نگرش آماری متن‌کاوی

مهدیه بیاتی [۱]

تاریخ دریافت: ۱۴۰۳/۰۹/۰۶
تاریخ پذیرش: ۱۴۰۳/۱۰/۲۴

چکیده:
ا در عصر اطلاعات زندگی می‌کنیم و همواره در حال درک و دریافت داده‌های زیادی از دنیای اطراف خود هستیم که برای استفاده از این اطلاعات لازم است آن‌ها را به کمک آمار و به‌صورت ریاضی بیان کنیم. آمار در همه‌ی زمینه‌ها نقش مؤثری ایفا می‌کند. یکی از مواردی که جدیداً موردتوجه قرارگرفته و از فنون آماری کمک می‌گیرد، متن‌کاوی است. متن‌کاوی یک روش تحقیقی برای شناسایی الگوهای موجود در متون است که می‌تواند نوشتاری، گفتاری و یا تصویری باشد. متن‌کاوی کاربردهای گسترده‌ای مانند رده‌بندی، خوشه‌بندی، وب‌کاوی، تحلیل احساسات و ... دارد. متن‌کاوی بسیار گسترده است همانند طبقه‌بندی متون، خوشه‌بندی متون، وب‌کاوی و عقیده کاوی و ... . از تکنیک‌های متن‌کاوی برای تخصیص مقادیر عددی به داده‌های متنی استفاده می‌شود تا امکان تجزیه‌وتحلیل آماری فراهم شود. ازآنجاکه کار با داده‌ها مستلزم یک پایه محکم در آمار است، از ابزارهای آماری در تجزیه‌وتحلیل متن برای پیش‌بینی‌هایی مانند پیش‌بینی تغییرات در قیمت سهام یا نرخ ارز بر اساس داده‌های متنی فعلی استفاده می‌شود. به‌کارگیری روش‌های آماری می‌تواند حقایق موجود در متن را کشف، تائید و یا رد کند. امروزه این مبحث در یادگیری ماشین بسیار پرکاربرد است. در این مقاله سعی کردیم تا آشنایی ابتدایی با ابزارهای آماری در روش متن‌کاوی داشته باشیم و از این ابزار قدرتمند برای تحلیل وقایع استفاده کنیم.

واژه‌های کلیدی: کدگذاری، آمار توصیفی، سری زمانی، رگرسیون، استنباط بیزی.

۱ مقدمه

ازنقطه‌نظر علمی آمار به مجموعه‌ای از روش‌ها برای جمع‌آوری، تنظیم و خلاصه کردن اطلاعات عددی و غیر عددی و انجام استنباط و نتیجه‌گیری به‌وسیله‌ی تجزیه‌وتحلیل آن‌ها اطلاق می‌شود. لذا در یک بررسی آماری با دو بخش آمار توصیفی و آمار استنباطی مواجه هستیم. در آمار توصیفی با تهیه و تنظیم و خلاصه کردن اطلاعات عددی و غیر عددی سروکار داریم و حال‌آنکه تجزیه‌وتحلیل و نتیجه‌گیری آماری به آمار استنباطی برمی‌گردد. (نعمت الهی[۷])

در آمار توصیفی خلاصه‌سازی داده‌ها با تکنیک‌ها یا ابزارهایی همانند جداول آماری، نمودارها و شاخص‌ها انجام می‌شود. برای به‌کارگیری این ابزارها داده‌ها به دو دسته‌ی کمی و کیفی طبقه‌بندی می‌شود. داده‌های کمی داده‌هایی هستند که کمیت‌پذیرند یعنی به‌صورت عددی گزارش می‌شوند همانند وزن محصولات تولیدی، سود شرکت، مقاومت شکست بطری، تعداد آرا به نماینده‌ی خاص و غیره و داده‌های کیفی یا رسته‌ای داده‌هایی هستند که نمی‌توان آن‌ها را به‌صورت عددی گزارش کرد همانند کیفیت محصول، رنگ پوست، سطح علوم دانش‌آموز سوم دبستان، گروه خونی و غیره؛ اما داده‌های کیفی می‌توانند فرمت‌های دیگری مانند فیلم، عکس، صوت، متن و غیره داشته باشند. تمرکز متن‌کاوی بیشتر بر روی این گروه از داده‌های کیفی است.

پژوهش‌ها یا بر مبنای داده‌های کمی هستند یا داده‌های کیفی. از دیدگاه مایکوت و مورهاس [۲۱] ، چهار تفاوت بنیادین بین پژوهش‌ها بر مبنای داده‌های کیفی و پژوهش‌های مبتنی بر داده‌های کمی وجود دارد:

۱- کلمات (مفاهیم) در برابر اعداد
۲- دیدگاه بینش مدار در مقابل دیدگاه عینیت(حقیقت) مدار
۳- کشف در برابر اثبات
۴- انسان‌محوری در برابر ابزار محوری

سؤالی که مطرح می‌شود این است که چطور می‌توان یک متن مصاحبه یا صوت خبرنگاری را توصیف آماری کرد؟ چطور می‌توان این قبیل

[۱] عضو هیئت‌علمی گروه آمار دانشگاه قم (نویسنده مسئول: bayati_2006@hotmail.com)

داده‌ها را با ابزارهای آماری گزارش داد؟ برای پاسخ به این سؤالات روش‌های معروف به روش اسنادی ارائه‌شده است. متمرکز بودن علوم انسانی و علوم اجتماعی بر دیدگاه‌ها و اندیشه‌ها موجب محدود شدن آن می‌گردد. همزمان با رشد تکنولوژی روش‌های تحقیق هم رشد یافته و روش‌های نوینی ارائه‌شده است. هدف از روش‌های اسنادی پژوهش، مطالعه و تحلیل بر روی متون است. یکی از این تکنیک‌ها متن‌کاوی است.

تحلیل متون بر روی کتاب مقدس مسیحیان آغاز شد و سپس در تحلیل روزنامه‌ها، شیوه‌هاي مربوط به خط‌شناسی و حتی تعبیر خواب توسط زیگموند فروید[۱۱] ادامه یافت. تقریباً نخستین تحلیل بر اساس مطالعه‌ی تجربی روزنامه‌ها در سال ۱۸۹۳ درباره‌ی روزنامه‌هاي منتشرشده در نیویورک بود که در پی پاسخ به این پرسش برآمد: آیا روزنامه‌هاي امروز مطالب ارزشمند ارائه می‌دهند؟ این تحلیل نشان داد که از ۱۸۸۱ تا ۱۸۹۳، روزنامه‌هاي نیویورک افترا، شایعات و ورزش را تا حد بسیاري جایگزین مطالب مذهبی، علمی و ادبی کرده‌اند (ایمان و نوشادی[۳]). شاخص‌های گوناگون علوم اجتماعی همچون ارتباطات، جامعه‌شناسی، علوم سیاسی و روان‌شناسی آن را در پژوهش‌هاي خود به کار گرفته‌اند (پول و فوگر [۲۳]). این روند به‌ویژه در طول جنگ جهانی دوم در دهه‌ی ۱۹۴۰ رواج بیشتري یافت. در ابتدا این روش را برای تحلیل تبلیغات و سپس برای اهداف اطلاعاتی و نظامی به کار گرفتند. دیدگاه‌ها و جهت‌هاي نوینی براي به‌کارگیری این تکنیک، پس از جنگ جهانی دوم توسط هارولد لاسول و همکارانش [۱۵]با انتشار کتاب زبان سیاست معرفی گردید. برلسون [۸] کتاب «متن‌کاوی در تحقیقات مربوط به ارتباطات» را منتشر کرد. در دهۀ ۱۹۶۰ برنامه‌هاي کامپیوتري جدیدی که داراي سیستم فرهنگ واژگان مشخص بودند به شمارش واژگان در متن می‌پرداخت و بر اساس آن، کلمه‌هاي متن با توجه به فرهنگ واژگان رمزگذاري می‌شد. اما این روش به دلیل در نظر نگرفتن محتوای پنهان متن موردانتقاد قرار گرفت. لذا رویکرد کیفی در متن‌کاوی موردتوجه قرارگرفت. در حقیقت متن‌کاوی در رابطه با متنی معنی پیدا می‌کند که آن متن برای انتقال پیام‌ها و مفاهیم معینی نوشته‌شده باشد و دارای ماهیتی مشخص باشد. لذا تحلیل در مورد کلمات عامیانه‌ای که دارای مفاهیم ساده، بدیهی و مشخصی هستند به کار نمی‌رود. متن‌کاوی به‌منزلۀ تکنیکی علمی عمدتاً در قرن بیستم رایج شد. امروزه به‌طور گسترده در علوم مختلف از جمله علوم اجتماعی، جامعه‌شناسی و مدیریت استفاده می‌شود. به‌عنوان‌مثال رانه [۲۵] ابعاد جدید استراتژی‌های بازاریابی در شرایط رقابتی را با به‌کارگیری تکنیک تحلیل محتوا معرفی کرد. او ابتدا با ۱۵ نفر از خبرگان بازاریابی در کشور سوئد مصاحبه کرد و با تکنیک تحلیل محتوا موارد ارزشمند را استخراج کرد و بر اساس آن پرسشنامه‌هایی را طراحی و در اختیار مدیران بازاریابی قرار داد و عوامل ابعاد مختلف استراتژي‌هاي بازاریابی در شرایط رقابتی را شناسایی کرد. منگاراسینا و همکاران [۱۹] عوامل مختلف تدوین استراتژی توزیع را شناسایی و ارائه کردند. آن‌ها با بررسی مقالات معتبر از سال ۱۹۷۲ تا ۲۰۱۳ با تکنیک تحلیل محتوا تمامی متغیرهای حوزه توزیع را طبقه‌بندی کردند.

آمار و احتمال به‌طور گسترده و باقدرت در متن‌کاوی به کار می‌رود. نظریه احتمال زیربنای بسیاری از مدل‌های زبانی است که در متن‌کاوی استفاده می‌شود. برای مثال مدل n-gram از احتمال برای پیش‌بینی احتمال توالی کلمات استفاده می‌کند که برای کارهای مانند تولید زبان یا تصحیح املا مفید است. کلود شانون [؟] ایده استفاده از مدل‌های احتمال را معرفی کرد و سپس محققانی همانند فردریک جلینک [۱۸] برای تشخیص گفتار و تولید خودکار متن از n-gram استفاده کردند و روزبه‌روز این کاربرد وسیع‌تر و قدرتمندتر از قبل می‌شود. اخیراً لیو و همکاران [۱۶] Infini-gram را پیشنهاد دادند. تحلیل رگرسیون نیز یکی دیگر از ابزارهای آماری است که در متن‌کاوی از آن استفاده می‌شود. درگا و همکاران [۱۰] تکنیک‌های رگرسیون را در کنار سایر مدل‌های یادگیری ماشین برای پیش‌بینی روند رسانه‌های اجتماعی بر اساس تحلیل احساسات استفاده کردند. تجزیه‌وتحلیل سری‌های زمانی ابزار دیگر آمار در متن‌کاوی است که شامل بررسی داده‌های متنی در یک توالی زمانی برای شناسایی الگوها، روندها و تغییرات است. درزمینه پست‌های رسانه‌های اجتماعی یا مقالات خبری، این رویکرد می‌تواند بینش‌های ارزشمندی را آشکار کند. نتولیکی و همکاران [۲۲] سری‌های زمانی در متن‌کاوی را برای پیش‌بینی حرکت قیمت سهام بررسی کردند. این تحقیق با موفقیت الگوهایی را شناسایی می‌کند که داده‌های متنی را به نوسانات قیمت سهام مرتبط می‌کند و پتانسیل تجزیه‌وتحلیل متن سری‌های زمانی را برای پیش‌بینی مالی نشان می‌دهد. برای بهتر روشن شدن نقش آمار در متن‌کاوی ابتدا به مفهوم متن‌کاوی می‌پردازیم و سپس با بیان مثال‌هایی نقش آمار را در متن‌کاوی بیان خواهیم کرد.

۲ متن‌کاوی

در حقیقت متن‌کاوی تکنیکی پژوهشی برای استنباط تکرارپذیر و معتبر از داده‌ها در متن می‌باشد. یکی از تعاریف اولیه و رایج متن‌کاوی از برلسون [۸] است. به اعتقاد او در متن‌کاوی، ویژگی‌هاي ظاهري یک پیام (متن، مکالمه و...) را به شکل عینی (مستقل از برداشت شخصی

متن‌کاوی کمی (بعد آماری)	متن‌کاوی کیفی
جانشین جنبه‌های ذهنی و استنباطی بررسی‌ها و تحلیل‌های نظری	روش تحقیقی برای تفسیر ذهنی محتوایی داده‌های متنی
یافتن تعداد تکرارهای کلمات، مضامین و نمادهای ارزشمند در متن	یکی از ویژگی‌های بنیادی نظریه‌پردازی به‌جای آزمون نظریه
به‌کارگیری تحلیل همبستگی برای بررسی ارتباط مفاهیم و عناصر	آشکارسازی الگوهای پنهان در متون

جدول ۱: خلاصه‌ای از متن‌کاوی کمی و کیفی

محقق) و نظام‌مند (بر طبق قواعد معین و کمی بر اساس شاخه‌هاي آماري) توصیف می‌کنند. این تعریف برلسون تعریفی کاملاً کمی از متن‌کاوی است. از نظر کریپندورف [۱۴] متن‌کاوی فقط ابزار است و بس. کاپلان با یک نگاه آماری به تحلیل‌های سیاسی می‌گوید: «روش متن‌کاوی، معناشناسی آماری مباحث سیاسی است». متن‌کاوی به دو روش متن‌کاوی کمی و کیفی تقسیم‌بندی می‌شود.

تحلیل کمی، متن‌کاوی را از مطالعه‌ی معمولی متن جدا می‌کند و جایگزین روش‌های ذهنی می‌شود که با مقیاس دقیق قابل ارزیابی نیست. بر اساس این ویژگی تحلیل‌گر معلوم می‌کند که چه کلمات، مضامین و نمادهایی بیش از همه و به چه تعداد در متن تکرار شده است. حتی می‌توانیم از تحلیل همبستگی استفاده کنیم و با کمک آن ارتباط مفاهیم و عناصر را مشخص کنیم. آیا مفاهیم همدیگر را جذب می‌کنند یا دفع؟ آیا مفاهیم ارتباط لازم را دارند؟ در واقع روش تجزیه‌وتحلیل متن در متن‌کاوی کمی می‌تواند توصیفی یا توصیفی- تحلیلی باشد. بر اساس نوع واحد تحلیل، باید فراوانی موضوع موردبررسی در مقوله‌بندی‌ها تعیین و شمارش شود و بر اساس آن نتیجه‌گیري به عمل آید. در اینجا محقق تنها به تحلیل فراوانی می‌پردازد؛ درحالی‌که در توصیفی تحلیلی نه‌تنها تحلیل فراوانی (شامل شدت، وسعت و اهمیت عناصر موردنظر در محتوا) که تحلیل همبستگی عناصر و مفاهیم جهت انسجام و هم‌بستگی و یا گسستگی بررسی می‌شود و نتیجه‌گیري نهایی صورت می‌پذیرد.

تقلیل متن به اعداد در تکنیک کمی موجب از دست رفتن یک سری اطلاعات ترکیبی یا نهانی می‌شود لذا این دیدگاه موردنقد قرار می‌گیرد. بنابراین متن‌کاوی کیفی را می‌توان روش تحقیقی برای تفسیر ذهنی محتوایی داده‌های متنی دانست. همچنین یکی از ویژگی‌های بنیادین پژوهش‌های کیفی نظریه‌پردازی به‌جای آزمون نظریه‌ها هست. متن‌کاوی کیفی به محققان اجازه می‌دهد اصالت و حقیقت داده‌ها را به‌گونه‌ای ذهنی ولی با روش عملی تفسیر کنند. متن‌کاوی کیفی حتی الگوهایی که در متون پنهان هستند را آشکار می‌سازد. جدول ۱ خلاصه‌ای از دو روش کمی و کیفی متن‌کاوی را نشان می‌دهد.

حال می‌توانیم در تحلیل متون از هر دو شیوه کمک بگیریم. به این صورت که با استفاده از تحلیل کیفی کلمات، جملات یا عبارات مهم و معنی‌دار را انتخاب کنیم و با تحلیل کمی آن‌ها را توصیف کنیم. برای این کار به کدگذاری نیازمندیم.

۱.۲ کدگذاری

کد معمولاً کلمه یا عباراتی کوتاه است که به شکل نمادین حاکی از ویژگی برجسته و فشرده و دربرگیرنده‌ی ذات یک‌چیز و یادآور بخشی از یافته‌های زبان بنیاد یا دیداری است (سالدنیا [۲۷]). با هدف فهم بهتر مطلب مثالی را در ادامه بیان می‌کنیم.

مثال ۱.۲. *متوجه شدم که در مقابل اکثر خانه‌ها زنجیره‌ای از پرچین تعبیه‌شده است. تعداد زیادی سگ در آنجا بودند و نشانه‌هایی روی پرچین‌ها بود که می‌گفت مراقب سگ باشید. به نظر می‌رسد که واژه‌ی «امنیت» نکته‌ی برجسته‌ی این متن می‌باشد.*

مثال ۲.۲. *متن زیر مصاحبه‌ی دانش‌آموز سال آخر دبیرستان در توصیف معلم محبوبش است.*

او مراقب من است، این را هرگز به من نگفته است بلکه در عمل چنین می‌کند. *"احساس ارزشمندی"*

و همیشه برای کمک به من حاضر است."ثبات"

من واقعاً وقتی در کنار او هستم احساس راحتی می‌کنم. *"راحتی"*

«احساس ارزشمندی»، «ثبات» و «راحتی» کدهای مناسب برای این مثال است. آیا شما نیز با این کدها موافقید؟ عبارات دیگری در ذهن شما مجسم نشد؟ اشکالی ندارد اگر انتخاب شما با این انتخاب‌ها متفاوت باشد. کدگذاری عملی تفسیری است. این‌ها دال بر ادراک اولیه شما می‌باشد. قابلیت کدگذاری برای تحلیل داده‌های کیفی با برنامه‌های کامپیوتری امکان‌پذیر است؛ که تعداد آن روزبه‌روز افزایش می‌یابد. مشهورترین نرم‌افزار تحلیل کیفی عبارت‌اند از: مکس کیو دا[۲]، ان وی وو[۳]، اطلس تی آی[۴] و کیو دی ای مانیر[۵]. در این میان مکس کیودا به علت پشتیبانی از زبان‌های مختلف و سازگاری کامل با زبان فارسی و همچنین مزیت‌های دیگر از محبوبیت و کارایی بیشتری برخوردار است. برای انجام کدگذاری ابتدا باید مسئله‌ی پژوهش را تعریف کنیم. طرح نمونه‌گیری را مشخص کنیم. اینجا اولین جایی است که آمار با نقش نمونه‌گیری وارد حل مسئله‌ی متن‌کاوی می‌شود. برای انجام نمونه‌گیری ابتدا باید فهرست تمام اجزای اسنادی که قرار است نتایج را به آن‌ها تعمیم دهیم تهیه می‌شود. هنگامی‌که جامعه مشخص شد بر اساس طرح نمونه‌گیری، نمونه‌ای از اسناد انتخاب می‌شود. حال که نمونه داریم برای آن عمل کدگذاری انجام می‌دهیم. کدگذاری در سه مرحله انجام می‌شود شامل کدباز، کدگذاری محوری و کدگذاری انتخابی. در کدگذاری باز همان‌طور که از اسمش مشخص است همه داده‌های خام کدگذاری می‌شود و کدهای مشابه در داخل یک مقوله قرار می‌گیرد. در کدگذاری محوری مقوله‌ها دوباره موردبررسی قرار می‌گیرند و در واقع برای کدهای باز یک محور مشخص می‌شود و بدین ترتیب کدهای باز به دسته‌های مشخص تقسیم‌بندی می‌شود. کدگذاری انتخابی، غربالگری نهایی را انجام می‌دهد. اطلاعات نامناسب کنار گذاشته می‌شود و مقوله‌های اندکی باقی می‌مانند. در این مرحله نظریه به رشته‌ی تحریر در می‌آید.

کدگذاری دومین جایگاهی هست که آمار برای حل مسئله کمک می‌کند.

۳ آمار در متن‌کاوی

تن کاوی را بیش‌تر با رویکرد کمی آن می‌شناسند. متن‌کاوی بر کمی کردن متن موردبررسی استوار است. در این روش تجزیه‌وتحلیل داده‌ها بر پایه آمار، ارقام، فراوانی و درصدها انجام می‌شود. این کمیت‌ها در یک پژوهش، عنصري ارزشمند به شمار می‌روند. زمانی که کدگذاری خاتمه یافت، برای درک بهتر متن از ابزارهای آماری برای نمایش مفاهیم متن و تجزیه‌وتحلیل آن می‌پردازند.

۱.۳ آمار توصیفی

آمار توصیفی ابزارهای متفاوتی در متن‌کاوی دارد و اغلب شامل محاسبه معیارهایی مانند فراوانی جمله- فراوانی معکوس متن $(TF-IDF)$ [۶] برای تعیین کمیت اهمیت کلمات در مجموعه اسناد است. این معیار به شناسایی اصطلاحات و مضامین کلیدی در متون کمک می‌کند. فراوانی جمله-فراوانی معکوس متن دو مؤلفه فراوانی جمله (TF) و فراوانی معکوس متن (IDF) را ترکیب می‌کند که برای ارزیابی اهمیت یک کلمه در یک سند نسبت به مجموعه‌ای از اسناد استفاده می‌شود. اصطلاح فراوانی به تعداد دفعات ظاهر شدن یک کلمه در یک متن اشاره دارد و بر این فرض استوار است که کلماتی که بیشتر در یک متن اتفاق می‌افتد مهم‌تر یا مرتبط با محتوای آن متن هستند. از سوی دیگر، فراوانی معکوس متن، میزان رایج یا کمیاب بودن یک کلمه را در کل متون اندازه‌گیری می‌کند. کلماتی که در بسیاری از متون ظاهر می‌شوند کمتر آموزنده در نظر گرفته می‌شوند و نمرهIDF کمتری دریافت می‌کنند، درحالی‌که کلماتی که در متون کمتری ظاهر می‌گردند متمایزتر تلقی می‌شوند و نمرهIDF بالاتری دریافت می‌کنند. با ضرب این دو عامل $(TF\times IDF)$،$TF-IDF$ به هر کلمه در یک سند وزنی اختصاص می‌دهد. این وزن متناسب با تعداد افزایش می‌یابد. برای مثال متن زیر را در نظر بگیرید گربه روی تشک نشست. سگ گربه را تعقیب کرد. پرنده‌ی روی تشک پرواز کرد. این متن شامل سه جمله است. مقدار$TF-IDF$ را برای کلمه «گربه» به این صورت به دست می‌آید. جمله اول شامل ۴ کلمه است که یکی از آن‌ها گربه است پس مقدار $TF=\frac{1}{4}$، از بین سه جمله در دو جمله کلمه گربه دیده‌شده است

[2] Maxqda
[3] Nvivo
[4] Atlasti
[5] Qdaminner
[6] Term Frequency-Inverse Document Frequency

پس مقدار $IDF = \log(\frac{۳}{۲})$ است. در نتیجه

$$TF - IDF(\text{"}\textbf{گربه}\text{"}) = TF * IDF = ۰٫۲۵ \times ۰٫۱۷۶ = ۰٫۰۴۴$$

کاربرد دیگر آمار توصیفی پیدا کردن محور موضوعی متن است. به‌عنوان‌مثال فرض کنید می‌خواهند محور موضوعی سوره یونس را بیابند. ابتدا آیات این سوره را کدگذاری کرده سپس فراوانی کلمات (شمارش کلیدواژه‌های باز) را در شکل (۱) نشان دادند. همان‌طور که در نمودار فراوانی کلمات مشاهده می‌شود، می‌توان گفت در این سوره تأکید زیادی بر توحید و ربوبیت الهی صورت گرفته است (الماسی قمی و همکاران، [۱]). مفاهیم دیگری همانند اعتبار و پایایی در آمار وجود دارد که در متن‌کاوی نیز برای نشان دادن اعتبار و پایایی نتایج به‌دست‌آمده استفاده می‌شود. وقتی صحبت از اعتبار می‌شود یعنی روش یا ابزار به‌کاررفته تا چه حد می‌تواند خصوصیت موردنظر را درست اندازه‌گیری کند. در یک مطالعه‌ی کیفی، اعتبار اشاره دارد به میزانی که محقق توانسته است پدیده‌های موردمطالعه یا متغیرهای مربوط به آن را انعکاس دهد. اعتبار در کدگذاری توسط اساتید تعیین می‌شود.

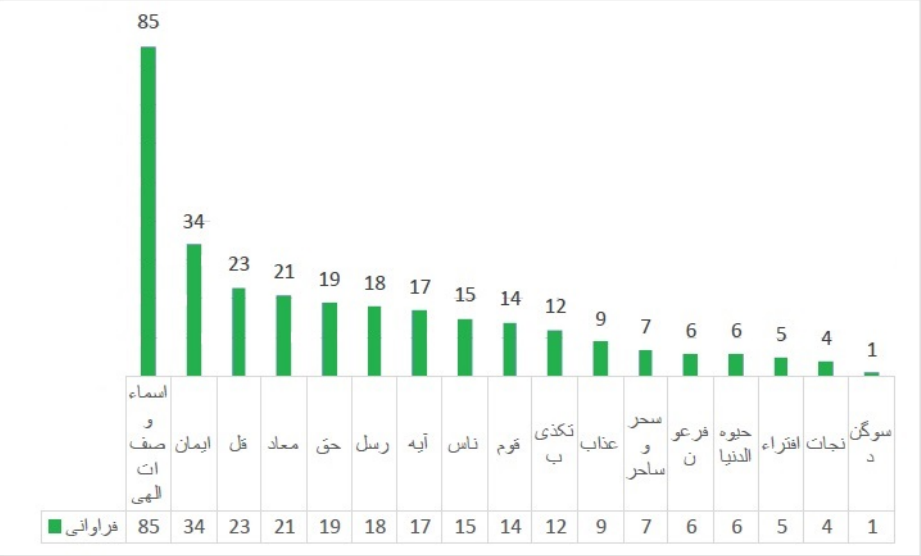
ابزار گردآوری داده‌ها باید ویژگی پایایی داشته باشد. به این معنا که اگر در چند زمان مختلف در یک جمعیت از آن استفاده کنیم در نتیجه‌ی به‌دست‌آمده اختلاف چندانی مشاهده نمی‌کنیم. پایایی بین کدگذاران واژه‌ای پراستفاده است که به میزان توافقی است که کدگذاران مستقل هنگام ارزیابی ویژگی‌های یک پیام یا متن به دست می‌دهند. پایایی در بین کدگذاران در متن‌کاوی موردنیاز است؛ زیرا کمیت مشابهی را که قضاوت‌های متفاوت به هر پدیده می‌دهند اندازه‌گیری می‌کند. در کدگذاری دو نوع پایایی وجود دارد:

۱- پایایی یک کدگذار (شاخص ثبات): شاخص ثبات به میزان سازگاری طبقه‌بندی داده‌ها در طول زمان اشاره دارد. این شاخص را می‌توان زمانی محاسبه کرد که یک کدگذار، یک متن را در دو زمان متفاوت کدگذاری کرده باشد. به دلیل این‌که در محاسبه‌ی این شاخص کدگذار و متن یکسان است، این نوع پایایی شامل کمترین احتمال دخالت عوامل کنترل نشده است (بوون و بوون [۹]). برای محاسبه‌ی پایایی، روش کار به‌این‌ترتیب است که از میان کل متون، چند نمونه، به‌صورت تصادفی انتخاب و هرکدام از آن‌ها، دو بار، در یک‌فاصله‌ی زمانی کوتاه و مشخص (بین پنج تا سی روز) کدگذاری می‌شوند. سپس کدهای مشخص‌شده، در دو فاصله‌ی زمانی، برای هرکدام از مصاحبه‌ها با یکدیگر مقایسه می‌شوند و از طریق میزان توافقات و عدم توافقات موجود، در دو مرحله کدگذاری، شاخص ثبات برای آن تحقیق محاسبه می‌گردد. در هرکدام از مصاحبه‌ها، کدهایی که در دو فاصله‌ی زمانی با هم مشابه هستند، با عنوان «توافق» و کدهای غیرمشابه با عنوان «عدم توافق» مشخص می‌شوند.

۲- پایایی بین دو کدگذار (شاخص تکرارپذیری): شاخص ثبات، سازگاری درک یا تفسیر یک فرد را در مورد یک متن خاص، در طی زمان اندازه می‌گیرد؛ درحالی‌که پایایی بین کدگذاران میزان سازگاری درک یا معنای مشترک متن را اندازه می‌گیرد. پایایی بین کدگذاران (تکرارپذیری) به درجه‌ای اشاره دارد که دو یا چند کدگذار نتایج یکدیگر را تکرار می‌کنند. فرایند کدگذاری، درصورتی‌که کدگذاران یک متن را به یک شیوه کدگذاری کنند، تکرارپذیر خوانده می‌شود. به این منظور از یک کدگذار و یک همکار پژوهش استفاده می‌کند. آموزش‌ها و تکنیک‌های لازم و استانداردشده برای کدگذاری متون به همکار پژوهش انتقال داده می‌شود. سپس محقق همراه این همکار پژوهش چند متن را به‌صورت تصادفی، انتخاب و کدگذاری می‌کنند.

مفاهیم دیگری همانند اعتبار و پایایی در آمار وجود دارد که در متن‌کاوی نیز برای نشان دادن اعتبار و پایایی نتایج به‌دست‌آمده استفاده می‌شود. وقتی صحبت از اعتبار می‌شود یعنی روش یا ابزار به‌کاررفته تا چه حد می‌تواند خصوصیت موردنظر را درست اندازه‌گیری کند. در یک مطالعه‌ی کیفی، اعتبار اشاره دارد به میزانی که محقق توانسته است پدیده‌های موردمطالعه یا متغیرهای مربوط به آن را انعکاس دهد. اعتبار در کدگذاری توسط اساتید تعیین می‌شود.

ابزار گردآوری داده‌ها باید ویژگی پایایی داشته باشد. به این معنا که اگر در چند زمان مختلف در یک جمعیت از آن استفاده کنیم در نتیجه‌ی به‌دست‌آمده اختلاف چندانی مشاهده نمی‌کنیم. پایایی بین کدگذاران واژه‌ای پراستفاده است که به میزان توافقی است که کدگذاران مستقل هنگام ارزیابی ویژگی‌های یک پیام یا متن به دست می‌دهند. پایایی در بین کدگذاران در متن‌کاوی موردنیاز است؛ زیرا کمیت مشابهی را که قضاوت‌های متفاوت به هر پدیده می‌دهند اندازه‌گیری می‌کند. در کدگذاری دو نوع پایایی وجود دارد:



شکل ۱: فراوانی کلمات در سوره یونس

۱- پایایی یک کدگذار (شاخص ثبات): شاخص ثبات به میزان سازگاری طبقه‌بندی داده‌ها در طول زمان اشاره دارد. این شاخص را می‌توان زمانی محاسبه کرد که یک کدگذار، یک متن را در دو زمان متفاوت کدگذاری کرده باشد. به دلیل این‌که در محاسبه‌ی این شاخص کدگذار و متن یکسان است، این نوع پایایی شامل کمترین احتمال دخالت عوامل کنترل نشده است (بوون و بوون [۹]). برای محاسبه‌ی پایایی، روش کار به‌این‌ترتیب است که از میان کل متون، چند نمونه، به‌صورت تصادفی انتخاب و هرکدام از آن‌ها، دو بار، در یک‌فاصله‌ی زمانی کوتاه و مشخص (بین پنج تا سی روز) کدگذاری می‌شوند. سپس کدهای مشخص‌شده، در دو فاصله‌ی زمانی، برای هرکدام از مصاحبه‌ها با یکدیگر مقایسه می‌شوند و از طریق میزان توافقات و عدم توافقات موجود، در دو مرحله کدگذاری، شاخص ثبات برای آن تحقیق محاسبه می‌گردد. در هرکدام از مصاحبه‌ها، کدهایی که در دو فاصله‌ی زمانی با هم مشابه هستند، با عنوان «توافق» و کدهای غیرمشابه با عنوان «عدم توافق» مشخص می‌شوند.

۲- پایایی بین دو کدگذار (شاخص تکرارپذیری): شاخص ثبات، سازگاری درک یا تفسیر یک فرد را در مورد یک متن خاص، در طی زمان اندازه می‌گیرد؛ درحالی‌که پایایی بین کدگذاران میزان سازگاری درک یا معنای مشترک متن را اندازه می‌گیرد. پایایی بین کدگذاران (تکرارپذیری) به درجه‌ای اشاره دارد که دو یا چند کدگذار نتایج یکدیگر را تکرار می‌کنند. فرایند کدگذاری، درصورتی‌که کدگذاران یک متن را به یک شیوه کدگذاری کنند، تکرارپذیر خوانده می‌شود. به این منظور از یک کدگذار و یک همکار پژوهش استفاده می‌کند. آموزش‌ها و تکنیک‌های لازم و استانداردشده برای کدگذاری متون به همکار پژوهش انتقال داده می‌شود. سپس محقق همراه این همکار پژوهش چند متن را به‌صورت تصادفی، انتخاب و کدگذاری می‌کنند.

برای محاسبه پایایی، متغیر تصادفی X را تعداد توافقات در n نمونه تصادفی مستقل در نظر می‌گیرند. در این صورت با یک توزیع دوجمله‌ای مواجه هستیم که موفقیت (p)، نسبت توافقات به تعداد کل کدها است. درصد پایایی را به‌صورت زیر محاسبه می‌کنند: (تات اسدی،[۵])

$$E(X) = ۱۰۰np$$

با بررسی پایایی کدگذاری، اعتماد به کلماتی که کدگذاری شده است افزایش می‌یابد و می‌توان از سایر روش‌های دیگر آمار نیز کمک گرفت.

۲.۳ رگرسیون

مدیریت و تجزیه‌وتحلیل مؤثر داده‌های متنی در مقیاس بزرگ، چالش‌های مهمی به‌ویژه به دلیل نیازهای بالای ذخیره‌سازی و پردازش را ایجاد می‌کند. تجزیه‌وتحلیل رگرسیون متن، شاخه خاصی از متن‌کاوی است و افراد، محققان و کسب‌وکارها را قادر می‌سازد تا بینش‌های معناداری را از حجم به‌سرعت در حال افزایش داده‌های متنی به دست

نظرات	امتیاز
محصول عالی است	۵
ارزش پول دادن ندارد.	۲
کاملاً دوستش داشتم.	۵
کیفیتش افتضاحه	۱
نسبت به قیمتش می ارزه	۴

جدول ۲: ارزش‌گذاری

آورند. رگرسیون را می‌توان در متن‌کاوی برای اهداف مختلف اعمال کرد. در تحلیل احساسات از رگرسیون برای پیش‌بینی نمرات احساسات بر اساس ویژگی‌های متنی استفاده می‌شود. می‌توان در طبقه‌بندی متن از رگرسیون لجستیک برای دسته‌بندی متون در کلاس‌های از پیش تعریف‌شده استفاده کرد. در مدل‌سازی موضوع از تکنیک‌های رگرسیون برای شناسایی و تعیین میزان روابط بین موضوعات و ویژگی‌های متن استفاده می‌شود. از رگرسیون در استخراج کلمات کلیدی استفاده می‌شود. تکنیک‌های رگرسیونی برای پیش‌بینی نویسنده بر اساس ویژگی‌های سبک نوشتن به کار می‌رود. با استفاده از رگرسیون می‌توان تولید متن کرد. از رگرسیون در مدل‌های زبانی برای پیش‌بینی کلمه یا دنباله بعدی در یک متن استفاده می‌کنند. به‌عنوان‌مثال شوبایو و همکاران [۲۹] نشان دادند که چگونه مدل‌های زبانی پیشرفته می‌توانند به درک داده‌های مالی پیچیده کمک کنند. نتایج آن‌ها نشان داد که رگرسیون لجستیک از FinBERT و GPT۴ در پیش‌بینی بازار سهام با استفاده از اخبار مالی بهتر عمل می‌کند.

برای اعمال رگرسیون، داده‌های متنی باید به فرمت عددی تبدیل شوند. مراحل متداول پیش‌پردازش به این صورت است که ابتدا متن به کلمات تقسیم می‌شود سپس کلمات رایج همانند «از»، «در» و ... حذف می‌گردد. سپس با استفاده از روش‌هایی همانند $TF-ITF$ و $BagogWords$ به عدد تبدیل می‌شوند. برای مثال در عبارت «محصول عالی است.» مقدار $TF-ITF$ ممکن است به‌صورت جدول (۳) باشد

حال به‌جای پرداختن به متن ساده، مدل رگرسیون با ویژگی‌های عددی کار می‌کند و می‌تواند امتیاز (ستاره) نظرات فاقد ستاره را با استفاده از مدل خطی

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

طاها[۳۰] با معرفی یک طبقه‌بندی روش‌شناختی که برای تحلیل رگرسیون متن طراحی‌شده است، به گروه‌بندی کلی الگوریتم‌ها در نظرسنجی‌های موجود می‌پردازد. این طبقه‌بندی، الگوریتم‌ها را به تکنیک‌های خاص و دسته‌های تفصیلی دسته‌بندی می‌کند که در دو سطح دسته روش‌شناسی و تکنیک روش‌شناسی سازمان‌دهی شده‌اند و برای تأیید صحت تکنیک‌ها و مقوله‌های مختلف و ارزیابی‌های تجربی از تکنیک‌های رگرسیون متن استفاده کردند. برای فهم بهتر مثالی ساده می‌آوریم.

مثال ۱.۳. *فرض کنید شما صاحب یک فروشگاه آنلاین هستید. مشتریان در مورد محصولات شما نظر می‌گذارند، اما برخی از نظرات رتبه‌بندی عددی (ستاره) به آن‌ها داده نشده است. شما می‌خواهید سیستمی بسازید تا به‌طور خودکار نمره احساسات (عددی بین ۱ تا ۵) را از بررسی متن‌ها پیش‌بینی کند. در جدول (۲) نمونه‌ای از نظرات آورده شده است.*

پیش‌بینی کند که در آن x_i مقادیر $TF-ITF$ می‌باشند.

۳.۳ سری زمانی

تجزیه‌وتحلیل سری‌های زمانی در متن‌کاوی برای درک الگوهای زمانی در داده‌های متنی بسیار مهم است. سری زمانی به بررسی داده‌های متنی می‌پردازد که دارای یک جزء زمانی هستند، مانند پست‌های رسانه‌های اجتماعی، مقالات خبری، یا نظرات مشتریان در طول زمان. این رویکرد چندین مزیت دارد که در ادامه به آن اشاره می‌کنیم:

کلمات	مقدار $TF-ITF$
محصول	۰/۴۵
عالی	۰/۳۲
است	۰/۱۲

جدول ۳: تبدیل متن به عدد

- شناسایی روند: موضوعات نوظهور، تغییرات در افکار عمومی یا تغییرات در استفاده از زبان در طول زمان را شناسایی می‌کند.
- مدل‌سازی و پیش‌بینی: روندها یا رویدادهای آینده را بر اساس الگوهای داده‌های متنی تاریخی پیش‌بینی می‌کند.
- تشخیص فصلی: الگوهای تکراری را در داده‌های متنی، مانند بررسی‌های فصلی محصول یا بحث‌های مربوط به تعطیلات را شناسایی می‌کند.
- تشخیص ناهنجاری: با شناسایی انحرافات از الگوهای متنی عادی، رویدادهای غیرمعمول یا غیرمنتظره را شناسایی می‌کند.
- تجزیه‌وتحلیل احساسات در طول زمان: ردیابی تغییرات در احساسات نسبت به موضوعات خاص، مارک‌ها یا محصولات در دوره‌های زمانی مختلف یکی از جذاب‌ترین کاربردهای سری زمانی در متن‌کاوی است.

تکنیک‌هایی همانند میانگین‌های متحرک، هموارسازی نمایی، مدل‌های ARIMA[۷] و رویکردهای مبتنی بر شبکه عصبی (به‌عنوان‌مثال LSTM و GRU) در تجزیه‌وتحلیل با استفاده از سری زمانی استفاده می‌شود. در متن‌کاوی، زمان می‌تواند تاریخ انتشار مقالات، مهرهای زمانی پست‌های رسانه‌های اجتماعی یا هر جنبه زمانی دیگری از داده‌های متنی باشد. با ترکیب این تکنیک‌های سری زمانی با روش‌های متن‌کاوی، محققان می‌توانند بینش عمیق‌تری در مورد جنبه‌های زمانی داده‌های متنی به دست آورند و توانایی آن‌ها را برای درک و پیش‌بینی روندها در زبان، موضوعات و احساسات در طول زمان افزایش دهند. یکی از حوزه‌های پرطرفدار در متن‌کاوی با استفاده از سری زمانی پیش‌بینی بازار سهام است که در برنامه‌ریزی فعالیت‌های تجاری نقش مهمی ایفا می‌کند. شه و همکاران [۲۸] به بازیابی، استخراج و تحلیل اثرات احساسات خبری بر بازار سهام دارویی پرداختند. یوسمانی و شمسی [۳۱] یک چارچوب جدید ارائه کردند تا با استفاده تکنیک‌های سری زمانی نمودارهای علیت را از متن رسانه‌های دیجیتال استخراج کنند.

مثال ۲.۳. *فرض کنید برای یک شرکت تجارت الکترونیک کار می‌کنید و می‌خواهید ببینید که چگونه احساسات مشتری در طول زمان بر اساس نظرات آن‌ها تغییر می‌کند. نظرات مشتریان شما در هر ماه سال گذشته جمع‌آوری‌شده است. این نظرات و مقدار امتیازهای (ستاره‌های) مرتبط با نظرات در جدول (۴) آورده شده است.*

عبارات	مقدار امتیاز
عالی بود	۵
دوست‌داشتنی هست	۵
هیچ‌وقت نخرید	۱
کیفیتش آنچه انتظار داشتم نبود	۳
نسبت به قیمتش می‌ارزد	۴
خدمات پس از فروش عالی هست	۵

جدول ۴: نظرات و امتیازهای مربوطه

[۷] Average Moving Integrated Autoregressive

می‌توان از سری‌های زمانی برای پیگیری چگونگی تغییر احساسات مشتری از ماه‌به‌ماه دیگر استفاده کرد. به‌طور مثال احساسات ممکن است در طول فصول تعطیلات به دلیل افزایش حجم سفارش و تأخیر کاهش یابد یا احساسات ممکن است پس از عرضه محصول جدید یا پس از حل مشکل مشتری بهبود یابد. متوسط امتیازها در ماه‌های دی، بهمن، اسفند، فروردین و اردیبهشت در جدول (۵) آورده شده است. مدل‌های سری زمانی مانند $ARIMA$ یا $LSTM$ می‌توانند برای پیش‌بینی روندهای احساسات آینده بر اساس داده‌های گذشته استفاده شوند. به‌عنوان‌مثال اگر مدل پیش‌بینی کند که احساسات در ماه‌های آینده کاهش می‌یابد، شرکت می‌تواند اقدامات پیشگیرانه‌ای برای بهبود خدمات مشتری و جلوگیری از بررسی‌های منفی انجام دهد.

ماه	متوسط امتیاز
دی	۴٫۵
بهمن	۳٫۸
اسفند	۲٫۱
فروردین	۴٫۲
اردیبهشت	۴٫۷

جدول ۵: امتیازها در طول ۵ ماه

۴.۳ آمار بیزی

مار بیزی نقش اساسی در متن‌کاوی ایفا می‌کند و می‌تواند یک چارچوب احتمالی برای مدیریت عدم قطعیت و استنتاج در مورد داده‌های متنی ارائه دهد. به‌ویژه در کارهایی که با در دسترس قرار گرفتن اطلاعات جدید، نیاز به برآورد احتمال، رده‌بندی داده‌ها یا به‌روزرسانی باورها دارند، تکنیک‌های آمار بیزی می‌تواند بسیار مؤثر باشد. در ادامه مختصراً به چند مورداشاره می‌کنیم.

- رده‌بند بیز: یکی از رایج‌ترین کاربردهای آمار بیزی در متن‌کاوی رده‌بندی متن است مانند فیلترکردن هرزنامه یا تجزیه. در رده‌بند بیز از قضیه بیز به این صورت استفاده می‌کنند که ویژگی‌ها (کلمات) با توجه به کلاس مستقل هستند. علیرغم سادگی، این رویکرد در حوزه‌هایی مانند رده‌بندی ایمیل و برچسب‌گذاری اسناد بسیار مؤثر است. به‌عنوان‌مثال در تشخیص هرزنامه، بیز یک مدل احتمال را برای تشخیص اسپم بودن یک ایمیل با توجه به وقوع کلمات کلیدی خاص ارائه می‌دهد. کومار و همکاران [۱۳] پیامدهای فرض استقلال را بر عملکرد مدل بررسی می‌کنند و استراتژی‌هایی را برای پرداختن به انحرافات دنیای واقعی ارائه می‌دهند. رانا و سینک [۲۴] با استفاده از تکنیک رده‌بندی بیز جهت‌گیری احساسات مثبت و منفی را با استفاده از نظرات کاربران فیلم بررسی کردند.
- تحلیل احساسات: چارچوب‌های بیزی امکان ادغام دانش احساسات قبلی را با داده‌های مشاهده‌شده فراهم می‌کند. به‌عنوان‌مثال تجزیه‌وتحلیل احساسات در داده‌های توییتر به موضوعی محبوب در سال‌های اخیر تبدیل‌شده است. رز و همکاران [۲۶] با استفاده از رده‌بندهای شبکه بیزی، به تحلیل احساسات در طول رویدادهای حیاتی مانند بلایای طبیعی یا جنبش‌های اجتماعی پرداختند.
- ابهام‌زدایی از کلمه و عبارت: مدل‌های بیزی وظیفه پردازش زبان طبیعی مانند ابهام‌زدایی از کلمه را دارند. به این معنی که یک کلمه ممکن است معانی متعددی داشته باشد. استنتاج بیزی، احتمال هر یک از معناها را با توجه به زمینه‌ای که کلمه در آن ظاهر می‌شود، ارزیابی می‌کند. لیو و همکاران [۱۶] از بیز برای ابهام‌زدایی از یک اصطلاح در حیطه زیست پزشکی استفاده کردند.
- تحلیل پویای متن: روش‌های بیزی برای مدل‌سازی پویای جریان‌های متنی، جایی که محتوای متنی در طول زمان تکامل می‌یابد (به‌عنوان‌مثال، رسانه‌های اجتماعی یا اخبار) استفاده می‌شود. مدل‌های بیزی سری زمانی می‌توانند تغییرات موضوع یا احساسات را در طول زمان ردیابی کنند. مدل‌های موضوعی

پویا از موارد پرکاربرد در این حوزه است. زوسا و ویلدینگ [۳۴]، یائو و وانگ [۳۲] و ژانگ و لو [۳۳] از این تکنیک استفاده کردند.

آمار بیزی ستون فقرات بسیاری از روش‌های مدرن متن‌کاوی است و یک رویکرد همه‌کاره و از لحاظ نظری صحیح برای مدل‌سازی داده‌های متنی ارائه می‌دهد. مدل‌های بیزی می‌توانند پیش‌بینی‌های خود را با رسیدن داده‌های جدید به‌روز کنند و بلافاصله از آن‌ها در کاربرد استفاده کنند. همچنین بیز امکان ترکیب اطلاعات قبلی را فراهم می‌کند که این موجب می‌شود در زمانی که داده‌ها کمیاب هستند عملکرد بهبود یابد. بیز ساده[8] یکی از محبوب‌ترین تکنیک‌ها در متن‌کاوی است، به‌ویژه برای کارهای رده‌بندی متن مانند تشخیص هرزنامه.

مثال ۳.۳. *تصور کنید در حال ساختن سیستمی هستید که به‌طور خودکار ایمیل‌ها را بر اساس محتوای متنی به دو قسمت هرزنامه یا غیر هرزنامه رده‌بندی می‌کند. برای این منظور می‌توانید مجموعه‌ای از کلمات را به‌عنوان هرزنامه قرار دهید. جدول (۶) نمونه‌ای از این عبارات را بیان کرده است.*

عبارات	رده‌بندی
شما برنده یک موبایل شده‌اید. برای دریافت جایزه کلیک کنید.	$Spam$
فردا ساعت ده صبح می‌بینمت	$Notspam$
تبریک می‌گویم. شما برنده‌شده‌اید	$Spam$
برنامه هفته بعد را برای من بفرست تا زمانی را مشخص کنم	$Notspam$

جدول ۶: ایمیل و رده‌بندی

بر اساس این جدول کلمات «برنده»، «موبایل»، «دریافت»، «جایزه»، «کلیک» و «تبریک» در هرزنامه قرار دارد و کلمات «فردا»، «ساعت»، «ده»، «می‌بینم»، «صبح»، «برنامه»، «هفته»، «بفرست»، «زمانی» و «مشخص» در غیرهرزنامه قرار می‌گیرد. حال از قضیه بیز برای محاسبه احتمال اینکه یک ایمیل هرزنامه هست یا خیر استفاده می‌گردد.

$$p(C|X) = \frac{p(X|C)p(C)}{p(X)}$$

که در آن $p(C|X)$ احتمال اینکه ایمیل متعلق به هرزنامه یا غیرهرزنامه (C) باشد با توجه به کلمات موجود در ایمیل، $p(X|C)$ احتمال حضور کلمات در C، $p(C)$ احتمال پیشین C و $p(X)$ احتمال پیشین برای کلمات ایمیل می‌باشد. توجه داشته باشید که در محاسبه احتمال شرطی از هموارساز لاپلاس (۱) استفاده می‌شود تا صفر رخ ندهد.

$$p(X|C) = \frac{\text{۱ + تعداد دفعات تکرار کلمه در } C}{\text{تعداد کل کلمات در } C + V} \qquad (۱)$$

که در آن V تعداد کلمات بدون در نظر گرفتن تکرارها می‌باشد. فرض کنید ایمیل «تبریک! کلیک کنید و جایزه‌بگیرید» دریافت شده است. با توجه به جدول (۶) احتمالات زیر را داریم

[8] Naive Bayes

$$p(\text{هرزنامه}) = \frac{\text{تعداد ایمیل‌های هرزنامه}}{\text{تعداد کل ایمیل‌ها}} = \frac{۲}{۴}, p(\text{غیرهرزنامه}) = \frac{\text{تعداد ایمیل‌های غیرهرزنامه}}{\text{تعداد کل ایمیل‌ها}} = \frac{۲}{۴}$$

$$p(\text{هرزنامه}|\text{ایمیل}) \propto p(\text{هرزنامه})p(\text{جایزه}|\text{هرزنامه})p(\text{کلیک}|\text{هرزنامه})p(\text{تبریک}|Spam)$$

$$= \frac{۱+۱}{۷+۶} \cdot \frac{۱+۱}{۷+۶} \cdot \frac{۱+۱}{۷+۶} \approx ۰/۰۰۴$$

$$p(\text{غیرهرزنامه}|\text{ایمیل}) \propto p(\text{غیرهرزنامه})p(\text{جایزه}|\text{غیرهرزنامه})p(\text{کلیک}|\text{غیرهرزنامه})p(\text{تبریک}|\text{غیرهرزنامه})$$

$$= \frac{۰+۱}{۱۰+۱۰} \cdot \frac{۰+۱}{۱۰+۱۰} \cdot \frac{۰+۱}{۱۰+۱۰} \approx ۰/۰۰۰۱$$

$$p(\text{هرزنامه}|\text{ایمیل}) \propto p(\text{هرزنامه})p(\text{ایمیل}|\text{هرزنامه}) = ۰/۰۰۴ \times ۰/۵ = ۰/۰۰۲$$

$$p(\text{غیرهرزنامه}|\text{ایمیل}) \propto p(\text{غیرهرزنامه})p(\text{ایمیل}|\text{غیرهرزنامه}) = ۰/۰۰۰۱ \times ۰/۵ = ۰/۰۰۰۰۵$$

بحث و نتیجه‌گیری

با توجه به مطالب عنوان‌شده درمیابیم که با به‌کارگیری تکنیک‌های آمار در واقع از نتایج خود با اطمینان بیشتری استفاده می‌کنیم. امروزه تلاش بر نزدیک کردن استنباط‌ها بر اساس تکنیک‌های آماری به‌طور گسترده موردتوجه است. در واقع این پژوهش نشان داد که استفاده از تکنیک‌های آماری در متن‌کاوی می‌تواند افق‌های جدیدی را برای تحلیل داده‌های کیفی باز کند و پژوهشگران را در دستیابی به درک عمیق‌تر از داده‌ها یاری دهد. بهره‌گیری از آمار نه‌تنها باعث افزایش دقت تحلیل‌ها می‌شود، بلکه امکان پاسخ به پرسش‌های پیچیده و چالش‌برانگیز در حوزه‌های مختلف را فراهم می‌کند.

تقدیر و تشکر

بدین‌وسیله مراتب قدردانی و سپاس خود را از داوران محترم این مقاله ابراز می‌دارم. نظرات دقیق، راهنمایی‌های ارزشمند و پیشنهادهای سازنده شما نقش بسزایی در بهبود کیفیت علمی این پژوهش داشته است. همچنین از ویراستاران گرامی که با دقت و دانش خود در ویرایش علمی و زبانی مقاله تلاش کردند، صمیمانه سپاسگزارم. تلاش‌های شما در اصلاح ساختار، انسجام محتوا و رعایت اصول نگارشی به ارتقای هرچه بیشتر کیفیت این اثر انجامید. بدون شک، همکاری و همدلی شما، گامی مؤثر در جهت تقویت محتوای علمی این مقاله و پیشبرد مرزهای دانش در این حوزه بوده است. از زحمات بی‌شائبه و همراهی بی‌دریغ شما صمیمانه سپاسگزاری می‌کنم.

مراجع

[۱] الماسی قمی، م. رضی مهابادی، ب. علایی رحمانی، ف. (۱۳۹۹)، کشف محور موضوعی سوره یونس با روش متن‌کاوی، مرکز نشر آثار علمی دانشگاه الزهرا، تهران.

[۲] ایمان، م.ت. (۱۳۸۸)، مبانی پارادایمی روش‌هاي تحقیقی کمی و کیفی در علوم انسانی، قم، پژوهشگاه حوزه و دانشگاه.

[۳] ایمان، م.ت. نوشادي، م.ر. (۱۳۹۰). متن‌کاوی کیفی، فصلنامه پژوهش، سال سوم، شماره دوم، پائیز و زمستان.

[۴] باردن، ل. (۱۳۷۴). متن‌کاوی، ترجمه ملیحه آشتیانی و محمد یمنی دوزي سرخابی، تهران: انتشارات دانشگاه شهید بهشتی

[۵] تات اسدی، م. رنگریز، ح. جعفری نیا، س. (۱۳۹۷)، جبران خدمات از دیدگاه اسلام به روش متن‌کاوی، *دانشگاه خوارزمی*

[۶] قائدی، م. گلشنی، ع. (۱۳۹۵). روش متن‌کاوی از کمی گرایی تا کیفی گرایی، فصلنامه روش‌ها و مدل‌های روان‌شناختی، سال هفتم، شماره ۲۳، بهار.

[۷] نعمت الهی، ن. (۱۳۸۰)، آمار و احتمال مهندسی، دانشگاه علامه طباطبایی، تهران.

[8] Berleson, B. (1952), *Content Analysis in Communication Research*, The Free press.

[9] Bowen, W.M. and Bowen, C.C. (2008), Content Analysis: in Kaifeng Yang and Gerald J. Miller (Eds.), *Handbook of Research Methods in Public Administration*, Taylor & Francis.

[10] Durga, S. V., Singh, N., Rana, A., Kalra, R., and Al Khidhir Abdullah, S. A. (2024), Mining Social Media Data for Sentiment Analysis and Trend Prediction, *10th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON),* Gautam Buddha Nagar, India, 1557-1562.

[11] Freud Sigmond, D. (1900), *Die Traumdeutung,* Franz Deuticke, Austria.

[12] Kent, J. T. K. and Mardia, K. V. (2002), Modelling Strategies for Spatial-Temporal Data, In Spatial Cluster Modelling, *Chapman and Hall,* 214–226.

[13] Kumar, A. J., Bigit Krishna Goswami, Soham Motiram Mhatre, Sneha Agrawal (2024), Naive Bayes in Focus: A Thorough Examination of its Algorithmic Foundations and Use Cases, *International Journal of Innovative Science and Research Technology (IJISRT),* 2078-2081.

[14] Krippendorff, K., (1980), *Content Analysis. An introduction to its Methodology*, The Sage Commtext Series, Sage Publications Ltd., London.

[15] Lasswell, H. D., associates (1949), *Language of Politics*, George, W. Stewart, Publisher, Inc., New York.

[16] Liu Hongfang, Virginia Teller, Carol Friedman, (2004) A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation, *Journal of the American Medical Informatics Association,* **11(4)**, 320–331.

[17] Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi (2024), Infini-gram: Scaling Unbounded n-gram Language Models to a Trillion Tokens, *arXiv*, 2401-17377.

[18] Jelinek, F., Bahl, L. R., & Mercer, R. L. (1975), Design of a linguistic statistical decoder for speech recognition. In *Proceedings of the 6th International Conference on Computational Linguistics (COLING)*, 139-145.

[19] Mangiaracina, R., Song, G., Perego, A., (2015), Distribution network design: A literature review and a research agenda, *International Journal of Physical Distribution and Logistics Management*, **45(5)**, 506-531.

[20] Michael Q. P. & Moharis, M. (2002), *Qualitative Research & Evaluation Methods* (3rd ed.), Sage Publications.

[21] Maykut, P., & Morehouse, R. (2002). *Beginning Qualitative Research: A Philosophical and Practical Guide*. Routledge.

[22] Netolicky, P., Petrovsky, J., and Darena, F. (2018), Text-Mining in Streams of Textual Data Using Time Series Applied to Stock Market, *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, **66**, 1573-1580.

[23] Poole, M. S., Folger, J. P. (1981), Modes of observation and the validation of interaction analysis schemes, *Small Group Behavior*, **12**, 477-493.

[24] Rana, S., and Singh, A. (2016), Comparative analysis of sentiment orientation using SVM and Naive Bayes techniques, *2nd International Conference on Next Generation Computing Technologies (NGCT)*, 106-111.

[25] Rundh, B. (2003), Rethinking the international marketing strategy, new dimensions in a competitive market, *Marketing Intelligence and Planning*, **21**, 249-257.

[26] Ruz Gonzalo A., Pablo A. Henríquez, Aldo Mascareño, (2020), Sentiment analysis of Twitter data during critical events through Bayesian networks classifiers, *Future Generation Computer Systems*, **106**, 92-104.

[27] Saldana, J., (1995), Is theatre necessary?: Final exit interviews with sixth grade participants from the ASU longitudinal study, *Youth Theater Journal*, **11**, 25-46.

[28] Shah, D., Isah, H., and Zulkernine, F. (2018), Predicting the Effects of News Sentiments on the Stock Market, *IEEE International Conference on Big Data*, 4705-4708.

[29] Shannon, C. E. (1948), A mathematical theory of communication, *The Bell System Technical Journal*, **27(3)**, 379-423.

[30] Shobayo, O., Adeyemi-Longe, S., Popoola, O., and Ogunleye, B. (2024), Innovative Sentiment Analysis and Prediction of Stock Price Using FinBERT, GPT-4, and Logistic Regression: A Data-Driven Approach. *Big Data and Cognitive Computing*, **8(11)**, 143.

[31] Taha, K. (2024), Text Regression Analysis: A Review, Empirical, and Experimental Insights, *IEEE*, **12**, 137333-137344.

[32] Usmani, S., Shamsi, J. A. (2021), News-sensitive stock market prediction: literature review and suggestions, *PeerJ Computer Science*, **7**, 490.

[33] Yao, Fang and Wang, Yan. (2020), Tracking urban geo-topics based on dynamic topic model, *Computers, Environment and Urban Systems*, **79**.

[34] Zhang, D.C., and Lauw, H. (2022), Dynamic Topic Models for Temporal Document Networks, *Proceedings of the 39th International Conference on Machine Learning*, **162**, 26281-26292.

[35] Zosa, Elaine and Granroth-Wilding, Mark. (2019), Multilingual Dynamic Topic Model. *In Proceedings of the International Conference on Recent Advances in Natural Language Processing*, 1388–1396.

Statistical approaches of text mining

Bayati[1], M.

[1]Department of Statistics, University of Qom, Qom, Iran.

Abstract: We live in the information age, constantly surrounded by vast amounts of data from the world around us. To utilize this information effectively, it must be mathematically expressed and analyzed using statistics. Statistics play a crucial role in various fields, including text mining, which has recently garnered significant attention. Text mining is a research method used to identify patterns in texts, which can be in written, spoken, or visual forms.

The applications of text mining are diverse, including text classification, clustering, web mining, sentiment analysis, and more. Text mining techniques are utilized to assign numerical values to textual data, enabling statistical analysis.

Since working with data requires a solid foundation in statistics, statistical tools are employed in text analysis to make predictions, such as forecasting changes in stock prices or currency exchange rates based on current textual data. By leveraging statistical methods, text mining can uncover, confirm, or refute the truths hidden within textual content. Today, this topic is widely used in machine learning. This paper aims to provide a basic understanding of statistical tools in text mining and demonstrates how these powerful tools can be used to analyze and interpret events.

Keywords: Coding, Descriptive Statistics, Time series, Regression, Bayse inference.