علوم و مهندسی کامپیوتر
نشریه علمی پژوهشی انجمن کامپیوتر ایران
مجلد ۸، شماره ۱ و ۳ (الف)، ۱۳۸۹
صفحات ۸-۱۳
مقاله کوتاه

کاوش متون فارسی بر مبنای روش طبقه‌بندی

محمد حسین سرایی آذر شاهقلیان

دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی اصفهان، اصفهان، ایران

چکیده

سازمان‌دهی اطلاعات On-line در قالب طبقه‌بندی مستندات زبان طبیعی در دسته‌های از قبل مشخص شده یکی از روش‌های مهم مدیریت اطلاعات محسوب می‌شود. با توجه به اهمیت موضوع و کاری که در این زمینه برای زبان‌های دیگر دنیا انجام گرفته است، نیاز به طبقه‌بندی متون فارسی به خوبی احساس می‌شود. در این مقاله از یادگیری ماشین برای ارائه روشی در طبقه‌بندی متون فارسی استفاده می‌شود. روش ارائه شده، تحت سیستم نرم‌افزاری طبقه‌بند متون فارسی، طراحی و پیاده‌سازی شده است. سیستم طبقه‌بند متون فارسی در فاز یادگیری، مجموعه‌ای از متون آموزشی را برای استخراج ویژگی‌های دسته‌ها بررسی می‌کند تا خصوصیات اصلی هر دسته را بدست آورد. به‌طوریکه در فاز تست سیستم طبقه‌بند متون فارسی، این ویژگی‌های مختص دسته، برای طبقه‌بندی متون دسته‌بندی نشده به‌کار می‌روند. از ریشه‌یابی برای کاهش بعد بردارهای ویژگی استفاده می‌شود. دقت روش پیشنهادی روی مجموعه جمع‌آوری شده‌ای از اخبار فارسی در هفت دسته مورد آزمایش قرار گرفته است. نتایج حاصله نشان می‌دهد که طبقه‌بند پیشنهادی برای دسته‌های اقتصادی، سیاسی و ورزشی دقت بسیار خوبی دارد. در سایر دسته‌ها نیز نتایج مورد قبول می‌باشد.

کلمات کلیدی: یادگیری ماشین، Text Mining، الگوریتم KNN، زبان فارسی.

۱- مقدمه

واضح است که در سازمان‌های جدید، دانش[1] فاکتور کلیدی برای رقابت سالم می‌باشد. دستیابی به موفقیت و باقی‌ماندن در صحنه رقابت به صورت گسترده به قابلیت یافتن اطلاعات مفید در زمان مناسب بستگی دارد. از طرفی رشد اطلاعات و تکنولوژی‌های ارتباطی به صورت بیش از اندازه باعث افزایش اطلاعات دردسترس شده است. لذا، فاکتور کلیدی، تنها خود اطلاعات نیست؛ بلکه قابلیت مدیریت و اعمال نفوذ بر این اطلاعات به صورت موفق، از اهمیت بالایی برخوردار است [۱].

توجه به این نکته ضروری است که بخش قابل توجهی از اطلاعات موجود در پایگاه‌های داده‌ای متنی، ذخیره شده‌اند. لذا از Text Mining برای مقایسه متون مختلف، رتبه‌بندی مهمترین و مرتبط‌ترین متون و یا یافتن الگوها و رفتارهای بین متون مختلف استفاده می‌باشد. متن‌کاوی عهده‌دار این وظایف می‌باشد [۲]. یکی از این وظایف طبقه‌بندی مستندات زبان طبیعی به طبقه‌های از قبل مشخص شده می‌باشد که از جمله روشهای مهم برای سازماندهی اطلاعات On-line می‌باشد. این وظیفه بطور معمول تحت عنوان دسته‌بندی متون[2] مورد بررسی قرار گرفته و به عنوان زیرمجموعه‌ای متن‌کاوی محسوب می‌شود. ازجمله کاربردهای طبقه‌بندی می‌توان به طبقه‌بندی صفحات وب، شاخص گذاری آیتم‌های خبری در منابع مختلف اینترنتی و دسته‌بندی موضوعی در زمینه‌های تجارت، پزشکی و بیوانفورماتیک اشاره نمود.

روش یادگیری ماشین[3] با ساختن قوانین طبقه‌بندیمتن می‌تواند بر همه این کمبودها فائق آید. برمبنای این روش، تعدادی از مستندات آموزشی طبقه‌بندی شده به صورت دستی، داده می‌شود. قوانین طبقه‌بندی متن باید به نحوی آموزش داده شوند که بتوانند مسأله یادگیری بانظارت (supervised) را به راحتی حل کنند [۳]. فرایند استنتاج عمومی به صورت اتوماتیک، قوانین را بوسیله یادگیری مشخصه‌های دسته‌ها از مجموعه مستندات از قبل برچسب‌گذاری شده، می‌سازد.

طبقه‌بندی متن که اغلب کلاسه‌بندی[4] نامیده می‌شود، متون زبان طبیعی را به یک یا بیشتر دسته‌های از قبل معرفی شده بر اساس محتوی نسبت می‌دهد[۳][۴]. مراحل انجام کار طبقه‌بندی به این صورت است که ابتدا پردازش

زبان طبیعی[۵] انجام شده، سپس با به کارگیری روش یادگیری اتوماتیک، این نتایج تفسیر می‌شود. وظایفی نظیر بازیابی متن و طبقه‌بندی متن از این تفسیر استفاده می‌کنند [۳] [۵]. هدف نهایی، طبقه‌بندی متون در تعداد ثابتی از دسته‌های از قبل معرفی شده می‌باشد. با توجه به این تفاسیر طبقه‌بندیمتون به عنوان زیرمجموعه‌ای از Text Mining محسوب می‌شود.

۲- مروری بر دیدگاه‌ها و روش‌های موجود طبقه‌بندی متون

هنگام بررسی طبقه‌بندی متون، اشاره به انواع طبقه‌بندیهای موجود لازم است. در کل طبقه‌بندیبه دو دسته انحصاری و غیر انحصاری تقسیم می شود. در طبقه‌بندی انحصاری، هر شیء دقیقاً به یک دسته وابسته می‌شود. در حالیکه در طبقه‌بندی غیر انحصاری می‌تواند به چند دسته اختصاص یابد و اصطلاحاً در این مورد گفته می‌شود که همپوشانی دارد. برای مثال، دسته بندی گروهی از افراد با وزن یا قد یکسان، از نوع انحصاری است و دسته‌بندی گروهی از افراد دارای بیماری، غیرانحصاری می‌باشد. زیرا یک شخص می‌تواند بطور همزمان به چند بیماری دچار شود.

طبقه‌بندی انحصاری می‌تواند به دو زیر شاخه تقسیم شود:

۱- ذاتی: intrinsic یا Unsupervised

۲- خارجی: extrinsic یا Supervised

تفاوت میان این دو دسته در این است که در دسته دوم، دسته‌های از قبل مشخص شده را برای طبقه‌بندی اشیاء (یعنی مستندات) مورد استفاده قرار می‌دهد. در حالیکه اولین گروه، یک راه پیش‌بینی شده را برای طبقه‌بندی مشخص می‌کند.

زیرشاخه ذاتی[۶] خود به دو زیر شاخه سلسله‌مراتبی و partitional تقسیم می‌شود. سلسله‌مراتبی را "nested sequence of partitions" می‌نامند، در حالیکه partitional یک single partition است. لازم به ذکر است که اصطلاحات خوشه‌بندی و خوشه‌بندی سلسله‌مراتبی به ترتیب برای Unsupervised partitional و طبقه‌بندی سلسله‌مراتبی به کار رفته است. در این مقاله، روش Supervised استفاده شده است.

۳- روش پیشنهادی برای طبقه‌بندی متون فارسی

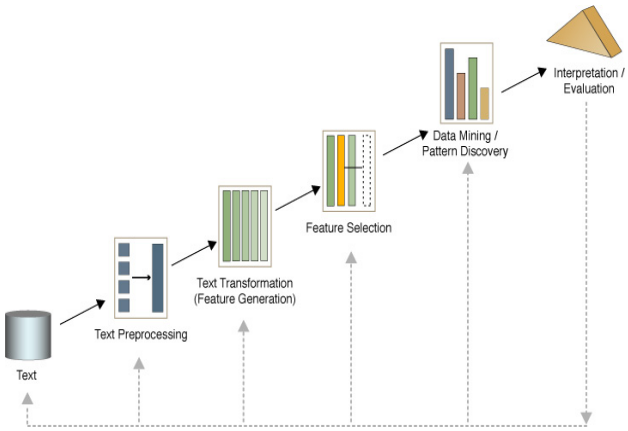
روش پیشنهادی برای طبقه‌بندی متون فارسی بر اساس روش یادگیری ماشین استوار است. در روش یادگیری ماشین، دو فاز آموزش و تست وجود دارد. در فاز آموزش دسته‌های از قبل مشخص شده‌ای را برای یادگیری ماشین استفاده می‌کنند و معنای هر طبقه برای سیستم یادگیری ماشین مشخص می‌شود. در فاز تست، مستندات شناخته نشده‌ای به سیستم داده می‌شود. سیستم به طور اتوماتیک آن متن را به طبقه‌ای که بیشتر شباهت دارد نسبت می‌دهد.

قانون اصلی یادگیری ماشین این است که یک فضای وسیع از فرضیات ممکن و دانش پیشین نگه‌داشته شده بوسیله یادگیر را جستجو نموده و بهترین طبقه ممکن را برای متن تست بدست آورد. وظیفه یادگیر جستجو در این فضا برای جایگذاری فرضیاتی است که بیشترین سازگاری را با نمونه‌های آموزشی موجود دارد.

عموماً در طراحی الگوریتم یادگیری ماشین باید انتخابهایی انجام شود. این انتخاب‌ها شامل انتخاب نوع داده آموزشی، تابع هدف، نمایش آن و یک الگوریتم برای یادگیری این تابع از نمونه های آموزشی می‌باشد [۶].

فازهای یادگیری ماشین برای طبقه‌بندی متن مطابق شکل ۱ می‌باشد. همانگونه که در شکل مشاهده می‌شود، طی ۵ فاز، سیستم یادگیری ماشین برای انجام طبقه‌بندی به کار گرفته می‌شود. این فازها عبارتند از:

- پیش‌پردازش (Preprocessing)
- تولید ویژگی (Feature Generation)
- انتخاب ویژگی (Feature Selection)
- اعمال الگوریتم‌های داده‌کاوی / متن کاوی (Data Mining/Pattern Discovery)
- تفسیر و ارزیابی (Interpretation /Evaluation)



شکل ۱- فازهای سیستم یادگیری ماشین برای طبقه‌بندی

۴- آماده‌سازی متون سیستم طبقه‌بند متون فارسی

داده‌های ورودی به برنامه طبقه‌بند متون، عموماً متونی هستند که از سایت‌های خبری فارسی جمع‌آوری شده‌اند. بررسی کلی روی هفت طبقه زیر می‌باشد:

۱- اجتماعی ۲- اقتصادی ۳- پزشکی ۴- سیاسی ۵- فرهنگی ۶- مذهبی ۷- ورزشی

برای هر یک از این طبقه‌ها ۱۰۰ متن در نظر گرفته شده است. از هر کدام از این ۱۰۰ متن، ۸۰ متن برای فاز یادگیری و ۲۰ متن دیگر برای فاز تست استفاده خواهد شد. با توجه به اینکه این متون در ویرایشگرهای فارسی متفاوتی تایپ شده است و نیز در هنگام بارگذاری روی اینترنت ممکن است دچار تغییراتی شده باشد، لازم است در ابتدا بازبینی کلی روی متون انجام شود. قسمتی از این بازبینی به شیوه دستی انجام شده است که اهم آنها عبارتند از:

- دسته‌بندی متون در گروه‌های مربوطه و تعیین اسم متون بر مبنای "شماره + حرف اختصاری نمایانگر آن گروه".
- برطرف نمودن غلط‌های املایی تا حد امکان.
- چک کردن رعایت فاصله صحیح بین کلمات.
- یکسان نویسی حروف و علائم ریاضی.
- معرفی افعال، پیشوندها، پسوندها، حروف ربط،اضافه و نشانه، افعال ربطی، علائم نقطه‌گذاری و ضمایر به صورت فایل‌های جداگانه به عنوان ورودی سیستم [۷].

۵- پیش‌پردازش سیستم طبقه‌بند متون فارسی

پس از آماده‌سازی اولیه متون، فاز پیش‌پردازش انجام می‌شود. در واقع پیش‌پردازش، اولین گام در جهت تطابق مستندات متنی با نمایش آنها در یک

قالب مناسب می‌باشد. ثابت شده است که تنها ۳۳٪ کلمات در یک متن مفید هستند و می‌توان از آنها برای استخراج اطلاعات استفاده نمود [۸]. اغلب کلمات در راستای رساندن منظور و هدف اصلی استفاده می‌شوند و بعضاً تکراری می‌باشند. در نتیجه هدف از این فاز، یافتن کلمات مفید[۷] و چشم‌پوشی از کلمات بی‌فایده می‌باشد. در این فاز، عملیات کلی زیر انجام می‌شود:

الف: تبدیل هریک از متون به برداری از کلمات

ب: پیدا کردن پایان جملات با استفاده از تشخیص افعال و حروف ربط

ج: اطمینان از یکسان بودن کاراکترهای الفبای فارسی (یکسان‌سازی برخی از کاراکترها مانند "ی" و "ک" از لحاظ کد اسکی)

د: تفکیک جملات از یکدیگر

ه: از بین بردن کلمات بی‌فایده (حروف ربط، اضافه، نشانه، علایم نقطه‌گذاری، ضمایر، افعال ربطی، شبه‌جمله‌ها و ...)

و: انجام عملیات کاهش و ریشه یابی[۸] با توجه به دو گروه کلمات ، افعال و اسامی:

- ✓ بررسی افعال

- بررسی پیشوندهای فعل در دو قسمت با تغییر معنی و بدون تغییر معنی. مثال: فراگرفتن= فرا+گرفتن

- حذف شناسه افعال و بررسی امکان پذیر بودن کاهش با استفاده از جدول بن و مصدر.

- جایگزین نمودن فعل با مصدرش با استفاده از جدول افعال.

مثال: می‌گفتم← گفتم← گفت← گفتن

- ✓ بررسی اسامی

- کاهش علامات جمع (مانند ها، های،...)

- حذف ضمایر متصل

- حذف پیشوند

- حذف پسوندهای اسم

- مشتق

- مرکب و ساده

لازم به ذکر است که در کاهش اسامی از منطق فازی استفاده شده است و بهترین کاهش بدست آمده برای کلمه به عنوان جایگزین کلمه استفاده می‌شود.

۶- تولید ویژگی متون در سیستم طبقه‌بند متون فارسی

پس از انجام فاز پیش‌پردازش در سیستم طبقه‌بند متون فارسی، متون از رشته‌ای از کلمات به بردارهایی از کلمات تبدیل می‌شوند. در این بردارها کلمات بی‌فایده وجود ندارند. البته این بردارها، شامل کلمات یکتایی نیستند [۹]. برای برطرف کردن این مشکل و آماده‌سازی یک ورودی مناسب برای الگوریتم یادگیری طبقه‌بندی متون، از متد تولید ویژگی استفاده شده است. در این متد، کلمات تکراری دسته‌بندی شده و فرکانس آنها محاسبه می‌شود. اینکار بوسیله پیاده‌سازی یک hash table انجام می‌گیرد. Hash table یک نگاشت از کلیدها (key) به ارزش‌ها (value) را فراهم می‌آورد. در اینجا کلمات به عنوان کلید و فرکانس آنها ارزش محسوب می‌شوند.

لازم به ذکر است که همه کلماتی که از فاز پیش‌پردازش به دست آمده‌اند، برای طبقه‌بندی متن لازم نیستند. در این قسمت باید ویژگی‌های شاخص هر متن را استخراج نموده تا بتوان به کمک آنها متون جدید را به بهترین کلاس تطبیق داد.

از فرمول‌های تولید ویژگی IDF، LTC، Entropy برای تولید ویژگی‌های هر متن استفاده می‌شود. البته فرکانس خام هم برای مقایسه و ارزیابی به‌کارگیری این فرمولهای تولید ویژگی نسبت به کار با فرکانس تکرار کلمات در نظر گرفته می‌شود. اگر a_{ik} به عنوان وزن کلمه iام در متن kام باشد، راههای متعددی برای تصمیم‌گیری در مورد آن وجود دارد. اما بیشتر روش‌ها برمبنای یکی از دو روش زیر استوار است:

بیشترین تعداد باری که کلمه در متن اتفاق می‌افتد.

بیشترین تعداد باری که کلمه در همه متون موجود در مجموعه اتفاق می‌افتد.

Tf*IDF Weighting: $$a_{ik} = f_{ik} * \log\left(\frac{N}{n_i}\right) \quad (۱)$$

Ltc- Weighting:

$$a_{ik} = \frac{\log(f_{ik}+1.0) * \log\left(N/n_i\right)}{\sqrt{\sum_{j=1}^{M}\left[\log(f_{ik}+1.0) * \log\left(N/n_i\right)\right]^2}} \quad (۲)$$

Entropy-Weighting:

$$a_{ik} = \log(f_{ik}+1.0) * \left(1 + \frac{1}{\log(N)} \sum_{j=1}^{N}\left[\frac{f_{ij}}{n_i} \log\left(\frac{f_{ij}}{n_i}\right)\right]\right) \quad (۳)$$

به طوریکه $\frac{1}{\log(N)} \sum_{j=1}^{N} \frac{f_{ij}}{n_i} \log\left(\frac{f_{ij}}{n_i}\right)$، Average uncertainty یا entropy کلمه i می‌باشد.

در این فرمول‌های وزن‌دهی، تعاریف زیر وجود دارند:

f_{ik}: فرکانس کلمه i در متن k

N: تعداد متون در مجموعه

M: تعداد کلمات مجموعه پس از انجام عملیات کاهش و حذف کلمات اضافی

n_i: مجموع تعداد دفعاتی که کلمه i در هر مجموعه اتفاق افتاده است

۷- انتخاب ویژگی متو ن در سیستم طبقه‌بند متون فارسی

در قسمت قبل، ویژگی‌های هر کدام از کلمات موجود در متون پیدا شد. با توجه به اینکه هنوز تعداد این ویژگی‌ها بسیار زیاد است باید از بین آنها، تعدادی انتخاب شوند به طوریکه این ویژگی‌ها، ویژگی‌های مؤثری در طبقه‌بندی باشند.

انتخاب ویژگی‌ها باعث کاهش سربار فضا و زمان برای پیاده‌سازی الگوریتم‌های طبقه‌بندی خواهد شد. برای انتخاب ویژگی در سیستم طبقه‌بند متون فارسی از یک حد آستانه[۹] استفاده می‌شود که این حد آستانه قابل تنظیم می‌باشد. البته با توجه به اینکه ثابت شده است تنها ۳۳٪ کلمات موجود در متن برای طبقه‌بندی مناسب هستند، پیشنهاد می‌شود که این حد آستانه بین ۰ تا ۳۰ انتخاب شود. انتظار می‌رود در این محدوده بهترین جواب‌ها بدست آیند.

۸- پیاده‌سازی الگوریتم KNN

پس از انجام مراحل پیش‌پردازش، تولید ویژگی و انتخاب ویژگی در سیستم طبقه‌بند متون فارسی که در قسمت‌های قبلی تشریح شد، نوبت به پیاده‌سازی الگوریتم‌های طبقه‌بندی می‌رسد.

KNN به خاطر سادگی و مؤثر بودن آن در طبقه‌بندی متون بسیار به‌کار برده می‌شود. مبنای کار این الگوریتم، مقایسه متن تست داده شده با متون آموزشی داده شده و بدست آوردن میزان شباهت بین آنها می‌باشد [۱۰]. متون آموزشی با n ویژگی موجود می‌باشد. هر متن به عنوان یک نقطه در یک فضای n بعدی نمایش داده می‌شود. هنگامی که متن ناشناخته‌ای را الگوریتم دریافت می‌کند، فضای الگو را برای یافتن متون آموزشی که شبیه آن متن ناشناخته باشند، جستجو می‌کند. در اینجا از فاصله اقلیدسی به عنوان معیار شباهت استفاده شده است [۱۱].

فاصله اقلیدسی بین دو متن یا دو نقطه $X=(x_1,x_2,x_3,\ldots,x_n)$ و $Y=(y_1,y_2,y_3,\ldots,y_n)$ با استفاده از فرمول (۴) محاسبه می‌شود:

$$D=\sum_{i=1}^{n}\sqrt{(x_i-y_i)^2} \qquad (۴)$$

برای ارزیابی شباهت با کمک گرفتن از فاصله اقلیدسی از سه روش زیر استفاده شده است:

۱- Smallest Distance:

در این روش فاصله متن تست با نزدیک‌ترین متن در هر کلاس، بدست آمده و از بین فاصله‌های بدست‌آمده مینیمم بین آنها انتخاب می‌شود. متن تست به این کلاس تعلق خواهد داشت.

۲- Average Distance:

در این روش برای هر دسته، میانگین مقادیر بردارهای ویژگی بدست آمده و بردار میانگین ویژگی کلاس محاسبه می‌شود. متن تست به کلاس با کمترین فاصله با بردار میانگین کلاس تعلق دارد.

۳- K_ MinMax:

در این روش فاصله متن را با تمام متون موجود محاسبه کرده و مجموعه‌ای با k متن دارای کمترین فاصله با متن تست موردنظر را انتخاب نموده و سپس کلاسی که دارای بیشترین متن در این مجموعه است، به عنوان کلاس آن متن تست انتخاب می‌شود.

۹- ارزیابی الگوریتم KNN

پس از پیاده‌سازی الگوریتم KNN باید با استفاده از متون تست، کارایی این الگوریتم مورد ارزیابی قرار گیرد. همانگونه که در قسمت سه گفته شد این متون از روی منابع اینترنتی مختلفی جمع‌آوری شده است و سعی شده است که از نظر پراکندگی به‌طور مناسبی انتخاب شوند. داده‌های ارزیابی شامل ۱۴۰ متن برای تست می‌باشد که برای هر دسته ۲۰ متن تست درنظر گرفته شده است. این متون به عنوان ورودی تست به برنامه داده می‌شود. ارزیابی الگوریتم طبقه‌بند براساس پارامترهای زیر صورت می‌گیرد [12][13]:

(۵) Recall= a / (a + c)

Precision=a / (a + b)

Fall out= b / (b + d)

Error rate= (b + c)/ (a + b + c + d)

بطوریکه

a: تعداد نمونه‌های عضو کلاس و درست تشخیص داده شده

b: تعداد نمونه‌های عضو کلاس و اشتباه تشخیص داده شده

c: تعداد نمونه‌های غیرعضو کلاس و درست تشخیص داده شده

d: تعداد نمونه‌های غیر عضو کلاس و اشتباه تشخیص داده شده

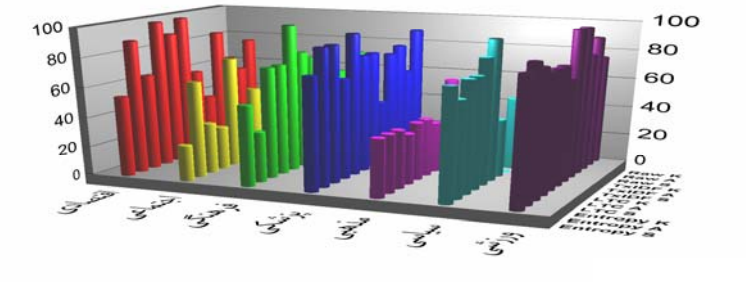
n=a +b +c +d تعداد کل متن‌های تست

با اجرای برنامه روی داده‌های تست، دقت برای الگوریتم‌های فوق بدست آمده است. برای هر دسته، بهترین الگوریتم به‌صورت پررنگ مشخص شده‌است. نتایج در جدول ۱ نشان داده شده است. دقت الگوریتم و نمودار کیفیت به‌ترتیب در شکل‌های ۲ و ۳ نمایش داده شده است. همانطور که در جدول ۱ ملاحظه می شود دقت الگوریتم KNN برای هر هفت دسته نمایش داده شده است. برای دسته های اقتصادی، اجتماعی، فرهنگی، پزشکی و سیاسی بر اساس فرمول تولید ویژگی LTC بهترین دقت را بدست می‌آورد. دقت الکوریتم در بهترین حالت برای دسته ورزشی ۱۰۰٪ است که فرمول تولید ویژگی TxIDF آنرا تولید کرده است. دسته مذهبی در مقایسه با سایر دسته‌ها دقت خوبی را نشان نمی‌دهد و دقت الگوریتم با کاربرد TxIDF برابر با ۶۰٪ می‌باشد.

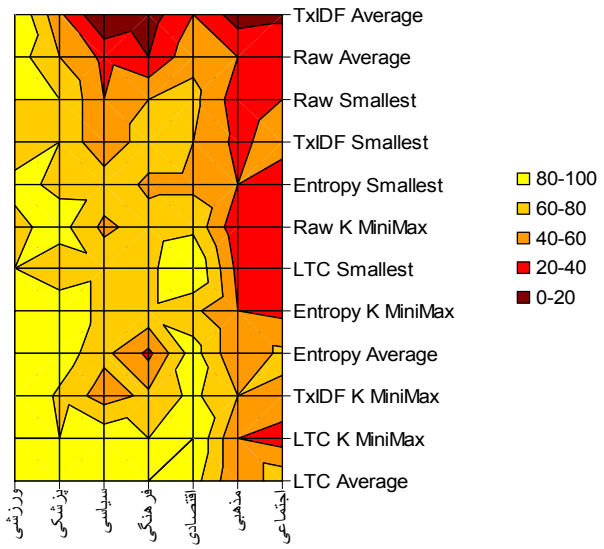
در شکل ۲، نمودار میله ای دقت الگوریتم KNN برای هر دسته آورده شده است. همانطور که در شکل قابل مشاهده است، همانطور که مشاهده می شود دقت الگوریتم برای دسته ورزشی برای کلیه روشهای تولید ویژگی تقریبا تغییرات چندانی ندارد و در دسته مذهبی تغییرات چندانی محسوس نمی‌شود.

جدول ۱- دقت الگوریتم KNN برحسب فرمول‌های تولید ویژگی

	Entropy			LTC			TxIDF			Raw			
دسته	Smallest	Average	K MiniMax	Smallest	Average	K MiniMax	Smallest	Average	K MiniMax	Smallest	Average	K MiniMax	میانگین
اقتصادی	۵۵	۹۰	۶۵	**۱۰۰**	۹۰	**۱۰۰**	۶۰	۴۰	۸۵	۶۵	۵۵	۷۵	۷۳.۳۳
اجتماعی	۲۵	۶۵	۳۵	۳۰	**۷۵**	۳۰	۵۰	۱۵	۵۵	۴۰	۴۰	۴۰	۴۱.۶۷
فرهنگی	۵۵	۳۵	۷۵	۷۵	**۱۰۰**	۸۰	۷۵	۱۰	۶۵	۶۰	۲۰	۷۰	۶۰.۰۰
پزشکی	۷۶	۹۲	۹۲	۶۸	**۹۶**	۸۰	۸۰	۴۴	۷۶	۸۰	۶۰	۸۸	۷۷.۶۷
مذهبی	۴۰	۴۰	۴۰	۳۵	۴۰	۴۰	۳۵	۱۰	**۶۰**	۳۰	۴۰	۲۵	۳۶.۲۵
سیاسی	۷۵	۶۵	۷۵	۷۵	۸۵	**۹۵**	۴۰	۰	۵۰	۴۰	۳۵	۵۵	۵۷.۵۰
ورزشی	۸۵	۹۰	۸۵	۸۰	۸۰	۸۰	۷۰	**۱۰۰**	**۱۰۰**	۸۰	۹۰	۷۵	۸۴.۵۸
جمع	۴۱۱	۴۷۷	۴۷۶	۴۶۳	۵۶۶	۵۰۵	۴۱۰	۲۱۹	۴۹۱	۳۹۵	۳۴۰	۴۲۸	۴۳۱



شکل ۲- نمودار دقت دسته برحسب الگوریتم KNN



شکل ۳- نمودار کیفیت دقت الگوریتم و دسته

همانگونه که در شکل ۳ مشاهده می‌شود، روش LTC Average برای دسته‌های ورزشی، پزشکی سیاسی، فرهنگی و اقتصادی بهترین جواب را برمی‌گرداند. این ناحیه روی شکل با رنگ زرد (۱۰۰-۸۰) مشخص شده است. بدترین جواب هم توسط رنگ قرمز (۲۰-۰) برای الگوریتم TxIDF برای کلاس‌های سیاسی، فرهنگی، مذهبی و اجتماعی می‌باشد.

با توجه به شکل ۲ می‌توان نتیجه‌گیری کرد که روش LTC Average بهترین جواب‌ها را تولید کرده است. جدول کیفیت دقت دسته بر حسب الگوریتم برای روش LTC Average به صورت جدول ۲ محاسبه شده است.

جدول ۲- کیفیت دسته بر حسب الگوریتم LTC Average

کلاس	Precison	Recall	FallOut	ErrorRate
اجتماعی	۷۵	۷۱.۴۳	۴.۰۳	۷.۵۹
اقتصادی	۹۰	۹۴.۷۴	۱.۵۹	۲.۰۷
پزشکی	۹۶	۹۲.۳۱	۰.۸۴	۲.۰۷
سیاسی	۸۵	۸۵	۲.۴	۴.۱۴
فرهنگی	۱۰۰	۵۸.۸۲	۰	۹.۶۶
مذهبی	۴۰	۸۸.۸۹	۸.۸۲	۸.۹۷
ورزشی	۸۰	۱۰۰	۳.۱	۲.۷۶

۱۰- نتیجه‌گیری

در گذشته بیشتر کاری که برای طبقه‌بندی متون انجام گرفته است بر روی متون زبانهای انگلیسی و چینی بوده است. این مقاله روشی برای طبقه‌بندی متون فارسی ارائه داده است. روش کلی برمبنای روش یادگیری ماشین استوار است که دو فاز یادگیری و تست را در بر گرفت. الگوریتم طبقه‌بند، اطلاعات در مورد طبقه‌ها را پس از انجام بررسی‌های لازم در فاز یادگیری بدست آورد. با توجه به جدید بودن تجزیه و تحلیل متون فارسی، نیاز به کار بر روی الگوریتم‌هایی لازم بود که زبان فارسی را پردازش نمایند تا در نهایت بتوان بردار ویژگی متن را برای الگوریتم‌های طبقه‌بندی معمول فراهم آورد. آنچه در این مقاله مورد تاکید قرار گرفته‌شد، الگوریتم‌هایی بودند که برای زبان فارسی ساخته شده‌اند بطوریکه با حفظ معنی مفید کلمات در عملیات کاهش آنها را به نحو مناسب برای مراحل بعدی طبقه‌بند آماده کنند. پس از ساخته شدن بردار ویژگی، الگوریتمی که بتواند با این داده‌ها، که از متون فارسی بدست آمده‌اند طبقه‌بندی مناسبی را انجام دهد، مطرح شد. در این مقاله از این نمونه الگوریتم‌ها می‌توان به K MinMax اشاره کرد، که با توجه به سربار تقریبا مساوی با دیگر الگوریتم‌ها از بازده بسیار خوبی برخوردار بوده است.

طبقه‌بند نه تنها برای پیدا کردن موضوع متن، که در فیلتر کردن متون با توجه به محتوای نسبی آنها نیز کاربرد دارد. از مواردی که به عنوان ادامه کار می‌توان پیشنهاد داد بهبود روش پردازش زبان طبیعی جهت تولید بردار کلمات

دقیق تر، بررسی الگوریتم‌های SVM و شبکه عصبی بر روی زبان فارسی و همچنین بررسی روش‌های تقویت بردار ویژگی مناسب برای زبان فارسی اشاره کرد.

مراجع

[1] T. Joachimes, *Learning to classify Text using support Vector Machines: Methods, Theory, and Algorithms*, Kuuwer Academic Publishers, Boston-Dordrecht-London, 2002.

[2] J. Han, and M. Kamber, *Data Mining Concepts and Techniques*, Elsevier, 2006.

[3] R. Feldman, and J. Sanger, *The Text Mining Handbook, Advanced Approach in Analyzing Unstructured Data*, Cambridge University Press, 2007.

[4] D. Chiang, H. Keh, H. Huang, and D. Chyr, "The Chinese text categorization system with association rule and category priority," Expert System with Applications vol. 35, no. 1-2, pp. 102-110, 2008.

[5] T. M. Valdivia, M. G. Vega, and A. U. Lopez, "LVQ for text categorization using a multilingual linguistic resource," *Neurocomputing*, vol. 55, no. 3-4, pp. 665-679. 2003.

[6] M. T. Mitchell, *Machine Learning*, McGraw-Hill Companies Inc., USA, 1997.

[۷] ح. انوری و ح. احمدی، **دستور زبان فارسی**. چاپ ششم، چاپخانه بهرام، تهران، ۱۳۷۰.

[8] L. Liu, J. Kang, J. Yu, and Z. Wang, "A Comparative Study on Unsupervised Feature Selection Methods for Text Clustering," *Proc. of the Conference on Natural Language Processing and Knowledge* Engineering, pp. 597- 601, 2005.

[9] Z. Li, Z. Xiong, Y. Zhang, C. Liu, K. Li, "Fast text categorization using concise semantic analysis," *Pattern Recognition Letters*, vol. 32, no. 3, pp. 441-448, 2011.

[10] S. Tan, "An effective refinement strategy for KNN text classifier," *Expert Systems with Application*, vol. 30, no. 2, pp. 290-298, 2006.

[11] R. J. Roiger, and M. W. Geatz, *Data Mining: A Tutorial-based Primer*, Addison-Wesley, 2003.

[12] S. Alvarez, *An exact analytical relation among recall, precision, and classification accuracy in information retrieval*, Technical Report BCCS-02-01, Computer Science Department, Boston College, 2002.

[13] Y. Yang, *An evaluation of Statistical Approaches to Text Categorization*, Kluwer Academic Publishers, Netherlands, 2000.

محمد حسین سرایی مسئول آزمایشگاه پایگاه داده‌ای هوشمند، داده‌کاوی و بیوانفورماتیک در دانشکده برق و کامپیوتر دانشگاه صنعتی اصفهان می‌باشد. حوزه اصلی تحقیقاتی ایشان پایگاه داده‌ای هوشمند، متن‌کاوی، temproral داده‌کاوی، بیوانفورماتیک و تجارت الکترونیک می‌باشد. ایشان دارای مقالات متعددی در کنفرانس‌ها و مجلات معتبر داخلی و خارجی می‌باشند. دکتر سرایی مدرک دکتری خود را از دانشگاه منچستر، انگلستان در علوم کامپیوتر و مدرک کارشناسی‌ارشد خود را از دانشگاه Wyoming آمریکا در رشته مهندسی کامپیوتر و مدرک کارشناسی خود را از دانشگاه شهید بهشتی تهران در رشته علوم کامپیوتر اخذ نموده‌اند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

saraee@cc.iut.ac.ir

آذر شاهقلیان کارشناس گروه کامپیوتر و IT دانشکده برق و کامپیوتر دانشگاه صنعتی اصفهان از سال ۱۳۸۲ می‌باشد. همچنین ایشان عضو آزمایشگاه پایگاه داده‌ای هوشمند، داده‌کاوی و بیوانفورماتیک در دانشکده برق و کامپیوتر دانشگاه صنعتی اصفهان می‌باشد. حوزه اصلی تحقیقات ایشان در زمینه داده‌کاوی و شبکه‌های اجتماعی است. آذر شاهقلیان مدرک کارشناسی‌ارشد خود را از دانشگاه آزاد نجف‌آباد در زمینه مهندسی کامپیوتر نرم‌افزار و مدرک کارشناسی خود را از دانشگاه صنعتی اصفهان در رشته مهندسی کامپیوتر اخذ نموده‌اند.

آدرس پست‌الکترونیکی ایشان عبارت است از:

shahgholian@gmail.com

اطلاعات بررسی مقاله:

تاریخ ارسال: ۸۷/۱۱/۲۶

تاریخ اصلاح: ۹۰/۴/۲۶

تاریخ قبول شدن: ۹۰/۵/۳

نویسنده مرتبط: دکتر محمد حسین سرایی، دانشکده مهندسی برق و کامپیوتر، دانشگاه صنعتی اصفهان، اصفهان، ایران.

[1] Knowledge
[2] Text Categorization
[3] Machine Learning Approach
[4] Text Classification
[5] Natural Language Processing
[6] Intrinsic
[7] Keyword
[8] Stemming
[9] Threshold