محققان ایرانی نیز ازین امر مستثنی نبوده‌اند. متنکاوی، روشی برای استخراج اطلاعات غیرساخت‌یافته و نیمه‌ساخت‌یافته از این حجم اطلاعات موجود در اینترنت و نیز، فرایند استخراج دانش و الگوهای ناشناخته و غیرقابل فهم و بالقوه، از میان انبوه مجموعه‌های داده‌های متنی است. متن‌کاوی به همراه تمامی روش‌ها و تکنیک‌های آن، کوششی است که پژوهشگران را در استخراج دانش و اطلاعات مفید و باارزش از انبوه متون غیرساخت‌یافته‌ای که در محیط اینترنت پراکنده‌اند، یاری می‌کند.

یافته‌های پژوهش نشان می‌دهد که متن‌کاوی، کاربردی از داده‌کاوی است و تفاوت اصلی این دو، استخراج الگوها از متنی با زبان طبیعی در متن‌کاوی است، درحالیکه داده‌کاوی بر روی پایگاه داده‌های ساخت‌یافته عمل می‌کند. آشنایی با فرایندهای متن‌کاوی و شناسایی تکنیک‌های آن، از جمله اهداف این پژوهش است. فرایندهای متن‌کاوی، دارای دو

۱. دانشجوی کارشناسی ارشد مدیریت اطلاعات، دانشگاه پیام نور تنکابن، مازندران. (نویسنده مسئول)
f.jalali85@yahoo.com
۲. استادیار گروه علم اطلاعات و دانش شناسی، دانشگاه پیام نور، تهران، ایران.
mshm1362@yahoo.com

فاز اصلی پیش‌پردازش مستندات و استخراج دانش هستند. تاکنون هشت تکنیک نیز برای متن‌کاوی معرفی شده است که عبارتنداز: استخراج اطلاعات، بازیابی اطلاعات، خلاصه‌سازی متن، طبقه‌بندی، خوشه‌بندی، بصری‌سازی، پردازش زبان طبیعی وعقیده‌کاوی. با اینکه روش‌های متن‌کاوی اکثراً بر روی منابع لاتین انجام گرفته‌اند، اما با جست‌وجو در پایگاه‌های اطلاعاتی فارسی، درمی‌یابیم طی یک دهه گذشته، موضوع متن‌کاوی برای محققان ایرانی به‌خصوص دانشجویان رشته‌های علوم کامپیوتر و فناوری اطلاعات، اهمیتی دوچندان پیدا کرده است؛ به طوری که بخش قابل توجهی از مقالات کنفرانس‌های مربوط به علوم و فنون کامپیوتر را مقالات مربوط به این حوزه تشکیل می‌دهند.
کلیدواژگان: متن‌کاوی، کشف دانش، دسته‌بندی متن، فناوری اطلاعات، داده‌کاوی.

مقدمه و بیان مسئله

امروزه اسناد و مقالات، حجم عظیمی از اطلاعات در اینترنت را تشکیل می‌دهند. می‌توان گفت استفاده و بهره‌مندی کامل از تمامی این اطلاعات تقریباً غیرممکن است. متن‌کاوی، روشی برای استخراج اطلاعاتی است که به صورت غیرساخت‌یافته نامرتب و نیمه‌ساخت‌یافته از این حجم اطلاعات می‌باشد.

در سال‌های اخیر به علت گسترش روزافزون اطلاعات، که درصد قابل توجهی از این اطلاعات تولید شده به صورت متنی غیرساخت‌یافته و نیمه‌ساخت‌یافته می‌باشد و نیز به علت رشد روزافزون داده‌های متنی در وب، اهمیت متن‌کاوی افزایش یافته است. همچنین این آنالیز مفهومی متن، که معادل محاسبات معنایی است، در زمینه متن‌کاوی در حال رشد می‌باشد. تکنیک‌های بازیابی اطلاعات قدیمی، برای مقدار زیاد داده‌های متنی که به طور فزاینده افزایش می‌یابند، ناکارآمد هستند و به همین دلیل، اهمیت به کارگیری متدهای جدیدی جهت استخراج و پردازش این متون، مانند برچسب‌زنی موضوعی، که محتمل‌ترین موضوعی که محتوای متن بدان اشاره دارد را مشخص می‌کند، تجزیه و تحلیل معنایی پنهان، که روشی برای استخراج معنایی اطلاعات است، مسائل مهم مربوط به چند زبان‌ها و خلاصه‌سازی از نوع چکیده، تولید خلاصه متن پس از فهم مطالب موجود در متن اصل ، انتخاب بهترین وب‌سایت از نظر محتوا، وب‌سایتی که با توجه به زمینه فعالیت خود بهترین و غنی‌ترین اطلاعات را دارد، ارتباط بین این متون و دسته‌بندی متون که می‌تواند کاربردهای زیادی داشته باشد و ... لازم و ضروری است(نظری؛حبیبی،۱۳۹۵). رشد فزاینده پایگاه داده‌ها، تقریباً در هر ناحیه از فعالیت انسان باعث شده است که نیاز به ابزارهای قدرتمند جدید برای تغییر دادن داده به دانش مفید افزایش یابد. بنابراین یکی از جدیدترین زمینه‌های مورد تحقیق در داده‌کاوی، متن‌کاوی برای این منظور گسترش یافت(رجب‌پور؛ طلعتیان آزاد،۱۳۹۴).

امروزه متون، یکی از رایج‌ترین قالب داده‌ها محسوب می‌شوند. شبکه‌های اجتماعی، سرویس دهنده‌های ایمیل، کتابخانه‌های دیجیتال و از همه مهم‌تر وب، بسترهایی هستند که در آن، نه هر روز، که هر ثانیه اطلاعات به صورت متن به آن افزوده می‌شود. در میان این حجم انبوه داده‌ها که هر لحظه نیز به اندازه آن اضافه می‌شود، جای خالی تکنیک‌های تحلیلی به خوبی احساس می‌شود. متن‌کاوی کوششی برای به دست‌آوردن دانش مفید از میان این داده‌ها است و رشد داده‌ها به شکل

متن، اهمیت متن‌کاوی را دو چندان کرده است(اسماعیلی؛ زاهد،۱۳۹۴).

هدف از متن‌کاوی، استخراج دانش مفید(مانند الگوها و روندها) از داده‌های خام است. متن‌کاوی نوعی خاص از داده‌کاوی است که بر کار با مجموعه داده‌های نیم‌ساخت‌یافته یا ساخت‌نیافته مانند اسناد متنی متمرکز است. متن‌کاوی می‌تواند با جست‌وجوی واژه‌هایی با بیشترین ارجاع، یا از طریق استفاده از واژه‌هایی معینی از فرهنگ لغت، به صورتی تصادفی مورد استفاده قرار گیرد. متن‌کاوی را می‌توان به عنوان یک روش میان رشته‌ای برای بازیابی اطلاعات، یادگیری ماشین، آمار، دسته‌بندی متون، زبان‌شناسی محاسباتی و مخصوصاً داده‌کاوی در نظر گرفت(نظری؛حبیبی،۱۳۹۵).

داده ها در سه دسته ساخت‌یافته، نیمه‌ساخت‌یافته وساخت‌نیافته قرار می‌گیرند و علیرغم وجود یک ساختار پنهان در متون، این قالب در دسته ساخت‌نیافته گروه‌بندی می‌شود. جدول‌ها در مدل رابطه‌ای، نمونه‌ای از داده ساخت‌یافته و اسنا XML گونه‌ای از داده‌های نیم‌ساخت‌یافته تلقی می‌شوند. به همین دلیل متن‌ها به شکل مجموعه‌ای از واژه‌ها یا عبارات دیده می‌شوند. مجموعه داده‌های ساخت‌نیافته یکی از ساده‌ترین شکل داده‌ها محسوب می‌شوند که می‌تواند در هر حوزه کاربردی تولید شود. بنابراین لازم است تا روش‌ها و الگوریتم‌هایی طراحی شوند تا بتوانند به صورت مؤثری این داده‌ها را کاوش کنند(اسماعیلی؛ زاهد،۱۳۹۴).

متن‌کاوی می‌تواند به صورت تکنیکی تعریف شود که برای استخراج اطلاعات سودمند یا دانش از اسناد متنی مورد استفاده قرار گیرد. برای بررسی متون، لازم است اول داده‌ها را بشناسیم.

داده‌ها، نخستین شکل اطلاعات هستند که به منظور ایجاد دانش، مدیریت و کاویده می‌شوند. داده‌ها دارای چندین مشخصه هستند:

۱ .حجم: که به مقدار زیاد داده‌ها برمی‌گردد،

۲ .سرعت بر حسب زمان: نرخ تولید داده‌ها در هر واحد زمانی را نشان می‌دهد،

۳ .تنوع: بر شکل‌های مختلف داده دلالت دارد؛ مانند متن، عدد، تصاویر، صوت، ویدئو و هر فرم و شکل دیگری که بتوان تصور کرد؛

۴ .صحت: با انحرافات، اختلالات و نویز در داده‌ها سر و کار دارد،

۵ .دوام‌پذیری: به معنای بررسی ارتباط یک متغیر در ارائه وسیعی از متغیرهای مربوط به داده‌های چند بعدی و ارتباطات میان متغیرهاست؛

۶ .اعتبار: این پرسش را درباره داده‌ها مطرح می‌کند که آیا آن داده، برای استفاده و کاربرد در نظر گرفته شده، قابل اعتماد و دقیق است؟

۷ .ارزش: حاکی از اهمیت کلیدی داده‌هاست. برخی از داده‌ها می‌توانند بسیار مهم باشند، در حالی که بعضی دیگر از ارزش کمتری برخوردارند؛

۸ .مدت اعتبار: داده‌ها چه مدت اعتبار دارند و باید ذخیره شوند.

هدف از داده‌کـاوی، کشف ضمنی الگوها و روند ناشناخته قبلی از پایگاه داده‌هااست. داده‌کاوی شامل تکنیک‌های بسیاری چون طبقه‌بندی، خوشه‌بندی، شبکه‌های عصبی و درخت‌های تصمیم است. متن، ممکن است در اندازه زیاد و فرم‌های متفاوتی همچون زبان‌های مختلف، با استفاده از نمادهای مختلف و قالب‌های متفاوت موجود باشد. از این رو، این پرسش ایجاد می‌شود که چگونه اطلاعات را می‌توان از این متن خارج کرد. در اینجاست که متن‌کاوی به ایفای نقش می‌پردازد.

متن‌کاوی، کاربردی از داده‌کاوی است. تفاوت اصلی این دو، آن است که در متن‌کاوی، الگوها از متنی با زبان طبیعی استخراج می‌شوند، این در حالی است که داده‌کاوی بر روی پایگاه داده‌های ساخت‌یافته عمل می‌کند (شمسی؛ دیوانی،۱۳۹۵).

چندین تکنیک برای متن‌کاوی پیشنهاد شده است که عبارتند از ساختار مفهومی، درخت تصمیم‌گیری، روش‌های استنتاج قوانین، همچنین تکنیک‌های بازیابی اطلاعات برای کارهایی مانند تطبیق دادن سندها، مرتب کردن، خوشه‌بندی و ... .به استفاده از روش‌های متن‌کاوی برای حل مسائل تجاری یا کسب و کار text analystics می‌گویند. متن‌کاوی به سازمان‌ها این امکان را می‌دهد که بینش تجاری ارزشمندی از محتواهای مبتنی بر متن خود مانند اسناد ورد، ایمیل و پست‌هایی که در استریم رسانه‌های اجتماعی مانند فیسبوک و توییتر و لینکداین وجود دارد به دست آورند.

روش‌های زیادی در فاز استخراج دانش وجود دارد. در عین حال تمامی این روش‌ها را شاید بتوان به دو دسته اصلی تقسیم کرد. این دو دسته اصلی، روش‌های مبتنی بر کارایی و روش‌های مبتنی بر دانش هستند

در روش اول، طراحان نگران کارایی سیستم هستند و طوری سیستم را طراحی می‌کنند که بهترین کارایی و سرعت را داشته باشد. روش‌های رایج‌تر در این نوع نگرش، روش‌های آماری و نیز شبکه‌های عصبی هستند. روش‌های آماری بر پایه هر نوع اطلاعات آماری است که از متون قابل استخراج است. دلیل اصلی به کار بردن روش‌های داده‌کاوی برای استناد متنی، ساختاربندی کردن آنهاست. ساختارهای دستیابی معروف عبارتند از: کاتالوگ‌های کتابخانه یا ایندکس‌های کتاب. مشکل ایندکس‌های طراحی شده به صورت دستی، زمان موردنیاز برای نگهداری آنها است. بنابراین برای منابع اطلاعاتی که خیلی تغییر می‌کنند، مثل وب مناسب نیستند و پیشنهاد نمی‌شوند. روش‌های موجود برای ساختاربندی کردن مجموعه‌ها عبارتند از: روش‌های رده‌بندی و روش‌های خوشه‌بندی. برخی از روش‌های مهم و جدید برای

رده‌بندی کردن متون وجود دارد و روش‌هایی برای به‌طور خودکار استخراج کردن الگوهای مفید از اسناد متنی، ترکیب این روش‌ها با روش‌های ساختاربندی(خوشه‌ای و رده‌بندی)، ابزارهای قدرتمندی برای کاوش الگوهای مفید در مجموعه‌های متنی فراهم می‌کند که برای یک کار اصولی که نتیجه مطلوب بدهد، پیشنهاد می‌شود.

علاوه بر داده‌های آشنای کمی و کیفی برای آمارشناسان، داده‌هایی که به عنوان ورودی الگوریتم‌های استخراج اطلاعات استفاده می‌شوند می‌توانند هر شکلی داشته باشند، مانند تصویر، فیلم، صدا یا متن. در متن‌کاوی ما بر منابع داده‌ای که به شکل متن هستند تمرکز می‌کنیم. منابع داده متنی برای استخراج اطلاعات می‌توانند از free form text (متن‌هایی به شکل‌های آزاد) تا متن‌های semi formatted مانند xml، html و ... را شامل شوند و آن دسته از منابعی را هم دربرمی‌گیرند که به فرمت‌های اسناد کد باز یا open source رمزگذاری شده‌اند(open document) و همچنین سایر فرمت‌های اختصاصی یک شرکت (برای مثال مایکروسافت ورد یا مایکروسافت پاورپوینت). استخراج اطلاعات از این منابع اطلاعاتی چالش بزرگی برای جامعه آماری و فناوری اطلاعات بوده است.

در متن‌کاوی، الگوها از متن‌های به زبان طبیعی استخراج می‌شوند و ورودی آن، یک متن غیرساخت‌یافته و آزاد است. ولی مثلاً در وب‌ماینینگ، منابع وب، اغلب ساخت‌یافته هستند. در تعریف متن‌کاوی گفته می‌شود کشف اطلاعات جدید و اطلاعاتی که از قبل ناشناخته بوده‌اند، توسط کامپیوتر به کمک استخراج خودکار اطلاعات از منابع متنی غیرساخت‌یافته‌ی اغلب بزرگ.

در بازیابی اطلاعات برخلاف متن‌کاوی هیچ اطلاعات جدیدی پیدا نمی‌شود و اطلاعات موردنظر و مطلوب به ندرت با اطلاعات مشابه دیگری به طور همزمان وجود دارند. متن‌کاوی، ترکیبی از فناوری‌های آماری، بازیابی اطلاعات، وب‌کاوی، داده‌کاوی و پردازش زبان طبیعی است.

به طور کلی روش‌هایی که در متن‌کاوی استفاده می‌شوند عبارتند از: استخراج اطلاعات، طبقه‌بندی، خوشه‌بندی، خلاصه‌سازی، ردیابی موضوع، ارتباط‌دهنده مفاهیم، نمایش اطلاعات، پرسش و پاسخ، کاوش مبتنی بر متن، تجزیه و تحلیل گرایش‌ها (فاخری،۱۳۹۶).

فرایند متن‌کاوی

دو فاز اصلی برای فرایند متن‌کاوی داریم: پیش پردازش مستندات و استخراج دانش.

اولین فاز، پیش پردازش مستندات است. خروجی این فاز می‌تواند دو شکل مختلف داشته باشد: ۱.مبتنی بر سند، ۲.مبتنی بر مفهوم.

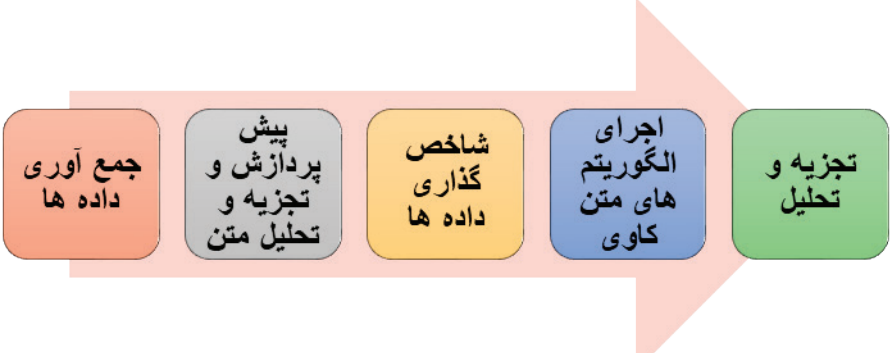
در شکل اول، نحوه نمایش بهتر مستندات مهم است، برای مثال تبدیل اسناد به یک شکل میانی و نیمه‌ساخت‌یافته یا به کار بردن یک ایندکس بر روی آنها یا هر نوع نمایش دیگری که کار کردن با اسناد را کاراتر می‌کند. هر موجودیت در این نمایش، در نهایت باز هم یک سند خواهد بود.

در شکل دوم، نمایش اسناد بهبود بخشیده می‌شود، مفاهیم و معانی موجود در سند و نیز ارتباط میان آنها و هر نوع اطلاعات مفهومی دیگری که قابل استخراج است، از متن استخراج می‌شود. در این حالت، نه با خود موجودیت بلکه با مفاهیمی که از این مستندات استخراج شده‌اند، مواجه هستیم (فاخری،۱۳۹۶).

کاوش در متن پیش‌پردازش شده، در مقایسه با اسناد به زبان طبیعی آسان‌تر است. از این جهت، پیش‌پردازش اسناد از منابع مختلف، قبل از اعمال هرگونه تکنیک متن‌کاوی، امر بسیار مهمی در طول فرایند متن‌کاوی به شمار می‌رود. به منظور کاهش حجم کلمات سندها، از روش‌های خاصی چون فیلتر کردن و ریشه‌یابی کردن استفاده می‌شود. روش‌های فیلتر کردن، آن دست از کلمات که اطلاعات مرتبط و ارزشمندی را فراهم نمی‌آورند، از مجموعه کل کلمات حذف می‌کند. روش‌های ریشه‌یابی نیز، برای تولید ریشه جمع‌ها یا افعال مورد استفاده قرار می‌گیرد (شمسی؛ دیوانی،۱۳۹۵).

قدم بعدی، استخراج دانش از این فرم‌های میانی است که براساس نحوه نمایش هر سند متفاوت می‌باشد. نمایش مبتنی بر سند، برای گروه‌بندی، طبقه‌بندی و تجسم‌سازی استفاده می‌شود، در حالی‌که نمایش مبتنی بر مفهوم، برای یافتن روابط میان مفاهیم، ساختن خودکار آنتولوژی و ... به کار می‌رود. متن‌کاوی برای آن قسمت از کشف دانش از متن به کار می‌رود که مربوط به استخراج الگوها از داده‌های متنی است (فاخری،۱۳۹۶).

به جزء این دو فرایند، دو روش دیگر نیز وجود دارد؛ روش مبتنی بر الگو و روش مبتنی بر عبارت. در روش مبتنی بر الگو، اسناد بر اساس الگو تحلیل می‌شوند. الگوها می‌توانند با استفاده از روش‌های داده‌کاوی مانند استخراج قوانین ارتباط، استخراج مجموعه موارد مکرر، استخراج الگوی متوالی و استخراج الگوی بسته کشف شوند. در روش مبتنی بر عبارت نیز، عبارات، سمانتیک‌های زیادی مانند اطلاعات دارند و دچار ابهام کمتری هم هستند. در این روش، سند براساس عبارت تحلیل می‌شود زیرا عبارات ابهام کمتری دارند و نسبت به اصطلاحات فردی، تمایز بیشتری دارند. اما این روش دارای مشکلاتی است که عملکرد آن را دچار اشکال می‌کند. این دلایل عبارتند از: عبارات نسبت به اصطلاحات، ویژگی‌های آماری کمتری دارند، عبارات فراوانی رخداد کمتری دارند و عبارات، حشویات زیادی دارند و بین آنها عبارات زاید وجود دارد (گیکواد،۲۰۱۴).



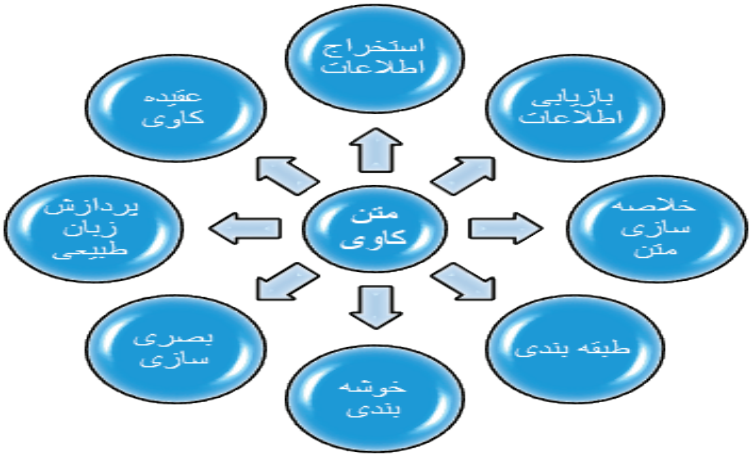
مراحل متن‌کاوی (Text mining)

تکنیک‌های متن‌کاوی

۱. استخراج اطلاعات

استخراج اطلاعات(IE)، وظیفه یافتن اطلاعات ساخت‌یافته را از درون متن غیرساخت‌یافته یا نیمه‌ساخت‌یافته برعهده دارد. نقطه شروع کامپیوترها برای تحلیل داده‌های غیرساخت‌یافته، استفاده از استخراج اطلاعات است. استخراج اطلاعات، امر مهمی در متن‌کاوی به‌شمار می‌رود و در جوامع تحقیقاتی مختلف مانند پردازش زبان طبیعی، بازیابی اطلاعات و وب‌کاوی به طور گسترده مورد مطالعه قرار می‌گیرد. دو وظیفه اساسی استخراج اطلاعات، شناخت موجودیت و استخراج ارتباط است.

نرم‌افزار استخراج اطلاعات، عبارات کلیدی و ارتباطات و وابستگی‌های درون متن را شناسایی می‌کند. این کار به وسیله جست‌وجوی دنباله‌های از پیش تعریف شده، در درون متن انجام می‌شود؛ فرایندی که تطبیق الگو نامیده می‌شود. این فناوری در هنگامی که با حجم زیادی از متن روبرو هستیم، بسیار مفید می‌باشد(شمسی؛ دیوانی،۱۳۹۵).



۲. بازیابی اطلاعات

بازیابی اطلاعات(IR)، فرایندی برای استخراج الگوهای مرتبط براساس مجموعه مشخصی از واژه‌ها یا عبارات است. متن‌کاوی و بازیابی اطلاعات برای داده‌های متنی، ارتباط نزدیکی دارند. در سیستم‌های IR، از الگوریتم‌های متفاوتی برای ردیابی رفتار کاربر و جست‌وجوی داده‌های مرتبط استفاده می‌شود. موتورهای جست‌وجوی گوگل و یاهو برای استخراج اسناد متنی مرتبط براساس عبارت یا وب، بیشتر از سیستم بازیابی اطلاعات استفاده می‌کنند. همچنین اطلاعات مرتب‌تر و مناسب‌تری به کاربر می‌دهند که نیازهای کار را تأمین می‌کند (تألیب،۲۰۱۶).

۳. خلاصه‌سازی متن

برای اینکه بتوانیم متوجه شویم آیا یک سند طولانی، متناسب با نیازها و خواسته‌های کاربر هست یا خیر و آیا ارزش مطالعه برای کسب اطلاعات بیشتر را دارد، خلاصه‌سازی متن فوق‌العاده مفید است. نکته اساسی در خلاصه‌سازی، کاهش دادن طول و جزئیات یک سند، با حفظ نکات اصلی و معنی کلی آن است. مسئله اینجاست که گرچه کامپیوترها می‌توانند مردم، مکان‌ها و زمان را شناسایی کنند، اما اینکه بتوانیم به نرم‌افزار، تجزیه، تحلیل و تفسیر معنا را بیاموزیم، همچنان کار دشواری است.

به طور کلی، زمانی که انسان‌ها متنی را خلاصه می‌کنند، نیاز دارند تا تمامی قسمت انتخاب شده را مطالعه کنند تا بتوانند درک کاملی از آن داشته باشند، آنگاه خلاصه‌ای از نکات اصلی و برجسته آن را بنویسند. از آنجا که کامپیوترها هنوز قابلیت زبان انسان‌ها را ندارند، می‌بایست روش‌های جایگزینی برای آن در نظر گرفته شود. یکی از استراتژی‌هایی که ابزارهای خلاصه‌سازی متن به صورت گسترده از آن استفاده می‌کنند، استخراج جملات است؛ که جملات مهم یک مقاله را با وزن‌دهی آماری به جملات، انتخاب می‌کنند. روش‌های هوشمند دیگری چون موقعیت اطلاعات نیز، برای خلاصه‌سازی مورد استفاده قرار می‌گیرد. خلاصه‌سازی، می‌تواند از روی یک سند یا از گروهی از اسناد باشد. مجموعه‌ای از سندها، با یک خلاصه جایگزین می‌شوند(شمسی؛ دیوانی،۱۳۹۵).

٤. طبقه‌بندی

به علت وجود تعداد زیادی سند به صورت آنلاین، توانایی سازماندهی و اختصاص‌دهی خود کار چنین اسنادی که کلاس‌ها به‌منظور تسهیل بازیابی و تحلیل‌های بعدی ، امری ملالت‌آور و در عین حال ضروری است،از این‌رو طبقه‌بندی خودکار اسناد، عملیات مهمی در متن‌کاوی است.طبقه‌بندی اسناد، در تگ کردن عنوان‌ها به صورت خودکار(برای مثال اختصاص دادن برچسب به سندها)، ساخت دایرکتوری موضوع، شناسایی شیوه‌های نگارش سند(که می‌تواند در محدود کردن نویسندگان اسناد ناشناس کمک کند) و... استفاده شده است.

روند کلی انجام طبقه‌بندی سندها، به این صورت است که ابتدا، مجموعه‌ای از اسناد طبقه‌بندی شده به عنوان مجموعه آموزشی در نظر گرفته می‌شوند. آنگاه مجموعه آموزشی به منظور به‌دست آوردن یک طرح طبقه‌بندی، مورد تجزیه و تحلیل قرار می‌گیرند. چنین طرح طبقه‌بندی شده‌ای، اغلب باید طی فرایند آزمون تصحیح شوند. حال طرح طبقه‌بندی حاصل، می‌تواند برای طبقه‌بندی دیگر اسناد آنلاین استفاده شود.

طبقه‌بندی ساده بیزی و طبقه‌بندیk نزدیکترین همسایه، از جمله روش‌های طبقه‌بندی هستند(شمسی؛ دیوانی،۱۳۹۵).

۵. خوشه‌بندی

خوشه‌بندی، فرایند گروه‌بندی اشیایی با خواص و ویژگی‌های مشابه است. در هر خوشه باید دو ویژگی اصلی به چشم بخورد: شباهت بین کلاسی کم و شباهت درون کلاسی زیاد. تجزیه و تحلیل خوشه‌بندی می‌تواند در کاربردهای زیادی چون تحلیل داده‌ها، پردازش تصویر، تجزیه و تحلیل بازار و ... مورد استفاده قرار گیرد.

خوشه‌بندی اسناد، یکی از مهم‌ترین تکنیک‌ها برای سازمان‌دهی اسناد با استفاده از شیوه نظارت نشده است. یا به عبارتی، خوشه‌بندی، تکنیکی برای گروه‌بندی اسناد مشابه است اما با طبقه‌بندی متفاوت است؛ در این روش، به جای آنکه از عناوین و موضوعات از پیش تعیین شده استفاده شود، اسناد به صورت پویا خوشه‌بندی می‌شوند. در این روش، متون مشابه درون یک خوشه فرا می‌گیرند. هر خوشه، شامل تعدادی سند می‌شود. الگوریتم‌های خوشه‌بندی می‌تواند به دو دسته تقسیم شود: خوشه‌بندی سلسله‌مراتبی و خوشه‌بندی افرازبندی.۴

٦. بصری‌سازی

متن‌کاوی بصری یا بصری‌سازی اطلاعات، منابع متنی بزرگی را در قالب یک نقشه یا سلسله‌مراتب بصری قرار می‌دهد و قابلیت‌های مروری و جست‌وجو را به جست‌وجوی ساده اضافه کرده و فراهم می‌آورد. این روش، اطلاعات قابل درک و فهم سریع‌تر و بهتری را فراهم می‌کند که به ما کمک می‌کند تا بتوانیم در میان مجموعه سندهای بزرگ، به کاوش و تجزیه و تحلیل بپردازیم. این تکنیک کاملاً مبتنی بر متن است. کاربران با استفاده از آن، می‌توانند میان رنگ‌ها، روابط، فاصله‌ها و ... تفاوت قائل شوند. مجموعه سندها می‌توانند به شکل ساختاریافته با استفاده از ایندکس کردن، مدل فضای برداری و ... نشان داده شوند (شمسی؛ دیوانی،۱۳۹۵).

۷. پردازش زبان طبیعی

پردازش زبان طبیعی(NLP)، در مورد پردازش و تحلیل خودکار اطلاعات متنی ساخت‌نیافته است. این روش از تحلیل‌های متفاوتی مانند تشخیص موضوع مشخص(NER) برای استخراج اختصارات و هم‌معناهای آنها جهت یافتن روابط آنها استفاده می‌کند.NER همه مصادیق موضوع مشخص را از بین گروهی از اسناد شناسایی می‌کند. این موضوعات و مصادیق آنها امکان شناسایی رابطه و سایر اطلاعات را برای کسب مفهوم اصلی آنها فراهم می‌کنند. البته این روش هم برای همه موضوعات معین به کار رفته برای شناسایی، فرهنگ لغتی کامل ندارد.

در NLP اغلب برای استخراج مترادف و اختصار از داده‌های متنی، از روش ارجاع مشترک استفاده می‌شود. زبان‌های طبیعی(NL) پیچیدگی‌های زیادی دارند، چنان‌که یک متن استخراج شده از منابع مختلف، کلمات یا اختصار یکسانی ندارد. باید این مسائل شناسایی شوند و قوانینی برای شناسایی یکسانی آنها ایجاد شود (تألیب،۲۰۱۶).

۸. عقیده‌کاوی

یکی از حوزه‌هایی که در این اواخر بسیار مورد توجه محققان قرار گرفته است عقیده‌کاوی است. امروزه شما بر روی اینترنت و به روش‌های متفاوت نظرات، احساسات و عقیده‌های مردم را مشاهده می‌کنید. سازمان‌ها و شرکت‌ها در حوزه کسب‌وکار خود بسیار علاقه‌مند هستند تا نظرات مردم که همانند مشتریان‌شان هستند را بدانند. وب‌سایت‌های این سازمان‌ها دارای حجم بسیار وسیعی از نظرات و توضیحات است و تحلیل این مجموعه داده‌های حجیم به صورت دستی که اغلب به صورت متنی است، کاری پرهزینه و زمان‌بر است. از این‌رو همگی آنها مایل هستند تا به صورت خودکار رویکردی را جهت تحلیل این داده‌ها در اختیار داشته باشند. به صورت کلی نظرات و عقیده‌ها می‌تواند در مورد هر چیزی باشد، مانند یک محصول یا کالا، یک نوع خدمات، یک فرد، یک سازمان، یک رخداد و یا هر چیز دیگر. موردی که درباره آن صحبت می‌شود به عنوان یک موجودیت شناخته می‌شود. این موجودیت‌ها نیز خود از مؤلفه‌ها و مشخصاتی تشکیل شده‌اند، مانند رنگ، وزن و ... . علاوه بر موجودیت و مؤلفه‌های مربوط به موجودیت، در یک نظر می‌توان ویژگی‌های دیگری را نیز پیدا کرد که در برخی از موارد عدم وجود آنها، می‌تواند توضیح و عقیده کاربر را خدشه‌دار کند. درواقع می‌توان گفت که موجودیت، مؤلفه‌های موجودیت، گرایش نظر، شخص اظهارنظرکننده و زمان ثبت نظر، پنج مؤلفه اساسی در یک نظر، احساس و یا عقیده محسوب می‌شوند. یکی از مزایای مهم این تعریف این است که پایه‌ای برای تبدیل متن ساخت‌نیافته(نظرات) به داده‌های ساخت‌یافته (این پنج مؤلفه) مهیا می‌سازد.

بنابر توضیحات، گام‌هایی که در عقیده‌کاوی دنبال می‌شوند عبارتند از: تشخیص موجودیتی که در مورد آن نظر داده شده است، استخراج ویژگی‌های موجودیت‌ها و خوشه‌بندی آنها و تعیین جهت نظرات و دسته‌بندی آن برای تجزیه و تحلیل نتایج.

این حوزه از متن‌کاوی هنوز با چالش‌های بسیاری روبرو است، زیرا تنوع استفاده از واژه میان مردم بسیار زیاد است و زبان طبیعی و به خصوص محاوره‌ای مردم، دارای پیچیدگی‌های خاصی می‌باشد (اسماعیلی؛ زاهد،۱۳۹۴).

با اینکه شاید عنوان «متن‌کاوی» برای بسیاری از ما ناشناخته باشد اما باید بدانیم که استفاده از ابزارها و تکنیک‌های متن‌کاوی با زندگی روزمره ما عجین شده و حوزه‌های مختلف زندگی بشر را احاطه کرده است. در ادامه برخی از این کاربردها را معرفی می‌کنیم:

حوزه تجارت و کسب‌وکار: تحلیل میزان رضایت مشتریان از خدمات ارائه شده و یا محصولات خریداری شده، واکنش بازار نسبت به ویژگی‌های هر محصول و بررسی میزان استقبال مشتریان که این خود موجب شناسایی سلیقه و درخواست مشتریان بالقوه خواهد شد.

حوزه سیاست: بررسی نظرات کاربران شبکه‌های اجتماعی در خصوص یک پدیده مهم سیاسی مثل انتخابات و پیش‌بینی نتایج از طریق تجزیه و تحلیل این نظرات، که موجب پاسخگویی به نیاز جامعه

و نیز جهت‌دهی مطالبات و افکار عمومی خواهد شد.

حوزه اقتصاد: بررسی شاخص‌های اقتصادی از جمله تحلیل بازار بورس و شناسایی روابط میان وضعیت اقتصادی جامعه با تغییرات شاخص‌های اقتصادی.

حوزه جامعه و روان: تجزیه و تحلیل وضعیت روانی جامعه با استفاده از نتایج بررسی علایق و خلقیات افراد جامعه و شناسایی مؤثرترین شیوه بیان از طریق رسانه‌های جمعی جهت تأثیرگذاری بر اجتماع.

حوزه حقوق و پیشگیری از جرم: شناسایی نظرات مخرب و توهین‌آمیز در شبکه‌های اجتماعی، افزایش میزان اعتماد به مقالات منتشر شده از طریق شناسایی سرقت علمی.

حوزه کتابخانه و نمایه‌سازی: سهولت در دسته‌بندی موضوعی متون و مشابهت‌یابی در بین مستندات، افزایش دقت جست‌وجوی عبارات در میان حجم عظیمی از اطلاعات و منابع اطلاعاتی.

حوزه آموزش: تسریع و تدقیق ویرایش متون، الزام و راهنمایی برای یادگیری زبان های غیربومی.

نتیجه‌گیری

در چند سال اخیر، توجه بسیار زیادی به متن‌کاوی شده است. مجموعه داده‌های متنی، گونه‌ای از داده‌ها هستند که نسبت به انواع دیگر مانند صوت و تصویر از محبوبیت بیشتری برخوردارند. وجود و افزایش چشم‌گیر داده‌های متنی، پژوهشگران را بر آن داشته است تا به دنبال روش‌هایی جهت کاوش این داده‌ها باشند (اسماعیلی؛ زاهد،۱۳۹۴). وقتی حجم زیادی داده متنی وجود دارد، باید برای استخراج اطلاعات ارزشمند تلاش شود. فنون متن‌کاوی برای تحلیل مؤثر اطلاعات مرتبط و مهم از انبوهی از داده‌های بدون ساختار به کار می‌رود(تألیب،۲۰۱۶). متن‌کاوی، که با عناوین کاوش در داده متنی یا کشف دانش در متن، نیز شناخته می‌شود، زمینه تازه و جدیدی است که ما را در استخراج دانش و اطلاعات مفید و با ارزش از متنی که به صورت غیرساخت‌یافته است، یاری می‌دهد (شمسی؛ دیوانی،۱۳۹۵). برای این منظور، روش‌ها و تکنیک‌های بسیاری بکار بسته می‌شوند که در این مجال کوشیدیم تا بخشی از آن را برای درک بیشتر این مفهوم، به تصویر بکشیم. با تمام این توضیحات، این شیوه نیز دارای مزایا و معایبی است.

مزایا این روش عبارتند از:

۱ .از آنجا که پایگاه داده، توانایی ذخیره مقدار اطلاعات کمی را دارد، این مشکل از طریق متن‌کاوی حل شده است.

۲ .با استفاده از روش‌هایی چون استخراج اطلاعات، می‌توان اسامی موجودیت‌های مختلف و رابطه میان آنها را از میان مجموعه اسناد دریافت.

۳ .متن‌کاوی، مشکل مدیریت مقادیر زیاد اطلاعات غیرساخت‌یافته برای استخراج الگوها را به

سادگی حل کرده است، در غیر این‌صورت این مسئله مشکل بزرگی بوده است .

و همچنین معایب آن عبارتند از:

۱ .هیچ برنامه‌ای نمی‌تواند به منظور تحلیل متن غیرساخت‌یافته به صورت مستقیم، کاویدن متن برای استخراج اطلاعات یا دانش تهیه شود.

۲ .اطلاعاتی که در آغاز کار نیاز است، در جایی نوشته نشده است.

از آنجا که در هر پژوهش علمی، بررسی پیشینه تحقیق اهمیت زیادی دارد، لازم است به بررسی نمونه‌هایی از پژوهش‌های صورت گرفته در زمینه متن‌کاوی بپردازیم. رجبپور و طلعتیانآزاد(۱۳۹۴)، در پژوهش خود به بررسی روش‌های متن‌کاوی با استفاده از یادگیری ماشین پرداخته‌اند. پژوهش آنان بر توسعه فناوری‌های یادگیری ماشینی برای کشف دانش حاصل از توصیفات متنی تشخیص خطای یک محیط خاص، تأکید دارد و به‌طور ویژه، بر روی الگوریتم‌های یادگیری ماشینی برای آشکارسازی مدارکی که شامل توصیفات شکست‌های سیستماتیک و ریشه‌هایی که باعث این اشتباهات و خطاها می‌شوند، تحقیق کرده و پیشنهادها و راهکارهایی ارائه می‌دهد.

اسماعیلی و زاهد(۱۳۹۴)، عقیده دارند در متن‌کاوی، دانش مفید از میان اسناد ناساختمند متنی استخراج می‌شود و از آنجا که بدون تردید، رایج‌ترین و محبوب‌ترین شکل داده‌ها، متون هستند، استفاده از شیوه‌های متن‌کاوی در رسانه‌های اجتماعی مانند وبلاگ‌ها و شبکه‌های اجتماعی، اهمیت زیادی پیدا می‌کند.

عظیمی‌همت و شمس‌عزت(۱۳۹۴) نیز با وجود رشد روزافزون حجم اطلاعات متنی، وجود ابزارهایی جهت سازماندهی، بازیابی و استخراج دانش مفید از این داده‌ها را ضروری می‌دانند. گرچه متن‌کاوی به عنوان داده‌کاوی متن، کشف دانش در متن و همچنین تجزیه و تحلیل هوشمند متن شناخته می‌شود، اما این تحقیقات در زمینه متون فارسی، به دلیل پیچیدگی‌های ساختاری آن، کمتر از زبان‌های دیگر انجام شده است. این پژوهش، روش‌های اصلی متن‌کاوی و کاربرد آن در پردازش متون زبان فارسی را مورد بحث و بررسی قرار داده است.

دیوانی و شمسی(۱۳۹۵)، در پژوهش خود به این نکته می‌پردازند که به دلیل افزایش روزافزون مقدار اطلاعات ذخیره شده غیرساخت‌یافته در دنیای امروز، تکنیک‌هایی مانند خلاصه‌سازی، طبقه‌بندی، خوشه‌بندی، استخراج اطلاعات و بصری‌سازی، که زیر مجموعه‌ی متن‌کاوی هستند، جهت استخراج اطلاعات سودمند یا دانش از اسناد متنی، اهمیت زیادی پیدا کرده‌اند.

نظری و حبیبی(۱۳۹۵)، متن‌کاوی را روشی برای استخراج اطلاعاتی که به صورت غیرساخت‌یافته نامرتب و نیمه ساخت‌یافته از این حجم اطلاعات می‌باشد، تعریف کرده‌اند. آنان دسته‌بندی متون را یکی از نمودهای داده‌کاوی متون می‌دانند و تشخیص طبقه، رده یا موضوع یک متن ناشناخته و تخصیص آن به دسته‌ی تشخیص داده شده را دسته‌بندی متون می‌گویند. این پژوهش، با بررسی متن‌کاوی، روش‌هایی که کار کاوش متن را تسریع می‌کند، بررسی کرده و به این نتیجه رسیده است

که این روش‌ها، با توابع ریاضی، کار کاوش متن را انجام می‌دهند.

فاخری(۱۳۹۶)، در پژوهش خود با عنوان «بررسی و تأمل در متن‌کاوی، روش‌های نوین و ابزارها»، متن‌کاوی را عموماً درگیر در فرایند ساختاردهی به ورودی‌های متنی، استخراج الگوهای درون داده‌های ساختاریافته و در نهایت ارزیابی و تفسیر خروجی‌ها می‌داند. این پژوهش سعی در ارائه توضیحات پیرامون متن‌کاوی و نیز روش‌ها، فناوری‌ها و تکنیک‌های مورد استفاده در آن و همچنین معرفی برخی از کاربردها و ابزارهای متن‌کاوی موجود در وب دارد.

در بحث کتاب و منابع چاپی، متاسفانه هنوز اثر قابل توجهی در زمینه متن‌کاوی به زبان فارسی تألیف نشده است و اکثر منابع موجود در این حوزه را آثار ترجمه‌ای تشکیل می‌دهند. در همین راستا، مهم‌ترین و اثرگذارترین کتب لاتین در این حوزه را معرفی می‌کنیم:

نخستین کتاب، کتاب «مبانی پردازش آماری زبان طبیعی[1] »، نوشته هانریش شوتز[2] است که در سال ۱۹۹۹ توسط انتشارات MIT به چاپ رسید. از این کتاب به عنوان یکی از اصلی‌ترین منابع پردازش زبان طبیعی یاد می‌شود. کتاب دارای چهار بخش اصلی است. در بخش اول، به بیان مقدمات، مفاهیم و پیش نیازهای درک پردازش زبان طبیعی پرداخته شده است. بخش دوم، درباره مفاهیم پردازش زبان طبیعی در سطح واژه و ابهام‌زدایی معنایی کلمات است. بخش سوم، به بیان مفاهیم دستور زبان از قبیل مدل مارکوف و تجزیه‌گر آماری می‌پردازد. بخش پایانی نیز، تکنیک‌ها و کاربردهای پردازش زبان طبیعی شامل مترجم‌های ماشینی، خوشه‌بندی، بازیابی اطلاعات و دسته‌بندی را تشریح می‌کند.

کتاب «پردازش گفتار و زبان[3] » نوشته دانیل ژورافسکی و مارتین[4] است که ویرایش اول آن در سال ۲۰۰۲ توسط انتشارات Pearson Education و ویرایش دوم آن در سال ۲۰۰۸، توسط انتشارات Prentice Hall به چاپ رسید. این کتاب دارای ۵ بخش است و به بررسی مفاهیمی چون پردازش زبان طبیعی در سطح واژه، سطح آوایی، سطح نحوی پردازش زبان طبیعی همچون گرامر زبان، سطح معنایی و کاربردگرایی و کاربردهایی برای پردازش زبان طبیعی پرداخته است.

کتاب «متن‌کاوی عملی و تجزیه و تحلیل آماری برای کاربردهای داده متنی غیرساختار یافته [5]» نوشته جان ادلر[6] و دیگران، در سال ۲۰۱۲ توسط انتشارات Elsevier به چاپ رسید. این کتاب به بیان مفاهیم مرتبط با کاربردهای مختلف متن‌کاوی و مطالعات موردی عملی در حوزه‌های مختلف پرداخته است.

1. Foundations of Statistical Natural Language Processing
2. Hinrich Schütze
3. Speech and Language Processing
4.Daniel Jurafsky & James H. Martin
5. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications
6. John Elder

کتاب «استخراج داده‌های متنی[1] » نوشته چارو و چنگ ژیانگ [2] در سال ۲۰۱۲ توسط انتشارات اشپرینگر[3] به چاپ رسید. کتاب دارای ۱۴ فصل است. در فصل اول این کتاب، به بیان مقدمات و مفاهیم متن‌کاوی پرداخته شده و از فصول بعدی، هر فصل به نظرسنجی درباره یکی از زمینه‌ها یا کاربردهای متن‌کاوی اختصاص داده شده است.

کتاب «تجزیه و تحلیل متن با پایتون: یک رویکرد عملی در دنیای واقعی برای به دست آوردن بینش عملی از داده‌های خود[4] » نوشته دیپنجان سرکار[5] ، در سال ۲۰۱۶ توسط انتشارات اَپرس [6] و در قالب ۷ فصل منتشر شد. مفاهیم و لایه‌های پردازش زبان طبیعی، مروری بر برنامه‌نویسی زبان پایتون، پیش‌پردازش و ابزارهای پایه پردازش متن، مفاهیم و شبه کد پایتون مرتبط با دسته‌بندی متون، تکنیک‌ها و ابزارهای استخراج کلیدواژه و خلاصه‌سازی متن، خوشه‌بندی متون و مفاهیم شبکه واژگان، تحلیل معنایی متن و مفاهیم و روش‌های تحلیل احساسات از موضوعاتی است که در این کتاب، بدانها پرداخته شده است.

و در نهایت، کتاب «پردازش زبان طبیعی در عمل: درک، تجزیه و تحلیل و تولید متن با پایتون [7]» نوشته هابسون لین[8] ، در سال ۲۰۱۹ توسط انتشارات منینگ[9] در سه بخش منتشر شد. در بخش اول، مفاهیم و اصول پردازش زبان طبیعی و متن‌کاوی، بخش دوم، مفاهیم یادگیری عمیق و شبکه‌های عصبی و کاربرد آنها در پردازش متن و بخش سوم، چالش‌ها و مسائل دنیای واقعی حوزه پردازش زبان طبیعی را بررسی می‌کند.

بررسی پیشینه پژوهش‌ها حاکی از آن است که بیشتر موضوعاتی که در پژوهش‌های صورت گرفته در این حوزه بدان‌ها پرداخته شده است عبارتنداز بررسی مفاهیم، روش‌ها و تکنیک‌های متن‌کاوی و ابزارهای مورداستفاده در داده کاوی و متن‌کاوی.

در پایان لازم به تأکید است گرچه متن‌کاوی به عنوان یک زمینه در حال رشد و پرکاربرد، به دنبال کشف دانش از متون غیرساخت‌یافته است، ولی به دلیل مشکلات ساختاری زبان فارسی، تحقیقات کمتری در این زمینه صورت گرفته است. اکثر سیستم‌های متن‌کاوی، برای متون زبان انگلیسی

1.Mining Text Data

2.Charu C. Aggarwal & ChengXiang Zhai

3.Springer

4.Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from your Data

5.Dipanjan Sarkar

6.Apress

7. Natural Language Processing in Action: Understanding, analyzing, and generating text with Python

8.Hobson Lane

9.Manning Publications

طراحی شده‌اند و قابل استفاده برای متون فارسی نیستند. توسعه این سیستم‌ها جهت کاربرد برای متون فارسی، به دلیل ماهیت زبان فارسی و در دسترس نبودن مجموعه‌ای شامل ریشه کلمات و کلمات پرکاربرد، کاری نسبتاً دشوار است و نیاز به پژوهش و همتی مضاعف در این زمینه دارد (عظیمی‌همت؛شمس عزت،۱۳۹۴).

منابع:

-اسماعیلی، مهدی؛ زاهد، عطیه(۱۳۹۴). مروری بر متن‌کاوی؛ مفاهیم؛ تکنیک‌ها و چالش‌ها، سومین کنفرانس بین‌المللی پژوهش‌های کاربردی در مهندسی کامپیوتر و فنآوری اطلاعات، تهران.

-رجب‌پور، نازنین؛ طلعتیان آزاد، سعید (۱۳۹۴). بررسی روش‌های متن‌کاوی با استفاده از یادگیری ماشین. کنفرانس بین‌المللی پژوهشهای کاربردی در فناوری اطلاعات، کامپیوتر ومخابرات، تربت حیدریه.

-شمسی، محبوبه؛ دیوانی، مرضیه (۱۳۹۵). مروری بر متن‌کاوی و روش‌های آن. سومین همایش ملی مهندسی رایانه و مدیریت فناوری اطلاعات، تهران.

-عظیمی همت، منیره؛ شمس عزت، فاطمه (۱۳۹۴). مروری بر متن‌کاوی متون فارسی، دومین کنفرانس بین‌المللی و سومین همایش ملی کاربرد فناوری‌های نوین در علوم مهندسی، مشهد.

-فاخری، سهیل (۱۳۹۶). بررسی و تأمل در متن‌کاوی، روش‌های نوین و ابزارها. سومین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و برق، رودسر.

-نظری، مهدی و حبیبی، مریم (۱۳۹۵). بررسی روش‌های LDA و LSA و PLSA در متن‌کاوی. چهارمین کنفرانس بین‌المللی مهندسی برق و کامپیوتر، تهران.

-Gaikwad ,S .V ,.Chaugule ,A & ,.Patil ,P .(2014) .Text mining methods and techniques .International Journal of Computer Applications.(17)85 ,

-Talib ,R ,.Hanif ,M .K ,.Ayesha ,S & ,.Fatima ,F .(2016) .Text mining :techniques ,applications and issues .International Journal of Advanced Computer Science and Applications.414-418 ,(11)7 ,

Investigating the Validity of the Model of Growth and Development of Social Capital in Government Organizations

Mohammad Ali Javanpour[1],**Farhad Emam Jomeh**[2],**Abdolrahim Rahimi**[3]

Abstract

Human capital includes competencies, knowledge, social and personality traits including creativity, embodiment of the ability to do It is a job to produce economic value. Social capital can bring benefits in terms of internal and functional order and cohesion The system brought in order to achieve the desired goals and benefits for organizations. So it deserves attention and development The development of social capital requires attention to the requirements and proper bedding in this field. This research examines the validity of the model deals with the growth and development of social capital in government organizations. The present study is an applied descriptive research.

The statistical population includes all experts and staff of the Martyrs and Veterans Affairs Foundation of Tehran. The research was informed and their volume was 1090 people and based on Cochran's formula, the sample size was 482 people which was determined by stratified and Simple random selections were selected. Analysis of information using descriptive and inferential statistics and with the help of SPSS23 software and PLS took place. To check the validity of the model, the tests of confirmatory factor analysis, structural equations and path analysis and others Relevant statistical tests were used. In the discussion of validation, the research results indicate a proper fit of the conceptual model, based on Indices of acceptable and significant values of path coefficients, factor loads, explained variance, cross-validity of redundancy index was. The results showed that the growth and development model of the foundation's social capital includes: the nature of the foundation (organizational strategies, approaches Personality and cultural approach); factor of social participation in the foundation; (social activities and political activities); factor of networks Current social in the foundation (social relations, social engagement); cause of social harm in the foundation social norms and Social un trust factor of social awareness (Social knowledge and development of social values) factor of social responsibility (Cohesion Social and social trust.)The model is also well-structured based on fit indicators.

Keywords: Validation, Growth and Development Model of Social Capital.

---

1. PhD Student in Cultural Management and Planning, Islamic Azad University, South Tehran Branch

2. Assistant Professor, Islamic Azad University, Arak Branchf. Correspondent Author: emamjome@yahoo.com

3. Assistant Professor, Islamic Azad University, South Tehran Branch