فصلنامه پژوهش‌های حفاظتی ـ امنیتی
دانشگاه جامع امام حسین (علیه‌السلام)
سال دوازدهم، شمارهٔ ۴۵، بهار ۱۴۰۲

مدیریت مستندات محتوای فارسی رسانه های آنلاین خبری در جامعه اطلاعاتی

● حمید میرزائی دهنوی ●
گروه مدیریت فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران.
● محمد علی کرامتی ●
گروه مدیریت صنعتی، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران. (نویسنده مسئول)
● محمدعلی عامری حبیب آبادی ●
گروه مدیریت، پژوهشگاه علوم انتظامی و مطالعات اجتماعی، تهران، ایران.

تاریخ دریافت: ۱۴۰۱/۱۱/۱۵ تاریخ پذیرش: ۱۴۰۱/۰۲/۲۱

چکیده

عصری که ما در آن زندگی می‌کنیم، عصر اطلاعات است و برای سازمان ها مهم‌ترین مسئله، اشراف بر همین اطلاعات است. با رشد روزافزون اخبار در دنیای دیجیتال و اینترنت، موضوعی که مهم و حائز اهمیت می‌شود، دسته‌بندی این اطلاعات و دسترسی سریع و ارزان ما به آن‌ها می‌باشد. این مهم به دست نمی‌آید مگر با انجام روش‌هایی که از آن به‌عنوان دسته‌بندی متون یاد شده است. هدف این پژوهش دسته‌بندی متون خبری در دسته‌های از قبل مشخص‌شده می باشد که با استفاده از ابزار مدل اتوماتیک که یکی از زیرمجموعه‌های متن‌کاوی محسوب می‌شود صورت می پذیرد. با توجه به اهمیت موضوع و کاری که در این زمینه برای زبان‌های دیگر دنیا انجام گرفته است، نیاز به طبقه‌بندی متون فارسی به خوبی احساس می‌شود. این نکته قابل توجه می باشد که تحقیقات برای متون انگلیسی توسعه داده شده و به کارگرفته می شود اما از آنجایی‌که زبان فارسی پیچیدگی های ساختاری نسبت به سایر زبان ها دارد و همچنین تحقیقات کمتری در این زمینه انجام گرفته است، این پژوهش از نوع کاربردی، توسعه ای می باشد که برای انجام آن می توان به روش پژوهش آزمایشی و استفاده از ابزار متن کاوی اشاره کرد، به این صورت که در یک محیط کاملاً تحت کنترل و با توانایی ثابت نگه داشتن سایر متغیرها انجام می گردد. در جامعه اطلاعاتی دسته‌بندی متون به‌وسیله افراد نخبه و به‌صورت دستی انجام می‌گیرد. دسته‌بندی متون آن‌هم با این حجم و به‌صورت دستی غیرممکن به نظر می‌رسد، بنابراین ما ناگزیریم که به دنبال روش‌هایی برای دسته‌بندی خودکار متون باشیم. از سوی دیگر ذخیره‌سازی، پردازش و تحلیل این حجم از اطلاعات تبدیل به چالشی جدی شده است. با توجه به حجم بالای اخبار، داده ها، اطلاعات، اسناد و پیچیدگی حفظ و نگهداری آنها، لازم است سیستمی جهت مدیریت دریافت، حفظ و نگهداری اخبار موجود، بکار گرفته شود. پیچیدگی سازمان ها نیاز به متمرکز بودن اخبار، اسناد، طبقه بندی درست، گردش صحیح اخبار و سهولت در دسترسی به آنها را ایجاد می نماید. مدیریت مستندات این امکان را برای سازمان های اطلاعاتی فراهم می آورد که اخبار و اسناد دریافتی یا موجود را به درستی طبقه بندی نموده، آنها را حفظ، نگهداری و بازیابی نماید. با بررسی، تحلیل و پردازش در این پژوهش به این نتیجه می رسیم که دقت و نتایج حاصل روش پیشنهادی روی متون خبری برخط نشان می‌دهد؛ مدل ماشین بردار پشتیبان دارای دقت ۹۳/۲۹، صحت ۹۳/۳۲، فراخوانی ۹۲/۹۶ و خطای ۶/۷۱ است.

واژگان کلیدی: طبقه‌بندی، مدیریت مستندات، محتوای فارسی، متن‌کاوی، اطلاعات و دسته‌بندی

مقدمه

یکی از زیرمجموعه‌های اسناد متنی که هر روزه با آن روبه‌رو هستیم، اخبار است. برخلاف چند دهه گذشته امروزه دیگر حجم غالب اخبار در روزنامه‌ها و مجلات چاپ نمی‌شود، بلکه بیشتر حجم اخبار منتشرشده در سایت‌های خبری و شبکه‌های اجتماعی دیده می‌شود. وجود حجم بسیار زیاد اخبار در این پایگاه‌های خبری، نیاز به طبقه‌بندی موضوعی اخبار را ایجاد می‌کند. به‌عبارت‌دیگر، پیش از انتشار اخبار در سایت‌های مختلف خبری نیاز است که دسته‌ی موضوعی هر متن خبری مشخص شود. با دسته‌بندی متون خبری که بر روی سایت‌های خبری منتشر می‌شوند، کاربران این سایت‌ها می‌توانند اخبار مورد نظر خود را آسان‌تر و سریع‌تر پیدا کنند. از طرفی به دلیل حجم بالای اخباری که هر روزه منتشر می‌شوند، طبقه‌بندی اخبار به‌صورت دستی فرآیندی پرهزینه و نیازمند تعداد زیادی نیروی انسانی متخصص است و علاوه بر این، معیار زمان که در انتشار اخبار، عامل بسیار مهمی است، در طبقه‌بندی اخبار به‌صورت دستی بسیار زیاد است. این به آن معنی است که هر سایت خبری تلاش می‌کند اخبار روزانه را سریع‌تر از دیگر سایت‌ها در صفحه‌ی خود بارگذاری کند. به همین دلیل است که طبقه‌بندی متون خبری به‌صورت خودکار امری بسیار کاربردی و لازم است. با طبقه‌بندی خودکار متون خبری، نویسندگان سایت‌ها و شبکه‌های خبری می‌توانند صدها متن خبری را سریع‌تر، آسان‌تر و با صرف هزینه و نیروی متخصص کمتری منتشر نمایند.

جامعه اطلاعاتی داده‌های متنی بسیاری را رصد می‌کنند، چگونه می‌توان این حجم از داده‌ها را مدیریت کرد؟ چگونه می‌توان اطلاعات مفیدی از این داده‌های متنی بی‌شمار بدست آورد؟ آیا این داده‌ها ارزشی دارند؟ متن‌کاوی فناوری مورداستفاده برای چنین مواردی است. با فناوری‌های متن‌کاوی می‌توان داده‌های متنی را بررسی و تحلیل کرده و از نتایج حاصل از این تحلیل، اطلاعات ارزشمندی کسب نمود. داده‌های متنی هیچ‌گونه ارزشی ندارند مگر اینکه متن‌کاوی شوند. متن‌کاوی به شناسایی الگوها، کلمات کلیدی، موضوعات و دیگر ویژگی‌های موجود در متن می‌پردازد. الگوریتم‌ها و روش‌های مختلفی برای متن‌کاوی وجود دارد و همچنان نیز درحال‌توسعه می‌باشد، اما مشکل اصلی این روش‌ها برای دستگاه‌های اطلاعاتی حفظ محرمانگی اطلاعات است. باید در کنار استفاده از علوم و فناوری‌های نوین اصل محرمانگی اطلاعات نیز حفظ شود. بسیاری از برنامه‌های موجود در موضوع داده‌کاوی و متن‌کاوی در بستر اینترنت کار می‌کنند که عملاً برای جامعه اطلاعاتی بدون استفاده است. خوشبختانه چند سالی است که شرکت‌های داخلی شروع به فعالیت در

زمینه متن‌کاوی نموده‌اند و پیشرفت‌های قابل‌توجهی هم در این زمینه داشته‌اند، اما کماکان نمی‌توان برای مجموعه‌های امنیتی و نظامی ازاین‌گونه نرم‌افزارها استفاده نمود. لذا لازم است مجموعه‌های اطلاعاتی امنیتی این موضوع را به‌صورت بومی پیاده‌سازی کنند و یک بستر مناسب برای تحلیل اطلاعات داشته باشند. لذا بر آن شدیم تا در این مقاله به مدیریت مستندات محتوای فارسی رسانه های آنلاین خبری بپردازیم که خروجی آن برای سازمان های اطلاعاتی این است که بتوانند از سایت‌های خبری، متون مختلف را جمع آوری و سپس دسته‌بندی و بازیابی نمایند.

روش‌شناسی پژوهش

امروزه سامانه‌های طبقه‌بند متون گستردگی بیشتری پیدا کرده‌اند و وجود یک سیستم طبقه‌بند متون فارسی که بهینه باشد بسیار احساس می‌شود. طبقه‌بندی متون، یکی از زیرمجموعه‌های متن‌کاوی است. متن‌کاوی بر روی پردازش متون تمرکز دارد. در پردازش متون سعی می‌شود دانشی از متون خام استخراج گردد. در طبقه‌بندی متون هدف مشخص کردن دسته متن است. به‌طور مثال: خبری که در یک پایگاه خبری درج شده است یک خبر سیاسی یا ورزشی است. در این مثال سیاسی و ورزشی بودن خبر دو دسته برای طبقه‌بندی اخبار را تشکیل می‌دهند. طبقه‌بندی متون قدمت بسیار زیادی در حوزه متن‌کاوی دارد. این موضوع از سال ۱۹۶۰ میلادی مورد توجه محققین بوده است ولی با رونق کامپیوتر و نرم‌افزارها، مانند دیگر موضوعات هوش مصنوعی و داده‌کاوی در دهه ۹۰ میلادی توجه به این موضوع رشد چشمگیری پیدا کرد و مورد توجه قرار گرفت. از آن جایی که در انجام پژوهش‌های مرتبط با هوش مصنوعی و زبان شناسی رایانشی، از علوم زبان شناسی و علوم رایانه استفاده می‌شود، ماهیت بین رشته‌ای علم زبان شناسی رایانشی در تعیین روش و ابزار انجام پژوهش حاضر نقش به سزایی دارد. در این بخش در پی آشنایی با گام‌های پیشنهادشده برای انجام پژوهش می‌باشیم. برای انجام تحقیق باید گام‌های ذیل مشخص شود:

الف) جمع‌آوری متون خبری فارسی (اخبار حوزه ایران، اخبار حوزه افغانستان و اخبار سطح جهانی)

ب) پیش‌پردازش و آماده‌سازی اخبار

ج) انتخاب ویژگی

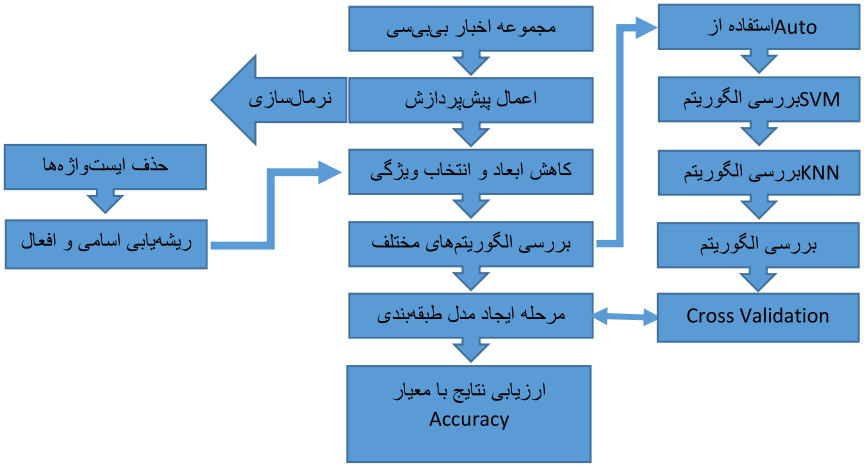
د) مقایسه الگوریتم‌های یادگیری ماشین برای دسته‌بندی (استفاده از مدل اتوماتیک[1])

1 Auto Model.

ه) ایجاد مدل موردنظر برای دسته‌بندی[1]

و) ارزیابی مدل ارائه‌شده

به‌طور خلاصه اجزای اساسی سیستم پیشنهادی در تصویر زیر قابل مشاهده می‌باشد.



شکل(۱): مراحل مختلف ایجاد مدل طبقه‌بندی متون خبری

با توجه به اینکه این پژوهش از داده‌های جمع‌آوری‌شده پایگاه خبری فارسی بی بی سی استفاده می‌کند، در گروه پژوهش‌های موردی قرار می‌گیرد. همچنین به دلیل اینکه به بررسی و طبقه‌بندی متون خبری فارسی می‌پردازد، یک تحقیق کاربردی به‌حساب می‌آید. به‌علاوه چون از داده‌های بی‌بی‌سی آرشیو شده استفاده می‌کند، مقطعی است. طبقه‌بندی اسناد در این پژوهش، مبتنی بر الگوریتم‌های یادگیری ماشین است. الگوریتم‌های مورداستفاده در این پژوهش عبارت است از: ماشین بردار پشتیبان، بیزین ساده و کی-نزدیک‌ترین همسایه. شکل زیر نمایی کلی از فرآیند پیشنهادی پژوهش را ارائه می‌دهد.

مبانی نظری

متن‌کاوی

متن‌کاوی شاخه‌ای از داده‌کاوی یا همان کشف دانش است. ازنظر فییاد و همکارانش، کشف دانش، فرآیند غیربدیهی تشخیص الگوهای معتبر، نو، مفید و درنهایت قابل‌درک در داده‌هاست. متن‌کاوی از

1 Classification.

فناوری‌های بازیابی اطلاعات و استخراج اطلاعات استفاده می کند (Usama, et.al ۱۹۹۶,).

آماده‌سازی متون سیستم طبقه‌بندی اخبار فارسی

داده‌های ورودی به برنامه طبقه‌بند اخبار، متونی هستند که از سایت بی‌بی‌سی جمع‌آوری شده‌اند، با توجه به اینکه این متون در ویرایشگرهای فارسی متفاوتی تایپ شده است و نیز هنگام بارگذاری روی اینترنت ممکن است دچار تغییراتی شده باشد لازم است در ابتدا بازبینی کلی روی متون انجام شود. قسمتی از بازبینی به شیوه دستی انجام می‌شود، ولی در حجم تعداد بالا ویرایش دستی ممکن نیست و باید از روش‌های مرتبط استفاده نمود.

پیش‌پردازش متن

پس از آماده‌سازی اولیه متون، پیش‌پردازش انجام می‌شود. درواقع پیش‌پردازش، اولین گام برای تطابق مستندات متنی با نمایش آن‌ها در یک قالب مناسب می‌باشد. ثابت شده است که تنها ۳۳ درصد کلمات در متن مفید هستند و می‌توان از آن‌ها برای استخراج اطلاعات استفاده نمود (هاشمی، ۱۳۹۴).

بااینکه مجموعه گسترده‌ای از روش‌ها در پردازش زبان طبیعی به کار می‌روند، فناوری‌های به‌کار رفته را می‌توان به سه دسته کلی تقسیم نمود. روش‌های آماری، روش‌های ساختاری مبتنی بر الگو و روش‌های مبتنی بر استنتاج. باید توجه داشت که این راهکارها لزوماً از هم جدا نیستند. درواقع، جامع‌ترین مدل‌ها از ترکیب هر سه این روش‌ها استفاده می‌کنند. تفاوت این راهکارها در نوع عملیات پردازشی است که قادر به انجام آن هستند و میزان قواعدی که در مقابل آموزش یادگیری خودکار از روی داده‌های زبانی نیاز دارند (بهرام پور و همکاران، ۱۳۹۴).

برای کاوش کردن حجم قابل‌توجهی از اسناد، ضروری است که اسناد پیش‌پردازش شوند و اطلاعات در یک ساختار داده‌ای مناسب برای پردازش‌های بعدی ذخیره گردد. زمانی که داده‌های ورودی در دسترس است باید آن‌ها را برای ورود به الگوریتم‌های یادگیری ماشین آماده کنیم. این مرحله برای داده‌های متنی به این معناست که آن‌ها را از حالت غیر ساخت‌یافته به فرمت ساختاریافته و قابل‌تشخیص برای ماشین تبدیل کنیم (پرئی و همکاران، ۱۳۹۵).

پیش‌پردازش

شامل پیش‌پردازش متون و پیش‌پردازش ادبی است .پیش‌پردازش متون، بر روی مجموعه اسناد

جمع‌آوری‌شده، نشانه‌گذاری[1]، حذف کلمه توقف یافته[2] و ریشه‌یابی[3] انجام می‌شود. پیش‌پردازش ادبی برای افزایش اطلاعات عبارات به کار می‌رود. بدین منظور دیدگاه‌ها، برچسب‌گذاری ادات سخن، تکه کردن متن، ابهام‌زدایی حس کلام[4] و تجزیه کردن[5] را مکرر اعمال می‌کند (آقا کاردان و همکاران، ۱۳۹۱).

طبقه‌بندی متن

محوری‌ترین موضوع در حوزه متن‌کاوی، طبقه‌بندی متون است. اساس کار طبقه‌بندی متون بر پایه کلمات کلیدی و مهمی است که از مستندات استخراج می‌شود (کریمی منش و همکاران، ۱۳۹۲).

برای دسته‌بندی متن از فناوری‌های استخراج اطلاعات[6]، پردازش زبان طبیعی[7] و یادگیری ماشین[8] به‌طور وسیع استفاده می‌شود. به‌طورکلی هدف یک دسته‌بند متون، دسته‌بندی اسناد در قالب تعداد معینی از دسته‌های از پیش تعیین‌شده می‌باشد. هر سند می‌تواند در یک، چند و یا هیچ دسته‌ای قرار بگیرد. در مورد هر سند به این سؤال پاسخ داده خواهد شد که این سند در کدام‌یک از دسته‌ها قرار می‌گیرد. این موضوع می‌تواند در قالب یک یادگیری خودکار قرار گیرد تا با استفاده از آن بتوان هر سند را به‌طور خودکار به دسته‌ای نسبت داد. بعد از انتخاب مجموعه داده و پاک‌سازی متون به کمک روش نرمال شده فرکانس کلمه-معکوس فرکانس سند به ویژگی‌ها وزن داده می‌شود و در طی دو مرحله ویژگی‌ها با استفاده از روش فرکانس سند و مربع‌کای[9] انتخاب می‌شوند (سیاحی و همکاران، ۱۳۹۴).

هدف از طبقه‌بندی متون نسبت دادن کلاس‌های از پیش تعریف‌شده به اسناد متنی است. در طبقه‌بندی، یک مجموعه آموزشی از اسناد با کلاس‌های معین وجود دارد. با استفاده از این مجموعه، مدل طبقه‌بندی معین‌شده و کلاس سند جدید مشخص می‌گردد. برای کارایی مدل طبقه‌بندی، یک مجموعه آزمودن، مستقل از مجموعه آموزشی در نظر گرفته می‌شود. برچسب‌های تخمین زده‌شده

1. tokenization
2.Stop-word removal
3. temming
4.Word Sense Disambiguation(WSD)
5.Text Chunking
6.Information Extraction
7. Natural Language Processing(NLP)
8.Machine Learning
9.Chi Square

با برچسب واقعی اسناد مقایسه می‌شود. نسبت اسنادی که به‌درستی طبقه‌بندی شده‌اند به تعداد کل اسناد، دقت[1] نامیده می‌شود.

مراحل اصلی طبقه‌بندی متن

مراحل اصلی طبقه‌بندی متن در دو فاز که فاز اول شامل دو مرحله انتخاب ویژگی و وزن‌دهی و فاز دوم شامل تخمین عملکرد و انتخاب یک الگوریتم طبقه‌بندی می‌باشد دسته بندی می شود. (کریمی منش و همکاران، ۱۳۹۲).

کاربردهای طبقه‌بندی متن

دسته‌بندی متون کاربردهای فراوانی می‌تواند داشته باشد که ازجمله آن می‌توان به دسته‌بندی گفتاری که ترکیبی از دسته‌بندی متون و تشخیص گفتار است، دسته‌بندی متون چندرسانه‌ای از طریق عنوان‌های متنی، تشخیص نویسنده برای متون نامشخص با موردبحث، تشخیص زبان برای متونی که زبان آن‌ها نامشخص است، تشخیص خودکار جنس متن، بایگانی اسناد، دسته‌بندی صفحات وب، یادگیری خودکار علایق مطالعاتی و پژوهشی کاربران، فیلتر کردن خودکار پست الکترونیکی بر اساس محتوا و... اشاره نمود (جمالی و همکاران، ۱۳۹۶).

پردازش زبان طبیعی[2]

پردازش زبان طبیعی در دهه ۵۰ میلادی به‌عنوان فصل مشترک علوم هوش مصنوعی و زبان شناسی به وجود آمد و در دهه ۸۰ دچار تحولات و تغییرات اساسی شد (NadKarni, et.al,۲۰۱۱). پردازش متن تا دهه ۸۰ میلادی با استفاده از روش‌های قانون‌محور انجام می‌گرفت. اگرچه روش‌های قانون‌محور ویژگی‌های مثبتی از قبیل قابل درک بودن برای انسان و قابلیت ارتقا کیفیت در طول زمان را دارا هستند، مشکلات و محدودیت‌هایی نیز دارند. لازمه‌ی نوشتن قوانین در روش‌های قانون‌محور داشتن دانش عمیقی از آن حوزه است و از طرفی این روش‌ها بسیار زمان‌بر بوده و ازنظر محاسباتی پیچیده هستند (Moulinier, et.al,۱۹۹۵).

این مشکلات و محدودیت‌ها باعث شد در دهه ۹۰ میلادی روش‌های یادگیری ماشین یا به عبارتی روش‌های آماری محور در پردازش متون طبیعی محبوب‌تر از روش قانون‌محور ظاهر شوند. درواقع در دهه ۹۰ میلادی بود که پژوهشگران به روش‌های یادگیری ماشین توجه بسیاری نشان دادند که این

1.precision
2.Natural language processing(NLP)

به دلیل توسعه‌ی سخت‌افزاری و نرم‌افزاری علم رایانه اتفاق افتاد (Eyheramendy, et.al,۲۰۰۳).

پژوهش‌ها نشان می‌دهند که طبقه‌بندی متن از نخستین تحقیقات انجام گرفته در حوزه‌ی متن‌کاوی است و نقش مهمی در آن ایفا می‌کند. به گفته‌ی فرانسیس تاریخچه پردازش متن به اواخر دهه ۵۰ میلادی بازمی‌گردد (Francis, ۲۰۰۶). همچنین به گفته‌ی گلباخ و بلشاکو طبقه‌بندی متن از دهه ۶۰ میلادی آغاز شده و در دهه‌های ۸۰ و ۹۰ میلادی توجه محققان بیشتری را به خود جلب کرده است(BolshaKov, et.al,۲۰۰۴).

هدف کلی پردازش زبان طبیعی رسیدن به یک درک بهتر از زبان طبیعی توسط کامپیوترهاست. کامپیوترها فناوری‌های مستحکم و ساده‌ای برای پردازش سریع متن به کار می‌برند. هم‌چنین از فناوری‌های آنالیز زبان‌شناسی نیز برای پردازش متن استفاده می‌کنند. نق ش پردازش زبان طبیعی در متن‌کاوی فراهم کردن یک سیستم در مرحله استخراج اطلاعات با داده‌های زبانی است.

پیشینه پژوهش

پیشینه پژوهش در قالب جدول شماره ۱، شامل نام پژوهشگر، عنوان پژوهش، یافته های پژوهشی و سال پژوهشی قابل دسترسی می باشد.

جدول شماره (۱) پیشینه پژوهش

پژوهشگر	سال	عنوان پژوهش	یافته‌های پژوهش
سمیه عربی نرئی و همکاران	۱۳۸۶	استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی	از ترکیب روش‌های Wordnet و Porter استفاده کرده‌اند. برای استخراج کلمات کلیدی از ساختار درهم‌ریزی Trie و دادگان مورداستفاده، دادگان همشهری بوده است.
محمد احسان بصیری	۱۳۸۶	مقایسه دسته‌بندی متون فارسی با استفاده از الگوریتم‌های KNN و FKNN و انتخاب ویژگی‌ها بر اساس بهره اطلاعات و فرکانس سند	در این مقاله دو الگوریتم KNN و FKNN با دو روش استخراج ویژگی IG و DF آزموده می‌شود که بهترین ترکیب استفاده از روش FKNN و IG است. با این ترکیب به میانگین دقت ۸۰٪ می‌رسیم.
مسلم محمدی و همکاران	۱۳۸۷	استفاده از شبکه‌های عصبی CC٤ برای رده‌بندی اسناد فارسی	در مرحله پیش‌پردازش کلمات عمومی حذف و سایر کلمات ریشه‌یاب می‌شوند. سپس با استفاده از روش مبتنی بر فراوانی کلمه، اسناد با اندازه‌های مختلف به یک فضای K بعدی با اندازه‌ی ثابت نگاشت می‌شوند. دقت روی دادگان ایسنا حدود ۹۰٪ گزارش شده است.
ایوب باقری و همکاران	۱۳۸۷	دسته‌بندی متون خبری فارسی با استفاده از الگوریتم بیز ساده	در این مقاله با استفاده از بیز ساده و کاهش فضای حالت با روش واریانس فرکانس مدت دسته‌بندی قابل قبولی ارائه شد.
بهاره بیتا و همکاران	۱۳۸۶	طبقه‌بندی خودکار متون فارسی	از روش‌های tri-gram و quad-gram با معیارهای اندازه‌گیری فاصله منهتن، اندازه‌ی دایس و ضرب نقطه‌ای، به همراه روش یادگیری KNN موردبررسی قرارگرفته و بهترین نتیجه از ترکیب quad-gram و ضرب نقطه‌ای به دست آمد.
نوشین مقبودی و همکاران	۱۳۸۷	ارائه روشی جدید در طبقه‌بندی متون فارسی با استفاده از دانش معنایی	با استفاده از معیارهای ICF و Uni ، ویژگی‌ها را انتخاب کرده و با استفاده از گنج‌واژه و وزن‌دهی TFIDF و SVM به دسته‌بندی متون پرداخته شده و به نتایج پایداری رسیده است.

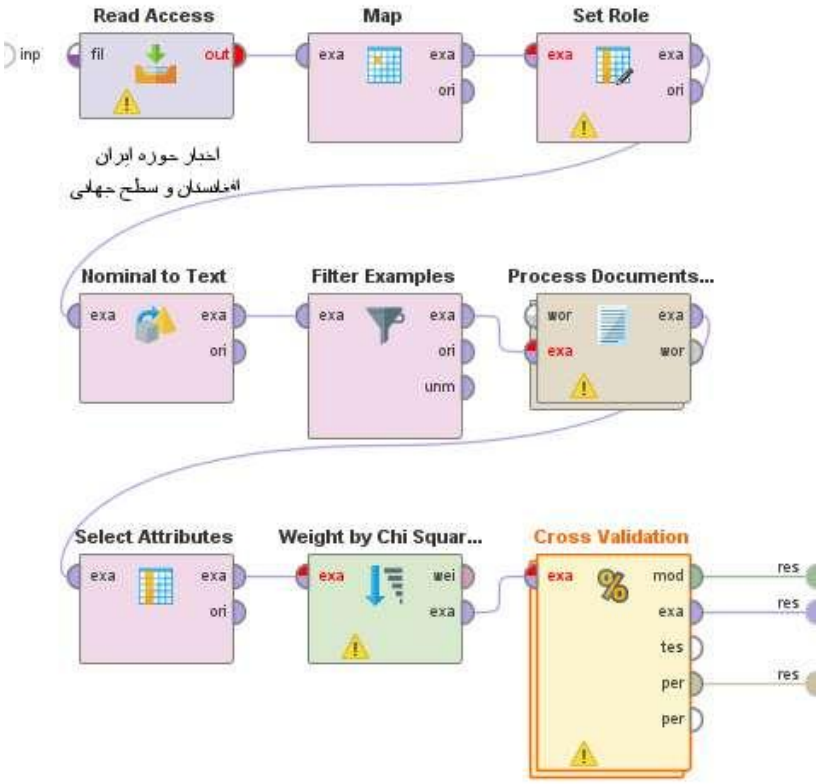
پژوهشگر	سال	عنوان پژوهش	یافته‌های پژوهش
علی قنبری سرخی	۱۳۹۰	بهبود عملکرد طبقه‌بندی متون فارسی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی با کمک معیار میانگین یادآوری و دقت	نتایج به‌دست‌آمده نشان داده است که با در نظر گرفتن روش‌های دسته-بندی KNN و Bayesian در روش پیشنهادی بهبود قابل‌توجهی در طبقه‌بندی متون فارسی و کاهش مدت‌زمان آزمون با ویژگی‌های استخراج‌شده به دست خواهد آمد.
حمید حسن پور	۱۳۹۱	استخراج بهترین ویژگی از متون فارسی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی با کمک میانگین یادآوری و الگوریتم ژنتیک	روش وزن‌دهیTFCRF را با استفاده از معیارهای ارتباط مثبت و منفی، به‌کاربرده‌اند. آن‌ها با دو دسته‌بند KNN و بیزین به دسته‌بندی متون پرداختند و نسبت به کارهای قبلی نتایج بهتری گرفتند
محسن طاهری نیا	۱۳۹۱	دسته‌بندی متون فارسی با استفاده از یادگیری نیمه نظارت‌شده	این پژوهش از ترکیبی از مثال‌های آموزشی برچسب دار و بدون برچسب برای یادگیری استفاده می‌کند و از این فن برای دسته‌بندی متون فارسی استفاده کرده و نتایج خوبی بدست آمد.
محسن زمانی و همکاران	۱۳۹۲	دسته‌بندی متون فارسی با استفاده از روش آنالیز معنایی پنهان احتمالاتی	از روش فاصله‌یابی اقلیدسی در فضای ماتریس‌های کاهش بعد استفاده کرده است. نتایج حاصل‌شده نشان می‌دهد که این روش در بهبود عملکرد سیستم نقش مؤثری دارد.
مهدی برقامی و همکار	۱۳۹۲	استفاده از ترکیب شبکه‌های عصبی جهت دسته‌بندی متون فارسی مبتنی بر الگوریتم‌های GA، کی-نزدیکترین همسایه، PCA جهت کاهش ویژگی	از ترکیب دو شبکه عصبی پرسپترون MLP در دسته‌بندی مستندات XML بر روی پایگاه داده روزنامه همشهری استفاده کرده‌اند.از روش TFCRF برای وزن‌دهی و PCA برای ارزش‌دهی ویژگی‌ها و از الگوریتم KNN برای دسته بندی استفاده شده است.
مائده عابدینی نیا و همکاران	۱۳۹۲	کاوش متون فارسی در وب با استفاده از تحلیل معنایی و روش طبقه‌بندی چندلایه‌ای اطلاعات	از مدل فضای برداری برای دسته‌بندی متون فارسی در وب استفاده شده است. روش پیشنهادی این امکان را می‌دهد که با دقت و سرعت بیشتری متن خود را در وب بیابند.
حمید حسن پور و همکار	۱۳۹۳	بهبود دقت سیستم دسته‌بندی خودکار اسناد فارسی به کمک هستان‌شناسی فارس‌نت	معیار χ^2 برای انتخاب ویژگی و روش وزن‌دهی TFIDF در وزن‌دهی به کار گرفته شده است. نتایج نشان داده است که این روش یک گام مؤثر در بهبود کارایی دسته‌بند است.
پیمان جلالی و همکاران	۱۳۹۴	ارائه‌ی روش دسته‌بندی متون با تکنیک کاهش ویژگی فیلتری و یادگیری ماشین	یک روش جدید انتخاب ویژگی بر اساس الگوریتم ژنتیک ارائه می‌شود. اثربخشی روش پیشنهادی با استفاده از الگوریتم بیزین ساده و طبقه‌بندی انجمنی در سه مجموعه مختلف از داده‌های متن عربی ارزیابی می‌شود.

پژوهشگر	سال	عنوان پژوهش	یافته‌های پژوهش
عارف سیاحی و همکاران	۱۳۹۴	بهبود یک روش مبتنی بر انتخاب ویژگی به‌منظور دسته‌بندی متون با الگوریتم‌های متن‌کاوی	از الگوریتم انتخاب ویژگی فیلتری، برای کاهش پیچیدگی دسته‌بندی و از الگوریتم‌های یادگیری بیز ساده، درخت تصمیم و ماشین بردار پشتیبان برای ارزیابی و بهبود کارایی استفاده شده است.
ایمان ابراهیمی و همکاران	۱۳۹۴	رده‌بندی متون فارسی با استفاده از ماشین بردار پشتیبان مبتنی بر روش‌های انتخاب ویژگی PCA و الگوریتم ژنتیک	ابتدا با استفاده از روش TFCRF به کلمات وزن‌دهی انجام می‌شود سپس از روش‌های PCA و الگوریتم ژنتیک استفاده شده است. نتایج تجربی نشان می‌دهد که این روش می‌تواند با دقت ۷۸ ٪ عمل رده‌بندی را انجام دهد.
محبوبه صبایی	۱۳۹۴	الگوریتمی جدید برای طبقه‌بندی مستندات، مبتنی بر وزن‌دهی به ویژگی‌ها و فایل‌ها	با استفاده از روش PCA ابعاد ویژگی‌ها کاهش داده می‌شود سپس با استفاده از SVM به پیاده‌سازی مدل پیشنهادی می‌پردازیم و درنهایت صحت مدل با روش اعتبار سنجی ۱۰ مرحله‌ای به صحت ۹۱،۸۶ می‌رسد که نسبت به کارهای پیشین انجام گرفته صحت بالاتری دارد.
عبدالحمید کهرازهی	۱۳۹۷	بهینه‌سازی نتایج موتورهای جستجو برای مفاهیم تخصصی حوزه پزشکی با به‌کارگیری روش‌های پردازش متن و داده‌کاوی	ابزار مورداستفاده تانگرا و متلب بوده و دقت روش پیشنهادی نسبت به الگوریتم KNN ۰۰۰۵،۱ ٪ ، نسبت به شبکه عصبی ۰۰۱۳،۱ ٪ و نسبت به الگوریتم نایوبیز ۰۰۳۱،۱ ٪ بهبود داشته است.
محمد بهروزیان نژاد و همکاران	۱۳۹۳	استفاده از تکنیک‌های داده‌کاوی در دسته‌بندی خودکار اسناد متنی	روش پیشنهادی از روش فیلتری بهره می‌برد و با استفاده از دسته‌بندهای بیز ساده و درخت تصمیم پیاده‌سازی شده است. نتایج نشان‌دهنده برتری روش ترکیبی نسبت به دسته‌بندهای منفرد می‌باشد.
سیده زهرا سجادی و همکاران	۱۳۹۷	دسته‌بندی روش‌های توصیف متن در عقیده کاوی از شبکه‌ی اجتماعی توییتر	برای استفاده از روش‌های یادگیری ماشین نیازمند استخراج ویژگی از متون می‌باشیم که در این مقاله به دسته‌بندی انواع روش‌های مختلف برای توصیف متن در عقیده کاوی از شبکه‌ی اجتماعی توییتر پرداخته شده است.
سعید حمدالهی اسکویی و همکار	۱۳۹۷	نظر کاو، کاوش نظرات و تحلیل احساسات در زبان فارسی	با استفاده از برچسب‌گذاری POS برای جملات زبان فارسی و سپس دادن امتیاز به هر ویژگی و تعیین مثبت، منفی و یا خنثی بودن ویژگی و در نهایت تعیین جهت‌گیری معنایی کل نظر با استفاده از امتیاز تمام ویژگی‌ها ارائه شده است.
سید محمد جوادی مقدم و همکار	۱۳۹۷	وب کاوی و متن‌کاوی با رویکرد ماشین یادگیری	در ابتدا با مفاهیم اساسی و کاربردی وب کاوی و متن‌کاوی و آنالیز احساس پرداخته شده و در ادامه این مفاهیم را با رویکرد و فناوری‌های یادگیری ماشین بررسی می‌نماید.

پژوهشگر	سال	عنوان پژوهش	یافته‌های پژوهش
ملیکا تجاری و همکار	۱۳۹۷	ارائه یک روش جدید مبتنی بر ماشین بردار پشتیبان در جهت طبقه‌بندی تراکنش‌های موبایل	در این مقاله تمرکز بر روی استفاده از SVM در طبقه‌بندی تراکنش‌های موبایل است. برای این منظور عمل انتخاب ویژگی با استفاده از الگوریتم PCA انجام می‌شود. برای مقایسه روش پیشنهادی، از شبکه عصبی پرسپترون چندلایه استفاده شده است. نتایج نشان می‌دهد ماشین بردار پشتیبان با میانگین مربع خطای ۰،۲۳۶ و دقت ۹۴،۱ درصد به ازای کل داده‌ها طبقه‌بندی تراکنش‌های موبایل را انجام می‌دهد.
شقایق سوداگر طالقانی و همکار	۱۳۹۷	تشخیص ویژگی‌های مؤثر برای شناسایی بیماری قلبی با استفاده از فیلترهای انتخاب ویژگی و طبقه‌بندی آن‌ها به کمک تکنیک‌های داده‌کاوی	در این مقاله از دسته‌بند Naive Bayes استفاده شده است. هم‌چنین از دو فیلتر انتخاب ویژگی Relief و Wrapper استفاده شده است.
الهام صدری و همکار	۱۳۹۷	ارائه الگوریتمی برای بازشناسی ارقام دست‌نویس فارسی با استفاده از ترکیب طبقه بندها	نتایج آزمایش‌ها بر روی پایگاه داده‌ای شامل ۶۰۰۰۰ تصویر اعداد نشان می‌دهد که روش همجوشی بر اساس این الگو بهتر از سایر طرح‌ها می‌باشد.
حدیث پورعباسی و همکار	۱۳۹۶	دسته‌بندی متون لاتین با روش‌های انتخاب ویژگی فیلتری، پوششی و الگوریتم‌های بیز ساده	در مرحله یادگیری از الگوریتم‌های خانواده بیز ساده استفاده شده است. روش پیشنهادی در مقایسه با کارهای گذشته در این زمینه بهبود بسیار خوبی داشته است.

تجزیه و تحلیل داده‌ها

به منظور فهم بیشتر روش پیشنهادی در ابتدا نمای کلی از فرآیند پیشنهادی در نرم افزار رپید ماینر را مشاهده می نمایید که ابتدا مجموعه داده اکسس خوانده، سپس برای ستون‌های آن نقش تعریف می گردد و محتوای خوانده‌شده به‌صورت متن تبدیل می‌شود تا قابل توکنیزه شدن توسط رپیدماینر باشد. در ادامه مراحل انجام کار شرح داده می‌شود.



شکل (۲): نمایی کلی از فرآیند پیشنهادی در نرم‌افزار رپیدماینر

جمع‌آوری اخبار

پایگاه خبری مورد نظر، آرشیو اخبار بی‌بی‌سی می‌باشد که این مجموعه داده به‌وسیله یک خزنده که به زبان پایتون نوشته شده است جمع آوری شده است. این خزنده با یک حلقه while مدام در حال جمع آوری اخبار از سایت BBC بوده و اخبار مربوطه را در محل از قبل تعیین شده ذخیره می‌نماید. اخبار جمع آوری شده در سه دسته خبری حوزه ایران، حوزه افغانستان و سطح جهانی توسط برچسب مشخص شده است و آموزش و آزمایش روی آن صورت خواهد گرفت. شکل زیر نمایی از مجموعه داده اخبار بی‌بی‌سی را ارائه می‌دهد.

	A	B
1	category	body
2	اخبار سطح جهانی	یک نظرسنجی تازه جهانی نشان دهنده عمق اضطرابی است که بسیاری جوانان درباره تغییر اقلیم احساس می کنند. تقریبا ۶۰ درصد جوانانی که در این بررسی مورد پرسش قرار گرفتند گفتند که خیلی یا شدیدا احساس نگرانی می کنند.
3	اخبار سطح جهانی	جمع شدن حدود ۱۰ هزار پناهجو در اطراف پل مرزی میان مکزیک و آمریکا به شدت گرفتن بحران انسانی در این منطقه منجر شده است. این پل شهر دل ریو، تگزاس را به شهر آکونیا در مکزیک متصل می‌کند و اردوگاه موقتی که در کنار آن شکل گرفته، در روزهای اخیر به شدت گسترش یافته است.
4	اخبار حوزه ایران	اعتراض‌ها در بهبهان؛ انتقاد از سیاست منطقه‌ای ایران در شهری کوچک تازه ترین اعتراض‌ها در ایران در شهری کوچک در استان خوزستان نمود پیدا کرد، در بهبهان. اتفاقا یکی از میدانهای مرکزی بهبهان در اعتراض‌های خونین آبان سال گذشته هم کانونی ملتهب بود.
5	اخبار حوزه ایران	آیت الله علی خامنه ای با صدور حکمی سرتیپ خلبان حمید واحدی را به عنوان فرمانده جدید نیروی هوایی ارتش منصوب کرده است. او در این حکم خواستار "ارتقای توان و آمادگی‌های رزمی نیروی هوایی شده است. "
6	اخبار سطح جهانی	نیم قرن پس از آن که آمریکا به عنوان نخستین کشور جهان، پرچم خود را روی کره ماه نصب کرد، پرچم چین به عنوان دومین کشور جهان روی این تنها قمر کره زمین به اهتزاز در آمد.تصاویری که اداره ملی فضانوردی چین منتشر کرده پرچم سرخ و ۵ ستاره این کشور را نشان می‌دهد.
7	اخبار سطح جهانی	دادگاهی در لندن می‌گوید شیخ محمد آل مکتوم، حاکم دبی، در جریان دعوای حقوقی با هیا بنت حسین، همسر سابق خود، برای گرفتن حق حضانت فرزندانشان دستور هک کردن تلفن او را داده است. به گفته دادگاه برای این کار از نرم‌افزار اسرائیلی "پگاسوس" استفاده شده است.
8	اخبار سطح جهانی	در چارچوب طرح کوواکس سازمان بهداشت جهانی، شب گذشته یک پرواز ویژه "سامان ایر" اولین محموله واکسن‌های "آسترازنکا" را از هند وارد دوشنبه، پایتخت تاجیکستان، کرد.به گفته مسئولان وزارت بهداشت و تأمین اجتماعی تاجیکستان، در مرحله نخست ۱۹۲ هزار دوز از این واکسن وارد شده است.
9	اخبار حوزه افغانستان	عمران خان، نخست وزیر پاکستان، در یک مصاحبه اختصاصی با بی‌بی‌سی گفته است که پاکستان در تعامل با همسایگان افغانستان و بطور "جمعی" درمورد به رسمیت شناختن یا نشناختن دولت طالبان تصمیم خواهند گرفت.
10	اخبار حوزه ایران	سعید خطیب‌زاده سخنگوی وزارت خارجه ایران در نشست هفتگی با رسانه‌ها گفته است که مذاکرات ایران و عربستان "به مسیر جدی‌تری وارد شده" و "نقشه راه و مسیری در حال گفت‌وگو است"به گفته وی "توافقات هیات اعزامی ایران با عربستان این بوده که در فضای عمومی این مذاکرات منتشر نشود"
11	اخبار حوزه ایران	وزارت خزانه‌داری آمریکا اعلام کرد که سه نفر به نام‌های دانیل بهزاد فردوس، مانوئل مهرزاد فردوس و محمد رضا دزفولیان را از لیست افراد تحریم‌شده توسط این کشور خارج شده‌اند، و در عین حال تاکید کرده است که تغییر در لیست تحریم‌ها ارتباطی با مذاکرات وین ندارد.
12	اخبار حوزه افغانستان	کمیسیون دسترسی به اطلاعات افغانستان از ریاست جمهوری این کشور خواسته است که اسدالله خالد، سرپرست وزارت دفاع و احمد جاوید رسولی، رئیس عمومی اداره ملی احصائیه و معلومات(مرکز آمار) این کشور را به دلیل عدم همکاری با رسانه‌ها برای یک ماه کسر حقوق کند.
13	اخبار حوزه ایران	ده ماه بعد از اعدام خضر قویدل در زندان ارومیه، دیوان عالی کشور حکم اعدام او را لغو کرده است. یکی از اعضای خانواده قویدل به بی‌بی‌سی فارسی گفت از شنیدن خبر لغو حکم او چند ماه پس از اعدامش جا خورده‌اند و غمشان بسیار افزونتر شده و با شکایت از قاضی شیخلو اعتراض خود را شروع کرده‌اند.

شکل(۳) نمایی از مجموعه داده اخبار بی‌بی‌سی

آماده‌سازی و پیش‌پردازش

در فاز آماده‌سازی متون، فرمت فایل از جیسون[1] به اکسس[2] تبدیل می‌شود، سپس متن سند که شامل کاراکترهای پشت سر هم است به نمایشی که برای الگوریتم‌های یادگیری و طبقه‌بندی مناسب باشد تبدیل می‌شود. این کار توسط پلاگین‌های[3] رپیدماینر[4] و همچنین قطعه کدی که به زبان پایتون نوشته شده است انجام می‌گیرد. در این کد از کتابخانه هضم استفاده شده است. این فرآیند شامل موارد ذیل است:

حذف تگ‌های اچ‌تی‌ام‌ال[5] و یا ایکس‌ام‌ال[6]

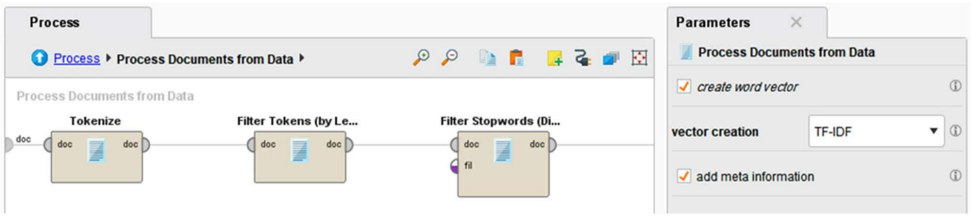
کدگذاری متون به یونی‌کد[7]

به دست آوردن ریشه کلمات و حذف پیشوندها و پسوندها

1.json
2.Access(*.accdb)
3.Plugin
4.Rapid Miner
5.html
6.xml
7 UTF-8.

توکنیزه کردن اطلاعات

به‌منظور توکنیزه کردن اطلاعات و تولید بردارهای قابل‌پردازش، از روش وزن‌دهی کلمات معکوس فراوانی متن استفاده شده است. این کار به کمک ابزار پردازش متن رپیدماینر به نام توکنایز انجام می‌شود. تصویر زیر خروجی مورد نظر را نمایش می‌دهد. دقت کنید که تمامی کنترل‌های موجود در تصویر در کامپوننت پردازش سند روی داده می‌باشند. همان‌طور که در تصویر قابل‌مشاهده می‌باشد از الگوریتم پردازش متن معکوس فراوانی متن استفاده شده است.



شکل(۴): کاربرد الگوریتم پردازش متن TF-IDF

در جدول زیر فهرستی از فراوانی هر لغت در دسته مربوطه پس از عملیات توکنیزه کردن نمایش داده شده است.

جدول(۲) نمونه‌ای از فراوانی هر کلمه در سه دسته خبری

اخبار سطح جهانی	اخبار حوزه ایران	اخبار حوزه افغانستان	کلمه توکن شده
۳۷	۳۱	۷	ترامپ
۷	۱۷	۱۱	تروریستی
۹	۹	۱۳	تروریست
۳۲	۱۴	۱۹	ترک

کاهش ابعاد

به علت سنگینی پردازش‌های موجود بر روی متون و حذف حشویات و مواردی که باعث عدم خوانایی برنامه و کاهش دقت کار می‌شود، از روش کاهش ابعاد استفاده می‌شود. برای این منظور از ابزار فیلتر توکنز استفاده می‌شود. این فیلتر یک ورودی متن می‌گیرد و خروجی آن هم فیلتر شده همان متن می‌باشد. برای کاهش ابعاد کار، توکن‌های کمتر از ۴ و بیشتر از ۲۵ را حذف کردیم. این کار به افزایش صحت کار کمک بیشتری می‌کند.

همان‌طور که در جدول ۳ مشاهده می‌کنید تعداد بار تکرار هر یک از کلمات پردازش‌شده توسط

الگوریتم مربوطه در هر سه دسته ایران، افغانستان و سطح جهانی قابل مشاهده می‌باشد.

همچنین برای کاهش اندازه ویژگی‌ها و بهینه‌تر شدن انتخاب ویژگی‌ها از عملگر قدرتمند وزن‌دهی با مربع کای استفاده شد. خروجی این عملگر یک مجموعه از اوزان کلمات می‌باشد که با توجه به فناوری‌های مربع کای[1] تولید و مورد استفاده قرار می گیرد.

جدول (۳): نمونه‌ای از کلمات پس از وزن‌دهی با روش مربع‌کای

آزاد	آذربایجان	آخرین	آثار	آنتن	آبان	دسته
۰	۰	۰	۰	۰,۰۰۷	۰	اخبار سطح جهانی
۰	۰	۰	۰	۰	۰	اخبار سطح جهانی
۰	۰	۰	۰	۰	۰	اخبار سطح جهانی
۰	۰	۰	۰	۰,۰۳۹	۰,۰۳۹	اخبار حوزه ایران
۰	۰	۰	۰	۰	۰	اخبار حوزه ایران
۰	۰	۰,۰۰۸	۰	۰	۰	اخبار سطح جهانی

ایجاد مدل طبقه‌بندی به کمک الگوریتم‌های یادگیری ماشین

بیشتر سامانه‌های طبقه‌بندی خودکار متون، برای متون زبان انگلیسی طراحی شده‌اند و معمولاً برای متون فارسی به دلیل ماهیت زبان فارسی و در دسترس نبودن مجموعه ریشه کلمات قابل‌استفاده نیستند (بینا و همکاران، ۱۳۸۶). اغلب روش‌های استفاده شده در این زمینه، روش‌های مبتنی بر یادگیری هستند. از مهم‌ترین روش‌های استفاده شده در طبقه‌بندی متون که در این پژوهش نیز استفاده شده است می‌توان به بیزین ساده، ماشین بردار پشتیبان و کی نزدیک‌ترین همسایه اشاره کرد.

ارزیابی

معمولاً برای ارزیابی صحت دسته‌بندی کننده اسناد از معیار صحت[2]، معیار فراخوانی[3]، معیار دقت[4] و معیار ارزیابی F[5] استفاده می‌شود. دقت مربوط به‌درستی نسبت دادن داده‌های آزمون به کلاس

1 Weight By Chi Squared.
2 Precision.
3 Recall.
4. Accuracy
5 F-measure (F-score)

مربوطه است و معیار ارزیابیF ترکیب هماهنگ از دقت و فراخوانی است. معادلات مربوطه در جدول زیر که موسوم به ماتریس درهم‌ریختگی[1] است آمده است. برای درک دو معیار ارزیابی دقت و فراخوانی و همچنین معیار صحت، ابتدا باید با عبارات مثبت واقعی[2]، مثبت کاذب[3]، منفی واقعی[4] و منفی کاذب[5] آشنا بود. این چهار عبارت برای مقایسه‌ی برچسب کلاس‌ها در ماتریس درهم‌ریختگی استفاده می‌شوند. جدول ۴ این عناوین را نشان می‌دهد.

جدول (۴) ماتریس درهم‌ریختگی

برچسب		برچسب واقعی	
		مثبت	منفی
برچسب پیش‌بینی شده	مثبت	(TP)مثبت واقعی	(FP)مثبت کاذب
	منفی	(FN)منفی کاذب	(TN)منفی واقعی

صحت: نسبت مقدار موارد صحیح طبقه‌بندی‌شده توسط الگوریتم از یک کلاس مشخص، به‌کل تعداد مواردی که الگوریتم چه به‌صورت صحیح و چه به‌صورت غلط، در آن کلاس طبقه‌بندی کرده است که به‌صورت زیر محاسبه می‌شود:

فرمول ۱:

$$\Pr ecision = \frac{TP}{TP + FP}$$

فراخوانی: نسبت مقداری موارد صحیح طبقه‌بندی‌شده توسط الگوریتم از یک کلاس به تعداد موارد حاضر در کلاس مذکور که به‌صورت زیر محاسبه می‌شود:

فرمول ۲:

$$\mathrm{Re}call = \frac{TP}{TP + FN}$$

دقت: درواقع این معیار عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های طبقه‌بندی است که نشان می‌دهد، طبقه‌بند طراحی شده چند درصد از کل مجموعه اسناد را به‌درستی دسته‌بندی کرده است.

1 Confusion Matrix
2. True Positive
3. False Positive
4. True Negative
5. False Negative

فرمول ۳:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

معیار ارزیابی F: با توجه به محاسبات انجام‌گرفته برای معیارهای دقت و فراخوانی، در این مرحله می‌توان مقدار کمیت وزن‌دار معیار ارزیابی F را محاسبه نمود. معیار ارزیابی F، پارامتر مناسبی برای ارزیابی کیفیت کلاس‌بندی می‌باشد و همچنین توصیف‌کننده میانگین وزن‌دار مابین دو کمیت دقت و فراخوانی می‌باشد. برای یک الگوریتم کلاس‌بندی کننده در شرایط ایده آل، مقدار این کمیت برابر با ۱ می‌باشد و در بدترین وضعیت برابر با صفر می‌باشد. این پارامتر با توجه به رابطه زیر محاسبه می‌شود:

فرمول ۴:

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall}$$

ابزار مورد استفاده

در این تحقیق از مجموعه داده اخبار بی‌بی‌سی فارسی به‌عنوان مجموعه آموزش و مجموعه آزمون استفاده شده است. این پیکره شامل ۱۷۰۰ خبر بین سال‌های ۱۴۰۰ تا ۱۴۰۱ می‌باشد. برای افزایش دقت، فراخوانی و هم‌چنین کارایی کلی الگوریتم‌های طبقه‌بندی از روش‌های متعدد پیش‌پردازش از قبیل حذف کلمات اضافه و اصلاح فاصله به نیم‌فاصله در ساخت مجموعه آموزش استفاده کرده و از سه روش طبقه‌بندی ماشین بردار پشتیبان، بیز ساده و کی-نزدیکترین همسایه برای طبقه‌بندی اسناد بهره برده‌ایم.

برای انجام مراحل پیش‌پردازش، وزن‌دهی، انتخاب ویژگی و طبقه‌بندی از نرم‌افزار رپیدماینر نسخه ۹,۶ استفاده کرده‌ایم.

استفاده از ابزار مدل اتوماتیک

روش استفاده از مدل اتوماتیک به این صورت است که ابتدا داده مورد نظر را به عنوان ورودی معرفی می‌کنیم، سپس از پلاگین پیش‌پردازش برای تمیز کردن داده‌ها استفاده می‌شود و معیارهای ارزیابی هر یک از الگوریتم‌های یادگیری ماشین را به تفکیک ارائه کرده و بهترین الگوریتم را به لحاظ دقت، صحت، زمان اجرا و ... پیشنهاد می‌دهد. در ادامه نتایج هر یک از الگوریتم‌ها آورده شده است.

بررسی معیار دقت الگوریتم بیز ساده

در این بخش تاثیر روش مربع کای و نرمال‌سازی کلمات بر روی عملکرد این الگوریتم را به صورت جداگانه بررسی می کنیم. فرآیند انتخاب ویژگی در بهبود عملکرد الگوریتم طبقه‌بند به خصوص در الگوریتم طبقه‌بند بیز ساده که به فرآیند انتخاب ویژگی بسیار وابسته است، تاثیر زیادی دارد بنابراین پیش‌بینی می‌شود که عملکرد الگوریتم طبقه‌بند بیز ساده با اعمال روش مربع کای بهبود یابد. جدول ۵ میانگین معیار دقت را برای الگوریتم طبقه‌بند بیز ساده در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرآیند انتخاب ویژگی مربع کای نشان می‌دهد.

جدول(۵):بررسی‌تاثیر نرمال‌سازی کلمات و روش مربع‌کای بر الگوریتم بیز ساده

معیار ارزیابی	بیز ساده و عدم نرمال‌سازی	بیز ساده و اعمال نرمال‌سازی کلمات	بیز ساده و مربع‌کای
میانگین دقت	٪۵۲,۴	٪۵۳,۷	۰/۵۶,۲

در شکل زیر نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم بیزساده و مربع کای آمده است.

Performances

Criterion	Value	Standard Deviation
Accuracy	56.2%	± 3.0%
Classification Error	43.8%	± 3.0%

Confusion Matrix

	true اخبار سطح جهانی	true اخبار حوزه ایران	true خبار حوزه افغانستان	class precision
pred. اخبار سطح جهانی	68	14	12	72.34%
pred. اخبار حوزه ایران	143	203	44	52.05%
pred. اخبار حوزه افغانستان	0	0	2	100.00%
class recall	32.23%	93.55%	3.45%	

شکل(۵) سایر معیارها برای الگوریتم بیز ساده

بررسی معیار دقت الگوریتم کی-نزدیکترین همسایه

در این بخش‌تاثیر نرمال‌سازی کلمات و روش مربع کای بر روی عملکرد الگوریتم کی-نزدیکترین همسایه را به صورت جداگانه بررسی می کنیم. جدول ۶ میانگین معیار دقت را برای الگوریتم کی-نزدیکترین همسایه در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرآیند انتخاب ویژگی مربع کای نشان می‌دهد.

جدول(۶) : بررسی تاثیر نرمال‌سازی کلمات و روش مربع کای بر الگوریتم کی-نزدیکترین همسایه

معیار ارزیابی	کی-نزدیکترین همسایه و عدم نرمال‌سازی کلمات	کی-نزدیکترین همسایه و اعمال نرمال‌سازی کلمات	کی-نزدیکترین همسایه و مربع کای
میانگین دقت	۸۶,۱٪	۸۷,۳٪	۸۹,۱٪

در شکل زیر نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم کی-نزدیکترین همسایه و مربع کای آمده است.

Criterion	Value	Standard Deviation
Accuracy	89.1%	± 2.7%
Classification Error	10.9%	± 2.7%

Confusion Matrix

	true اخبار سطح جهانی	true اخبار حوزه ایران	true اخبار حوزه افغانستان	class precision
pred. اخبار سطح جهانی	182	24	9	84.65%
pred. اخبار حوزه ایران	15	188	1	92.16%
pred. اخبار حوزه افغانستان	2	2	64	94.12%
class recall	91.46%	87.85%	86.49%	

بررسی معیار دقت الگوریتم ماشین بردار پشتیبان

در این بخش تاثیر نرمال‌سازی کلمات و روش مربع کای بر روی عملکرد الگوریتم ماشین بردار پشتیبان را به صورت جداگانه بررسی می کنیم. جدول ۷ میانگین معیار دقت را برای الگوریتم ماشین بردار پشتیبان در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرآیند انتخاب ویژگی مربع کای نشان می‌دهد.

جدول (۷) : بررسی تاثیر نرمال‌سازی کلمات و روش مربع کای بر الگوریتم ماشین بردار پشتیبان

معیار ارزیابی	ماشین بردار پشتیبان و عدم نرمال‌سازی کلمات	ماشین بردار پشتیبان و اعمال نرمال‌سازی کلمات	ماشین بردار پشتیبان، اعمال نرمال‌سازی کلمات و مربع‌کای
میانگین دقت	۸۷،۵٪	۸۸،۸٪	۹۰،۸٪

در شکل زیر نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم ماشین بردار پشتیبان و مربع کای آمده است.

Performances

Criterion	Value	Standard Deviation
Accuracy	90.8%	± 2.6%
Classification Error	9.2%	± 2.6%

Confusion Matrix

	true اخبار سطح جهانی	true اخبار حوزه ایران	true اخبار حوزه افغانستان	class precision
pred. اخبار سطح جهانی	183	18	8	87.56%
pred. اخبار حوزه ایران	14	195	2	92.42%
pred. اخبار حوزه افغانستان	2	1	64	95.52%
class recall	91.96%	91.12%	86.49%	

شکل (۷) : سایر معیارها برای الگوریتم ماشین بردار پشتیبان

محدودیت‌های تحقیق

جمع‌آوری اخبار از سایت‌های مختلف یا سایت واحد نیاز به یک خزنده با قابلیت جمع‌آوری متون فارسی دارد که خزنده‌های آنلاین نیز با توجه به تحریم روی سایت‌های فارسی، بدون فیلتر اعمال نمی‌شوند و استفاده از پراکسی باعث کند شدن روند فعالیت خزشگر می‌شود. همچنین یک مجموعه واحد از ریشه لغات زبان فارسی وجود ندارد به همین دلیل عملیات ریشه‌یابی تاثیر زیادی در نتیجه عملکرد ما ندارد.

نتیجه گیری و پیشنهادات

متن‌کاوی به‌عنوان یک زمینه در حال رشد و پرکاربرد، به دنبال کشف دانش از متون غیر ساخت‌یافته است. به دلیل مشکلات ساختاری زبان فارسی، تحقیقات کمتری در این زمینه صورت گرفته است. با توجه به اهمیت پردازش داده‌های متنی فارسی در کشور و در مجموعه‌های اطلاعاتی، در این پژوهش به مسئله طبقه‌بندی اخبار که قابلیت بسط به حوزه‌های مختلف متن‌کاوی را دارد پرداخته شده است. در

گذشته بیشتر کاری که برای طبقه‌بندی متون انجام گرفته است بر روی متون زبان انگلیسی و چینی بوده است. در این پژوهش روشی برای طبقه‌بندی متون فارسی ارائه شده است. روش کلی بر مبنای یادگیری ماشینی استوار است که در دو فاز آموزش و آزمون ارائه شده است. با توجه به تجزیه‌وتحلیل پیچیده متون فارسی از اکستنشن‌های[1] متن موجود در رپیدماینر استفاده شده که نتایج خوبی نیز در برداشته است. پس از ساخته شدن بردار ویژگی توسط ابزار قدرتمند پردازش متن از داده‌ها[2] به کمک ابزار مدل اتوماتیک الگوریتم‌های مختلف بررسی شد و مشخص گردید بهترین نتیجه مربوط به الگوریتم ماشین بردار پشتیبان می‌باشد سپس با استفاده از ابزار اعتبارسنجی متقابل طبقه‌بندی اسناد صورت گرفت و نتایج این ابزار نیز نشان می‌دهد مدل ماشین بردار پشتیبان بهره‌وری[3] بهتری داشته است و استفاده از روش‌های بیزین ساده و نزدیک‌ترین همسایه نتوانسته نتیجه بهتری به دست آورد با اینکه ما در مدل پیشنهادی موضوع زمان اجرا[4] را نیز به‌عنوان معیار مهم در اولویت قرار ندادیم. همچنین به‌عنوان کارهای آتی می‌توان به استفاده از روش‌های متفاوت وزن‌دهی مثل کلمه به بردار[5]، استفاده از روش‌های جدید انتخاب ویژگی مثل الگوریتم کرم شب‌تاب[6]، بهبود روش پردازش زبان طبیعی، بهبود و تقویت الگوریتم‌های ریشه‌یابی و توکنیزه کردن و بررسی روش‌های تقویت بردار ویژگی مناسب برای زبان فارسی، مورد بررسی قرار داد.

1.Extension
2.process document from data
3.performance
4.Run time
5.Word2vec
6.Firefly algorithm

منابع

۱. باقری، ایوب؛ فرزانه فر، حامد؛ سرایی، محمدحسین و احمدزاده، محمدرضا (۱۳۸۷). دسته‌بندی متون خبری فارسی با استفاده از الگوریتم بیز ساده، دومین کنفرانس داده‌کاوی ایران، تهران، https://civilica.com/doc/۷۰۵۲۴

۲. برفامی، مهدی و فاطری، سهیل (۱۳۹۲). استفاده از ترکیب شبکه‌های عصبی جهت دسته‌بندی متون فارسی مبتنی بر الگوریتم‌های GA، کی-نزدیکترین همسایه، PCA جهت کاهش ویژگی، اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات، رشت، https://civilica.com/doc/۲۲۵۸۸۷

۳. بصیری، محمد احسان؛ نعمتی، شهلا و قاسم آقایی، ناصر (۱۳۸۶). مقایسه دسته‌بندی متون فارسی با استفاده از الگوریتم‌های کی-نزدیکترین همسایه و fkNN و انتخاب ویژگی‌ها بر اساس بهره اطلاعات و فرکانس سند، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، جزیره کیش، https://civilica.com/doc/۴۱۷۸۶

۴. بینا، بهاره؛ رهگذر، مسعود و ده موبد، آذین (۱۳۸۶). طبقه‌بندی خودکار متون فارسی سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، جزیره کیش، انجمن کامپیوتر، دانشگاه صنعتی شریف.

۵. حسن‌پور، حمید؛ قنبری سرخی، علی و پارسی، اشکان (۱۳۹۱). استخراج بهترین ویژگی از متون فارسی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی با کمک میانگین یادآوری و الگوریتم ژنتیک، نخستین کنفرانس بین‌المللی پردازش خط و زبان فارسی.

۶. حسن‌پور، حمید و مدنی، صبا سادات (۱۳۹۳). بهبود دقت سیستم دسته‌بندی خودکار اسناد فارسی به کمک هستان‌شناسی فارس‌نت، مجله‌ی علمی پژوهشی، رایانش نرم و فناوری اطلاعات، جلد ۳، شماره ۱

۷. زمانی، محسن؛ دیانت، روح‌الله و صادق زاده، مهدی (۱۳۹۲). دسته‌بندی متون فارسی با استفاده از روش آنالیز معنایی پنهان احتمالاتی، همایش ملی کاربرد سیستم‌های هوشمند (محاسبات نرم) در علوم و صنایع، قوچان، https://civilica.com/doc/۲۰۶۲۵۱

۸. طاهری نیا، محسن (۱۳۹۱). دسته‌بندی متون فارسی با استفاده از یادگیری نیمه نظارت‌شده، چهارمین کنفرانس مهندسی برق و الکترونیک ایران، گناباد، https://civilica.com/doc/۱۶۴۲۲۶

۹. عابدینی نیا، مائده؛ الله دادی، لاله و شیخی، فاطمه (۱۳۹۲). کاوش متون فارسی در وب با استفاده از تحلیل معنایی و روش طبقه‌بندی چندلایه‌ای اطلاعات، اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات، رشت، https://civilica.com/doc/۲۲۵۳۷۷

۱۰. عربی نرئی، سمیه؛ وحیدی اصل، مجتبی و مینایی بیدگلی، بهروز (۱۳۸۶). استخراج کلمات کلیدی جهت

طبقه-بندی متون فارسی، اولین کنفرانس داده‌کاوی ایران، تهران، https://civilica.com/doc/۳۳۰۹۴

۱۱. قنبری سرخی، علی و ابراهیمی، فاطمه (۱۳۹۰). بهبود عملکرد طبقه‌بندی متون فارسی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی با کمک معیار میانگین یادآوری و دقت، چهاردهمین کنفرانس دانشجویی مهندسی برق کشور، کرمانشاه،https://civilica.com/doc/۱۲۱۵۵۸

۱۲. مقصودی، نوشین و همایون‌پور، محمدمهدی (۱۳۸۸). ارائه روشی جدید در طبقه‌بندی متون فارسی با استفاده از دانش معنایی "، پانزدهمین کنفرانس بین‌المللی سالانه انجمن کامپیوتر ایران.

۱۳. آقا کاردان، احمد و کیهانی نژاد، مینا (۱۳۹۱). ارائه مدلی برای استخراج اطلاعات از مستندات متنی، مبتنی بر متن‌کاوی در حوزه یادگیری الکترونیکی. فصلنامه علمی-پژوهشی فناوری اطلاعات و ارتباطات ایران، سال چهارم، شماره‌های ۱۱ و ۱۲، ص ۴۷ تا ۵۴

۱۴. بازقندی، مهدی؛ تدین تبریزی، قمرناز و وفایی جهان، مجید (۱۳۹۱). نخستین کنفرانس بین‌المللی پردازش خط و زبان طبیعی، دانشگاه سمنان.

۱۵. بهرام پور، اکبر؛ بهشتی، همایون و لاکتراشی، طیبه (۱۳۹۴). بررسی روش‌ها و الگوریتم‌های دسته‌بندی اخبار با استفاده از پردازش زبان طبیعی (NLP)، دومین کنفرانس ملی توسعه علوم مهندسی، تنکابن، موسسه آموزش عالی آیندگان.

۱۶. پرئی، اعظم‌السادات و حمیدی، حجت اله (۱۳۹۵). ارائه رویکردی برای مدیریت و سازماندهی اسناد متنی با استفاده از تجزیه تحلیل هوشمند متن. فصل‌نامه علمی- پژوهشی پژوهشگاه علوم و فناوری اطلاعات ایران، دوره ۳۲، شماره۴، ص۱۱۷۱ تا ۱۲۰۲

۱۷. جمالی، ایمان؛ میرعابدینی، سید جواد و هارون‌آبادی، علی (۱۳۹۶). ارائه‌ی یک مدل جهت دسته‌بندی متون فارسی با استفاده از ترکیب روش‌های دسته‌بندی، مجله مهندسی مخابرات، سال هفتم، شماره ۲۳

۱۸. سیاحی، عارف؛ هاشمی، سید محسن و مزرعه، سعید (۱۳۹۴). بهبود یک روش مبتنی بر انتخاب ویژگی به‌منظور دسته‌بندی متون با الگوریتم‌های متن‌کاوی، دومین کنگره سراسری فناوری‌های نوین ایران با هدف دستیابی به توسعه پایدار، تهران، مرکز راهکارهای دستیابی به توسعه پایدار، موسسه آموزش عالی مهر اروند.

۱۹. شیخی، مریم؛ اکبرپور، شاهین و فرزان، علی (۱۳۹۱). متن‌کاوی متون فارسی در راستای طبقه‌بندی آن. چهارمین کنفرانس مهندسی برق و الکترونیکی ایران.

۲۰. کریمی منش، مصطفی و شیرازی، حسین (۱۳۹۲). مقایسه روش‌های وزن‌دهی ویژگی در فرایند

طبقه‌بندی مستندات. اولین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات ایران

۲۱. هاشمی، سید محسن (۱۳۹۴). بهبود دسته‌بندی متون فارسی با ترکی روش دو مرحله‌ای انتخاب ویژگی و الگوریتم‌های یادگیری ماشین، کنفرانس بین‌المللی یافته‌های نوین پژوهشی در مهندسی برق و علوم کامپیوتر، تهران، موسسه آموزش عالی نیکان.

22. BolshaKov, I. A. & GelbuKh, A. (2004). Computational linguistics: models, resources, applications: Instituto Politecnico Nacional.
23. Eyheramendy, S. GenKin, A. Ju,W.H. Lewis, D. D. & Madigan, D. (2003). Sparse bayesian classifiers for text categorization. Journal of Intelligence Community Research and Development, 13.
24.Francis, L. A. (2006). Taming Text: An Introduction to Text Mining. Paper presented at the Casualty Actuarial Society Forum
25.McCallum, A. & Nigam, K. (1998). A comparison of event models for naive bayes text classification. Paper presented at the AAAI-98 worKshop on learning for text categorization.
26.Moulinier, I. & Ganascia, J. G. (1995). Applying an existing machine learning algorithm to text categorization. Paper presented at the International Joint Conference on Artificial Intelligence.
27.P. MulaK, &N. Talhar (2015). Analysis of distance measures using K-nearest neighbor algorithm on KDD dataset. International Journal of Science and Research, 4(7), 2101-2104.
28.NadKarni, P. M. Ohno-Machado, L. & Chapman, W. W. (2011). Natural language processing: an introduction. Journal of the American Medical Informatics Association, 18(5), 544-551.