تاریخ دریافت مقاله: ۹۷/۱۲/۱۶
تاریخ پذیرش مقاله: ۹۸/۰۸/۰۴

ارائه روش ترکیبی مبتنی بر یادگیری ماشین برای دسته‌بندی خودکار متون اینترنتی

محمد رستمی*
دانشجوی دکتری مهندسی نرم افزار و الگوریتم دانشگاه کاشان، اصفهان، ایران
پست الکترونیکی: mohamad.rostami10@yahoo.com

حسین ابراهیم پور کومله
استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه کاشان، اصفهان، ایران
پست الکترونیکی: ebrahimpour.kashanu@gmail.com

چکیده:

با افزایش حجم اطلاعات در دسترس بر روی اینترنت و پایگاه‌های داده، نیاز به ابزارهایی که بتوانند در جستجو، پالایش و مدیریت منابع کمک کنند، ضروری است. برای رسیدن به این منظور در این پژوهش، از دسته‌بندی متون با استفاده از الگوریتم‌های یادگیری ماشین استفاده شده است. دسته‌بندی یا رده‌بندی متون، به اختصاص یک سند متنی به یک طبقه مناسب از پیش تعیین شده گفته می‌شود. چالش اصلی دسته‌بندی متون، بزرگی فضای ویژگی‌ها در این گونه مسائل است. در بسیاری از الگوریتم‌های موجود چنین فضای بزرگی منجر به کند شدن بسیار زیاد دسته‌بند و ناکارآمدی آن خواهد شد. علاوه بر این ویژگی‌هایی وجود دارند که نه تنها باعث دسته‌بندی بهتر متون نمی‌شوند بلکه دقت دسته‌بندی را نیز کاهش می‌دهند. در این پژوهش جهت دست یافتن به کارایی مناسب ابتدا آماده‌سازی متون یا مجموعه داده انجام شده است. برای این منظور ابتدا اسناد متنی را به شکل یکسان (حروف کوچک) تبدیل کرده و سپس نتایج در دو حالت، بدون حذف کلمات متوقف کننده و با حذف کلمات متوقف کننده به‌دست آمده است.

این سیستم شامل دو مرحله، پردازش متن و دسته‌بندی متن می‌باشد. در مرحلهٔ اول برای استخراج ویژگی‌ها از معیارهای شاخص‌گذاری مختلفی نظیر trigram، bigram و quadgram استفاده شده، سپس در مرحلهٔ دوم برای آموزش سیستم از الگوریتم یادگیری ماشین W-SMO استفاده شده است. به منظور ارزیابی و مقایسه نتایج دو معیار دقت و بازخوانی، Macro-F1 و Micro-F1 برای روش‌های مختلف شاخص‌گذاری محاسبه شده‌اند. نتایج آزمایش‌ها که بر روی ۷۶۷۶ سند متنی استاندارد خبرگزاری رویترز انجام گرفت، نشان داد که روش پیشنهادی بهترین کارایی را نسبت به الگوریتم‌های K-NN، Naïve Bayes، W-j48 و W-LADTREE دارد. بررسی نتایج نشان می‌دهد که روش پیشنهادی باعث بهبود دقت میکرو تا ۹۵٫۱۷٪ در دسته‌بندی متون می‌گردد.

واژه‌های کلیدی: دسته‌بندی متون، یادگیری ماشین، N-gram،W-SMO

* نویسندهٔ مسئول

۱-مقدمه

امروزه اطلاعـــات ارزش زیادی دارند. با افزایش حجم اطلاعـــات در دســـترس روی اینترنت، نیـــاز فوق‌العاده به ابزارهایی که بتوانند در جســـتجو، پالایش و مدیریت منابع کمک کنند،کاملا محسوس است.

دســـته‌بندی متـــون[1] (رده‌بنـــدی متـــون) بـــه عمـــل برچسب‌گذاری موضوعی متون زبان طبیعی بر مبنای یک مجموعه از پیش تعیین شـــده، اطلاق می‌شـــود. هم اکنون دسته‌بندی متون در بسیاری از زمینه‌ها از شاخص‌گذاری[2] متون بـــر مبنای یـــک لغت‌نامه کنترل شـــده[3] تـــا پالایش متون، تولید خودکار فـــراداده، ابهام‌زدایی از کلمه[4]، تولید فهرست‌های سلســـله‌مراتبی از منابع وبی[5] و به طور کلی در هر کاربردی که نیاز به سازماندهی مستندات یا توزیع انتخابی و تطبیقی خاصی از مستندات مد نظر باشد، کاربرد دارد [۱]. از کاربردهای دیگر دســـته‌بندی متون می‌توان به سیستم‌های خودکار پاســـخ به سوالات، پالایش اطلاعات، تشـــخیص موضوعیـــت داده‌هـــا، نامه‌هـــای الکترونیکی بـــی‌ارزش، تشـــخیص عنـــوان و دیگر زمینه‌هـــای مرتبط اشـــاره نمود[۲]. چالش اصلی دســـته‌بندی اسناد، بزرگی فضای ویژگی‌ها[6] در این گونه مســـائل است. در بسیاری از الگوریتم‌هـــای موجود چنین فضای بزرگی منجر به کند شـــدن بسیار زیاد دســـته‌بند و ناکارآمدی آن خواهد شد. علاوه بـــر این، ویژگی‌هایی وجود دارنـــد که نه تنها باعث دسته‌بندی بهتر اســـناد نمی‌شوند، بلکه دقت دسته‌بندی را نیز کاهش می‌دهند [۳].

یک متن نمی‌تواند به‌صورت مستقیم توسط یک دسته‌بند یا یک الگوریتم دسته‌بندســـاز تفسیر شود. بلکه با استفاده از یـــک فرآیند شـــاخص‌بندی که متن را به یـــک نمایه (که محتویات آن را ابعاد بیان می‌کند) نگاشـــت می‌کند، تفسیر می‌شود. این مهم کمک می‌کند تا یکنواختی و یک‌شکلی لازم برای متون مجموعه آموزشی[7]، آزمایشی[8] و اعتبارسنجی[9] فراهم آورده شود [۱].

ابعاد بزرگ فضای کلمه‌ها در دسته‌بندی متون معمولاً دردسرســـاز می‌باشـــد. در حقیقت، با بزرگ شدن فضای کلمه‌ها، تعداد ویژگی‌هـــا افزایش می‌یابد و باعث پیچیدگی بیشـــتر (صرف هزینه زمانی و فضای حافظه بیشـــتر) و عدم وابســـتگی بین داده‌ها می‌گردد که ارزش دسته‌بندی ندارنـــد. به عبارت دیگـــر، در کاهش ابعاد عمومی ســـعی می‌شـــود تا با تحلیل پیکره زبانی، تمامی متون موجود در مجموعه آموزشـــی کلمه‌هایی را که ارزش کمی در کاربرد مد نظر دارند، تعیین و این دســـته از کلمه‌ها به‌صورت یک لیست ثابت، معین گردند. کلمه‌های متن ورودی به‌صورت خودکار توســـط این لیســـت پالایش می‌گردنـــد. در حوزه محلی، همین کار برای هر یک از دســـته‌ها به‌صورت مجزا انجام می‌شود. لذا مسئلهٔ کاهش ابعاد خود یکی از زمینه‌های مفید در بازیابی اطلاعات و مخصوصاً دســـته‌بندی متون می‌باشد[۲].

اولیـــن رهیافـــت بـــرای کاهش ابعـــاد با اســـتفاده از انتخـــاب کلمه‌هـــا، رهیافـــت پالایـــش نامیده می‌شـــود. با اســـتفاده از ابزارهایـــی که نظریهٔ آمار یـــا اطلاعات فراهم نموده اســـت کلمه‌های بی‌ربط از کلمه‌های استخراج شده پالایش می‌شـــوند. در نهایت دســـته‌بندها، مستقل از تابع پالایه‌ســـاز، با اســـتفاده از فضای کلمه کاهش یافته تولید می‌شوند [۲].

بر اســـاس این الگوریتم‌ها اکثـــراً ویژگی‌هایی انتخاب می‌شـــوند که حاوی بیشـــترین اطلاعات مفید هستند. در این راســـتا، در این پژوهش با بررســـی نقاط قوت، ضعف و کارایی الگوریتم‌های دســـته‌بندی متن و انتخاب ویژگی، مدلی پیشنهادی ارائه شده است.

در ایـــن مقاله، به منظور دســـته‌بندی خودکار متون از سه روش شاخص‌گذاری bigram، trigram و quadgram و الگوریتـــم یادگیری ماشـــین W-SMO اســـتفاده شـــده

1- Text Classification
2- Indexing
3- Controlled Dictionary
4- Word Sense Disambiguation
5- Population Of Hierarchical Catalogues Of Web Resources
6- Features Space
7- Training set
8- Experimental
9- Validation

است. نتایج به‌دست آمده نشان می‌دهد که بهترین روش شاخص‌گذاری متون bigram می‌باشد.

برای طبقه‌بندی متون از مجموعه داده ۷۶۷۶ مقاله خبرگزاری رویترز استفاده شده است. این مجموعه داده به‌نام Reuters-21578 گردآوری شده است که در وبگاه پایگاه داده روش ساخت و اطلاعات آماری این مجموعه داده توضیح داده شده است[۲۸].

از مجموعه مقالات انتخاب شده با توجه به محتوای آن‌ها که در ۸ دسته طبقه‌بندی شده‌اند، هر بار یک دسته برای اعتبارسنجی و بقیه دسته‌های دیگر برای آموزش به‌کار می‌روند. این روال ۸ بار تکرار می‌شود و همه داده‌ها دقیقاً یک بار برای آموزش و یک بار برای اعتبارسنجی به‌کار می‌روند. در نهایت میانگین نتیجه این ۸ بار اعتبارسنجی به‌عنوان یک تخمین نهایی برگزیده می‌شود که طبقه‌بندی خودکار روی آن‌ها انجام می‌پذیرد.

ساختار ادامه مقاله به این صورت می‌باشد: در بخش دوم پیشینه پژوهش مورد بررسی قرار می‌گیرد. در بخش سوم روش پیشنهادی تشریح می‌گردد. بخش چهارم به ارزیابی و تفسیر نتایج اختصاص دارد. در نهایت بخش پنجم نیز به نتیجه‌گیری و کارهای آینده می‌پردازد.

۲- پیشینه پژوهش

با توجه به گستردگی حجم اطلاعات متنی الکترونیکی که به طور قابل توجهی از طریق اینترنت و سایر منابع قابل دسترس می باشند، در صورت عدم شاخص‌گذاری و دسته‌بندی مناسب، کار بازیابی و پردازش اطلاعات متنی دسته‌بندی نشده با مشکلات زیادی مواجه می‌گردد. دسته‌بندی متون، کاربردهای زیادی از جمله پیگیری اسناد، مدیریت اسناد، گسترش اسناد و کاهش حجم اطلاعات دارد. بر اساس تعداد رده‌ها، دسته‌بندی می‌تواند به دسته‌بندی دودویی[10] و دسته‌بندی چند رده‌ای[11] تقسیم شود. دسته‌بندی دودویی، نمونه‌ها را دقیقاً به یکی از دو ردهٔ موجود دسته‌بندی اختصاص می‌دهد، حال آن‌که دسته‌بندی چند رده‌ای با بیش از دو رده سر و کار دارد. لازم به ذکر است در این مقاله به دسته‌بندی چند رده‌ای پرداخته شده است[۴,۵].

اکثر روش‌های یادگیری ماشین در زمینه دسته‌بندی متن‌ها که در سال‌های اخیر به‌کار برده شده است، شامل مدل‌های برگشتی، دسته‌بندی نزدیکترین همسایه[12][۶]، شبکه‌های بیزین[13][۷] و درخت تصمیم‌گیری[14] می‌باشد [۸] که هر کدام از این روش‌ها دقت و محاسبات متفاوتی دارند.

در مطالعه روش گوران و همکاران [۹] دسته‌بندی متون به زبان ترکی با استفاده از N-gram مورد بررسی قرار گرفته است. در این پژوهش با استفاده از trigram، unigram، bigram و quadgram متون دسته‌بندی شده‌اند. آزمایش‌ها در این مقاله بر روی ششصد سند متنی که به شش دسته تخصیص یافته‌اند انجام شده و کارایی این تحقیق برابر ۹۵,۸۳٪ گزارش شده است.

در مطالعه روش وان و لی [۱۰] دسته‌بندی متون با استفاده از یک الگوریتم ترکیبی ارائه شده است که این الگوریتم ترکیبی از الگوریتم K-NN و SVM می‌باشد. نتایج این پژوهش که بر روی مجموعه داده خبرگزاری رویترز انجام شده، نشان می‌دهد که در بهترین حالت کارایی این روش ترکیبی ۸۱,۴۸٪ و در بدترین حالت ۵۴,۵۵٪ می‌باشد.

در مطالعه دیوی و همکاران [۱۲] یک روش انتخاب ویژگی مبتنی بر روش PSO را با استفاده از اصول تکاملی بهبود داده‌اند. سپس ویژگی‌های استخراج شده را در مجموعه داده WebKB [۳۰] با استفاده از یک شبکه عصبی موازی به منظور کاهش هزینه‌های محاسباتی مورد آزمایش قرار داده‌اند.

لان و همکاران [۱۶] با استفاده از طرح مدل نمایش فضای برداری و وزن‌دهی واژه‌ها اقدام به دسته‌بندی اسناد با استفاده از ماشین بردار پشتیبان و k–نزدیک‌ترین

10- Binary Classification
11- Multiclass Classification

12-K-nearest neighbor(K-NN)
13-Naïve Bayes Network
14-Decision Tree

همسایه بر روی مجموعه داده گروه خبری ۲۰ [۲۹] نموده و بـــه ترتیب به نتایج ۰/۸۰۸ و ۰/۶۹۱ دســـت پیدا کرده‌اند. محدودیت این روش، ابعاد بالا جهت نمایش و از بین رفتن همبســـتگی و بافت هر کلمه که در تفهیم ســـند مهم است، می‌باشد.

در روش فورن کرانز [۱۷] نتایج نشـــان داده است که پس از حذف کلمات متوقف شـــده، توالی کلمات با طول ۲ یا ۳ مفیدترین می‌باشند و استفاده از توالی‌های طولانی‌تر کارایی طبقه‌بنـــدی را کاهش می‌دهد. برخی نتایج از روش unigram، bigram و trigram و بـــا اعمـــال روش مجذور کای توسط دسته‌بندهای مختلف بر روی مجموعه خبری ۲۰ در مقاله بررسی شده است.

در روش باکاس و کامل [۱۸] یک دسته‌بند اسناد مبتنی بر بیز ارائه شـــده اســـت که از عبارات به‌عنوان ویژگی‌ها استفاده می‌کند. این عبارات با استفاده از یک دستور زبان که پیاپی قوانین توالی و ترتیب واژگان در ســـند را اعمال می‌کند، استخراج می‌شـــود. این قوانین بر اساس مجموعه داده آموزشی تولید می‌شوند.

در روش مونتانز و همکاران [۱۹] پس از نمایش سند با استفاده از روش کیسه کلمات، مولفه‌های بردار با استفاده از tf پر شده است. همچنین پس از اعمال رپر ماشین بردار پشتیبان جهت انتخاب ویژگی بر روی مجموعه داده رویترز (مجموعه متنی)، نشان داده شده است که این کار از کارایی خوبی نســـبت به پالایه‌ها برخوردار اســـت. از مزایای این روش می‌توان به نیاز نداشـــتن به تعیین حد آستانه جهت انتخاب تعداد ویژگی‌ها در مقایسه با پالایه‌ها اشاره کرد.

در مطالعه روش وانگ و همـــکاران [۲۰] یک الگوریتم ترکیبی انتخاب ویژگی مطرح شـــده اســـت که در ابتدا یک مســـئله C رده‌ای را به C مســـئلهٔ دو رده‌ای تبدیل می‌کند. ســـپس الگوریتم انتخاب ویژگی، ویژگی‌ها را با استفاده از جستجوی رو به جلو برای آموزش ماشین بردار پشتیبان انتخاب می‌کند.

سایس و همکاران [۲۱] مطالعه‌ای بر روی مجموعه‌ای از تکنیک‌هـــای انتخاب ویژگی انجـــام داده‌اند. روش آن‌ها اثبات کرده اســـت که الگوریتم‌های ترکیبی در مقایســـه با الگوریتم‌های منفـــرد در داده‌های با ابعاد بالا نتایج بهتری را از خود نشان می‌دهند.

در مطالعـــه روش رافی و شـــیخ [۲۲] به مقایســـه دو الگوریتم SVM و RSVM در ســـه مجموعـــه داده با تعداد نمونه‌هـــا و رده‌های متفاوت پرداخته شـــده که مشـــخص می‌کند SVM از ســـرعت بالایی جهت دسته‌بندی و RSVM از ســـرعت پایین اما اندازه میانگیـــن F1 بهتری برخوردار است.

در مطالعه روش پرادهان [۲۳] یک روش ترکیبی برای دســـته‌بندی متون ارائه شده که در این روش از ترکیب دو الگوریتم k– نزدیک‌ترین همســـایه و دسته‌بندی کننده بیز استفاده شـــده است. نتایج به‌دســـت آمده در این مقاله که بر روی ۲۱۵۷۸ ســـند متنی خبرگزاری رویترز انجام شده دارای کارایی برابر ۹۵٫۵٪ میکرو گزارش شده است.

بصیری و همکاران [۲۴] به بررســـی دسته‌بندی متون فارسی با استفاده از الگوریتم‌های KNN و FKNN پرداخته‌اند. آزمایش‌ها بر روی ششصد ســـند متنی که به شش دسته تقسیم شـــده‌اند، انجام شده است. هدف اصلی این بررسی، مقایسه دو الگوریتم مذکور برای دسته‌بندی متن فارسی و ترکیب آن‌ها با روش‌هـــای انتخاب ویژگی بهره اطلاعات و فرکانس سند DF است. از این دو روش برای انتخاب ویژگی‌ها و کاستن از ابعاد فضای ویژگی‌ها استفاده شده است. نتایج نشان می‌دهد که دقت الگوریتم FKNN از الگوریتم KNN بهتر اســـت. همچنین دقت دسته‌بندی با استفاده از ترکیب FKNN و IG از سایر ترکیب‌ها بیشتر می‌باشد. دقت دسته‌بندی در بهترین حالت به ۰/۸۰۴ دقت میکرو F1 و ۰/۷۵۵دقت ماکرو F1 گزارش شـــده است. در بین دســـته‌های موجود بهترین دسته‌بندی در مورد بزرگ‌ترین دسته یعنی اسناد مربوط به دسته اقتصادی انجام گرفته است که دقت دسته‌بندی برای این دسته تا ۰/۹۱۰ دقت ماکرو F1 و ۰/۹۴۵ دقت میکرو F1 گزارش شده است.

باقـری و همکارانش [۲۵] پژوهشـی به نـام ارائه یک روش انتخاب ویژگی ترکیبی برای دسته‌بندی متون به نام PSA ارائه داده‌اند که در پژوهش خود با اسـتفاده از روش پیشـنهادی انتخاب ویژگی و روش بیز به دسته‌بندی هفت دسته خبری با روش اعتبارسنجی پنج مرحله‌ای پرداخته‌اند و توانسـته‌اند روش خـود را با دقت ۸۸,۲٪ پیاده‌سـازی کنند.

یوسـفیان و فولادوند [۳۱] یک روش دسته‌بندی برای دسته‌بندی اسناد معرفی نمودند. تمرکز اصلی این پژوهش، بـر روی روش‌های پیش‌پردازش و اسـتفاده از آن‌ها در افزایش کارایی سیستم دسته‌بندی بود. از جمله روش‌های مورد اسـتفاده بـرای نیل به هـدف مذکـور می‌توان به الگوریتم ریشه‌یابی، حذف کلمات توقف و همچنین انتخاب زیرمجموعه‌ای از ویژگی‌ها با روش اطلاعات متقابل اشاره کرد. مطرح کردن بردار نماینده برای غنی‌تر کردن مجموعه ویژگـی از مفاهیمی بود که در این پژوهش به آن پرداخته شد و هدف از استفاده از این مفهوم، افزایش دقت سیستم دسـته‌بندی اسناد و همچنین یافتن کلماتی بود که از لحاظ معنایی با هم ارتباط دارند. از دیگر اهداف استفاده از بردار نماینده می‌توان به حل چالش متنوع بودن منابع اطلاعاتی اشـاره کرد. از آنجایی که یک منبـع نمی‌تواند منبع کاملی برای دسته‌بندی اسناد باشد، از این‌رو، استخراج کلمات از منابع مختلف به کمک بردار نماینده و افزودن این کلمات به مجموعه ویژگی اولیه، باعث افزایش دقت روش دسته‌بندی تا ۸۰٪ می‌باشـد. در ارزیابی ایـن روش از مجموعه داده همشهری استفاده شده است [۳۲].

جلالـی و همکارانش [۳۳] بعد از انتخاب مجموعه داده و پاکسازی متون به کمک روش نرمال شده فرکانس کلمه معکوس فرکانس سـند (norm TF-IDF) به ویژگی‌ها وزن داده می‌شـود و در طـی دو مرحله ویژگی‌ها با اسـتفاده از روش‌های فرکانس سـند (DF) و مجـذور کای انتخاب می‌شـوند و بعد با اسـتفاده از روش تحلیل مؤلفه اصلی (PCA) ابعاد ویژگی‌ها کاهش داده می‌شود و در مرحله بعد با اسـتفاده از ماشین بردار پشتیبان(SVM) به پیاده‌سازی مدل پیشـنهادی می‌پردازد. نتایج این مدل عمل دسته‌بندی متون را برای هفت دسته با صحت ۹۱,۸۶٪ نشان می‌دهد.

باقری و همکاران [۳٤] یک روش انتخاب ویژگی ترکیبی برای دسته‌بندی متون فارسی به نام PSA ارائه دادند که در پژوهش خود با استفاده از روش پیشنهادی انتخاب ویژگی و روش بیز به دسته‌بندی هفت دسته خبری با روش اعتبار سـنجی پنج مرحله‌ای پرداخته‌اند و توانسته‌اند روش خود را با دقت ۸۸,۲ ٪ پیاده‌سازی کنند.

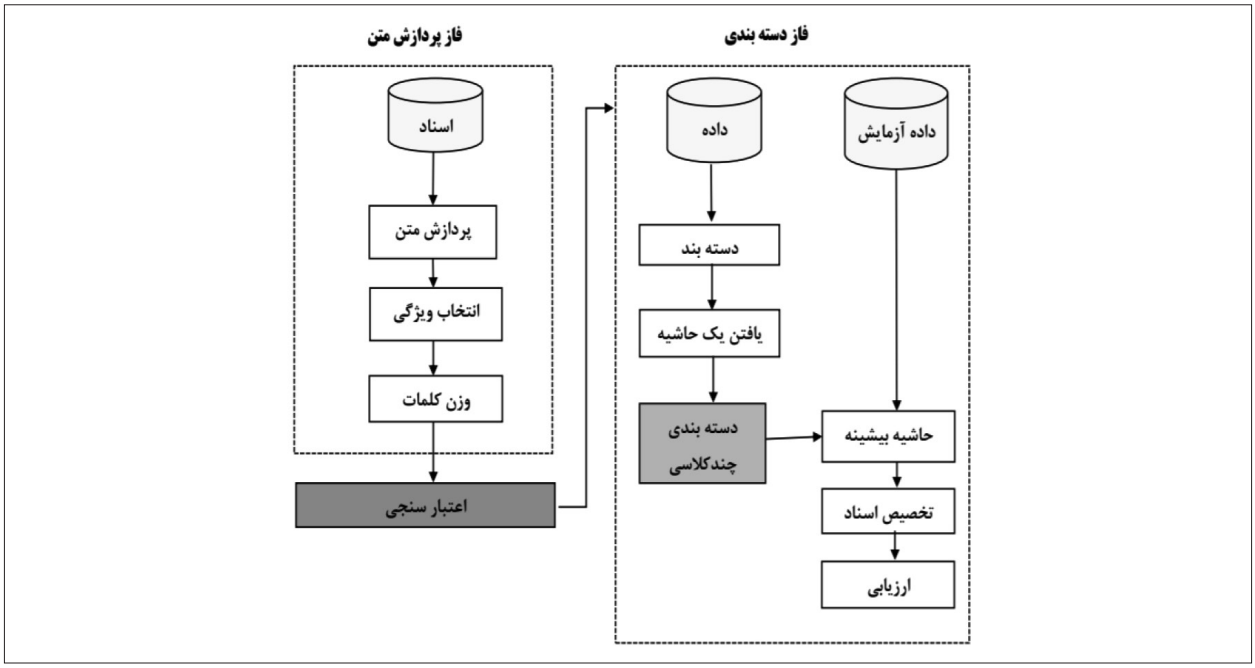
۳- روش پیشنهادی

در این پژوهش روش ترکیبی برای دسـته‌بندی متون ارائه شده است. جزئیات این روش به این صورت می‌باشد که قبل از بررسـی یک سـند، موضوع آن ناشـناخته در نظر گرفته می‌شـود، سپس روش پیشـنهادی با استفاده از دسـته‌های از پیش تعریف شده، که در مرحله یادگیری آموزش دیده است، اسناد را دسته‌بندی می‌نماید. به‌عنوان مثال یک مجموعه خبر وجود دارد که دارای چندین دسـته می‌باشد و هر دسته یک موضوع خاص را پوشش می‌دهد و در صورت وارد شـدن یک خبر جدید به مجموعه داده، سیستم پیشـنهادی موضوع خبر را تشخیص داده و خبر را در دسـته مربوط به خبر قرار می‌دهد. در نهایت هدف، طراحی یک سیستم ترکیبی جهت دسته‌بندی متون است تا بتواند تابع هدف f را نمایش دهد. یعنی مشخص می‌کند که سند p مربوط به کدام دسته c_i است. فرمول (۱) بیانگر تابع f می‌باشد.

$$f : p \longrightarrow \{c_1, c_2, \ldots, c_i\} \qquad i>=2 \qquad (۱)$$

در شکل (۱) مراحل کلی روش پیشنهادی نشان داده شده اسـت که در ادامه هر کدام از این مراحـل به همراه نتایج به‌دست آمده، توضیح داده می‌شود.

جمـع آوری داده‌هـا از انواع مختلـف قالب‌ها از قبیل html، doc، pdf می‌باشد.



شکل ۱: روش پیشنهادی برای دسته‌بندی متون

۳-۱- مرحلهٔ پردازش متن[15]

مجموعه عملیاتی را که منجـر به تولید مجموعه‌ای از داده‌های پالایش شده، جهت دستیابی به ویژگی‌های مناسب متون می‌باشـد، اصطلاحاً پردازش متن می‌گویند[۲۶]. این عملیات شامل مراحل آماده‌سـازی متون، شاخص‌گذاری مستندات و وزن‌دهی شـاخص‌ها می‌باشد که در ادامه به شرح هر کدام از این مراحل پرداخته می‌شود.

۳-۱-۱- مرحلـه آماده‌سـازی متـون و یـک شکل‌سـازی حروف

داده‌کاوی فرآیند استخراج الگوهای مخفی از مجموعه داده‌های بزرگ می‌باشـد. داده‌ها در دنیـای واقعی اغلب ناکامل و ناسازگار هسـتند و ممکن است دارای خطاهای بسیاری باشـند. در این مرحله برای پالایش داده‌ها اعمال مختلفی بر روی آن‌ها انجام می‌شود.

اولیـن و مهم‌تریـن مرحلـه در فرآینـد دسـته‌بندی آماده‌سـازی داده می‌باشـد. هدف در ایـن مرحله تامین ورودی مناسب برای مرحله حیاتی یادگیری مدل است. در این مرحله داده پردازش نشده از کل منابع داده موجود (که ممکن است توزیع شده نیز باشـد) استخراج شده، سپس در مرحله‌ای مسـتقل مورد پردازش اولیـه قرار می‌گیرد. خروجی در مرحله آماده‌سـازی داده عبارت است از داده پیش‌پردازش شـده کـه امکان یادگیری مـدل از روی آن وجود دارد. در این مرحله، عملیات شـامل یک‌شکل‌سازی حروف، حذف کلمات متوقف کننده و وزن‌دهی شاخص‌ها می‌باشد که در ادامه به شرح هر کدام از این مراحل خواهیم پرداخت.

در مرحله آماده‌سازی متون، متن که شامل نویسه‌های پشت سـر هم اسـت به نمایشـی که برای الگوریتم‌های یادگیری و طبقه‌بندی مناسب باشد تبدیل می‌شود.

این مرحله در روش پیشنهادی معمولاً شامل موارد زیر است:

- به دست آوردن ریشه کلمات.
- حذف پیشوندها و پسوندها.
- حذف کلمات متوقف کننده و علائم نگارشی.
- یک شکل‌سازی حروف

ریشـه‌یابی کلمـات[16]: در ایـن مرحله کلمـات به فرم ریشه‌شان تبدیل می‌شوند و کلماتی که به خاطر پیشوندها و پسوندهایشـان از یکدیگر متمایز شـده‌اند ولی ریشـه

15-Text Processing Phase

16- Term Stemming

یکسانی دارند، در یک گروه قرار داده می‌شوند (اصطلاحا کلماتی که هم‌خانواده هستند).

کلمات متوقف‌کننده کلماتی هستند که معمولاً از فراوانی بالایی برخوردار بوده و اطلاعات خاصی را حمل نمی‌کنند. جدول کلمات متوقف‌کننده عموماً حاوی انواع کلمات به طول سه نویسه، حروف ربط، اضافه و افعال کمکی است که عمومی بوده و به دسته خاصی تعلق ندارند. از کلمات متوقف‌کننده پر استفاده می‌توان به مواردی مانند the،why، is،am و ... اشاره کرد. همچنین در این مرحله تمام حروف به‌صورت حروف کوچک نمایش داده شده تا در هنگام پردازش از بروز کلمات تکراری جلوگیری شود. در این مقاله تمرکز بر روی کلمات متوقف کننده می‌باشد.

۳-۱-۲- مرحله شاخص‌گذاری

در این مرحله یک سند از متن به بردار سند تبدیل می‌شود. یکی از معمول‌ترین روش‌ها برای این کار مدل فضای بردار نام دارد که در آن اسناد به‌صورت برداری از کلمات نمایش داده می‌شوند. پس از انجام مراحل فوق متن به‌صورت برداری از کلمات که ویژگی‌های متن هستند می‌تواند نمایش داده شود. حال می‌توان این ویژگی‌ها را با توجه به میزان اهمیت آن‌ها نسبت به سند متنی و دسته آن وزن‌دهی کرد. هرچه ویژگی‌ها دارای اهمیت بیشتری باشند وزن بیشتری خواهند داشت.

در این مرحله، یک متن d_j، با برداری از وزن عبارت‌هایش نشان داده می‌شود. به عبارت دیگر $d_j = <w_{1j}, w_{2j}, ..., w_{|T|j}>$ به طوری که T مجموعه عبارت‌هایی است که دست‌کم یک‌بار در سرتاسر مجموعه آموزشی آمده باشند (در بعضی اوقات به آن ویژگی نیز گفته می‌شود) و $0 \leq wk_j \leq 1$ است. معمولاً تفاوت رهیافت‌ها در این زمینه به یکی از دلایل ذیل می‌باشد:

تفاوت در تعریف چیزی که «عبارت» نامیده می‌شود.

تفاوت در طریقه محاسبه وزن کلمه‌ها.

در این مقاله از روش شاخص‌گذاری متن به‌صورت کلمات ساده و روش N-gram (bigram، trigram و quadgram) استفاده شده که در ادامه روش N-gram توضیح داده میشود.

• روش N-gram

در این روش، شاخص‌گذاری به‌صورت ترتیبی از N حرف پشت سر هم می‌باشد. یک کلمه متن به‌صورت مجموعه‌ای از N-gramها که با هم همپوشانی دارند نشان داده می‌شود [۱۱]. به‌عنوان مثال کلمه "TEXT" از N-gramهای زیر تشکیل شده است:

Bigram: _T, TE, EX, XT, T_
Trigram: _TE, TEX, EXT, XT_, T_ _
Quadgram: _TEX, TEXT, EXT_, XT_ _, T_ _ _

"_" نشان‌دهنده فاصله می‌باشد. در این روش کلمات به‌صورت ترکیبی از نویسه‌ها که پیوسته در کنار هم قرار دارند، در نظر گرفته می‌شوند. همچنین نویسه‌های () ? ! ' " , : ; . به‌عنوان جدا کننده مشخص شده‌اند که جهت ترکیب نادیده گرفته می‌شوند. مزیت N-gram با توجه به طبیعت آن می‌باشد، چون هر رشته از تعداد محدودی از کلمات تشکیل شده است، خطاها منتشر نمی‌شوند و روی تعداد محدودی از رشته‌ها اثر می‌گذارند.

۳-۱-۳- مرحله وزن‌دهی ویژگی‌ها

برای وزن‌دهی به ویژگی‌ها می‌توان از رویکردهای متفاوتی بهره برد. در ساده‌ترین حالت این وزن‌دهی می‌تواند به‌صورت دودویی انجام شود. انتخاب دیگر وزن‌دهی به هر کلمه با توجه به تعداد تکرار هر کلمه می‌باشد. اما یکی از راهکارهای مناسب و مورد توجه استفاده از tf-idf [۱۵] است. حاصلضرب فرکانس هر کلمه در معکوس فرکانس سند معمولاً به‌صورت زیر تعریف می‌شود:

$$tf - idf(t_k, d_j) = tf(t_k, d_j) \times log \frac{|N|}{N(t_k)} \quad (۲)$$

که N نماینده تعداد کل اسناد و t_k تعداد اسنادی از مجموعه آموزشی است که کلمه t_k در آن حداقل یک بار رخ داده است. $tf(t_k,d_j)$ نیز نشان دهنده تعداد تکرارهای کلمه kام در سند jام است. به این ترتیب رخداد بیشتر یک کلمه

در صورتـــی در افزایـــش وزن آن مؤثر اســـت که در همه متون دیگر تکرار نشـــده باشد. با توجه به این‌که این روش بازنمایی و صورت‌های دیگر آن در کارهای زیادی مورد اســـتفاده قرار گرفتـــه و کارایی خـــوب آن روی مجموعه داده‌های متفاوتی به اثبات رســـیده است، در این مقاله نیز این روش بازنمایی انتخاب می‌گردد.

۳-۲- فاز دسته‌بندی

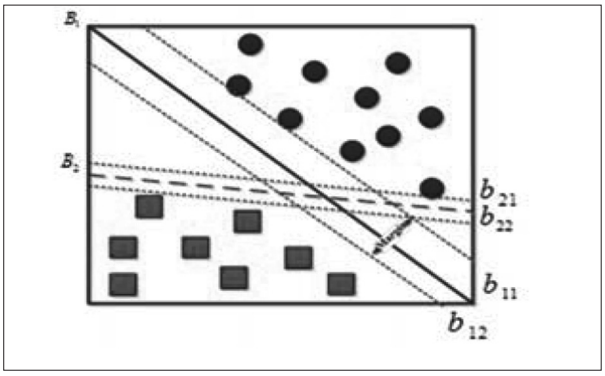
این فاز شامل دو مرحله یادگیری و آزمایش می‌باشد، که در ادامه هر کدام از این مراحل شرح داده می‌شود.

۳-۲-۱- الگوریتم طبقه‌بندی خودکار متون[17]

پس از انجام مراحل فوق، اکنون متن به‌صورت برداری از ویژگی‌هـــا بـــا وزن‌های متفـــاوت درآمده اســـت که به الگوریتم یادگیری داده می‌شـــود تا مدل دسته‌بندی تولید شـــود. پس از تولید مدل دسته‌بندی می‌توان دسته مربوط به متن‌های جدید را تشخیص داد.

در این مقاله از الگوریتم W-SMO [۱۲,۱۳] استفاده شده اســـت. استفاده از الگوریتم‌های ماشـــین بردار پشتیان در مسائل دسته‌بندی متون، رویکرد جدیدی است که در چند ســـال اخیر مورد توجه بسیاری قرار گرفته است. رویکرد W-SMO به این صورت اســـت که در فاز یادگیری، سعی دارد که مرز تصمیم‌گیـــری را به گونه‌ای انتخاب نماید که حداقل فاصله آن با هر یک از دسته‌های مورد نظر بیشینه شـــود. این نوع انتخاب باعث می‌شود که تصمیم‌گیری در عمل، شـــرایط نوفه‌ای را به خوبی تحمل نموده و همچنین پاسخ‌دهی مناسبی داشته باشـــد. این نوع انتخاب مرز، بر اساس نقاطی به نام بردارهای پشتیبان انجام می‌شود[٤].

الگوریتم‌هـــای مبتنی بر ماشـــین‌های بردار پشـــتیبان الگوریتم‌هایی هســـتند که ســـعی می‌کنند یک حاشـــیه[18] را بیشینه کنند. این الگوریتم‌ها برای پیدا کردن خط جدا کننده دسته‌ها، از دو خط موازی شروع کرده و این خطوط را در خلاف جهـــت یکدیگر حرکت می‌دهند تا هر کدام از خطوط به یک نمونه از یک دسته خاص در سمت خود برسد. پس



شکل۲: حاشیه خطوط دسته‌بندی نمونه

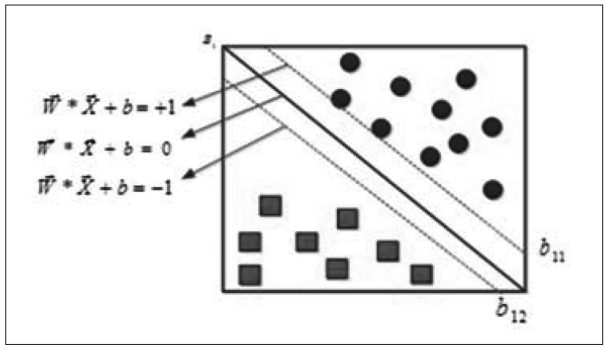
از انجام این مرحله، میان دو خط موازی یک نوار یا حاشیه شـــکل می‌گیرد. هر چه پهنای این نوار بیشتر باشد، به این معناست که الگوریتم توانسته حاشیه را بیشینه کند و هدف نیز بیشینه نمودن این حاشیه است. در واقع هدف این است که بیشترین مقدار ممکن برای این حاشیه انتخاب شود. در مرکز حاشیه، خط جدا کننده دسته‌ها یا همان خط مرکزی قـــرار می‌گیرد. حال از بین خطوطی که رســـم می‌شـــوند، الگوریتم، خطی را که حاشـــیه کناری آن بیشـــترین باشد، به‌عنوان جدا کننده دسته‌ها انتخاب می‌کند. حاشیه مربوط به دو خط b_1 و b_2 در شکل (۲) نمایش داده شده است.

رابطه محاسبه حاشیه به‌صورت رابطه (۳) است:

$$\text{Margin} = \frac{۲}{||\vec{w}||^۲} \qquad (۳)$$

w طـــول بـــردار وزن عمود بـــر صفحه می‌باشـــد. از محاســـبه حاشـــیه، الگوریتـــم، خـــط b_1 را به‌عنـــوان خط جداکننده انتخاب می‌کند. چرا که حاشـــیه کناری این خط، نسبت به حاشیه کناری خط b_2 بیشتر است. پس از انتخاب خط جداکننده، الگوریتم بر اســـاس مجموعه معادلات خط جداکننده و مجموعـــه معادلات خط موازی، تابعی را برای محاســـبه دســـته‌بندی رکوردهای جدید محاسبه می‌کند. مجموعه معادلات خط جداکننـــده b۱ و همچنین مجموعه معـــادلات خطوط موازی در شـــکل (۳) نمایش داده شـــده است. در این شکل $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_{\mathbf{1}} + \mathbf{b} = \mathbf{1}$ معادله خط b_{11} است. در نتیجه $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_{\mathbf{1}} + \mathbf{b} \geq \mathbf{1}$ اشـــاره به ســـمت راست این

17-Automatic Text Classification Algorithm
18- Margin



شکل۳: کمینه‌سازی حاشیه خط دسته‌بند در ماشین بردار پشتیبان

خط و در واقع اشـــاره به مناطقـــی دارد که رکوردهایی از نوع دســـته دایره در آن واقع شده‌اند. هنگامی که الگوریتم پس از قـــرار دادن مقادیـــر ویژگی‌های رکـــورد جدید در تابع، به رابطه $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_1 + \mathbf{b} \geq \mathbf{1}$ برســـد، مقدار ۱ را باز می‌گرداند، بدین معنا که رکورد جدید به دســـته دایره تعلق دارد. $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_1 + \mathbf{b} = -\mathbf{1}$ معادله خط b_{12} است. در نتیجه $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_1 + \mathbf{b} \leq -\mathbf{1}$ اشـــاره به ســـمت چپ این خط و در واقع اشـــاره به مناطقی دارد که رکوردهای از نوع دســـته مربع در آن واقع شـــده‌اند. در مواقعی که الگوریتم پس از قرار دادن مقادیر ویژگی‌های رکورد جدید در تابع، به رابطه $\vec{\mathbf{w}} \times \vec{\mathbf{x}}_1 + \mathbf{b} \leq -\mathbf{1}$ برســـد، مقدار ۱- را بـــر می‌گرداند، بدین معنا که رکورد جدید به دسته مربع تعلق دارد.

۳-۲-۲- آزمایش‌ها

در ایـــن بخش به بیـــان جزئیات پیاده‌ســـازی پرداخته شـــده و قبل از آن توضیحاتی در مورد مجموعه داده‌هایی که برای یادگیری و آزمایش دســـته‌بند استفاده شده، بیان می‌شود.

مجموعه داده‌ها[19]: در ایـــن مقاله مجموعه داده حقیقی اخبار رویترز ۲۱۵۷۸ [20] انتخاب شـــده است. این مجموعه داده در سال ۱۹۸۷ گردآوری و توسط گروهی از کارکنان خبرگزاری رویترز تهیه و شـــاخص‌گذاری شـــد. نســـخه اصلی این مجموعه داده شـــامل ۲۱۵۷۸ ســـند متنی شامل اخبار رویترز اســـت. مجموعه داده خبرگزاری رویترز یک مجموعه داده شناخته شده برای انجام آزمایش‌ها در زمینه

جدول۱: لیست اسناد مرحله آموزش و آزمایش برای مجموعه داده Reuters-21578

دسته‌ها	مرحله آموزش	مرحله آزمایش	جمع
Acq	۱۵۹۶	۶۹۶	۲۲۹۲
Trade	۲۵۱	۷۵	۳۲۶
Ship	۱۰۸	۳۶	۱۴۴
Interest	۱۹۰	۸۱	۲۷۱
Grain	۴۱	۱۰	۵۱
Crude	۲۵۳	۱۲۱	۳۷۴
Earn	۲۸۴۱	۱۰۸۳	۳۹۲۴
Money-fx	۲۰۶	۸۸	۲۹۴
Total	۵۴۸۶	۲۱۹۰	۷۶۷۶

متن از جمله دســـته‌بندی و خوشه‌بندی متن می‌باشد. این مجموعه داده با نام رویترز گردآوری شده است که شامل دو زیرمجموعـــه R8 و R52 می‌باشـــد. برای پیاده‌ســـازی سیستم پیشنهادی از مجموعه R8، که داده‌های به ۸ دسته و هر کدام به موضوع خاصی اشاره می‌کنند، استفاده شده است.

این مجموعه داده‌ها شامل ۷۶۷۶ سند متنی با اندازه‌های مختلف می‌باشد که در ۸ طبقه دسته‌بندی شده‌اند. هر سند به‌صورت دســـتی بر اســـاس محتویات و حوزه‌ای که در آن یافت می‌شد، برچسب‌گذاری شـــده و هر سند در فایل جداگانه‌ای که به وســـیله برچســـبی که دسته یا مجموعه دســـته را توصیف می‌کند، قرار داده می‌شـــود. جدول (۱) لیســـت مجموعه داده از خبرگـــزاری رویترز برای مرحله آموزش و آزمایش را نشان می‌دهد.

بـــرای جدا کـــردن مجموعـــه آمـــوزش و آزمایش از X-Validation اســـتفاده شده اســـت. تعداد زیرمجموعه‌ها برای این کار ۸ در نظر گرفته شده است و مجموعه اسناد به ۸ زیرمجموعه مســـاوی تقســـیم شـــده‌اند. هر بار یک زیرمجموعه به‌عنوان مجموعه آزمایش و زیرمجموعه‌های دیگر به‌عنوان مجموعه آموزش در نظر گرفته شـــده است. در نهایت میانگین نتایج به‌دست آمده محاسبه شده‌اند.

19-Data set
20-http://www.daviddlewis.com/resources/testcollections/reuters21578/

۳-۳- معیارهای ارزیابی

در مسایل دسته‌بندی متن معمولاً از معیارهای یادآوری[21]، دقت[22] و معیار F1 استفاده می‌شود که در زیر فرمول‌های آن آمده است [۱٤].

$$\text{Recall}=\frac{TP}{TP+FN} \quad (۴)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (۵)$$

$$F1=2\times\frac{\text{Recall}\times\text{Precision}}{\text{Recall}+\text{Precision}} \quad (۶)$$

TP: تعداد متونی که درست به یک طبقه منسوب شده‌اند.

FN: تعداد متونی که نادرست به یک طبقه منسوب شده‌اند.

FP: تعداد متونی که نادرست از یک طبقه رد شده‌اند.

و در نهایت برای ارزیابی کارایی روی تمام طبقات از روش میانگین استفاده شده است. در میانگین‌گیری کلان مقادیر دقت و یادآوری تمام طبقات محاسبه می‌شود. در این روش به همه طبقات وزن مساوی داده می‌شود .

پس از به‌دست آوردن دقت، یادآوری و F1 برای هر دسته، دو روش برای محاسبه میانگین این معیارها به‌کار می‌رود [۱٤]. در فرمول‌های (۷) و (۸) دقت ماکرو با precision^{M} و دقت میکرو با $\text{precision}\pi^{\mu}$ نمایش داده شده است.

$$\text{precision}^{M}=\frac{\sum_{i=1}^{|C|}\text{precision}_i}{|c|} \quad (۷)$$

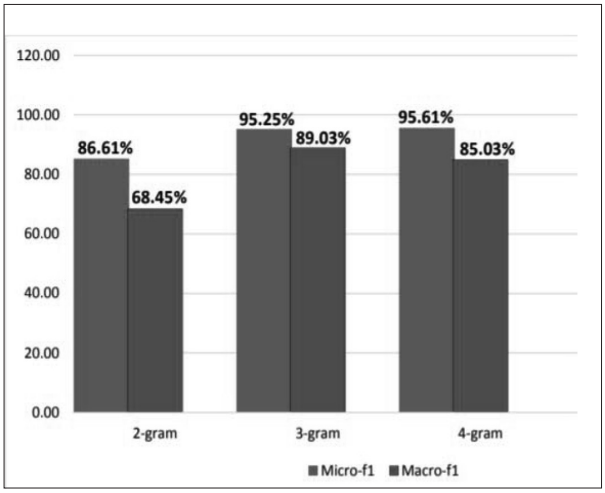
$$\text{precision}^{\mu}=\frac{\sum_{i=1}^{|C|}TP_i}{\sum_{i=1}^{|C|}(TP_i+FP_i)} \quad (۸)$$

در هر دو فرمول فوق منظور از |c| تعداد دسته‌ها (رده‌ها) است که در آزمایش فوق برابر ۸ است.

۴- نتایج، یافته‌های تجربی و ارزیابی

برای ارزیابی روش پیشنهادی از نرم‌افزار شبیه‌ساز RapidMiner استفاده شده است. در این مقاله، به منظور دسته‌بندی خودکار متون از سه روش شاخص‌گذاری

21- Recall
22- Precision



شکل۴: نتایج مدل پیشنهادی بدون حذف کلمات متوقف کننده

bigram، trigram و quadgram در دو حالت حذف کلمات متوقف کننده و بدون حذف کلمات متوقف کننده با استفاده از الگوریتم یادگیری ماشین W-SMO استفاده شده است.

نتایج ارزیابی دسته‌بند با استفاده از روش‌های شاخص‌گذاری و بدون حذف کلمات متوقف کننده در شکل (٤) نشان داده شده است. روش شاخص‌گذاری quadgram از لحاظ معیار میکرو-F1 و روش شاخص‌گذاری trigram از نظر معیار ماکرو-F1 بهترین کارایی را دارند.

همچنین برای هر کدام از دسته‌های خبری بدون حذف کلمات متوقف کننده معیارهای دقت و بازخوانی مورد ارزیابی قرار گرفته و نتایج حاصله در جدول (۲) نشان می‌دهد که بهترین کارایی الگوریتم پیشنهادی مربوط به دسته Earn با دقت ۹۷٫۸٦٪ و بازخوانی ۹۸٫۲٤٪ می‌باشد.

نتایج ارزیابی دسته‌بند با استفاده از روش‌های شاخص‌گذاری با حذف کلمات متوقف کننده با N-گرم‌های مختلف در شکل (٥) نشان داده شده است. روش شاخص‌گذاری bigram از لحاظ معیار ماکرو-F1 و میکرو-F1 نسبت به روش‌های trigram و quadgram کارایی بهتری دارد.

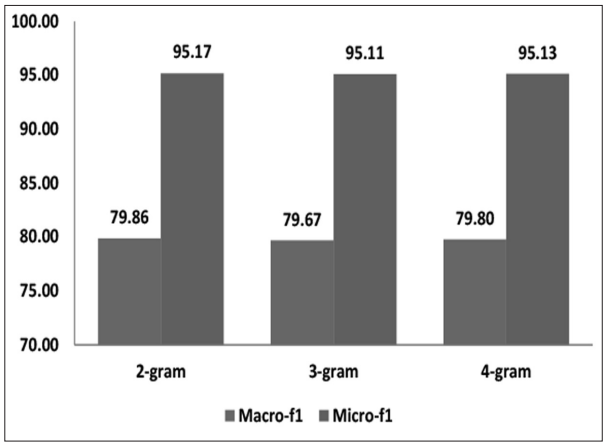
همچنین برای هر کدام از دسته‌های خبری با حذف کلمات متوقف کننده معیارهای دقت و بازخوانی مورد ارزیابی قرار گرفته، و نتایج حاصله در جدول (۳) نشان می‌دهد که بهترین کارایی الگوریتم پیشنهادی مربوط

جدول۲: جزئیات نتایج الگوریتم پیشنهادی برای هر دسته بدون حذف کلمات متوقف کننده

دسته	دقت	بازخوانی
Acq	٪۹۵,۴۹	٪۹۶,۸۷
Trade	٪۹۳,۷۵	٪۹۵,۶۲
Ship	٪۹۱,۸۴	٪۸۳,۳۳
Interest	٪۸۸,۸۹	٪۸۴,۲۱
Grain	٪۷۸,۲۶	٪۴۳,۹۰
Crude	٪۹۴,۸۰	٪۹۳,۶۸
Earn	٪۹۷,۸۶	٪۹۸,۲۴
Money-fx	٪۸۷,۰۲	٪۸۷,۸۶

جدول۳: جزئیات نتایج الگوریتم پیشنهادی برای هر دسته با حذف کلمات متوقف کننده

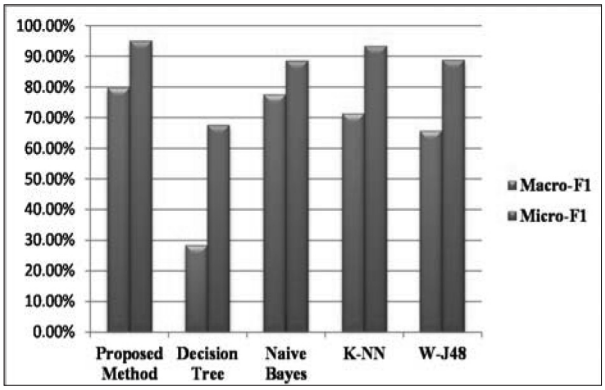
دسته	دقت	بازخوانی
Acq	٪۹۴,۰۶	٪۹۶,۲۶
Trade	٪۹۲,۳۱	٪۹۰,۶۷
Ship	٪۷۰,۰۰	٪۶۱,۱۱
Interest	٪۸۷,۴۳	٪۸۲,۷۲
Grain	٪۵۶,۵۲	٪۴۰,۰۰
Crude	٪۹۳,۴۲	٪۸۸,۴۳
Earn	٪۹۸,۷۳	٪۹۹,۲۶
Money-fx	٪۸۳,۹۴	٪۸۰,۶۴



شکل ۵: نتایج به‌دست آمده برای روش‌های شاخص‌گذاری با حذف کلمات متوقف کننده

به دسته Earn با دقت ٪۹۸,۷۳ و بازخوانی ٪۹۹,۲۶ می‌باشد.

با کمی تامل در نتایج به‌دست آمده در مدل پیشنهادی



شکل۶: بهترین نتایج به‌دست آمده بر اساس دو معیار ماکرو و میکرو روش پیشنهادی با روشهای K-NN، Naïve Bayes، W-j48و W-LADTREE

می‌توان نتیجه گرفت که مرحلهٔ پیش‌پردازش در بحث دسته‌بندی بسیار مهم می‌باشد. همانطور که نتایج به‌دست آمده نشان می‌دهد حذف کلمات متوقف کننده به بهبود نتایج می‌انجامد. همچنین می‌توان دلیل بالا بودن دقت و بازخوانی دسته Earn در مدل پیشنهادی را توزیع مناسب داده‌های آموزش و آزمایش در این دسته دانست. بنابراین یکی از مهم‌ترین پارامترهای موثر در کارایی الگوریتم‌های دسته‌بندی متون داشتن توزیع مناسب داده آموزش و آزمایش در یادگیری می‌باشد.

همچنین روش پیشنهادی با الگوریتم‌های یادگیری ماشین مثل K-NN، Naïve Bayes، W-j48و W-LADTREE ارزیابی شد که نتایج این ارزیابی در شکل (٦) نشان داده شده است. نتایج نشان می‌دهد کارایی دسته‌بندی با استفاده از ترکیب W-SMO و bigram از سایر ترکیب‌ها بیشتر می‌باشد. دقت دسته‌بندی در بهترین حالت به ٪۹۵,۱۷ دقت میکرو–F1 و ٪۷۹,۸۵ دقت ماکرو–F1 رسیده است.

٤-۱- مقایسه کارایی روش‌های پیشنهادی با کارهای دیگر

جدول (٤)، میانگین دقت میکرو و ماکرو F1 روش پیشنهادی را با کارهای دیگران در زمینه دسته‌بندی نشان می‌دهد. این جدول گویای برتری سیستم پیشنهادی نسبت به روش‌های قبلی می‌باشد.

۵- نتیجه‌گیری و پیشنهادهای آینده

امـــروزه بخش قابـــل توجهی از اطلاعـــات موجود در پایگاه داده‌های متنی یا اســـناد متنی ذخیره می‌شوند. یکی از مهم‌ترین مباحثی که مطرح اســـت بحث سازماندهی این اسناد می‌باشد. یکی از راهکارهای سازماندهی اسناد متنی، دسته‌بندی آن‌ها می‌باشد. دسته‌بندی متون به انتساب اسناد متنی به دسته‌های واقعی آن‌ها می‌باشد. دسته‌بندی اسناد متنی شـــامل دو مرحله اصلی انتخـــاب ویژگی و الگوریتم یادگیری می‌باشد.

در این مقاله روشی برای دسته‌بندی خودکار متون ارائه شده است. این روش با مجموعه داده استاندارد خبرگزاری رویتـــرز که شـــامل ۷۶۷۶ ســـند که در ۸ دســـته متفاوت طبقه‌بندی شده، ارزیابی گردید. با استفاده از آزمایش‌های مختلفـــی که روی روش‌های شـــاخص‌گذاری انجام شـــد، ترکیب روش شاخص‌گذاری 2-gram و الگوریتم یادگیری W-SMO با حذف کلمات متوقف کننده بهترین کارایی را دارد. همچنین روش پیشنهادی با الگوریتم‌های یادگیری ماشین K-NN، Naïve Bayes، W-j48 و W-LADTREE ارزیابی شـــد کـــه نتایج ارزیابی نشـــان داد روش پیشـــنهادی برای این مجموعه داده بهترین کارایی را نســـبت به این الگوریتم‌ها دارد. سرانجام روش پیشنهادی معیار ماکرو-F1 و معیار میکـــرو F1 به ترتیب مقادیر ۷۹,۸۵٪ و ۹۵,۱۷٪ را برای این مجموعـــه داده ارزیابی نمود. در زمینه دســـته‌بندی متون بررسی موارد زیر جهت پیشنهادهای آینده بیان می‌گردد.

- چگونگی توزیع بهتر مجموعه داده آموزش به نحوی کـــه بهترین کارایی را در دســـته‌بندی ایجاد نماید موضوع دیگری است که در جهت تخصصی کردن دسته‌بندها قابل پژوهش می‌باشد.

- آزمـــودن روش‌های پیشـــنهادی در ایـــن پژوهش با مجموعه دادههای اســـتاندارد دیگر (غیر از مجموعه داده رویترز)، جهت قطعی‌تر شـــدن نتایج حاصل با روش‌های پیشنهادی.

- اســـتفاده از پالایه‌های مختلف جهت دســـت یافتن به کارایی بهتر دســـته‌بندی متون می‌تواند مورد پژوهش قرار بگیرد.

- اســـتفاده از روش‌های پیشنهادی در این مقاله برای زبان‌های دیگر (غیر از زبان انگلیســـی) می‌تواند موضوع مناسبی برای پژوهش باشد.

جدول۴: مقایسه نتایج روش پیشنهادی با کارهای پیشین

مرجع	روش مورد استفاده	Data set	Micro-F1	Macro-F1	Accuracy
[10]	K-NN	Reuters-21578	-	-	٪۹۲.۵۵
	K-NN&SVM		-	-	٪۸۱.۴۸
[22]	SVM	20Newsgroup Reuters-21578 OHSUMED	٪۸۹	٪۹۰.۲۹	-
	RVM		٪۹۲.۲	٪۹۲.۳	-
[17]	JRIP	20Newsgroup	٪۸۹.۶	-	-
	SMO		٪۹۳.۲	-	-
	NAÏVE BAYES		٪۸۰.۲	-	-
[24]	KNN+IG	Reuters-21578	٪۷۸.۳	٪۷۵.۳	-
	KNN-DF		٪۷۰.۹	٪۶۷.۱	-
	FKNN+IG		٪۸۰.۴	٪۷۵.۵	-
	FKNN-DF		٪۷۶.۹	٪۷۳.۶	-
[12]	PSO	WebKB	٪۸۵.۹	٪۷۹	٪۹۴.۳۸
[31]	SVM R- Vector	Hamshahri	-	-	٪۹۰
[33]	PCA SVM	Hamshahri	-	-	٪۹۱,۸۶
[34]	PSA	7-NewsGroups	-	-	٪۸۸,۲
نتایج مدل پیشنهادی	W-SMO+N-gram (2-gram)	Reuters-21578	٪۹۵.۱۷	٪ ۷۹.۸۵	٪۹۵.۱۷

۶- مراجع

1. S. Eyheramendy, A. Genkin, W. H. Ju, D. D. Lewis, D. Madigan. (2003). Sparse Bayesian Classifiers for Text Categorization. Joint Statistical Meeting in San Francisco, California.
2. I. Guyon, A. Elisseeff. (2003). An Introduction to Variable and Feature Selection. Journal of Machine Learning Research,Vol. 3, pp(1157-1182).
3. Y. Lin, Y. Qu, Z. Wang. (2007). A Novel Feature Selection Algorithm for Text Categorization. Expert Systems with Applications, Vol. 33, pp(1–5).
4. A. Jain, R. D. Mishra. (2016). An Effective Approach for Text Classification. Proc. International Journal of Research in Engineering and Technology, Volume 05-06.
5. A. S. Tilve, S. N. Jain. (2017). A survey on machine learning techniques for text classification. international jour-

nal of engineering sciences & research technology. doi: 10.5281/zenodo.322477.
6. C. H. Wan, L. H. Lee , R. Rajkumar , D. Isa. (2012). A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-nearest neighbor and Support Vector Machine. Elsevir.
7. J. Sreemathy, P. S. Balamurugan. (2012). An Efficient Text Classification Using KNN and Naïve Bayesian. International Journal on Computer Science and Engineering (IJCSE), Vol. 4 No. 03.
8. Li Y. H. and Jain A. K. (1998). Classification of text documents. The Computer Journal 41(8), pp.537-546.
9. A. Guran, S. Akyokus, N. G. Bayazit, M. Zahidbgurbuz. (2009). Turkish Text Categorization Using n-gram word. International Symposium on Innovations in Intelligent Systems and Applications.
10. C. H. Wan, L. H. Lee. (2012). A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine. Expert Systems with Applications. Vol. 39, no. 15, pp. 11880-11888, Elsevir.
11. Cavnar, William B. (1993). N-Gram-Based Text Filtering For TREC-2. to appear in the proceedings of The Second Text Retrieval Conference (TREC-2), ed. by, Harman, D.K., NIST, Gaithersburg, Maryland.
12. B. L. Devi, & A. Sankar, (2015). Feature selection for web page classification using swarm optimization. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 9(1), 340-346.
13. Y. Huang. (2012). Support Vector Machines for Text Categorization Based on Latent Semanticindexing", Technical report, Electrical and Computer Engineering Department, Johns Hopkins University, 2012.
14. Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No.1, pp. 107-131.
15. M. H. Aghdam, N. Ghasem-Aghaee, M. E. Basiri. (2009). Text feature selection using ant colony optimization", Expert Systems with Applications,PP(6843–6853).
16. M.lan, C. L. Tan, J. Su, and Y. Lu. (2009). Supervised and traditional term weighting methods for automatic text categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 31, no. 4, pp. 72,735-1.
17. S. Chua, F. Coenen, G. Malcom, M. Fernando and G. Constatino. (2011). Using Negation and phases in Inducing Rules for Text classification. Research and Development in Intelligent Systems XXVIII. Pp.153-166: Springer.
18. J. Bakus and M. Kamel. (2002). Document classification using phases. Structural, Syntactic, and Statistical Pattern Recognition, pp. 557-565: Springer.
19. E. Montanes, J. R Quevedo and I. Daiz. (2003). A wrapper approach with support vector machines for text categorization. Computational Methods in Neural Modeling, pp 230-237: Springer.
20. L. Wang, N. Zhou and F. Chu. (2008). A general wrapper approach to selection of class-dependent features. Neural network, IEEE Transactions on, vol. 19, no. 7, pp. 1267-1278.
21. Y. Saeys, T. Abeel, and Y. Van de Peer. (2008). Robust feature selection using ensemble feature selection techniques. Machine learning and Knowledge Discovery in Databases, pp. 313-325:Springer.
22. M. Rafi and M. S. Shaikh. (2013). A comparison of SVM and RVM for Document classification. arXiv preprint arXiv:1301,2785.
23. A. Pradhan. (2012). Support vector machine -A Survey. International Journal of Emerging Technology and Advanced Engineering. Volume 2, Issue 8. pp 82-85.
24. M. A. Basiri, Sh. Nemati, N. Ghasemi. (2008). Comparison of Persian text classification using fknn, knn algorithms and selecting features based on information gain and document frequency. Thirteenth National Conference on Electrical Power Engineering of Iran.
25. A. Bagheri, M. Saraee, SH. Nadi. (2014). PSA:A Hybird Feature selection Approach for Persian Text Classification. Jurnal of Computing and Security, Vol.1, No.4, pp.261-272.
26. J. Hartmann, J. Hupperts, C. Schamp, M. Heitmann. (2018). Comparing automated text Classification methods. International Journal of Research in Marketing. IJRM-01274; No of Pages 26.
27. Data Science Platform [WWW Document]. RapidMiner(2018) Available online at: http://rapidminer.com/ (Accessed March 21, 2018).
28. http://www.daviddlewis.com/resources/testcollections/reuters21578/
29. http://www-2.cs.cmu.edu/afs/cs/project/theo-11/www/wwkb.
30. Sun, A., Lim, E. P., & Ng, W. K. (2002, November). Web classification using support vector machine. In Proceedings of the 4th international workshop on Web information and data management (pp. 96-99). ACM.
31. M. Yousefian, H. Fooladvand. (2018). Use the rbf coronel backup vector machine to improve the text categorization system. thesis of Khorramabad University.
32. http://ece.ut.ac.ir/dbrg/hamshahri/faindex.html.
33. I. Jalali, S. J. Mirabedini, A. Haronabadi. (2017). Provide a model for categorizing texts using a combination of categorization methods. Journal of Telecommunication Engineering. Vol. 7, no. 23, pp. 34-44.
34. A. Bagheri, M. Saraee, S. Nadi. (2014). PSA: A Hybrid Feature Selection Approach for Persian Text Classification. Journal of Computing and Security, Vol. 1, No. 4, pp. 261-272.