Electronic and Cyber Defens 2025/ Vol. 13/ No. 2/ 89-101

Electronic and Cyber Defens

ISSN: 2322-4347 , E-ISSN: 2980-8979

Organizational Data Security with Persian Content Document Management Model for News Media

Hamid Mirzaei Dehnavi[1], Mohammad Ali Keramati[2*], Mohammad Ali Ameri Habibabadi[3]

[1] *PhD student, Department of Information Technology Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran.. Email:* Msb66@chmail.ir)

[2] *Professor, Department of Information Technology Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran . Email:*(*Correspondence: Mohammadalikeramati@yahoo.com)

[3] *Professor, Department of Information Technology Management, Central Tehran Branch, Islamic Azad University, Tehran, Iran*) *Email:*M.ali.ameri.h@gmail.com

ARTICLE INFO

Article history:

Article Type: Research paper

Received: 18 April 2025

Revised: 27 May 2025

Accepted: 08 June 2025

Available online: 28 June 2025

Keywords:

Data Security

Text Mining

Classification

Document Management

Persian Content

Information and Classification

ABSTRACT

With the increasing growth of Persian news texts in the digital world and the Internet, an important and significant issue is the intelligent classification and management of news documents and our quick and cheap access to them. Given that there are various algorithms and methods for text mining and they are still being developed; but the main problem of these methods for organizations is to maintain the confidentiality of information, and also to maintain the principle of information confidentiality along with the use of modern science and technologies. Many existing programs in the field of data mining and text mining work on the Internet, which is practically a violation of the principle of information confidentiality for organizations. Therefore, our main question refers to how to ensure data security using the document management model, which is one of the main concerns of every organization. Existing programs in the field of data mining and text mining work on the Internet, which is practically useless for some organizations that have classified data; Fortunately, domestic companies have been working in the field of text mining for several years and have made significant progress in this field; but it is still not possible to use such software for organizations and military complexes. Therefore, it is necessary for organizations to implement the desired model natively in order to meet their needs in this field and have a suitable platform independent of the Internet network for analyzing information. News document management provides organizations with the opportunity to correctly classify received or existing news and documents, preserve, maintain and retrieve them. In this research, which is of an applied type and uses the experimental research method and text mining tools, the statistical population of the present research is Persian news for the first 6 months of 1402. The BBC news data is a collection of Persian news texts in a labeled form. BBC News includes 1700 news texts from this dataset, which only contains Persian news, Afghan news and global news. Also, this research reviews and describes the literature on document management, natural language processing and data analysis, and then introduces document management systems as a part of the knowledge management system that is responsible for acquiring, organizing and sharing knowledge in the organization. By reviewing, analyzing and processing in this research, we conclude that the Persian content document management model using the support vector machine model has a precision of 93.29, a precision of 93.32, a recall of 92.96 and an error of 6.71.

Cite this article: Mirzaei Dehnavi, Hamid©, Keramati, Mohammad Ali©, Ameri Habibabadi, Mohammad Ali© (2025). Organizational Data Security with Persian Content Document Management Model for News Media. Journal of Electronic and Cyber Defens. 2025; 13l(2):89-101.

DOR:https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.2.8.5

Publisher: Imam Hossein University.

پدافند الکترونیک و سایبری/ سال ۱۴۰۴/ دوره ۱۳/ شماره ۲/ صفحه ۱۰۱-۸۹

پدافند الکترونیک و سایبری

شاپا الکترونیکی: ۴۳۴۷- ۲۳۲۲۲ شاپا چاپی: ۸۹۷۹-۲۹۸۰

امنیت داده‌های سازمانی با مدل مدیریت مستندات محتوای فارسی رسانه‌های خبری

حمید میرزائی دهنوی[۱]، محمد علی کرامتی[۲*]، محمدعلی عامری حبیب آبادی[۳]

[۱] دانشجوی دکتری، گروه مدیریت فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران (Msb66@chmail.ir)

[۲] استاد، گروه مدیریت فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران (نویسنده مسئول: Mohammadalikeramati@yahoo.com)

[۳] استاد، گروه مدیریت فناوری اطلاعات، واحد تهران مرکزی، دانشگاه آزاد اسلامی، تهران، ایران (M.ali.ameri.h@gmail.com)

چکیده (استایل عنوان چکیده)

با رشد روزافزون متون خبری فارسی در دنیای دیجیتال و اینترنت موضوعی که مهم و حائز اهمیت است دسته‌بندی و مدیریت هوشمند مستندات خبری و دسترسی سریع و ارزان ما به آن‌ها است. باتوجه‌به اینکه الگوریتم‌ها و روش‌های مختلفی برای متن‌کاوی وجود دارد و همچنان نیز درحال‌توسعه هستند؛ اما مشکل اصلی این روش‌ها برای سازمان‌ها حفظ محرمانگی اطلاعات است همچنین 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 نقض اصل محرمانگی اطلاعات برای سازمان‌ها است. بنابراین سؤال اصلی ما اشاره به چگونگی تضمین امنیت داده با استفاده از مدل مدیریت مستندات که از دغدغه‌های اصلی هر سازمان به شمار می‌آید است. برنامه‌های موجود در موضوع داده‌کاوی و متن‌کاوی در بستر اینترنت کار می‌کنند که عملاً برای برخی از سازمان‌ها که داده‌های طبقه‌بندی‌شده دارند 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 سازمان‌ها و مجموعه‌های نظامی ازاین‌گونه نرم‌افزارها استفاده نمود؛ لذا لازم است سازمان‌ها به‌منظور رفع نیاز خود در این حوزه، مدل موردنظر را به‌صورت بومی پیاده‌سازی کنند و یک بستر مناسب و مستقل از شبکه اینترنت برای تحلیل اطلاعات داشته باشند. مدیریت مستندات اخبار این امکان را برای سازمان‌ها فراهم می‌آورد که اخبار و اسناد دریافتی یا موجود را به‌درستی طبقه‌بندی نموده آن‌ها را حفظ، نگهداری و بازیابی نماید. این پژوهش از نوع کاربردی و برای انجام آن از روش پژوهش آزمایشی و ابزار متن‌کاوی استفاده گردیده است. جامعهٔ آماری پژوهش حاضر اخبار فارسی ۶ماهه اول سال ۱۴۰۲ است. داده اخبار بی‌بی‌سی مجموعه‌ای از متون خبری فارسی به‌صورت برچسب خورده هستند. اخبار بی‌بی‌سی شامل ۱۷۰۰ عدد متن خبری از این مجموعه داده که فقط حاوی اخبار حوزه فارسی، اخبار حوزه افغانستان و سطح جهانی است. همچنین در این پژوهش به بررسی و تشریح ادبیات مدیریت مستندات، پردازش زبان طبیعی و تجزیه‌وتحلیل داده‌ها پرداخته‌شده است و سپس بامطالعه سیستم‌های مدیریت مستندات به‌عنوان بخشی از سیستم مدیریت دانش که وظیفه اکتساب، سازمان‌دهی و اشتراک دانش در سازمان را بر عهده‌دارند، معرفی می‌شود. با بررسی، تحلیل و پردازش در این پژوهش به این نتیجه می‌رسیم که مدل مدیریت مستندات محتوای فارسی با استفاده از مدل ماشین بردار پشتیبان دارای دقت ۹۳/۲۹، صحت ۹۳/۳۲، فراخوانی ۹۲/۹۶ و خطای ۶/۷۱ است.

مشخصات مقاله

تاریخچه مقاله:

نوع مقاله: علمی پژوهشی

دریافت: ۱۴۰۴/۰۱/۱۷

بازنگری: ۱۴۰۴/۰۲/۲۱

پذیرش: ۱۴۰۴/۰۳/۱۸

ارائه آنلاین: ۱۴۰۴/۰۴/۲۰

کلید واژه‌ها:

امنیت داده

متن‌کاوی

طبقه‌بندی

مدیریت مستندات

محتوای فارسی

اطلاعات و دسته‌بندی

استناد: میرزائی دهنوی، حمید©، کرامتی، محمد علی©، عامری حبیب آبادی، محمدعلی©. امنیت داده‌های سازمانی با مدل مدیریت مستندات محتوای فارسی رسانه‌های خبری. پدافند الکترونیک و سایبری. (۱۴۰۴)؛ ۱۳ (۲): ۱۰۱-۸۹.

DOR:https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.2.8.5

ناشر: دانشگاه جامع امام حسین(ع).

۱- مقدمه

بدیهی است اتخاذ تصمیم در مقاطع زمانی مشخص و ارزیابی عملکرد پیشرفت برنامه‌ها، نیاز به اطلاعات دارد و تصمیمات باید مبتنی بر اطلاعات قابل‌اعتماد باشد. با در اختیار داشتن اطلاعات کامل، روشن، دقیق و به‌موقع، مدیران این امکان را خواهند داشت که تصمیماتی مطمئن و قابل‌اجرا اتخاذ نمایند. در غیر این صورت اطلاعات ناقص، مبهم و نارسا و توأم با حدس و گمان مسیر منطقی برنامه‌ها را تغییر داده و درنهایت سازمان را از اهداف اصلی خود دور می‌کند. با افزایش حجم اطلاعات الکترونیکی، نیاز به سیستم جامع برای مدیریت مستندات فزونی یافته است [۱]. شاید بسیاری از مدیران سازمان‌ها تصور کنند با صرف هزینه‌های بسیار و خرید بسته‌های نرم‌افزاری آرشیو اسناد می‌توانند سیستم مدیریت مستندات هوشمند را در سازمان خود پیاده‌سازی و از آن بهره‌برداری نمایند. درحالی‌که لازمه پیاده‌سازی این سیستم در هر سازمان تنها خرید یک بسته نرم‌افزاری نبوده و مستلزم انجام اقدامات گوناگونی است [۲]. یکی از زیرمجموعه‌های اسناد متنی که هرروزه 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 دستهٔ موضوعی هر متن خبری مشخص شود. با دسته‌بندی متون خبری که بر روی سایت‌های خبری منتشر می‌شوند، کاربران این سایت‌ها می‌توانند اخبار موردنظر خود را آسان‌تر و سریع‌تر پیدا کنند. از طرفی به دلیل حجم بالای اخباری که هر روزه منتشر می‌شوند، طبقه‌بندی اخبار به‌صورت دستی فرایندی 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 این معنی است که هر سایت خبری تلاش می‌کند اخبار روزانه را سریع‌تر از دیگر سایت‌ها در صفحهٔ 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.

سازمان‌ها داده‌های متنی بسیاری را رصد می‌کنند چگونه می‌توان این حجم از داده‌ها را مدیریت و امنیت آن را تضمین نمود؟ چگونه می‌توان اطلاعات مفیدی از این داده‌های متنی بی‌شمار به دست آورد؟ آیا این داده‌ها ارزشی دارند؟ الگوی مناسب پیاده‌سازی مدیریت مستندات الکترونیکی در سازمان‌ها چیست؟ و آیا برای ایجاد آن نیاز به طی مراحل و فرایند خاصی است؟ این مراحل چیست و یک سیستم ایدئال طراحی‌شده چه ویژگی‌هایی را باید داشته باشد؟

متن‌کاوی فناوری مورداستفاده برای چنین مواردی است. با فناوری‌های متن‌کاوی و مدیریت مستندات می‌توان داده‌های متنی را بررسی و تحلیل کرده و از نتایج حاصل از این تحلیل، اطلاعات ارزشمندی کسب نمود [۳]. داده‌های متنی هیچ‌گونه ارزشی ندارند مگر اینکه متن‌کاوی شوند، متن‌کاوی به شناسایی الگوها، کلمات کلیدی، موضوعات و دیگر ویژگی‌های موجود در متن می‌پردازد [۲]. الگوریتم‌ها و روش‌های مختلفی برای متن‌کاوی وجود دارد و همچنان نیز درحال‌توسعه است اما به دلیل پیچیدگی‌ها موجود در زبان فارسی، سعی در بهینه نمودن روش متن‌کاوی در این حوزه را داشته‌ایم همچنین مشکل اصلی این روش‌ها برای سازمان‌ها 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 برخی از سازمان‌ها که داده‌های طبقه‌بندی‌شده دارند 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 سازمان‌ها و مجموعه‌های نظامی ازاین‌گونه نرم‌افزارها استفاده نمود. لذا لازم است سازمان‌های مذکور این موضوع را به‌صورت بومی پیاده‌سازی کنند و یک بستر مناسب و مستقل از شبکه اینترنت برای تحلیل اطلاعات داشته باشند؛ در این پژوهش به مدیریت مستندات محتوای فارسی رسانه‌های خبری پرداخته می‌شود که خروجی آن برای سازمان‌ها جمع‌آوری و دسته‌بندی متون مختلف از سایت‌های خبری و سپس دسته‌بندی و بازیابی(مدیریت مستندات) است . همچنین مدلی جامع و کارآمد جهت تأمین امنیت داده‌ها و پیاده‌سازی سیستم مدیریت مستندات آنلاین خبری در سازمان‌ها ارائه می‌گردد.

ساختار مابقی مقاله شامل ادبیات موضوع، پیشینه تحقیق، ارائه روش پیشنهادی، ایجاد مدل طبقه‌بندی، ارزیابی و تحلیل داده‌ها و درنهایت نتیجه‌گیری است.

۲- ادبیات موضوع و پیشینه تحقیق

۲-۱- طبقه‌بندی مستندات فارسی

محوری‌ترین موضوع در حوزه متن‌کاوی، طبقه‌بندی مستندات است. اساس کار طبقه‌بندی مستندات بر پایه کلمات کلیدی و مهمی است که از مستندات استخراج می‌شود [۳]. برای دسته‌بندی متن از فناوری‌های استخراج اطلاعات[1]، پردازش زبان

1 *Information Extraction*

طبیعی[1] و یادگیری ماشین[2] به‌طور وسیع استفاده می‌شود. به‌طورکلی هدف یک دسته‌بند متون، دسته‌بندی اسناد در قالب تعداد معینی از دسته‌های ازپیش‌تعیین‌شده است. 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 [۳]. 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[3] انتخاب می‌شوند [۴].

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 [۲].

۲-۲- پردازش زبان طبیعی[4]

پردازش زبان طبیعی در دهه ۵۰ میلادی به‌عنوان فصل مشترک علوم هوش مصنوعی و زبان‌شناسی به وجود آمد و در دهه ۸۰ دچار تحولات و تغییرات اساسی شد [۵]. پردازش متن تا دهه ۸۰ میلادی با استفاده از روش‌های قانون محور انجام می‌گرفت. اگرچه روش‌های قانون محور ویژگی‌های مثبتی از قبیل قابل‌درک بودن برای انسان و قابلیت ارتقا کیفیت در طول زمان را دارا هستند، مشکلات و محدودیت‌هایی نیز دارند. لازمهٔ نوشتن قوانین در روش‌های قانون محور داشتن دانش عمیقی از آن حوزه است و از طرفی این روش‌ها بسیار زمان‌بر بوده و ازنظر محاسباتی پیچیده هستند [۶]. 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 قانون محور ظاهر شوند. درواقع در دهه ۹۰ میلادی بود که پژوهشگران به روش‌های یادگیری ماشین توجه بسیاری نشان دادند که این به دلیل توسعه سخت‌افزاری و نرم‌افزاری علم رایانه اتفاق افتاد [۷]. پژوهش‌ها نشان می‌دهند که طبقه‌بندی متن از نخستین تحقیقات انجام‌گرفته در حوزه متن‌کاوی است و نقش مهمی در آن ایفا می‌کند. به گفته فرانسیس تاریخچه پردازش متن به اواخر دهه ۵۰ میلادی بازمی‌گردد [۸]. همچنین طبقه‌بندی متن از دهه ۶۰ میلادی آغازشده و در دهه‌های ۸۰ و ۹۰ میلادی توجه محققان بیشتری را به خود جلب کرده است [۹].

۲-۳- الگوریتم بیز ساده

الگوریتم‌های بیزین ساده در چند دهه اخیر عملکرد بسیار خوبی داشته و به همین دلیل موردتوجه هرچه بیشتر محققان قرارگرفته‌اند [۱۰]. الگوریتم نایوبیز که مدلی احتمالی بر اساس نظریه بیز است، یکی از روش‌های ساده اما قدرتمند طبقه‌بندی متن محسوب می‌شود. الگوریتم طبقه‌بندی نایوبیز یکی از ساده‌ترین و پرکاربردترین الگوریتم‌ها در میان طبقه‌بندی‌ها است. الگوریتم نایوبیز گاوسی[5]، برنولی[6] و چندجمله‌ای[7] از زیرمجموعه‌های الگوریتم نایوبیز هستند. الگوریتم نایوبیز گاوسی برای داده‌های گسسته استفاده می‌شود و دو الگوریتم نایوبیز برنولی و چندجمله‌ای برای داده‌های پیوسته استفاده می‌گردد. این الگوریتم از یک مدل احتمالی بر پایه قضیه بیز استفاده می‌کند و با مستقل درنظرگرفتن ویژگی‌ها، احتمال وقوع داده جدید را در هر یک از دسته‌ها پیش‌بینی می‌کند. درنهایت سند به دسته‌ای تعلق می‌گیرد که بیشترین احتمال را دارد. اگرچه فرض مستقل درنظرگرفتن ویژگی‌ها در واقعیت درست نیست، اما الگوریتم نایوبیز با درنظرگرفتن این فرض نتایج خیلی خوبی ارائه داده است [۱۰].

این روش متداول‌ترین روش در دسته‌بندی متون است. در این روش متن به‌صورت مجموعه‌ای از کلمات مستقل از یکدیگر و مستقل از محل قرار گرفتن در متن در نظر گرفته می‌شود. تعریف تابع احتمال هر متن از حاصل‌ضرب احتمال کلمات آن و احتمال رخداد متنی با آن طول به دست می‌آید. عمل یادگیری در این بخش تخمین پارامترهایی برای تولید مدلی است که فقط از متون بر چسب‌دار استفاده می‌کند [۱۱].

۲-۴- الگوریتم ماشین بردار پشتیبان[8]

ماشین بردار پشتیبان یک روش یادگیری ماشین است که اولین بار در سال ۱۹۹۵ توسط و اپتیک[9] برای حل مسئله تشخیص الگوی دو کلاسه مورداستفاده قرار گرفت. این روش در یک فضای برداری تعریف می‌شود و هدف مسئله یافتن یک جداکننده است که به بهترین شکل بتواند داده‌های دو کلاس داده‌شده را از هم جدا کند. مجموعه داده‌های آموزش داده‌شده به سیستم در صورتی جدایی‌پذیر به‌صورت بهینه خواهند بود که بدون خطا تعیین کلاس شده و فاصله بین نزدیک‌ترین نمونه آموزشی به ابر صفحه حداکثر باشد. وقتی‌که داده‌ها جدایی‌پذیر خطی نباشند،

[1] Natural Language Processing (NLP)
[2] Machine Learning
[3] Chi Square
[4] Natural Language Processing (NLP)
[5] Gaussian Naïve Bayes
[6] Bernoulli Naïve Bayes
[7] Multinomial Naïve Bayes
[8] Support Vector Machine
[9] Vapnik

ورودی‌های اولیه به‌صورت غیرخطی به فضای با ابعاد بالاتر نگاشت می‌شوند. به‌این‌ترتیب در فضای جدید داده‌ها با استفاده از تابع هسته می‌توانند با ابر صفحه مناسب، بدون افزایش پیچیدگی محاسبات و به‌صورت خطی از هم جدا شوند. در حقیقت تابع هسته از شباهت بین داده‌ها در فضای اولیه برای یافتن شباهت بین بردارها در فضای با ابعاد بالاتر استفاده می‌کند. تابع هسته می‌تواند تابع چندجمله‌ای، تابع تانژانت هیدرولیک و یا توابع مناسب دیگر انتخاب شود [۱۲]. این الگوریتم در برابر نویز داده‌ها مقاوم است، می‌تواند با تعداد زیادی ویژگی کار کند و در طبقه‌بندی متون عملکرد خوبی دارد [۱۳].

۲-۵- الگوریتم نزدیک‌ترین همسایه

الگوریتم طبقه‌بند نزدیک‌ترین همسایه، برخلاف الگوریتم طبقه‌بند نایوبیز، فرایند طبقه‌بندی را تا حضور داده آزمون به تعویق می‌اندازد. به همین دلیل است که این الگوریتم در گروه یادگیرنده‌های تنبل طبقه‌بندی می‌شود [۱۴]. در الگوریتم طبقه‌بند نزدیک‌ترین همسایه، میزان شباهت داده آزمون با داده‌های آموزش با استفاده از یک معیار شباهت اندازه‌گیری می‌شود. سپس از کلاس همسایه‌های آن داده برای تشخیص کلاس داده موردنظر استفاده می‌شود. برای اندازه‌گیری میزان شباهت بین دو سند (متن) می‌توان از معیارهای مختلفی ازجمله فاصله اقلیدسی، ضریب شباهت جا کارد و شباهت کسینوسی استفاده کرد. فاصله اقلیدسی روش مناسبی برای اندازه‌گیری میزان شباهت بین دو متن است. پس از اندازه‌گیری میزان شباهت داده آزمون با هر یک از داده‌های آموزش و مقایسه این مقادیر، تعداد K تا از نزدیک‌ترین همسایه‌های (مشابه‌ترین اسناد به) سند آزمون انتخاب می‌شوند و درنهایت سند آزمون به دسته‌ای تعلق می‌گیرد که بیشترین همسایه‌هایش به آن دسته تعلق دارند؛ بنابراین می‌توان گفت که روش نزدیک‌ترین همسایه یک روش مبتنی بر مورد است چراکه از همه داده‌های آموزشی استفاده می‌کند. مقدار K می‌تواند هر عدد صحیح باشد. اگر مقدار ۱ برای K در نظر گرفته شود، دسته شبیه‌ترین همسایه به داده موردنظر نسبت داده می‌شود. به دلیل وجود داده‌های پرت مقدار ۱ مقدار مناسبی برای K نیست و بهتر است از روش تجربی برای تخمین بهترین مقدار K بهره گرفته شود. در پژوهش حاضر از روش تجربی۱ به‌منظور تعیین مقدار بهینه K استفاده‌شده است. در این روش با تکرار الگوریتم با مقادیر مختلف K، مقداری که باعث کمترین نرخ خطا می‌شود به‌دست‌آمده و بهترین مقدار K مشخص می‌گردد.

۲-۶- پیشینه پژوهش

در پژوهش استخراج کلمات کلیدی جهت طبقه‌بندی متون فارسی [۱۵]. از ترکیب روش‌های Wordnet و Porter استفاده کرده‌اند. برای استخراج کلمات کلیدی از ساختار درهم‌ریزی Trie و دادگان مورداستفاده، دادگان همشهری بوده است.

مقایسه دسته‌بندی متون فارسی با استفاده از الگوریتم‌های KNN و FKNN و انتخاب ویژگی‌ها بر اساس بهره اطلاعات و فرکانس سند [۱۶].در این مقاله دو الگوریتم KNN و FKNN را 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.

در پژوهشی دیگر با عنوان بهبود عملکرد طبقه‌بندی متون فارسی با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی با کمک معیار میانگین یادآوری و دقت [۱۷].نتایج به‌دست‌آمده نشان داده است که با در نظر گرفتن روش‌های دسته‌بندی KNN و Bayesian درروش پیشنهادی بهبود قابل‌توجهی در طبقه‌بندی متون فارسی و کاهش مدت‌زمان آزمون با ویژگی‌های استخراج‌شده به دست خواهد آمد.

همچنین در پژوهش کاوش متون فارسی در وب با استفاده از تحلیل معنایی و روش طبقه‌بندی چندلایه‌ای اطلاعات [۱۸].از مدل فضای برداری برای دسته‌بندی متون فارسی در وب استفاده‌شده است. روش پیشنهادی این امکان را می‌دهد که با دقت و سرعت بیشتری متن خود را در وب بیابند.

ارائه روش دسته‌بندی متون با تکنیک کاهش ویژگی فیلتری و یادگیری ماشین [۱۹]. 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.

دسته‌بندی متون لاتین با روش‌های انتخاب ویژگی فیلتری، پوششی و الگوریتم‌های بیز ساده [۲۰]. در مرحله یادگیری از الگوریتم‌های خانواده بیز ساده استفاده‌شده است. روش پیشنهادی در مقایسه باکارهای گذشته در این زمینه بهبود بسیار خوبی داشته است.

بررسی تطبیقی پرکاربردترین الگوریتم‌های طبقه‌بندی در پیش‌بینی و اعتبارسنجی مشتری [۲۱].این مقاله به بررسی و مقایسه تطبیقی عملکرد آن‌ها در اعتبارسنجی و شناسایی مشتریان خوش‌حساب و بدحساب با استفاده از یک مجموعه داده واقعی آلمانی پرداخته است. نتایج این بررسی عملکرد بهتر درخت تصمیم را گزارش کرده است.

در کارهای مرتبط انجام‌شده باتوجه‌به اهمیت موضوع و کاری که در این زمینه برای زبان‌های دیگر دنیا انجام‌گرفته است، نیاز به طبقه‌بندی متون فارسی به‌خوبی احساس می‌شود. این نکته قابل‌توجه است که تحقیقات برای متون انگلیسی توسعه داده‌شده و به کار گرفته می‌شود؛ اما ازآنجایی‌که زبان فارسی پیچیدگی‌های ساختاری نسبت به سایر زبان‌ها دارد و همچنین تحقیقات کمتری در این زمینه انجام‌گرفته است بر آن شدیم که ابتدا روش

دسته‌بندی متون فارسی را بهبود ببخشیم و درنهایت با بومی‌سازی روش پیشنهادی و همچنین عدم وابستگی به پراکسی‌ها و شبکه اینترنت امنیت را برای سازمان‌ها تا حدودی در این زمینه تأمین نماییم. در تحقیقات اشاره‌شده در پیشینه تحقیق تلاش برای بهینه نمودن پارامترهای روش دسته‌بندی، بالابردن سرعت اجرای دسته‌بندی و همچنین کاهش درصد خطا است که در این پژوهش با استفاده از روش پیشنهادی موفق شدیم پارامترهای مذکور را تعمیم و درصد خطا را نیز کاهش بدهیم.

۳- روش پیشنهادی

امروزه استفاده از اینترنت و فناوری اطلاعات، درواقع در کنار مزایایی که دارد، خطر برملا شدن اسرار خصوصی را به همراه دارد. اتصال به اینترنت در سیستم‌ها و شبکه‌های کامپیوتری موجب شده است تا متخصصان امنیت، روزبه‌روز شاهد تهدیدات امنیتی جدیدی باشند. مکانیسم‌های رمزنگاری و احراز هویت در سیستم‌های مدیریت پایگاه داده به معنای در امان بودن داده‌ها از آسیب‌پذیری‌های احتمالی نبوده و این مکانیزم‌ها توانایی مقابله با انواع حملات را ندارند. باید سعی شود که امنیت داده‌ها و حفظ حریم خصوصی در هنگام داده‌کاوی، کشف روابط بین داده‌ها و ذخیره در پایگاه داده‌ها حفظ شود. هدف از داده‌کاوی عمومی‌سازی اطلاعات است، نه اینکه اطلاعات شخصی را در اختیار عموم قرار دهیم. لازم به ذکر است که استفاده از داده‌کاوی جهت ایجاد مدل و استخراج الگو، امری ضروری است. حفظ حریم خصوصی درداده کاوی یک موضوع مهم در زمینه داده‌کاوی و امنیت پایگاه داده است. صاحبان داده‌ها به علت ترس از افشای اطلاعات شخصی و محرمانه خود توسط دیگران، چندان تمایلی جهت انجام داده‌کاوی روی‌داده‌های خود نداشته؛ ولی این مطلب را نیز می‌دانند که بدون انجام داده‌کاوی به نتایج و دانش مفید از داده‌های یکدیگر دسترسی پیدا نمی‌کنند. داده‌ها مادامی‌که پردازش نشوند و اطلاعات مفید از آن‌ها جهت تصمیم‌گیری‌های راهبردی، مباحث رقابتی و رسیدن به سود بیشتر استخراج نگردد، برای مدیران و برنامه‌ریزان چندان کارساز نیست. داده‌کاوی روی‌داده‌های مشترک باید طوری انجام شود که حریم شخصی صاحبان داده‌ها حفظ گردد؛ بنابراین مشکل اصلی بحث حفظ حریم شخصی در استخراج دانش، داده‌کاوی نیست، بلکه نحوه انجام داده‌کاوی است. الگوریتم‌ها و روش‌های مختلفی برای متن‌کاوی وجود دارد و همچنان نیز درحال‌توسعه است؛ اما مشکل اصلی این روش‌ها برای سازمان‌ها 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 برخی از سازمان‌ها که داده‌های طبقه‌بندی‌شده دارند 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 سازمان‌ها و مجموعه‌های نظامی ازاین‌گونه نرم‌افزارها استفاده نمود؛ لذا لازم است سازمان‌های مذکور این موضوع را به‌صورت بومی پیاده‌سازی کنند و یک بستر مناسب و مستقل از شبکه اینترنت برای تحلیل اطلاعات داشته باشند؛ لذا بر آن شدیم تا در این پژوهش روشی برای دسته‌بندی هوشمند متون ارائه نمائی‌ام که بتواند از سایت‌های خبری، اخبار مختلف را دسته‌بندی نماید. امروزه سامانه‌های مدیریت متون گستردگی بیشتری پیداکرده‌اند و وجود یک سیستم مدیریت متون فارسی که بهینه باشد بسیار احساس می‌شود. مدیریت 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 درج‌شده است یک خبر سیاسی یا ورزشی است. در این مثال سیاسی و ورزشی بودن خبر دودسته برای طبقه‌بندی اخبار را تشکیل می‌دهند. طبقه‌بندی متون قدمت بسیار زیادی در حوزه متن‌کاوی دارد این موضوع از سال ۱۹۶۰ میلادی موردتوجه محققین بوده است ولی بارونق کامپیوتر و نرم‌افزارها، مانند دیگر موضوعات هوش مصنوعی و داده‌کاوی در دهه ۹۰ میلادی توجه به این موضوع رشد چشمگیری پیدا کرد و موردتوجه قرار گرفت. ازآنجایی‌که در انجام پژوهش‌های مرتبط باهوش مصنوعی و زبان‌شناسی رایانشی، از علوم زبان‌شناسی و علوم رایانه استفاده می‌شود، ماهیت بین‌رشته‌ای علم زبان‌شناسی 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 هستیم. برای انجام تحقیق باید گام‌های ذیل مشخص شود که در شکل ۱ سیر مراحل مشخص و گویا است.

مراحل روش پیشنهادی به شرح ذیل است.

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

ب- پیش‌پردازش و آماده‌سازی اخبار

ج- انتخاب ویژگی

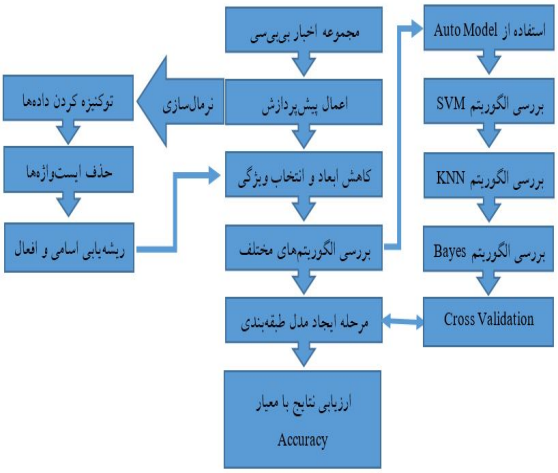
د- مقایسه الگوریتم‌های یادگیری ماشین برای دسته‌بندی (استفاده از مدل اتوماتیک [1])

ه- ایجاد مدل موردنظر برای دسته‌بندی [2]

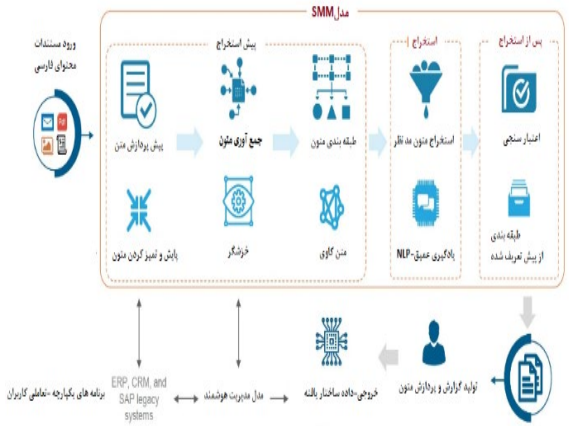
و- ارزیابی مدل ارائه‌شده

به‌طور خلاصه اجزای اساسی سیستم پیشنهادی در شکل ۲ قابل‌مشاهده است.

[1] Auto Model
[2] Classification



شکل (۱): روش پیشنهادی مدل طبقه‌بندی



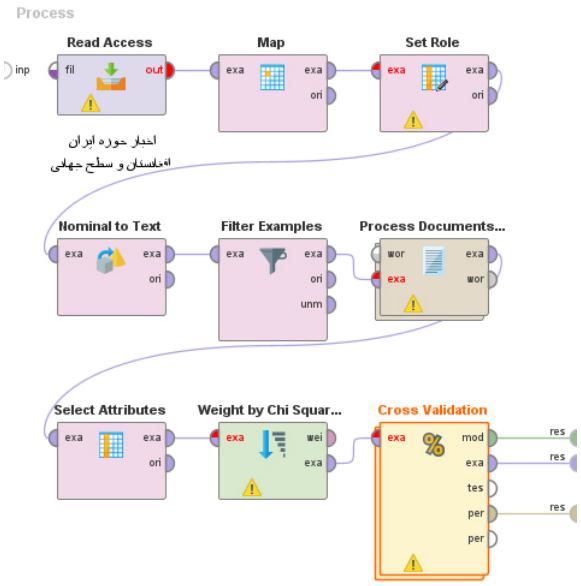
شکل (۲): مراحل مختلف ایجاد مدل طبقه‌بندی متون

۳-۱- تجزیه‌وتحلیل داده‌ها

جامعهٔ آماری پژوهش حاضر اخبار فارسی ۶ماهه اول سال ۱۴۰۲ است. داده اخبار بی‌بی‌سی مجموعه‌ای از متون خبری فارسی به صورتت برچسب خورده است. اخبار بی‌بی‌سی شامل ۱۷۰۰ عدد متن خبری از این مجموعه داده که فقط حاوی اخبار حوزه فارسی، اخبار حوزه افغانستان و سطح جهانی است. همچنین 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، وزن دهی، انتخاب ویژگی و طبقه‌بندی از نرم‌افزار رپید ماینر استفاده کرده‌ایم.

ابتدا مجموعه داده اکسس خوانده می‌شود. سپس برای ستون‌های آن نقش تعریف می‌شود. محتوای خوانده‌شده به‌صورت متن تبدیل می‌شود تا قابل نوک نیزه شدن توسط رپید ماینر شود. در ادامه مراحل انجام کار شرح داده می‌شود.



شکل (۳): نمایی کلی از فرآیند پیشنهادی در نرم‌افزار رپیدماینر

۳-۲- جمع‌آوری اخبار

پایگاه خبری موردنظر، آرشیو اخبار بی‌بی‌سی است که این مجموعه داده به‌وسیله یک خزنده[1] که به زبان پایتون نوشته‌شده است جمع‌آوری‌شده است جامعهٔ آماری پژوهش حاضر اخبار فارسی ۶ماهه اول سال ۱۴۰۲ است. داده اخبار بی‌بی‌سی مجموعه‌ای از متون خبری فارسی به‌صورت برچسب خورده است. اخبار بی‌بی‌سی شامل ۱۷۰۰ عدد متن خبری از این مجموعه داده که فقط حاوی اخبار حوزه فارسی، اخبار حوزه افغانستان و سطح جهانی است. این خزنده با یک حلقه while مدام در حال جمع‌آوری اخبار بوده و اخبار مربوطه را در محل از قبل تعیین‌شده ذخیره می‌نماید. اخبار جمع‌آوری‌شده در سه دسته خبری حوزه ایران، حوزه افغانستان و سطح جهانی توسط برچسب مشخص‌شده است و آموزش و آزمایش روی آن صورت خواهد گرفت. شکل زیر نمایی از مجموعه داده اخبار را ارائه می‌دهد.

در شکل ۵ صفحه مربوط به داشبورد مدیریتی از نحوه جمع‌آوری اخبار را از بستر اینترنت و فضای مجازی ارائه می‌دهد.

۳-۳- آماده‌سازی و پیش‌پردازش

در فاز آماده‌سازی متون، فرمت فایل از جیسون[2] به اکسس[3] تبدیل می‌شود، سپس متن سند که شامل کاراکترهای پشت سرهم است به نمایشی که برای الگوریتم‌های یادگیری و طبقه‌بندی مناسب باشد تبدیل می‌شود. این کار توسط پلاگین‌های[4] رپید ماینر[5] و همچنین قطعه کدی که به زبان

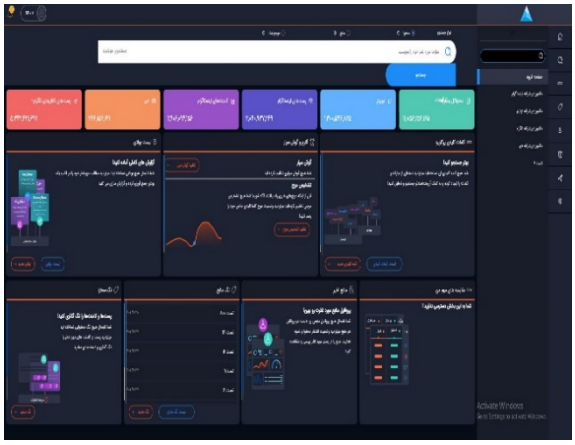
[1] Crawler
[2] Json
[3] Access(*.accdb)
[4] Plugin
[5] Rapid Miner

پایتون نوشته‌شده است انجام می‌گیرد. در این کد از کتابخانه هضم استفاده‌شده است. این فرایند شامل موارد ذیل است:

- حذف تگ‌های اچ‌تی‌ام‌ال[1] و یا ایکسم[2]
- کدگذاری متون به یونی‌کد[3]
- به دست آوردن ریشه کلمات و حذف پیشوندها و پسوندها

	A	B
	category	body
1	اخبار سطح جهانی	یک نظرسنجی تازه جهانی نشان دهنده عمق نگرانی 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.
2	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 ازدوگاه موقتی که در کنار آن شکل گرفته، در روزهای اخیر به شدت گسترش یافته است.
3	اخبار حوزه ایران	اعتراض‌ها در بهبهان، انتقاد از سیاست متقشای ایران در شهری کوچک تازه ترین اعتراض‌ها در ایران در شهری کوچک در استان خوزستان نمود پیدا کرده. در بهبهان، اتفاقا یکی از میدانهای مرکزی بهبهان در اعتراض‌های خونین آبان سال گذشته هم کانونی ملتهب بود.
4	اخبار حوزه ایران	آیت الله علی خامنه ای با صدور حکمی سرتیپ خیال حمید واحدی را به عنوان فرمانده جدید نیروی هوایی ارتش منصوب کرده است. او در این حکم خواستار "آرتقای توان و آمادگی‌های رزمی نیروی هوایی شده است."
5	اخبار سطح جهانی	سهم 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.
6	اخبار سطح جهانی	داد‌گاهی در لندن می‌گوید شیخ محمد آل مکتوم، حاکم دبی، در جریان دعوی 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.
7	اخبار سطح جهانی	در چارچوب طرح کوواکس سازمان بهداشت جهانی، طی گذشته یک پرواز ویژه "شاهان ایر" اولین محموله واکسن‌های "استرازنکا" را از هند وارد دوشنبه، پایتخت تاجیکستان، کرده به گفته مسئولان وزارت بهداشت و تأمین اجتماعی تاجیکستان، در مرحله نخست ۱۹۲ هزار دوز از این واکسن وارد شده است.
8	اخبار حوزه افغانستان	عصران 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 خواهد گرفت.
9	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 "گفته 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".
10	اخبار حوزه ایران	وزارت خزانه‌داری آمریکا اعلام کرده که سه نفر به نام‌های دانیل بهزاد فردوس، مانوئل بهزاد فردوس و محمد رضا دانفولیان 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.
11	اخبار حوزه افغانستان	کمیسیون دسترسی به اطلاعات افغانستان از ریاست جمهوری این کشور خواسته است که استدلله حاقد 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.
12	اخبار حوزه ایران	ده ماه بعد از اعدام خضر قبادی در زندان اروومیه، دیوان عالی کشور حکم اعدام او را لغو کرده است. یکی از اعضای خانواده قبادی به بی‌بی‌سی فارسی گفت از شنیدن خبر لغو حکم او چند ماه پس از اعدامش جا خورده‌اند و همشان بسیار افزونتر شده و با شکایت از قاضی شعبه اعتراض خود را شروع کرده‌اند.
13		

شکل (۴): نمایی از مجموعه داده اخبار بی‌بی‌سی



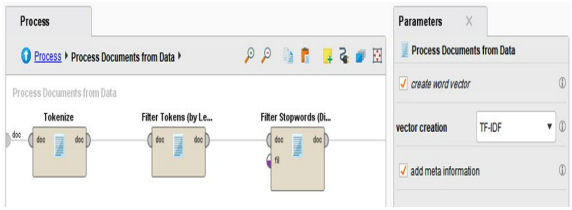
شکل (۵): نمایی از داشبورد مدیریتی جمع‌آوری خبر

۳-۴- نشان‌دار کردن اطلاعات (توکنیزه)

به‌منظور نوک نیزه کردن اطلاعات و تولید بردارهای قابل‌پردازش؛ از روش وزن دهی کلمات معکوس فراوانی متن استفاده‌شده است. این کار به کمک ابزار پردازش متن رپید ماینر به نام توکنایز انجام می‌شود. تصویر زیر خروجی موردنظر را نمایش می‌دهد. دقت کنید که تمامی کنترل‌های موجود در تصویر در کامپوننت پردازش سند روی‌داده هستند. همان‌طور که در تصویر قابل‌مشاهده است از الگوریتم پردازش متن معکوس فراوانی متن استفاده‌شده است.

در جدول ۱ فهرستی از فراوانی هر لغت در دسته مربوطه پس از عملیات توکنیزه کردن نمایش داده‌شده است.

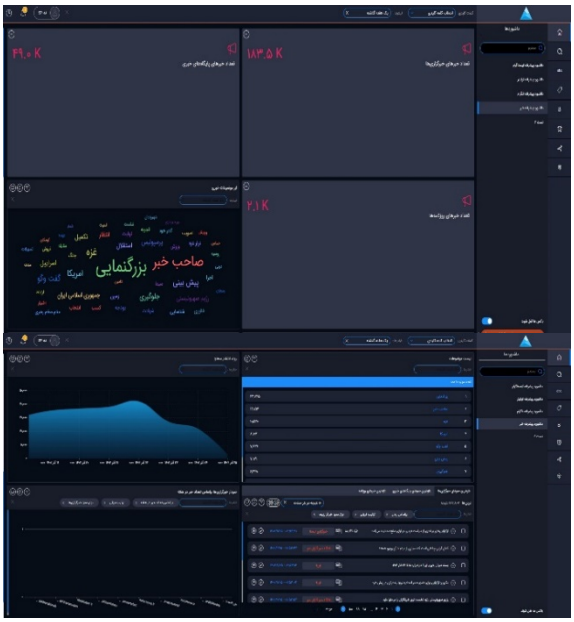
در شکل ۷ نمایی از داشبورد مدیریتی فراوانی کلمه‌ها را مشاهده می‌کنید.



شکل (۶): کاربرد الگوریتم پردازش متن TF-IDF

جدول (۱): نمونه‌ای از فراوانی هر کلمه در سه دسته خبری

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



شکل (۷): نمایی از داشبورد مدیریتی فراوانی کلمه

۳-۵- 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 توکن استفاده می‌شود. این فیلتر یک ورودی متن می‌گیرد و خروجی آن‌هم فیلتر شده همان متن است. برای

[1] html
[2] xml
[3] UTF-8

کاهش ابعاد کار؛ بهامُهرهای کمتر از ۴ و بیشتر از ۲۵ را حذف کردیم. این کار به افزایش صحت کار کمک بیشتری می‌کند.

همان‌طور که در جدول ۲ مشاهده می‌کنید تعداد بار تکرار هر یک از کلمات پردازش‌شده توسط الگوریتم مربوطه در هر سه دسته ایران، افغانستان و سطح جهانی قابل‌مشاهده است.

همچنین برای کاهش اندازه ویژگی‌ها و بهینه‌تر شدن انتخاب ویژگی‌ها از عملگر قدرتمند وزن دهی با مربع‌کاری[1] استفاده شد. خروجی این عملگر یک مجموعه از اوزان کلمات است که باتوجه‌به فناوری‌های مربع‌کاری تولید و مورداستفاده قرار می‌گیرد.

جدول (۲): نمونه‌ای از کلمات پس از وزن دهی با روش مربع‌کاری

دسته	آبان	آتش	آثار	آخرین	آذربایجان	آزاد
اخبار سطح جهانی	۰	۰.۰۰۷	۰	۰	۰	۰
اخبار سطح جهانی	۰	۰	۰	۰	۰	۰
اخبار سطح جهانی	۰	۰	۰	۰	۰	۰
اخبار حوزه ایران	۰.۰۳۹	۰.۰۳۹	۰	۰	۰	۰
اخبار حوزه ایران	۰	۰	۰	۰	۰	۰
اخبار سطح جهانی	۰	۰	۰	۰.۰۰۸	۰	۰

۴- ایجاد مدل طبقه‌بندی به کمک الگوریتم‌های یادگیری ماشین

بیشتر سامانه‌های طبقه‌بندی خودکار متون، برای متون زبان انگلیسی طراحی‌شده‌اند و معمولاً برای متون فارسی به دلیل ماهیت زبان فارسی و در دسترس نبودن مجموعه ریشه کلمات قابل‌استفاده نیستند [۱].اغلب روش‌های استفاده‌شده در این زمینه، روش‌های مبتنی بر یادگیری هستند. از مهم‌ترین روش‌های استفاده‌شده در طبقه‌بندی متون که در این پژوهش نیز استفاده‌شده است می‌توان به بیزین ساده، ماشین بردار پشتیبان و کی نزدیک‌ترین همسایه اشاره کرد. شکل زیر داشبورد مدیریتی مربوط به طبقه‌بندی‌ها را نمایش می‌دهد.

شکل (۸): نمایی از داشبورد مدیریتی طبقه‌بندی خبر

۵- ارزیابی روش پیشنهادی و تحلیل نتایج

معمولاً برای ارزیابی صحت دسته‌بندی 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 مربوطه است و معیار ارزیابی F ترکیب هماهنگ از دقت و فراخوانی است. معادلات مربوطه در جدول زیر که موسوم به ماتریس درهم‌ریختگی[6] آست آمده است. برای درک دو معیار ارزیابی دقت و فراخوانی و همچنین معیار صحت، ابتدا باید با عبارات مثبت واقعی[7]، مثبت کاذب[8]، منفی واقعی[9] و منفی کاذب[10] آشنا بود. این چهار عبارت برای مقایسه برچسب کلاس‌ها در ماتریس درهم‌ریختگی استفاده می‌شوند. جدول ۳ این عناوین را نشان می‌دهد.

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:

$$Precision = \frac{TP}{TP+FP} \quad (۱)$$

[1] Weight By Chi Squared

[2] Precision
[3] Recall
[4] Accuracy
[5] F-measure (F-score)
[6] Confusion Matrix
[7] True Positive
[8] False Positive
[9] True Negative
[10] False Negative

جدول (۳): ماتریس درهم‌ریختگی

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

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

$$Re\,c\,all = \frac{TP}{TP+FN} \quad (۲)$$

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

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (۳)$$

معیار ارزیابی F: باتوجه‌به 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 است و همچنین توصیف‌کننده میانگین وزن‌دار مابین دو کمیت دقت و فراخوانی است. برای یک الگوریتم کلاس‌بندی در شرایط ایدئال، مقدار این کمیت برابر با ۱ است و در بدترین وضعیت برابر با صفر است. این پارامتر باتوجه‌به رابطه زیر محاسبه می‌شود:

$$F - Measure = 2 \times \frac{precision \times recall}{precision + recall} \quad (۴)$$

۶- ابزار مورداستفاده

در این تحقیق از مجموعه داده اخبار فارسی به‌عنوان مجموعه آموزش و مجموعه آزمون استفاده‌شده است. این پیکره شامل ۱۷۰۰ خبر بین سال‌های ۱۴۰۰ تا ۱۴۰۲ است. 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، وزن دهی، انتخاب ویژگی و طبقه‌بندی از نرم‌افزار رپید مایند استفاده کرده‌ایم.

۷- استفاده از ابزار مدل اتوماتیک

روش استفاده از مدل اتوماتیک به این صورت است که ابتدا داده موردنظر را به‌عنوان 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.

۷-۱- بررسی معیار دقت الگوریتم بیز ساده

در این بخش تأثیر روش مربع‌کاری و نرمال‌سازی کلمات بر روی عملکرد این الگوریتم را به‌صورت جداگانه بررسی می‌کنیم. فرایند انتخاب ویژگی تأثیر زیادی در بهبود عملکرد الگوریتم طبقه‌بند به‌خصوص در الگوریتم طبقه‌بند بیز ساده که به فرایند انتخاب ویژگی بسیار وابسته است، دارد بنابراین پیش‌بینی می‌شود که عملکرد الگوریتم طبقه‌بند بیز ساده با اعمال روش مربع‌کاری بهبود یابد. جدول ۴ میانگین معیار دقت را برای الگوریتم طبقه‌بند بیز ساده در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرایند انتخاب ویژگی مربع‌کاری نشان می‌دهد.

در شکل ۹ نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم بیزساده و مربع‌کاری آمده است.

جدول (۴): بررسی نرمال‌سازی کلمات و روش مربع‌کاری بر الگوریتم بیز ساده

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

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 (۹): سایر معیارها برای الگوریتم بیز ساده

۷-۲- بررسی معیار دقت الگوریتم کی-نزدیکترین همسایه

در این بخش تأثیر نرمال‌سازی کلمات و روش مربع‌کاری بر روی عملکرد الگوریتم کی - نزدیک‌ترین همسایه را به‌صورت جداگانه بررسی می‌کنیم. جدول ۵ میانگین معیار دقت را برای الگوریتم کی - نزدیک‌ترین همسایه در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرایند انتخاب ویژگی مربع‌کاری نشان می‌دهد.

در شکل ۱۰ نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم کی-نزدیکترین همسایه و مربع‌کاری آمده است.

جدول (۵): بررسی نرمال‌سازی کلمات و روش مربع‌کاری 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 مربع‌کاری
میانگین دقت	٪۸۶,۱	٪۸۷,۳	٪۸۹,۱

Performances

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%	

شکل (۱۰): سایر معیارها برای الگوریتم کی-نزدیکترین همسایه

۷-۳- بررسی معیار دقت الگوریتم ماشین بردار پشتیبان

در این بخش تأثیر نرمال‌سازی کلمات و روش مربع‌کاری بر روی عملکرد الگوریتم ماشین بردار پشتیبان را به‌صورت جداگانه بررسی می‌کنیم. جدول ۶ میانگین معیار دقت را برای الگوریتم ماشین بردار پشتیبان در سه حالت بدون اعمال نرمال‌سازی کلمات، اعمال نرمال‌سازی کلمات و با اعمال فرایند انتخاب ویژگی مربع‌کاری نشان می‌دهد.

در شکل ۱۱ نیز سایر معیارهای ارزیابی برای ترکیب الگوریتم ماشین بردار پشتیبان و مربع‌کاری آمده است.

جدول (۶): بررسی‌تاثیر نرمال‌سازی کلمات و روش مربع‌کای 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 مربع‌کاری
میانگین دقت	٪۸۷,۵	٪۸۸,۸	٪۹۰,۸

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 (۱۱). سایر معیارها برای الگوریتم ماشین بردار پشتیبان

۸- محدودیت‌های تحقیق

جمع‌آوری اخبار از سایت‌های مختلف یا سایت واحد نیاز به یک خزنده باقابلیت جمع‌آوری متون فارسی دارد که خزنده‌های آنلاین نیز باتوجه‌به تحریم روی سایت‌های فارسی، بدون فیلتر اعمال نمی‌شوند و استفاده از پراکسی باعث کند شدن روند فعالیت خزش گر می‌شود. همچنین یک مجموعه واحد از ریشه لغات زبان فارسی وجود ندارد به همین دلیل عملیات ریشه‌یابی تأثیر زیادی درنتیجه عملکرد ما ندارد.

۹- نتیجه‌گیری

متن‌کاوی به‌عنوان یک زمینه درحال‌رشد 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. باتوجه‌به 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 پرداخته‌شده است. درگذشته بیشتر کاری که برای طبقه‌بندی متون انجام‌گرفته است بر روی متون زبان انگلیسی و چینی بوده است. در این پژوهش روشی برای طبقه‌بندی متون فارسی ارائه‌شده است. روش کلی بر مبنای یادگیری ماشینی استوار است که در دوفاز آموزش و آزمون ارائه‌شده است. باتوجه‌به تجزیه‌وتحلیل پیچیده متون فارسی از اکستنشن‌های[1] متن

[1] Extension

موجود در رپید ماینر استفاده‌شده که نتایج خوبی نیز در برداشته است. پس از ساخته‌شدن بردار ویژگی توسط ابزار قدرتمند پردازش متن از داده‌ها[1] به کمک ابزار مدل اتوماتیک الگوریتم-های مختلف بررسی شد و مشخص گردید بهترین نتیجه مربوط به الگوریتم ماشین بردار پشتیبان است سپس با استفاده از ابزار اعتبارسنجی متقابل طبقه‌بندی اسناد صورت گرفت و نتایج این ابزار نیز نشان می‌دهد مدل ماشین بردار پشتیبان بهره‌وری[2] بهتری داشته است و استفاده از روش‌های بیزین ساده و نزدیک‌ترین همسایه نتوانست نتیجه بهتری به دست آورد بااینکه ما در مدل پیشنهادی موضوع زمان اجرا[3] را نیز به‌عنوان معیار مهم در اولویت قرار ندادیم. همچنین به‌عنوان کارهای آتی می‌توان به استفاده از روش‌های متفاوت وزن دهی مثل کلمه به بردار[4]، استفاده از روش‌های جدید انتخاب ویژگی مثل الگوریتم کرم شب‌تاب[5]، بهبود روش پردازش زبان طبیعی، بهبود و تقویت الگوریتم‌های ریشه‌یابی و نوک نیزه کردن و بررسی روش‌های تقویت بردار ویژگی مناسب برای زبان فارسی، موردبررسی قرارداد.

۱۰- مراجع

[1] B.Bina, M.Rahgozar, A.DehMobad, "Automatic Classification of Persian Texts", 13th Annual Conference of the Iranian Computer Association, Kish Island, Computer Association, Sharif University of Technology, 2007.(In Persian)

[2] I.Jamali, S.J.Mir-Abedini, A.Haroonabadi, "Presenting a model for Persian text classification using a combination of classification methods", Journal of Telecommunications Engineering, Volume 7, Issue 23, 2017.(In Persian)

[3] M.Karimimanesh, H.Shirazi, "Comparison of feature weighting methods in the document classification process", First National Conference on New Approaches in Computer Engineering and Information Retrieval in Iran, 2013.(In Persian)

[4] A.Siyahi, S.M.Hashemi, S.Mazraa, "Improving a Feature Selection Method for Text Classification with Text Mining Algorithms", Second National Congress of Iranian New Technologies for Sustainable Development, Tehran, Center for Strategies for Sustainable Development, Mehr Arvand Institute of Higher Education, 2015. (In Persian)

[5] P.NadKarni, M.Ohno-Machado, L.Chapman, " Natural language processing: an introduction", Journal of the American Medical Informatics Association, 18(5), 544-551, 2011.

[6] I.Moulinier, J.G.Ganascia, "Applying an existing machine learning algorithm to text categorization", Paper presented at the International Joint Conference on Artificial Intelligence, 1995.

[7] S.Eyheramendy, A.GenKin, W.Ju, H.Lewis, D.Madigan, "Sparse bayesian classifiers for text categorization", Journal of Intelligence Community Research and Development, 13, 2003.

[8] L.A.Francis, "Taming Text: An Introduction to Text Mining", Paper presented at the Casualty Actuarial Society Forum, 2006.

[9] I.A.BolshaKov, A.GelbuKh, "Computational linguistics: models, resources, applications: Instituto Politecnico Nacional", 2004.

[10] A.McCallum, K.Nigam, "A comparison of event models for naive bayes text classification", Paper presented at the AAAI-98 worKshop on learning for text categorization, 1998.

[11] Kardan.A, M.Keyhani Nejad, "Presenting a Model for Extracting Information from Text Documents Based on Text Mining in the Field of E-Learning", Iranian Information and Communication Technology Scientific and Research Quarterly, Year 4, Issues 11 and 12, pp. 47-54, 2012.(In Persian)

[12] M.Sheikhi, S.Akbarpour, A.Farzan, "Text Mining of Persian Texts for Its Classification", Fourth Iranian Conference on Electrical and Electronic Engineering, 2012. (In Persian)

[13] M.Bazghandi, Q.Tabrizi, J.Vafaei, "First International Conference on Natural Language and Handwriting Processing", Semnan University, 2012.(In Persian)

[14] P. MulaK, N. Talhar, "Analysis of distance measures using K-nearest neighbor algorithm on KDD dataset", International Journal of Science and Research, 4(7), 2101-2104, 2015.

[15] S.Arabi-Narei, M.Vahidi-Asl, B.Minaei-Bidgoli, "Extracting Keywords for Persian Text Classification", First Iranian Data Mining Conference, Tehran, https://civilica.com/doc/33094, 2007.(In Persian)

[16] M.E.Basiri, S.Nemati, N.Ghasem Aghaei, "Comparison of Persian Text Classification Using K-Nearest Neighbor and fkNN Algorithms and Selecting Features Based on Information Gain and Document Frequency", Thirteenth Annual Conference of the Iranian Computer Association, Kish Island, https://civilica.com/doc/41786, 2007.(In Persian)

[17] A.Ghanbari-Sorkhi, F.Ebrahimi, "Improving the performance of Persian text classification using principal component analysis with the help of average recall and precision criteria", 14th National Electrical Engineering Student nConference, Kermanshah, https://civilica.com/doc/121558, 2011.(In Persian)

[18] M.Abedininia, L.Allahdadi, F.Sheikhi, "Exploring Persian texts on the web using semantic analysis and multi-layer information classification method", First National Conference on New Approaches in Computer Engineering and Information Retrieval, Rasht, https://civilica.com/doc/225377, 2013.(In Persian)

[19] P.Jalali, S.M.Hashemi, A.Siyahi, A.Heydari, "Presentation of a text classification method using filter feature reduction and machine learning techniques", International Conference on New Research Findings in Electrical Engineering and

[1] Process Document From Data
[2] Performance
[3] Run Time
[4] Word2vec
[5] Firefly Algorithm

Computer Science, Tehran, Nikan Institute of Higher Educationhttps, 2015.(In Persian)

[20] H.PourAbbasi, M.Abbasi-Dezfuli, "Latin Text Classification with Filtering Covering and Naive Bayes Algorithms", First National Conference on Computer and Information Technology, Sepidan, Islamic Azad University, Sepidan Branch, 2017.(In Persian)

[21] M.Golparvar, S.J.Mousavi-Rad, "A Comparative Study of the Most Used Classification Algorithms in Customer Prediction and Validation", Fifth National Conference on Applied Research in Management and Accounting, Tehran, Iranian Management Association., (2018). (In Persian)