برچسب‌زنی موضوعی متون فارسی

هادی عبدی قویدل[۱]، بهرام وزیرنژاد[۲]، محمد بحرانی[۳] و مهدی مرادی[۴]
[۱] دانشگاه صنعتی شریف، hadi_abdighavidel@mehr.sharif.ir
[۲] دانشگاه صنعتی شریف، bahram@sharif.ir
[۳] دانشگاه صنعتی شریف، bahrani@sharif.ir
[۴] دانشگاه صنعتی شریف، mehdi_moradi@mehr.sharif.ir

چکیده – برچسب‌زنی موضوعی، محتمل‌ترین موضوعی که محتوای متن بدان اشاره دارد را مشخص می‌کند. برای نیل به این هدف، در مقالهٔ حاضر از روش فاصله‌یابی در فضای بردار بسامدی بهره گرفته شده است. در این روش فاصله بین بردارهای بسامد وزن دار کلمات کلیدی محاسبه می شود. کلمات کلیدی شامل کلمات با بیشترین حاصل‌ضرب بسامد پیکره‌ای و فاکتور وابستگی به طبقه هستند که مجموعه این کلمات از پیکرهٔ متنی زبان فارسی استخراج شده‌اند. در مرحلهٔ هرس، علاوه بر کلمات دستوری مانند حروف اضافه و ربط، کلمات کلیدی زاید نیز به صورت دستی بر اساس دیدگاه‌های زبان‌شناختی و معنی‌شناختی حذف می‌شوند. در گام بعد، برای هریک از طبقات مورد نظر یک بردار بسامد وزن دار کلمات کلیدی تشکیل می‌گردد. از بین این بردارها، برداری که کمترین فاصله را با بردار بسامد وزن‌دار متن آزمون دارد، طبقهٔ مربوط به متن آزمون را مشخص می‌نماید. نتایج حاصل شده در این مقاله نشان می‌دهد که هرس دستی کلمات کلیدی تاثیر بسزایی در بهبود عملکرد سیستم دارد. بدین ترتیب، میانگین معیار F در حدود ۰/۶۵ به دست می‌آید.

کلید واژه‌ها– برچسب‌زنی موضوعی، بردار بسامد وزن‌دار، بسامد پیکره‌ای، فاکتور وابستگی به طبقه، کلمات کلیدی، معیار F

۱. مقدمه

برچسب‌زنی موضوعی متون امری مهم در حوزهٔ بازیابی اطلاعات می‌باشد. این امر به نوعی دسته‌بندی یا طبقه‌بندی متون در زبان طبیعی است. سیستم معمولاً روی یک دسته از متون از قبل برچسب‌زنی شده آموزش داده می شود و سپس با استفاده از مدل‌های حاصل از مرحلهٔ آموزش، طبقه‌بندی متون جدید صورت می‌گیرد.

امروزه دسته‌بندی متون در بسیاری از زمینه‌ها از جمله فیلتر کردن متون مخصوصاً نامه‌های الکترونیکی، تشخیص طبقه، ابهام زدایی از کلمات، سیستم‌های خودکار پاسخ به سوالات و یا حتی نمره‌دهی به مقالات در سیستم‌های آموزشی و به طور کلی در هر کاربردی که سازماندهی مستندات و یا توزیع انتخابی و تطبیقی خاصی از مستندات مد نظر باشد، کاربرد دارد. برچسب‌زنی موضوعی متون با مسائلی چون استخراج اطلاعات و دانش از متون و داده‌کاوی متون دارای ویژگی های فنی مشترک می‌باشد.

اگرچه این مسئله از سال ۱۹۶۰ میلادی به بعد مورد مطالعه قرار گرفته است، اما با شروع دههٔ ۹۰ به لطف پیشرفت‌های نرم‌افزاری و سخت‌افزاری دسته‌بندی متون پیشرفت چشم‌گیری داشته است. در تکنیک یادگیری ماشین، طبقه‌بندی‌کننده‌ها با استفاده از یادگیری از یک مجموعهٔ مستندات از پیش طبقه‌بندی شده مشخصات دسته‌ی جدید را معین می‌سازند. برای ساختن طبقه بندی‌کننده‌ها، نیاز مبرم به دانش مهندسی و زبان‌شناسی افراد خبره وجود دارد، اما اگر به جای استفاده از ماشین، طبقه‌بندی به صورت دستی انجام گیرد علاوه بر زمان‌بری و هزینه‌ی زیاد معایب زیر را با به همراه خواهد داشت:

۱. برای زمینه‌های تخصصی خاص نیاز به دانش افراد خبره دارد (مانند زمینه‌های پزشکی، مهندسی و غیره).

۲. برچسب‌زنی موضوعی متون به صورت دستی مبتنی بر دانش و تجربهٔ فرد می‌باشد، از این‌رو درصد خطاپذیری آن بسیار زیاد است.

۳. تصمیم دو فرد خبره در عمل برچسب‌زنی موضوعی متون می تواند متفاوت و یا حتی ناسازگار باشد.

دسته‌بندی متون با روش‌های مختلف برای زبان انگلیسی صورت گرفته است. Yang و Liu در سال ۱۹۹۹ دسته‌بندی متون را با استفاده از بردارهای فراوانی ریشهٔ کلمات انجام دادند [۱]. Joachims در سال ۱۹۹۸ دسته‌بندی متون را با استفاده از ماشین بردار پشتیبان انجام داد [۲]. Bellegarda در سال ۲۰۰۰ روش آنالیز معنایی پنهان (LSA) را برای دسته‌بندی به کار برد [۳]. Wood و Gedeon در سال ۲۰۰۱ از شبکه‌های عصبی هیبرید به منظور دسته‌بندی متون استفاده کردند [۴]. در همین سال Torkolla آنالیز تمایزی خطی را در دسته‌بندی به کار گرفت [۵]. Blei و همکاران در سال ۲۰۰۳ روش «تخصیص دیریکلهٔ پنهان» (LDA) را برای مدل‌سازی متون پیشنهاد دادند و از آن در دسته‌بندی متون نیز استفاده کردند [۶]. در سال ۲۰۰۵ نیز Guandong و همکاران روش تحلیل معنایی پنهان احتمالاتی (PLSA) را برای دسته‌بندی صفحات وب به کار گرفتند [۷].

تحقیقات انجام‌شده در زمینهٔ دسته‌بندی متون برای زبان فارسی تا کنون بسیار اندک بوده است. عرب‌سرخی و فیلی یک روش دسته‌بندی با استفاده از بردارهای فراوانی ریشهٔ کلمات و الگوریتم بیزین ساده پیشنهاد داده‌اند. سپس آنها با ترکیب روش بیزین با ایدهٔ نگهداری کلمات همنشین، روش خود را بهبود بخشیدند [۸]. حاجی‌حسینی و الماس‌گنج نیز یک روش بانظارت برای دسته‌بندی متون فارسی با استفاده از تحلیل معنایی پنهان (LSA) پیشنهاد دادند. روش LSA بردارهایی را در یک فضای برداری کاهش بُعد یافته برای هر متن در اختیار قرار می‌دهد. با استفاده از این بردارها آنها از روش شبکهٔ عصبی برای آموزش دسته‌بند و تعیین دستهٔ مربوط به متون جدید استفاده کردند [۹]. پیله‌ور و همکاران، با استفاده از یادگیری چندی‌سازی برداری دسته‌بندی مستندات متنی فارسی را برروی پیکرهٔ همشهری انجام دادند [۱۰]. در مقاله‌ای دیگر فرهودی و یاری، با استفاده از روش بهره‌جویی از گنج‌واژه و انتخاب ویژگی دو مرحله‌ای به دسته‌بندی متون فارسی پرداخته‌اند [۱۱].

در مقالهٔ حاضر از بردار بسامد وزن‌دار کلمات کلیدی در متون آموزشی هر کلاس به عنوان مدلی برای طبقه‌بندی استفاده شده است و برچسب‌زنی متون جدید با استفاده از تعیین میزان شباهت یا فاصلهٔ بین بردار متن جدید با بردار هر کلاس صورت می‌گیرد. عمده ایدهٔ به‌کاررفته در تهیهٔ بردار کلاس‌ها، بررسی دیدگاه‌های زبان‌شناختی اعم از صرف و معنی‌شناسی در استخراج کلمات مهم یا کلیدی و مخصوصا هرس کلمات زاید به صورت دستی می‌باشد.

۲. روش کار

۲,۱. نرمال‌سازی متن فارسی

تمامی حوزه‌های مرتبط با پردازش زبان طبیعی به نحوی از انحنا با متون واقعی سروکار دارند. صورت‌های غیر استاندارد نویسه‌ها وکلمات به وفور در این نوع متون دیده می‌شوند . قبل از اینکه بتوان از این متون به منظور استفاده در سیستم‌های تبدیل متن به گفتار، ترجمه ماشینی، بازشناسی حروف فارسی، خلاصه‌ساز فارسی، جستجو در متون فارسی و ... استفاده کرد و یا در پایگاه داده ذخیره نمود، باید ابتدا پیش‌پردازشی روی آنها انجام گیرد تا صورت‌های غیر استاندارد[1] به شکل استاندارد تبدیل گردند. اگر حروف، نشانه‌های نگارشی و کلمات فارسی به شکل یکسانی نوشته نشوند، متون مورد استفاده قابل تحلیل توسط سامانه‌های رایانه‌ای نخواهند بود. طی فرایند نرمال‌سازی، علایم نگارشی، حروف، فاصله‌های بین کلمات، اختصارات و غیره بدون ایجاد تغییرات معنایی در متن به شکل استاندارد تبدیل می‌گردند. در این مقاله از نرمال‌ساز متنی تهیه شده در آزمایشگاه پردازش گفتار و زبان دانشگاه صنعتی شریف استفاده می‌شود.

[1] . NSWS

۲,۲. دادگان

در این مقاله از پیکرهٔ متنی زبان فارسی [۱۲] به عنوان دادگان آموزشی استفاده می‌شود. مستندات مربوط به بخشی از این پیکره که شامل حدود ۱۰ میلیون کلمه است، با برچسب‌های موضوعی مشخص شده است. این برچسب‌ها در کل بیش از ۶۰ موضوع را شامل می‌شود که البته حجم متون مربوط به بعضی از موضوع‌ها بسیار اندک می‌باشد. در این مقاله از بین مستندات این پیکره، زیرپیکره‌ای شامل تعدادی از مستندات در ۱۰ موضوع ادبیات داستانی، اقتصادی، اجتماعی، فرهنگی، هنری، مذهبی، پزشکی، سیاسی، تاریخی و ورزشی انتخاب شده است. این زیرپیکره که دارای ۸۴۸۶ مستند می‌باشد به عنوان مجموعهٔ آموزش و آزمون و همچنین به منظور استخراج کلمات کلیدی به کار می‌رود.

دادگان دیگری که در این مقاله از آن به منظور هرس کردن کلمات کلیدی استفاده شده است، واژگان زایای زبان فارسی [۱۳] است. این مجموعه واژگان شامل بیش از شصت هزار کلمه همراه با ویژگی‌های زبان‌شناختی و صرفی آنهاست. کلمات نهفته در این واژگان شامل تقریباً تمامی تکواژهای رایج زبان فارسی اعم از تکواژهای آزاد و مقید می‌باشد.

۲,۳. الگوریتم بردار فاصله

الگوریتم پیشنهادی برای برچسب‌زنی موضوعی بدین صورت عمل می‌کند که ابتدا تعدادی کلمه از بین کلمات دادگان آموزشی به عنوان کلمهٔ کلیدی انتخاب می‌شود (نحوهٔ استخراج کلمات کلیدی در بخش بعد توضیح داده شده است). در مرحلهٔ بعد بسامد وزن‌دار کلمات کلیدی در متون آموزشی مربوط به هر موضوع استخراج می‌شود. به‌ازای هر موضوع، بردار بسامد وزن‌دار کلمات کلیدی به عنوان بردار نمایندهٔ آن کلاس در نظر گرفته می‌شود. در مرحلهٔ آزمون، برای برچسب‌زنی موضوعی به یک متن جدید، ابتدا بردار بسامد وزن‌دار کلمات کلیدی متن جدید استخراج گشته و سپس فاصلهٔ این بردار با تک‌تک بردارهای مربوط به کلاس‌ها سنجیده می‌شود. کلاس یا موضوعی که بردار مربوط به آن کمترین فاصله را از بردار ورودی داشته باشد، به عنوان محتمل‌ترین موضوع برای متن جدید در نظر گرفته می شود. در این مقاله از معیار کسینوسی به عنوان :معیار فاصلهٔ بین بردارها استفاده شده است. در این روش موضوع متن جدید از رابطهٔ زیر به‌دست می‌آید:

$$t = \arg\min_{w_i} \cos^{-1} \frac{v.w_i}{|v||w_i|} \qquad (1)$$

که v بردار مربوط به متن جدید و w_i بردار مربوط به موضوع i می‌باشد.

۲,۴. استخراج کلمات کلیدی:

مهم ترین و پایه‌ای‌ترین بخش سیستم طبقه بندی کننده متون، تهیهٔ کلمات کلیدی است. تهیهٔ این مجموعه لغات با در دست داشتن پیکرهٔ متنی زبان فارسی صورت پذیرفت. روش‌های گوناگونی برای استخراج کلمات کلیدی وجود دارد که معروفترین آنها حاصل‌ضرب tf-idf [۱۴] می‌باشد. tf متناسب با بسامد یک کلمه در مستند و idf یک فاکتور وزنی است که بیانگر معکوس میزان پراکندگی یک کلمه در سندهای مختلف است. برای استخراج کلمات کلیدی، tf-idf را برای هر کلمهٔ i به روش زیر به دست می‌آوریم:

$$tf\_idf_i = tf_i \times idf_i = \frac{n_i}{\sum_i n_i}.\log\frac{M}{df_i} \qquad (2)$$

در فرمول(۲)، n_i تعداد کلمهٔ i در زیرپیکره، M فراوانی کل سندها در زیر پیکره و df_i تعداد مستنداتی است که شامل کلمهٔ i می‌باشند. با در نظر گرفتن حد آستانه‌ای روی tf-idf می‌توان کلمات کلیدی را انتخاب کرد. کلمات کلیدی کلماتی هستند که دارای tf-idf بالایی باشند. با توجه به تعریف فاکتور tf-idf روشن است که به دلیل وجود فاکتور tf کلماتی که بسامد وقوع بالایی دارند امتیاز بیشتری برای انتخاب به عنوان کلمه کلیدی خواهند داشت. درعین‌حال این توضیح لازم است که تمام کلماتی که دارای tf بالایی هستند، ارزش استفاده در فرآیند طبقه بندی را ندارند. برای مثال می‌توان به کلمهٔ "است" اشاره داشت که تعداد وقوع بسیار بالایی دارد اما کلمهٔ کلیدی محسوب نمی‌شود. در اینجا فاکتور idf به کار می‌آید. این فاکتور در مورد کلماتی که در تعداد محدودی از مدارک دیده شده باشند بالاست. برای مثال در مورد کلمه "است" این فاکتور مساوی صفر است! بنابراین با توجه به اینکه معیار انتخاب کلمات کلیدی حاصل‌ضرب دو فاکتور احتمال وقوع یا بسامد در پیکره و فاکتور تمرکز idf است، کلماتی به عنوان کلمه کلیدی انتخاب می‌شوند که ضمن داشتن بسامد وقوع بالا در تعداد محدودی سند واقع شده باشند و به عبارتی در اسناد مربوط به طبقه خاصی به تعداد زیاد دیده شوند.

در مرحلهٔ بعدی، کلماتی که در رساندن معنای متن تاثیری ندارند حذف می‌گردند. نام متداول اینگونه کلمات در مقابل

کلمات محتوایی که دارای معنی هستند دستوری می‌باشد. عمده دلیل حذف این کلمات به خاطر عدم کارآمدی آنها در طبقه بندی متون و افزایش سرعت پردازش ماشین می‌باشد. در بسیاری از کارهای قبلی، مشکل عمده، نداشتن مجموعهٔ منسجم و منظم از این کلمات برای زبان فارسی بود. در این تحقیق کلمات دستوری به کمک واژگان زایای زبان فارسی و بر اساس دیدگاه‌های صرفی و معناشناختی جمع آوری شده‌اند که در جدول زیر به همراه مثال‌هایی قابل مشاهده می‌باشد.

جدول شماره ۱ – لیست انواع کلمات دستوری که چون دارای معنی و محتوا نمی باشند از لیست کلمات کلیدی حذف می‌شوند.

مقولهٔ نحوی	مثال	مقولهٔ نحوی	مثال
صفت اشاره	آن، همان، همین، هر ...	**ضمیر پرسشی**	آیا، چند، چرا، کی...
صفت مبهم	اندکی، بهمان، برخی...	**اعداد**	۱، ۲، ۳، ۴....
حروف الفبا	الف، ب، ج، د...	**عدد حرفی**	یک، اول، دومین، سه...
حروف ربط	اگر، اما، آنچنانکه، چه...	**حرف اضافه پسین**	را
حروف ربط گروهی	اگرچه، اگرنه، از همین‌رو...	**حروف اضافه**	از، به، در، بر...
ضمیر فاعلی	آنان، اینجانب، ما...	**فعل وجه نما**	باید، بایست، توان، بایستی...
ضمیر اشاره	این، آن، همان‌ها...	**فعل کمکی**	باش، داشت، خواه...
ضمیر تاکیدی انعکاسی	خودم، خودش...	**فعل اسنادی**	بودن، شدن، است، بود،...
ضمیر مشترک و متقابل	هم، همدیگر،		

جدول شماره ۲- تعداد مستندات به‌کارگیری‌شده در هر کلاس در بخش آموزش و ارزیابی

کلاس	تعداد کل مستندات	تعداد مستندات آموزش	تعداد مستندات آزمون
ادبیاتی داستانی	۲۲۵	۲۰۰	۲۵
اقتصادی	۱۲۲۴	۱۱۱۷	۱۰۷
اجتماعی	۱۲۸۷	۱۲۳۷	۵۰
فرهنگی	۱۰۰۶	۹۵۸	۴۸
هنری	۴۶۹	۴۲۲	۴۷
مذهبی	۳۴۹	۳۱۴	۳۵
پزشکی	۷۱۴	۶۳۹	۷۵
سیاسی	۲۴۲۷	۲۲۱۰	۲۱۷
تاریخی	۲۷۳	۲۵۹	۱۴
ورزشی	۴۶۰	۴۱۴	۴۶
کل	۸۴۸۶	۷۸۴۰	۶۴۶

علاوه بر این کلمات، علائم نگارشی که در واژگان زایا در قالب لیست منسجمی می‌باشند نیز حذف می‌گردند.

در مرحلهٔ بعدی تعدادی از واژه‌های دیگر که دارای اهمیت کمتری هستند، به صورت دستی حذف می‌شوند [۱۵] و برداری شامل هزار کلمهٔ کلیدی تشکیل می‌گردد. نکتهٔ قابل‌توجه این است که تعدادی از کلمات نیز که به صورت صرف‌های مختلف از افعال اسنادی در زمان‌های مختلف هستند، به دلیل نبود تحلیل گر صرفی، به صورت دستی هرس می‌گردند [۱۶] تا هزار کلمهٔ انتخاب شده به صورت صحیح و از میان کلمات محتوایی و معنی دار در بردار قرار گیرند.

۳. آزمایش و نتیجه

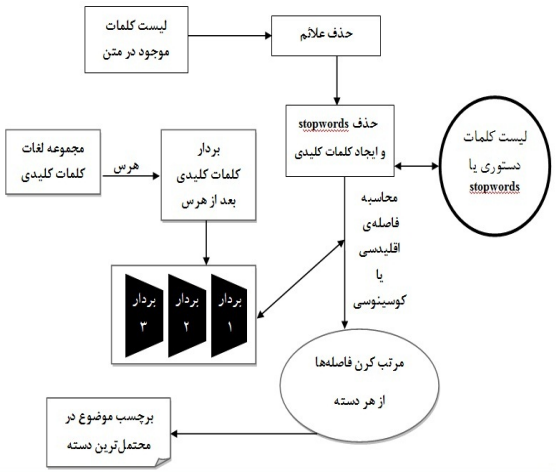
همان‌طور که گفتیم در این مقاله از زیرپیکره‌ای شامل ۸۴۸۶ مستند از پیکرهٔ متنی زبان فارسی به عنوان دادگان آموزش و آزمون استفاده می‌کنیم. این مستندات شامل ۱۰ موضوع است. جدول ۲ تعداد مستندات آموزش و آزمون را به تفکیک موضوع نشان می‌دهد.

متون پیکرهٔ متنی زبان فارسی نیاز به نرمال‌سازی دارد. پس متون را نخست باید با استفاده از نرمال‌ساز متنی، نرمال کنیم.

سپس با استفاده از روشی که در بخش قبل شرح دادیم، لیستی از هزار کلمهٔ کلیدی را از متون آموزشی استخراج می‌کنیم. بردار نمایندهٔ هر موضوع با استفاده از حاصل‌ضرب tf-idf مربوط به هر کلمهٔ کلیدی در متون همان موضوع به دست می‌آید.

در مرحلهٔ آزمون، ابتدا بردار مربوط به هر یک از متون آزمون را استخراج کرده و فاصلهٔ آن را با بردار مربوط به هر کلاس می سنجیم. موضوعی که بردار آن کمترین فاصله را با بردار متن آزمون دارد، محتمل‌ترین موضوع برای آن متن است.

شکل ۱ مراحل آموزش و آزمون سامانهٔ طبقه‌بندی متون را نمایش می‌دهد.



شکل ۱- مراحل آموزش و آزمون سامانهٔ طبقه بندی متون

معیارهای دقت و بازخوانی به صورت گسترده برای ارزیابی طبقه بندی متون به کار می‌رود. این معیارها به‌صورت زیر تعریف می شوند:

$$\text{دقت} = \frac{a}{a+b} \qquad (3)$$

$$\text{بازخواني} = \frac{a}{a+c} \qquad (4)$$

که a تعداد مستندات مرتبط با موضوع است که به طور صحیح برچسب‌زنی شده‌اند، b تعداد مستندات غیرمرتبط با موضوع است که به اشتباه برچسب موضوع موردنظر را خورده‌اند و c تعداد مستندات مرتبط با موضوع است که به اشتباه برچسبی غیر از موضوع موردنظر را خورده‌اند.

جدول ۳ نتایج ارزیابی طبقه‌بندی را برروی متون آزمون نشان می‌دهد. در این جدول معیارهای دقت و بازخوانی برروی متون آزمون به‌ازای هر دسته از موضوع‌ها آمده است. معیار F (ستون سمت چپ) میانگین توافقی معیارهای دقت و بازخوانی است.

لازم به ذکر است که متون اجتماعی چون همواره سایر متون را دربر می‌گیرند دارای دقت و بازخوانی کمتری نسبت به بقیهٔ مستندات می‌باشند. سایر متون دارای نتایج بهینه‌ای هستند که می‌تواند برای بسیاری از متونی که از سایت‌های مختلف خبری استخراج می‌شود، درست کار کند.

جدول شماره ۳- نتایج ارزیابی طبقه‌بندی برروی متون آزمون

نام کلاس	دقت	بازخوانی	معیار F
ادبیات داستانی	۰/۳۳	۰/۷۶	۰/۴۶
اقتصادی	۰/۸۴	۰/۸۸	۰/۸۶
اجتماعی	۰/۵۷	۰/۴۲	۰/۴۸
فرهنگی	۰/۷۶	۰/۴۵	۰/۵۶
هنری	۰/۵۰	۰/۷۲	۰/۵۹
مذهبی	۰/۵۷	۰/۷۵	۰/۶۴
پزشکی	۰/۸۸	۰/۹۱	۰/۸۹
سیاسی	۰/۴۶	۰/۸۹	۰/۶۱
تاریخی	۰/۴۰	۰/۵۰	۰/۴۴
ورزشی	۰/۹۰	۱	۰/۹۴

۴. نتیجه‌گیری و بحث

در این مقاله به ارائه روشی برای استخراج کلمات کلیدی با استفاده از معیار tf-idf پرداختیم که بخوبی می تواند با در نظر گفتن احتمال وقوع کلمات و نیز فاکتور تمرکز، کلمات مناسب برای طبقه بندی متون را شناسایی نماید. همچنین الگوریتمی بر پایهٔ فاصلهٔ بین بردارهای بسامد کلمات کلیدی، به منظور برچسب‌زنی موضوعی و طبقه‌بندی متون ارائه دادیم. مشاهده شد که هرس دستی کلمات غیر کلیدی که به اشتباه در زمرهٔ کلمات کلیدی قرار گرفته‌اند، سبب ایجاد تاثیر بسزایی در افزایش معیارهای دقت و بازخوانی در ارزیابی می‌گردد. کار هرس بر پایه دانش زبان‌شناختی و معنی‌شناختی انجام شد. هرچند نتیاج رضایت‌بخشی در قسمت ارزیابی سیستم مشاهده گردید ولی در آینده قصد داریم با استفاده از یک تحلیل گر صرفی زبان فارسی کارایی سیستم را افزایش دهیم.

مراجع

[1] Y. Yang and X. Liu, "A Re-examination of Text Categorization Methods", *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pp. 42-49, 1999.

[2] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features in Machine Learning", *10th European Conference on Machine Learning*, pp. 137-142, 1998.

[3] J.R. Bellegarda, "Exploiting Latent Semantic Information in Statistical Language Modeling", *Proceedings of IEEE*, Vol. 88, No. 8, pp. 1279-1296, 2000.

[4] S.A. Wood and T.D. Gedeon, "A Hybrid Neural Network for Automated Classification" *Proceedings of the 6th Australasian Document Computing Symposium*, 2001.

[5] K. Torkolla, "Linear Discriminant Analysis in Document Classification", *IEEE ICDM workshop on text mining*, 2001.

[6] D., Blei, A. Ng, M. Jordan, "Latent Dirichlet Allocation", Journal of Machine Learning Research, Vol. 3, pp. 993-1022, 2003.

[7] X. Guandong, Y. Zhang, Z. Zhou, "Using Probabilistic Latent Semantic Analysis for Web Page Grouping", *Proceedings of Research Issues in Data Engineering: Stream Data Mining and Applications*, pp. 29-36, 2005.

[۸] محسن عرب‌سرخی، هشام فیلی، "ارائه یک سیستم دسته بندی موضوعی متون فارسی بر اساس روش‌های احتمالاتی"، *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۱۵۱-۱۶۱، ۱۳۸۵.

[۹] آزاده حاجی‌حسینی، فرشاد الماس‌گنج، "دسته‌بندی موضوعی متون فارسی بر اساس روش آنالیز معنایی پنهان"، *مجموعه مقالات دومین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۱۹۰-۲۰۱، ۱۳۸۵.

[10] T. Pilehvar, H. Faili, M. Soltani, "Classification of Persian textual documents using Learning Vector Quantization", *4th IEEE Conference on Knowledge Engineering and Natural Language Processing*, NLP-KE, 2009.

[11] M. Farhoodi, A., Yari, M. Mahmoudi., "A Persian Web Page Classifi er Applying a Combination of Content-Based andContext-Based Features", *International Journal of Information Studies*, Vol. 1, No. 4, 2009.

[12] Bijankhan, M. & J. Seikhzadeghan & M. Bahrani & M. Ghayoomi, "Lessons from Creation of a Persian Written Corpus: Peykare". Language Resources and Evaluation Journal. Vol. 45, No. 2. 143-164, 2011.

[۱۳] محرم اسلامی، مسعود شریفی، صدیقه علیزاده، طاهره زندی، "واژگان زایای زبان فارسی"، *مجموعه مقالات اولین کارگاه پژوهشی زبان فارسی و رایانه*، صص ۶-۱۱، ۱۳۸۳.

[14] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval", Information Processing and Management, 24(5):513–523, 1988.

[۱۵] ویدا شقاقی، *مبانی صرف*، سازمان مطالعه و تدوین کتب علوم انسانی دانشگاه‌ها (سمت)، مرکز تحقیق و توسعهٔ علوم انسانی، تهران، ۱۳۸۹.

[۱۶] خسرو کشانی، *اشتقاق پسوندی در زبان فارسی امروز*، مرکز نشر دانشگاهی، تهران، ۱۳۷۱.