بنام خداوند جان و خرد

کز این برتر اندیشه برنگذرد

دانشکدهٔ مهندسی کامپیوتر و فناوری اطلاعات

گروه هوش مصنوعی

دسته‌بندی اخبار فارسی

با استفاده از تکنیک‌های هوش مصنوعی

زهرا رباطی

استاد راهنما:

دکتر مرتضی زاهدی

استاد مشاور:

مهندس مرضیه رحیمی

پایان‌نامه جهت اخذ درجه کارشناسی ارشد

شهریور ۱۳۹۳

دانشگاه صنعتی شاهرود

دانشکده: مهندسی کامپیوتر و فناوری اطلاعات

گروه: هوش مصنوعی

پایان‌نامه کارشناسی ارشد خانم زهرا رباطی

تحت عنوان: دسته‌بندی اخبار فارسی با استفاده از تکنیک‌های هوش مصنوعی

در تاریخ ..... توسط کمیته تخصصی زیر جهت اخذ مدرک کارشناسی‌ارشد مورد ارزیابی و با درجه ..... مورد پذیرش قرار گرفت.

استاد راهنما	امضاء	استاد مشاور	امضاء
جناب آقای دکتر مرتضی زاهدی		سرکار خانم مهندس مرضیه رحیمی	

اساتید داور	امضاء	نماینده تحصیلات تکمیلی	امضاء
نام و نام خانوادگی:		نام و نام خانوادگی:	
نام و نام خانوادگی:			

تقدیم به پدر و مادر عزیز و صبورم

به خانواده دوست داشتنی‌ام

و همه خوبان زندگی‌ام

تشکر و قدردانی

سپاس خدای را که سخنوران، در ستودن او بمانند و شمارندگان، شمردن نعمت های او ندانند و کوشندگان، حق او را گزاردن نتوانند.

از استاد راهنمای خود، جناب آقای دکتر زاهدی که علاوه بر راهنمایی های بی دریغشان در تدوین این پایان نامه، مشاوری دلسوز و دوستی گرانبها در زندگی اینجانب هستند، کمال تشکر را دارم.

از سرکار خانم مهندس رحیمی که استاد مشاورم در این پایان نامه بودند و در مراحل مختلف پایان نامه، راهنمایی هایشان را از اینجانب دریغ ننمودند نیز کمال تشکر را دارم.

همچنین قدردان زحمات سایر اساتیدم، جناب آقای دکتر حسن پور و جناب آقای دکتر پویان که در دوره ی تحصیلی کارشناسی ارشد، علاوه بر حمایت های علمی ایشان، درس زندگی از آنان آموختم، هستم.

در پایان از راهنمایی های استاد عزیزم جناب آقای مهندس نامور فرشاد صمیمانه تشکر و قدردانی می کنم.

تعهد نامه

اینجانب **زهرا رباطی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی کامپیوتر** دانشکده **کامپیوتر و فناوری اطلاعات** دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **دسته‌بندی اخبار فارسی با استفاده از تکنیک-های هوش‌مصنوعی** تحت راهنمایی **دکتر مرتضی زاهدی** متعهد می‌شوم.

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

چکیده

با توجه به رشد روزافزون متون الکترونیکی و از جمله متون خبری، معرفی یک دسته‌بند کارآمد در بسیاری از برنامه‌های کاربردی مرتبط با متون از جمله وبسایت‌های خبری، اهمیت بیشتری یافته است. از مهم‌ترین مسائل دامنه‌ی دسته‌بندی متون، استخراج ویژگی‌های کارا برای دسته‌بندی است. در کارهای گذشته معیارهای ارزش‌گذاری ویژگی متفاوتی ارائه شده و مورد استفاده قرار گرفته‌اند. در این پایان‌نامه نیز یک معیار ارزش‌گذاری ویژگی بنام E-Dominace ارائه شده است که با استفاده از آن، کاهش چشمگیری در تعداد ویژگی‌های انتخابی رخ می‌دهد. ویژگی‌های به کار رفته در این دسته‌بندی، ویژگی‌های همرخداد هستند که تابه‌حال در دسته‌بندی متون فارسی مورد استفاده قرار نگرفته‌اند.

در کارهایی که در زبان انگلیسی با استفاده از ویژگی‌های همرخداد انجام شده‌اند، روش وزن‌دهی باینری به کار گرفته شده‌اند. در این پایان‌نامه با توسعه‌ی روش وزن‌دهی TFIDF ، یک روش وزن-دهی برای ویژگی‌های همرخداد به نام Co-occur TFIDF معرفی و به کار گرفته شده است که دقت بهتری را در مقایسه با وزن‌دهی باینری نشان می‌دهد. ویژگی‌های توسعه‌گر نیز برای کم کردن اثر همپوشانی کلاس‌ها و توسعه‌ی متون، در این تحقیق به کار گرفته شده‌اند. نتایج آزمایشات، نشان-دهنده‌ی بهبود قابل توجهی در کارایی و دقت الگوریتم دسته‌بندی با بکارگیری معیار E-Dominance و استفاده از ویژگی‌های همرخداد با روش وزن‌دهیCo-occur TFIDF است.

واژه‌های کلیدی: دسته‌بندی متون فارسی، انتخاب ویژگی، استخراج ویژگی، معیار E-Dominance، ویژگی‌های همرخداد، ویژگی‌های توسعه‌گر، وزن‌دهی Co-occur TFIDF.

لیست مقالات مستخرج از پایان‌نامه

1. Zahra Robati, Morteza Zahedi, Najme Fayazi Far, *"Feature Selection and Reduction for Text Classification"*, International Journal of Computer Applications, IJCA (Accepted on September 2014).

۲. زهرا رباطی، مرتضی زاهدی، نجمه فیاضی‌فر، "استخراج ویژگی‌های همرخداد با وزن‌دهی TFIDF همرخدادی در دسته‌بندی متون فارسی"، سومین همایش ملی زبان‌شناسی رایانشی، دانشگاه صنعتی شریف (پذیرفته شده مهر ۱۳۹۳).

فهرست مطالب

عنوان **شماره صفحه**

فهرست شکل‌ها

عنوان شکل.....شماره صفحه

فهرست جداول

عنوان جدول.....شماره صفحه

فصل اول

مقدمه

۱-۱- پیش زمینه‌ای از دسته‌بندی متون

امروزه با گسترش روز افزون متون و نوشتار الکترونیکی، بخصوص رشد گسترده در وبسایت‌های خبری، وجود ابزار کارآمد دسته‌بندی خودکار متون و به طور ویژه متون خبری، اجتناب ناپذیر گردیده است. در حال حاضر دسته‌بندی متون در بسیاری از زمینه‌های تحقیقاتی و کاربردی، از شاخص‌گذاری متون بر مبنای یک لغت‌نامه کنترل شده تا فیلتر کردن متون، تولید خودکار فراداده، ابهام‌زدایی از کلمه، تولید کاتالوگ‌های سلسله مراتبی از منابع وبی و به طور کلی در هر کاربردی که به سازماندهی متون یا توزیع انتخابی و تطبیقی خاصی از متون نیاز دارد، کاربرد دارد[۱]. اما با اشراف بر این موضوع که دسته‌بندی متون امری لازم و ضروری است، باز هم حدود ٪۸۰ تا ٪۹۰ متون، بدون سازماندهی مشخص در ادارات و نهادها نگهداری می شوند[۲].

یکی از کاربردهای رو به رشد این حوزه، دسته‌بندی متون خبری است. در عصر حاضر، کار اصلی نرم-افزارهای خبری، تنها رساندن و نشر خبر نیست. یک وبسایت خبری هوشمند باید پاسخگوی تمامی نیازهای کاربر و بازدیدکننده، مثل دسته‌بندی و سازماندهی درست اخبار، امکان جستجوی سریع خبرها با واژه‌های کلیدی کارآمد، پیشنهادات مناسب برای کاربر بر اساس سوابق بازدید از صفحات مختلف، جذاب-سازی خبر جهت افزایش آمار بازدیدکنندگان و به طور کلی جوابگوی نیازهای روز افزون کاربر باشد، تا بتواند در عرصه‌ی رقابت باقی بماند. روش کلی کار در دسته‌بندی خبر و سایر متون یکسان است و در بیشتر کارهای انجام شده از متون خبری در آزمایشات استفاده شده است اما تاکیدی بر نوع متون نشده است و روش کار به صورت کلی بر روی همه‌ی متون بیان شده است.

دو دلیل عمده ما را بر آن داشته است که تمرکزی بر متون خبری داشته باشیم، اولا دادگان مورد استفاده در این تحقیق، متون خبری بوده و ثانیا هدف ما ارتقاء یک نرم افزار خبری تحت وب ساده به یک

نرم‌افزار خبری هوشمند بوده است. در این پایان‌نامه هر جا از واژه‌ی "متون" نام برده شده، منظور متون خبری است.

در حال حاضر تکنیک‌های یادگیری ماشین، داده‌کاوی و پردازش زبان طبیعی به کمک هم آمده و سعی در دسته‌بندی دقیق متون الکترونیکی دارند [۲, ۳]. روش‌های متداول موجود در دسته‌بندی متون، شامل الگوریتم‌های احتمالی (تئوری بیز و مدل‌های n-gram)، الگوریتم‌های خطی مثل SVM و نیز الگوریتم‌های مبتنی بر استخراج قوانین است [۴]. بردار ویژگی مورد استفاده در این الگوریتم‌ها، معمولا از روش مجموعه‌ی کیسه‌ی لغات[۱] یا مدل فضای بردار[۲] تولید می‌شود؛ یعنی متن اصلی با یک بردار از کلمات انتخابی و وزن این کلمات نمایش داده می‌شود [۵, ۶]. دلیل اصلی در گوناگونی روش‌های دسته‌بندی متون، شیوه‌ی ایجاد بردار ویژگی و نیز انتخاب ویژگی‌های مناسب، به دلیل اهمیت فوق‌العاده‌ی آن در نتایج دسته‌بندی می‌باشد. علاوه بر کیسه‌ی لغات، عبارات معنایی[۳] و N-gram های حروف و کلمات نیز از ویژگی‌های مطرح و پرکاربرد در حوزه‌ی دسته‌بندی متن هستند [۲]. از روش‌های دیگری که برای استخراج ویژگی وجود دارند، استخراج ویژگی‌های همرخداد[۴] است [۷-۱۱] . این ویژگی، تا بحال به عنوان یک ویژگی برای دسته‌بندی متون فارسی استفاده نشده است. ویژگی‌های همرخداد به دلیل ماهیتی که دارند، ویژگی‌های مطمئن‌تری از ویژگی‌های منفرد مثل کیسه‌ی لغات هستند و علاوه بر در نظر گرفتن ویژگی آماری ، از دانش معنایی استخراج شده از داده‌های آموزشی نیز بهره می‌گیرند. این ویژگی‌ها نیز با معضل حجم بالای تعداد ویژگی‌ها روبرو هستند [۷]. در این نوع ویژگی‌ها، انتخاب ویژگی‌های اولیه و منفرد برای تشکیل زوج ویژگی‌ها، یکی از موثرترین عوامل در تعداد ویژگی‌های نهایی و دقت دسته‌بندی است [۷].

[1] Bag of Word
[2] Vector Space Model
[3] Syntactic Phrases
[4] CO-occurance Features

مسئله‌ی وزن‌دهی و معیارهای ارزش‌گذاری به ویژگی‌ها نیز از مهم‌ترین مسائل دامنه‌ی دسته‌بندی متون هستند که هم در انتخاب و استخراج ویژگی‌ها نقش مهمی ایفا کرده و هم در دقت دسته‌بندی موثرند. از روش‌های متداول وزن‌دهی، وزن‌دهی باینری، بسامد عبارت[۱] (TF) ،بسامد متن[۲] (DF)، TFIDF و ... هستند که در کارهای مختلف استفاده شده‌اند. از معیارهای ارزش‌گذاری ویژگی نیز بهره‌ی اطلاعات[۳]، آنتروپی[۴]، مربع چی[۵] (χ^2) و Dominance و ... هستند که در مرحله‌ی انتخاب و استخراج ویژگی‌ها بسیار پرکاربرد هستند [۲, ۱۲]. گرچه این معیارها و متدهای وزن‌دهی همچنان در تحقیقات مورد استفاده قرار می‌گیرند، اما هیچ‌یک از آنها کارایی قابل ملاحظه‌ای در مقابل سایر روش‌ها ندارد و بسیاری از ویژگی‌های انتخابی بر اساس این معیار و شیوه‌های وزن‌دهی، نامناسب و ناکارآمد هستند.

صرف نظر از نحوه‌ی تشکیل بردار ویژگی، در میان کارهای ارائه شده با انواع مختلف دسته‌بندها مثل k- همسایه نزدیکتر[۶] ، بیزین[۷] ، مدل بیشترین آنتروپی[۸] ، SVM [۹] و ... ، به طور میانگین روش‌های مبتنی بر SVM از دقت بیشتری نسبت با سایر روش‌ها برخوردارند [۲, ۳, ۶, ۷, ۱۲] . در واقع این دسته‌بندها سعی می‌کنند ابرصفحه‌ای که بیشترین حاشیه‌ی مجاز بین ابرصفحه و نمونه‌های مثبت و منفی را در برداشته باشند، بیابند [۲, ۱۳]. اگرچه این دسته‌بندها به سبب سادگی ذاتی دارای پشتوانه‌ی نظری بسیار خوبی هستند، اما هزینه‌های محاسباتی، زمانی و حافظه‌ای آن به یک معضل تمام عیار تبدیل گشته است؛ بخصوص در این کاربرد که با داده‌های متنی و حجم عظیم کلمات و واژه‌ها و پردازش آنها سر و کار داریم، این معضل بیشتر دیده می‌شود. بعد از SVM نیز در کارهای انجام شده، دسته بند K- همسایه نزدیکتر دقت بالاتری را نسبت به سایر دسته‌بندها ارائه کرده است .

[1] Term Frequency
[2] Documet Frequency
[3] Information Gain
[4] Entropy
[5] Chi Square
[6] K-Nearest Neighbors
[7] Baysian
[8] Maximum Entropy Models
[9] Support Vector Machine

ویژگی که متون خبری نسبت به سایر متون دارند این است که معمولا از قالب مشخص‌تری در هر دسته برای ارائه متن استفاده می‌کنند و ویژگی‌های مشخصی از هر دسته در قالب هر کلاس وجود دارند؛ مثلا به طور معمول متون خبری ورزشی از کلمات متداولی مثل "رقابت"، "برنده" و "امتیاز" استفاده می‌کنند. در فرآیند آموزش به دنبال یادگیری این قالب در مجموعه متون هر کلاس هستیم. این ویژگی در سایر متون مثلا مقالات علمی کمتر دیده می‌شوند؛ مثلا دو مقاله در شاخه فیزیک که یکی درباره عدسی‌ها و دیگری درباره ذرات بحث می‌کنند، ویژگی‌های مشترک کمتری دارند و در زیرگروههای مربوط به موضوع اصلی، ویژگی‌های مشترک بیشتری خواهند داشت. در این پروژه با ارائه یک معیار اهمیت داده کارآمد، ویژگی‌های همرخداد انتخاب شده و از میان آنها ویژگی‌های پرتکرارتر به عنوان ویژگی‌های توسعه‌گر برای توسعه متون به کار رفته‌اند. در انتخاب ویژگی‌های توسعه‌گر، ویژگی نامبرده متون خبری کمک شایانی به انتخاب ویژگی‌های توسعه‌گر می‌کند و قالب فراگیری شده از متون آموزشی به کل متون آموزشی کلاس و متون آزمایشی مشابه القا می‌شود. سپس با استفاده از طرح توسعه‌ی داده‌ها و نیز به کارگیری روش‌وزن‌دهی TFIDF همرخدادی[1] برای ویژگی‌های همرخداد، علاوه بر کاهش تعداد ویژگی‌ها، دقت نسبتا خوبی در نتایج دسته‌بندی بدست آمده است.

۱-۲- چالش‌ها

گرچه تحقیقات گسترده‌ای در زمینه‌ی دسته‌بندی متون در زبان‌های مختلف صورت گرفته است، اما برخی از مشکلات هنوز لاینحل باقی مانده است و انگیزه برای ادامه‌ی کار در این زمینه وجود دارد. از مهم‌ترین این مشکلات می‌توان به موارد زیر اشاره کرد:

۱-۲-۱- همپوشانی متون در کلاس‌های متفاوت: متون غیر منفک داده‌هایی هستند که نمی‌توان به طور قطع نشان داد به کدامین کلاس تعلق دارند. برخی از متون به طور ذاتی به بیش از یک

[1] Co-Occur TFIDF

دسته تعلق دارند؛ یعنی ممکن است یک متن خبری با میزان تعلق متفاوت هم سیاسی باشد و هم اجتماعی . این امر باعث می‌شود که کلاس‌ها با هم همپوشانی داشته باشند که خود منجر به افزایش خطا در دسته‌بندی می‌گردد [۴].

۱-۲-۲- ریشه‌یابی: اکثر روش‌های تشکیل بردار ویژگی به منظور کاهش خطا، از ریشه‌یابی کلمات استفاده می‌کنند. برای ریشه‌یابی، از کلمات مورد پردازش، پیشوندها و پسوندها حذف شده و تبدیل به ریشه می‌شوند. از مشکلات ریشه‌یابی می‌توان به بیش ریشه‌یابی و کم ریشه‌یابی اشاره نمود. در بیش ریشه‌یابی، ریشه‌هایی تولید می‌شوند که هیچ معنایی ندارند و در کم ریشه‌یابی امکان تولید ریشه برای حالت‌ها استثنا وجود ندارد [۱۴].

۱-۲-۳- وزن‌دهی: اینکه هر یک از کلمات موجود در متن چه وزن و ارزشی خواهند داشت، یکی از اصلی‌ترین چالش‌هایی است که با آن روبرو هستیم. ثابت شده است که استفاده از عبارات بیشترین ارزش معنایی و کمتریی ارزش آماری و از طرف دیگر استفاده از کلمات، کمترین ارزش معنایی و بیشترین ارزش آماری را دارد [۴]. برای به کار بردن هر کدام از این دو نوع ویژگی، یک روش وزن‌دهی موثر که میزان تاثیر و کارایی آن ویژگی را بیان کند، بسیار قابل توجه است. طرح‌های وزن‌دهی متفاوتی وجود دارند که در کارهای مختلف مورد استفاده قرار می‌گیرند، اما هنوز فقدان یک روش قدرتمند که منجر به یک دسته‌بندی با دقت بالا شود، احساس می‌شود.

۱-۲-۴- برخورد با چالش‌ها

در مواجه با چالش داده‌های غیر منفک و همپوشانی کلاس‌ها در این پایان‌نامه، سعی شده است با طرح توسعه‌ی داده‌ها با استفاده از ویژگی‌های توسعه‌گر متون را غنی‌تر کرده و اثر همپوشانی کلاس‌ها به حداقل رسانده شود. این کار باعث بهبود نتیجه در دقت دسته‌بندی می‌گردد.

در رویارویی با چالش وزن‌دهی نیز در این پایان‌نامه، عبارات همرخداد به عنوان ویژگی استفاده شده‌اند که ارزش معنایی را در بر دارند و از طرفی با توسعه‌ی روشTFIDF به Co-Occur TFIDF برای ویژگی‌های همرخداد از ارزش آماری ویژگی‌ها نیز استفاده شده است.

۱-۳- دستاوردهای تحقیق

- ارائه‌ی یک مجموعه‌ی ۱۰۸۰ کلمه‌ای از Stop Words برای زبان فارسی
- ارائه‌ی معیار E-Dominance[1] به عنوان بهبودی از معیار Dominance برای اهمیت ویژگی‌ها
- کاربرد ویژگی‌های همرخداد برای اولین بار در دسته‌بندی متون فارسی
- کاربرد ویژگی‌های توسعه‌گر برای توسعه‌ی متون برای اولین بار در دسته‌بندی متون فارسی
- توسعه‌ی طرح وزن‌دهیTFIDF پایه به Co-Occur TFIDF برای ویژگی‌های همرخداد

۱-۴- ساختار پایان‌نامه

این پایان‌نامه در ۵ فصل تدوین شده است. پس از مقدمه، در فصل دوم به معرفی مفاهیم پایه در حوزه‌ی دسته‌بندی متون پرداخته شده است. همچنین در این فصل یک مرور کلی بر روی کارهای انجام شده و روش‌های موجود انجام شده است. فصل سوم به معرفی روش پیشنهادی اختصاص دارد و به صورت گام به گام آن را تشریح می‌کند. در فصل چهارم نیز نتایج حاصل از آزمایشات انجام شده و مقایسه روش‌پیشنهادی با روش‌های دیگر بیان می‌گردد. درفصل پنجم یک نتیجه‌گیری کلی از تحقیق و نیز ارائه‌ی پیشنهادات برای ادامه‌ی کار ارائه می‌گردد.

[1] Extended Dominance

فصل دوم

دسته‌بندی خودکار متون و

مروری بر کارهای پیشین

۲-۱- مقدمه

در این فصل ابتدا به تعریف مفهوم دسته‌بندی متون پرداخته و سپس اهمیت و لزوم دسته‌بندی و نیز کاربرد آن در زمینه‌ی یادگیری ماشین بیان می‌شود. دسته‌بندی متون دارای مشکلات و محدودیت‌هایی است که در این فصل به طور خلاصه در حوزه‌ی عمومی و نیز در حوزه‌ی نگارش زبان فارسی به آنها پرداخته می‌شود. انواع روش‌های دسته‌بندی نیز در این فصل بیان می‌شود و در ادامه مراحل مختلف سیستم دسته‌بندی خودکار متون مورد بررسی قرار می‌گیرد. در پایان نیز مروری بر کارهای گذشته در زبان فارسی و سایر زبان‌ها در حوزه‌ی دسته‌بندی متون انجام می‌شود.

۲-۲- دسته بندی متون

در صورتیکه مجموعه‌ای از متون $D=\{(d_1,y_2) ,\ldots, (d_i,y_i), \ldots, (d_n,y_n)\}$ داشته باشیم به طوریکه n تعداد متون و $d_i =[w_{i,1},\ldots, w_{i,k},\ldots,w_{i,|di|}]$ متن iام این مجموعه باشد، آنگاه $W_{i,k}$ به وزن کلمه‌ی kام در متن i ام و y_i به کلاسی که متن به آن متعلق است (یعنی $y_i \epsilon c$ به طوریکه $C=\{c_1,c_2,\ldots,c_{|C|}\}$ مجموعه کلاس‌های از پیش تعریف شده در سیستم باشد) اشاره می‌کند. هدف در دسته‌بندی متون، استنتاج یک تابع f است به نحوی که $y_i=f(d_i)$ باشد. دسته‌بندی متون تعیین یک مقدار بولی برای هر جفت $<d_j,c_i> \epsilon D*C$، در جایی است که D مجموعه‌ای از متون و C مجموعه کلاس‌های از پیش تعیین شده باشد. مقدار T تعیین می‌کند که متن d_j به کلاس c_i تعلق دارد و مقدار F نیز عدم تعلق را نشان می‌دهد. هدف در اینجا به‌دست آوردن تابع تخمین $\Phi: D * C \rightarrow \{T, F\}$ می‌باشد [۴].

۲-۲-۱-کاربردهای دسته‌بندی متون

کاربردهای دسته‌بندی متون به کار آقای ماروت[1] بر روی دسته‌بندی متون به صورت احتمالی باز می-گردد. از آن پس دسته‌بندی متون برای کاربردهای متنوعی استفاده شده است. لازم به ذکر است که این گروه‌بندی مرز مشخصی نداشته و برخی از این گروه‌ها با یکدیگر هم‌پوشانی دارند که به طور خلاصه عبارتند از [۴]:

- **شاخص‌بندی[2] برای سیستم‌های بازیابی اطلاعات**: در سیستم‌های بازیابی اطلاعات، برای هر متن یک یا چند کلید واژه (عبارت کلیدی) که محتوای آن متن را توصیف می‌کند، تعیین می‌شود. این کلید واژه‌ها متعلق به مجموعه لغت‌نامه‌ای کنترل‌شده با عنوان گنج‌واژه هستند. اغلب این گنج‌واژه‌ها ساختار سلسله مراتبیِ موضوعی[3] دارند (مانند گنج‌واژه ناسا[4] برای مرتب‌سازی مسائل فضایی و یا گنج‌واژه مش[5] برای مسائل پزشکی).
- **سازمان‌دهی متون**: سازماندهی متون و بایگانی برای مقاصد شخصی، سازمانی و... توسط روش-های دسته‌بندی خودکار متون امکان‌پذیر می‌باشد. برای مثال در ورودی تحریریه‌ی یک روزنامه، تبلیغات و آگهی‌ها برای انتشار باید تحت گروه‌های خدمات عمومی، فروش، خرید، درخواست کار و... دسته‌بندی‌شده باشند. [۱۵].
- **فیلترکردن متون:** فیلترکردن متون بر روی رشته‌ای از متن‌های ورودی که توسط یک تولیدکننده‌ی اطلاعات برای یک مصرف‌کننده‌ی اطلاعات بصورت غیرهمزمان توزیع می‌شود، اعمال می-گردد [۱۶]. به عنوان مثال در تلکس‌های خبری، تولیدکننده‌ی خبر، یک خبرگزاری است و اخبار را برای

[1] marot
[2] Indexing
[3] Thematic Hierarchical Thesaurus
[4] NASA Thesaurus
[5] MESH Thesaurus

یک مصرف‌کننده‌ی خبر که یک روزنامه است، می‌فرستد [۱۷]. در این حالت سیستم فیلترگذاری باید از دریافت خبرهایی که برای گیرنده جذاب نیستند، جلوگیری کند.

علاوه بر موارد ذکر شده، یک سیستم فیلتر می‌تواند متون را به دسته‌های موضوعی مورد نظر مصرف-کننده نیز دسته‌بندی کند. اگر سیستم فیلتر در سمت تولیدکننده باشد، متون را به نحوی که مورد نظر مصرف‌کنندگان است، توزیع می‌کند. بدین منظور سیستم برای هر مصرف‌کننده یک نمایه[1] منحصر به فرد تولید کرده و مرتبا آن را بروز می‌نماید. اگر سیستم فیلتر در سمت مصرف‌کننده باشد، از دریافت مطالب بی‌ربط به مصرف‌کننده جلوگیری می‌کند. در این حالت تنها یک نمایه مورد نیاز است. نمایه توسط کاربر مقداردهی اولیه شده و با استفاده از اطلاعات بازخوردی از کاربر (صریح و یا غیرصریح)، بروز می‌شود [۱۸]. این مسأله را فیلتر تطبیق‌پذیر می‌نامند.

- **رفع ابهام:** عملیات رفع ابهام در بسیاری از کاربردها، همچون پردازش زبان طبیعی و شاخص-بندی مستندات با استفاده از ترجیح نقش کلمه بر خود کلمه در حوزه بازیابی اطلاعات مورد توجه است. در دسته‌بندی، تکرار کلمه در زمینه، به عنوان متن و نقش کلمه به عنوان دسته دیده می‌شود [۱۹, ۲۰].
- **دسته‌بندی سلسله مراتبی صفحات وب:** دسته‌بندی سلسله مراتبی صفحات وب منجر به افزایش کیفیت پرس‌وجو برای موتورهای جستجو می‌گردد. در این کاربرد، روش‌های دسته‌بندی مبتنی بر دسته[2] انتخاب می‌شوند؛ چراکه اجازه اضافه‌کردن دسته‌های جدید و یا حذف دسته‌های بی‌مصرف را امکان‌پذیر می‌نمایند [۲۱, ۲۲].

از دیگر کاربردهای دسته‌بندی متون می‌توان به دسته‌بندی گفتاری که در واقع ترکیبی از دسته‌بندی متون و تشخیص گفتار است، دسته‌بندی متون چندرسانه‌ای[3] از طریق عنوان‌های متنی، تشخیص نویسنده

[1] Profile
[2] Category-based Text Classification
[3] Multimedia Document Categorization

برای متون ادبیاتی نامشخص، تشخیص زبان برای متونی که زبان آنها مشخص نیست، تشخیص خودکار جنس متن و رتبه‌بندی خودکار کیفیت نوشتار اشاره نمود [۱].

۲-۲-۲- اهمیت دسته‌بندی متون

دسته‌بندی متون بصورت دستی، فرآیندی زمانبر و نیازمند دانش افراد متخصص در زمینه‌های مختلف است و دسترسی به این افراد به طور مداوم امکان‌پذیر نخواهد بود. بدلیل کاربرد دانش انسانی در این فرآیند، دسته‌بندی متون عاری از خطا و اشتباه نخواهد بود. همچنین در صورت به‌کار گرفتن چند متخصص انسانی در حین فرآیند دسته‌بندی، سیستم بدلیل تفاوت تصمیم‌های اتخاذ شده افراد مختلف، دچار ناسازگاری می‌گردد. بخصوص در زمینه‌ی متون خبری، سرعت دسته‌بندی و سازمان‌دهی خبر برای نشر آن، امری ضروری است که به صورت دستی سرعت کار بسیار کند خواهد بود. بنابراین به دلیل حجم زیاد متون، دسته‌بندی بصورت دستی تقریبا غیرممکن است و به منظور صرفه‌جویی در زمان، هزینه، نیروی انسانی و از طرفی افزایش دقت و کارایی، در سال‌های اخیر به روش‌های دسته‌بندی خودکار روی آورده شده‌است.

۲-۲-۳- انواع روش‌های دسته‌بندی

در ادامه روش‌های مختلف دسته‌بندی متون مورد مطالعه قرار می‌گیرد [۴]. قابل توجه است که در اینجا فرض این است که دسته‌بندی متون بر مبنای دانش درون‌زا انجام می‌شود. یعنی یک متن فقط با استفاده از اطلاعات درونی اش و بدون استفاده از اطلاعات خارجی دسته‌بندی می‌شود.

۲-۲-۳-۱- دسته بندی تک‌برچسبی در مقابل چندبرچسبی

اگر مساله دسته‌بندی در حالتی باشد که هر متن فقط به یک دسته تعلق داشته باشد و کلاس ها با هم همپوشانی نداشته باشند، اصطلاحا دسته‌بندی تک‌برچسبی گفته می‌شود. اما زمانی‌که یک متن به بیش از

یک کلاس تعلق داشته باشد، یعنی بین کلاس‌ها همپوشانی داشته باشیم، دسته‌بندی از نوع چندبرچسبی خواهد بود.

۲-۲-۳-۲- دسته‌بندی متون مبتنی بر کلاس[1] در مقابل مبتنی بر متن[2]

در روش مبتنی بر متن، DPC، برای یک متن، باید تمام کلاس‌ها را که باید انتخاب شوند یافت. از طرف دیگر در روش مبتنی بر کلاس، CPC، برای هرکلاس، باید متونی که عضو آن کلاس هستند را یافت. این اختلاف بیشتر فلسفی است تا مفهومی، اما در جایی اهمیت پیدا می‌کند که مجموعه‌های C و D از ابتدا موجود نباشند. به طور عمده شرایط DPC از شرایط CPC متداول‌تر است و بیشتر مورد استفاده قرار می‌گیرد.

۲-۲-۳-۳- دسته‌بندی قطعی در مقابل دسته‌بندی رتبه‌ای

فرض کنید یک فرد خبره می‌خواهد مجموعه‌ای از متون را دسته‌بندی کند، سیستمی وجود دارد که متون را بدون هیچ‌گونه دسته‌بندی قطعی در کلاس‌های موجود رتبه‌بندی می‌کند. حالا فرد خبره، انتخاب کلاس برای هر متن را به بالای لیست به جای کل لیست محدود می‌کند و سریعتر و بهتر تصمیم می‌گیرد. از طرفی برای هر کلاس از مجموعه کلاس‌ها، سیستم توانایی رتبه‌دهی ساده متون را بر مبنای تناسب با آن کلاس خاص را دارد، یعنی برای دسته‌بندی تحت آن کلاس، کافی است فرد خبره متون با رتبه بالا را به جای کل متون آزمایش کند. [۴].

۲-۲-۴- مشکلات دسته‌بندی متون

دسته‌بندی خودکار متون، همانند سایر داده‌ها، دارای مشکلاتی است. برخی از این مشکلات مربوط به کلیات و شیوه‌های دسته‌بندی است و برخی دیگر مربوط به داده‌ها. علاوه بر مشکلات عمده‌ی دسته بندی

[1] Category-Pivoted Categorization (CPC)
[2] Document-Pivoted Categorization (DPC)

متون، دسته‌بندی متون به زبان فارسی نیز مشکلات خاص خود را دارد که در بخش‌های بعدی به آن اشاره خواهیم کرد.

۲-۲-۴-۱- مشکلات و محدودیت‌های دسته‌بندی متون

شیوه نمایش متون از اولین و مهم‌ترین محدودیت‌های دسته‌بندی متون است. هر سند متنی باید از نسخه‌ی متنی به یک بردار ویژگی، تبدیل گردد. این فرآیند منجر به کاهش پیچیدگی متون و سهولت در مدیریت آنها می‌گردد. رایج‌ترین روش در شیوه نمایش متون"مدل فضای برداری[1]" است، که بصورت گسترده‌ای در دسته‌بندی متون مورد استفاده قرار می‌گیرد [۵, ۶]. مدل فضای برداری، هر متن را به شکل یک بردار ویژگی با استفاده از عبارات درون متن که شامل وزن‌های آن عبارات است، نشان می‌دهد. این روش را از آنجائی‌که هر کلمه را به عنوان یک ویژگی در نظر می‌گیرد، روش کیسه کلمات[2] می‌نامند. این روش نمایش متن دارای عیوبی مانند نادیده گرفتن وابستگی میان عبارات و عدم توجه به ساختار عبارات درون متون است. در شکل ۲-۱ یک نمونه از این نمایش متن بصورت مدل فضای برداری مشاهده می‌شود. در این مدل هر سطر نماینده یک متن و هر ستون نماینده یک عبارت کلیدی است. مقدار هر سلول از این ماتریس میزان اهمیت و وزن آن عبارت در متن مربوطه را نشان می‌دهد.

	F_1	F_2	F_m
d_1	$W_{1,1}$	$W_{1,2}$	$W_{1,m}$
d_2	$W_{2,1}$	$W_{2,2}$	$W_{2,m}$
.			
.			
d_n	$W_{n,1}$	$W_{n,1}$	$W_{n,m}$

شکل (۲-۱) مدل فضای برداری

[1] Vector Space Model (VSM)
[2] Bag Of Words

به طور کلی طراحی سیستم دسته‌بندی متون ، با مشکلات زیر مواجه است:

- **بردار ویژگی با ابعاد زیاد** : بالا بودن بعد بردارهای نمایش ویژگی، یکی از مهم‌ترین مشکلات یک داده‌ی متنی است. این مشکل علاوه بر زمان‌بر بودن می‌تواند منجر به "بیش یادگیری[1]" در سیستم گردد [۲۳].

- **نیاز به دانش پیش‌زمینه و قالب معنایی متون**: آنچه مسلم است مفاهیم و معانی که بیان کننده مفهوم و موضوع کلی یک متن هستند، اهمیت زیادی در دسته‌بندی ایفا می‌کنند، بنابراین نمی‌توان صرفا به معیارهای آماری برای دسته‌بندی اکتفا نمود [۲۳].

- **اهمیت ارتباط میان کلمات**: برخی کلمات در کاربردهای متفاوت و در همراهی با سایرکلمات، معانی مختلفی خواهند داشت که این مسئله فرآیند دسته‌بندی را دچار مشکل خواهد کرد. همچنین ممکن است وابستگی کلمات به کلمات قبل و بعد از خود و همراهی این کلمات با سایر کلمات، معانی آنها را تغییر داده و کارایی دسته‌بندی را مختل کند.

به منظور حل مشکل بعد بالای مدل فضای برداری در دسته‌بندی متون، روشهای بسیاری به منظور کاهش بعد فضای مشخصه مورد بررسی قرار گرفته‌اند [۲۴, ۲۵].

۲-۲-۴-۲- مشکلات زبان فارسی با رویکرد دسته‌بندی

هدف ما در این تحقیق، پیاده‌سازی یک سیستم دسته‌بند برای متون و اخبار فارسی است. بدین جهت در این بخش به طور خلاصه به بررسی مشکلات مطرح در دسته‌بندی متون زبان فارسی می‌پردازیم.

[1] Overfitting

زبان فارسی از دسته‌ی زبانهای هندی-اروپایی است. گرچه زبان فارسی همانند زبان انگلیسی دارای ریخت‌شناسی[1] وندافزای است و وجود پس‌وندها و پیش‌وندها در معانی لغات تغییر ایجاد می‌کند؛ اما پیچیدگی‌های ذاتی بیشتری نسبت به زبان انگلیسی دارد.

در اینجا به دلیل حجم مطالب، از توضیح مشکلات زبان فارسی به طور کامل اجتناب شده است. برای مطالعه بیشتر به مرجع [۱] مراجعه شود. اما در کل اهم مشکلات پردازش متون فارسی در چند دسته زیر خلاصه می‌شود:

۱. مشکل تشخیص مرز کلمات (مسأله‌ی شیوه‌های نگارش متفاوت)
۲. مشکل تشخیص مرز گروه‌های اسمی (مسأله کسره‌ی اضافه نامرئی)
۳. از دست‌دادن اطلاعات گویشی
۴. مسأله ابهام
۵. افعال مرکب و اصطلاحات
۶. معناشناسی و مشکلات تحلیل معنایی

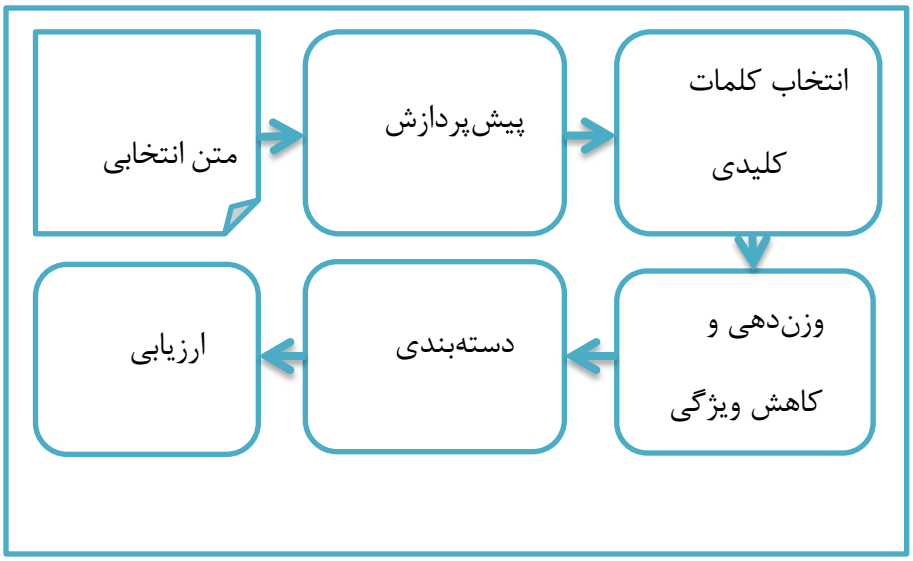
موارد ذکر شده از ساختار زبان فارسی تنها بخشی از پیچیدگی این زبان را نشان می‌دهد. توجه به تمام این حالات به منظور قطعه‌بندی، ریشه‌یابی و نیز دسته‌بندی متون فارسی منجر به افزایش بار محاسباتی و نیز زمان مصرفی می‌گردد.

۲-۳-مراحل مختلف دسته‌بندی متون به صورت کلی

در شکل ۲-۲ مراحل مختلف و اصلی فرآیند دسته‌بندی متون آورده شده است. در حالت‌های خاص و کاربردهای متفاوت ممکن است مرحله‌ی جدیدی به آن اضافه گردد، اما کلیت کار دسته‌بندی به صورت

[1] Morphology

زیر است. در بخش‌های بعدی این مراحل را به تفصیل بیان می‌کنیم. پس از شرح هر مرحله و یا به صورت همگام، مروری بر کارهای انجام شده در آن حوزه خواهیم داشت.



شکل (۲-۲) نمودار کلی فرآیند دسته‌بندی متون

۲-۳-۱-پیش پردازش

همانطور که می‌دانیم هر نوع داده‌ای که قرار است در سیستم دسته‌بندی قرار گرفته و برچسب مناسب کلاس خود را دریافت کند، نیاز به یک پیش‌پردازش اولیه دارد. این پیش‌پردازش اولیه باعث می‌شود که داده‌ی خام اولیه به یک داده‌ی مناسب برای فرآیند دسته‌بندی آماده گردد. به طور مثال در داده‌های مورد مطالعه ما وجود علائم نگارشی مانند نقطه و ویرگول و سایر علائم در متن مورد آزمایش، تاثیری در انتخاب برچسب کلاس آن متن ندارد. پس در گام نخست می‌توان این علائم را حذف کرد. اما به طور کلی پیش‌پردازش‌های عمده و معمولی وجود دارند که در کارهای مختلف انجام شده و متن را برای مرحله استخراج ویژگی‌ها و کلمات کلیدی آماده می‌سازد. قابل ذکر است که ممکن است در برخی مقالات بعضی

از مراحل ذکر شده جزء مرحله پیش‌پردازش نبوده و یکی از مراحل اصلی فرآیند دسته‌بندی مطرح گردند. صرف نظر از این نکته مراحل مختلف پیش‌پردازش به صورت زیر است:

۲-۳-۱-۱-پردازش لغوی و تعیین مرز کلمات

تعیین مرز کلمات و واژه‌ها در یک متن، از ابتدایی‌ترین مراحل در فرآیند دسته‌بندی اسناد است. این عملیات با توجه به پیچیدگی‌هایی که زبان فارسی دارد، یکی از مراحل دشوار در دسته‌بندی متون فارسی محسوب می‌گردد [۲۶-۲۸].

در بیشتر مواقع، تعیین مرز جمله‌ها از طریق بررسی علائم جداکننده انجام می‌شود. در این پایان‌نامه، شناسایی واژه‌ها و جمله‌ها با بررسی علائم نامبرده انجام گرفته است. علائمی که برای تعیین مرز جمله از آن‌ها استفاده می‌شود، عبارتند از: ".","؟","? ",":","!". باید توجه داشت که جستجو برای یافتن این علائم به تنهایی کافی نیست و در بسیاری از موارد مشکلاتی را به همراه دارد. به عنوان مثال، در زبان فارسی بعضی از واژه‌های اختصاری به صورت چند حرف که با نقطه از هم جدا شده‌اند، ظاهر می-شوند(مانند«ه.ق.» که مخفف «هجری قمری» است.) در صورتی که تعداد این حالات مشکل‌ساز در پیکره زیاد باشد، باید از روش‌های پیچیده‌تری استفاده شود، که قادر به شناسایی این موارد باشند. برای حل این مشکل از شمارش حروف وکلمه استفاده شده است؛ از آنجایی که کوچکترین جمله شامل دوکلمه است وکوچکترین کلمه از سه حرف تشکیل می‌شود (به غیر حروف اضافه مانند از، به) پس می‌توان واژه‌های اختصاری را مشخص کرد و به عنوان جمله در نظر نگرفت؛ از این رو فقط علائم در متن بررسی نشده‌اند بلکه اگر کلمات بین این علائم بیشتراز یک حد آستانه بود(که در اینجا دو در نظر گرفته شده) به عنوان جمله در آرایه قرار می‌گیرند.

شناسایی واژه‌ها نیز از طریق بررسی علائم قابل انجام است. این علائم عبارتند از: فضای خالی، علامت خط جدید، ", "، "،"، ">"، "<"، "]"، "["، "_"، "/"، "—"، "؛".

در [۲۹]، نویسنده علاوه بر بررسی روش‌های مختلف تعیین مرز واژه‌ها وجملات، مشکلات هر یک از آنها را مورد بررسی قرارداده است. [۳۰] نیز روشی برای شناسایی انتهای واژه‌ها معرفی کرده است که امکان جداسازی واژه‌ها را فراهم می‌آورد. در [۳۱] نیز یک روش آماری برای تعیین کسره اضافه معرفی شده است که با استفاده از آن می‌توان مرز گروه‌های اسمی را به طور دقیق‌تری شناسایی نمود.

۲-۳-۱-۲- حذف لغات غیرضروری

در حوزه‌ی پردازش متن، کلماتی را که هیچ بار معنایی خاصی در بر ندارند به "لغات ایست[1]" می-شناسند. لغات ایست شامل حروف اضافه، بسیاری از قیود و صفات، حروف ربط و برخی از افعال هستند. این واژه‌ها در همه‌ی متون و با تکرار زیاد حضور داشته، اما دربردارنده اطلاعات معنایی نبوده و به همین دلیل دقت دسته‌بندی را کاهش می‌دهند. به طور مثال در زبان انگلیسی، لغات ایست شامل , that, the, an a و... هستند. در زبان فارسی نیز کلماتی مثل "از" ، "برای"، " هر " و ...جزو کلمات ایست محسوب می-گردند. اگر این کلمات حذف نشوند، بخش زیادی از شاخص‌ها را شامل می‌شوند. این کلمات که در متن پرتکرار هستند، در فرآیند وزن‌دهی که موثر از تعداد تکرار کلمات است، وزن بالایی به خود گرفته و اثر وزن کلمات دیگر را که در دسته‌بندی مهم بوده ولی پرتکرار نیستند، کم می‌کنند. حذف این دسته از کلمات نه تنها منجر به کاهش بعد فضای مشخصه می‌گردد، بلکه اثر نویزی این کلمات در دسته‌بندی را رفع می‌نماید. در حین استخراج کلمات اضافی باید به دامنه کلمات در اسناد و نیز پیکره مورد استفاده، توجه گردد [۳۲]. چرا که ممکن است برخی کلمات که در یک دامنه، کلمات اضافی هستند در برخی اسناد با دامنه‌ی دیگر دارای بار معنایی بوده و در دسته‌بندی اثربخش باشند. مرجع [۳۳] لغات غیرضروری زبان فارسی را در دو جدول فعل و غیرفعل در دسترس قرار داده است. البته این مجموعه کامل نبوده و قابل گسترش است. جدول۲-۱ نمونه‌ای از کلمات ایست را نشان می‌دهد.

[1] Stop Words

جدول (۲-۱) نمونه‌ای از واژه‌های غیر ضروری(شامل اسم، صفت وحروف و وندها و افعال)

در	نیز	برای	یا	را	بلکه
به	تا	ها	دو	های	شاید
از	ما	آن	آنها	و	وقتی
که	باید	وی	اما	ممکن	اندک
این	اند	یک	دیگر	هر	کمی
با	هم	خود	اگر	ای	بسیار
شما	همچنین	بر	چه	اندکی	زیاد
خواسته	داشتند	گرفتید	گیرند	یافتیم	خواسته

۲-۳-۱-۳- یکسان‌سازی نگارشی

در برخی زبان‌ها مثل فارسی وجود نویسه‌های متفاوت از یک حرف با کد اسکی متفاوت، باعث می‌شود که هنگام ساخت فرهنگ لغات و شمردن آن‌ها، دو واژه‌ی یکسان با املاهای متفاوت به عنوان دو واژه‌ی مختلف در نظر گرفته شوند. به طور مثال به کد اسکی حرف "ک" و "ك" که دو فرم فارسی و عربی از یک حرف هستند دقت کنید، ممکن است در دو متن مختلف یک کلمه با دو فرم متفاوت از این حرف ظاهر شده باشد. گرچه این دو کلمه یکسان هستند اما سیستم آن‌ها را دو کلمه‌ی مختلف تشخیص می-دهد. برای جلوگیری از این معضل باید در زمان پردازش لغوی، حروف یک دسته را تبدیل به یک حرف یکسان نمود. به طور مثال در همه‌ی متون حرف "ك"به "ک" تبدیل گردد. علاوه بر حالت مذکور یک کلمه ممکن است به خاطر وجود پیشوند یا پسوند، املاهای متفاوتی داشته باشد. به عنوان مثال، پیشوند«می» و پسوند«ها» در ابتدا و انتهای واژه‌ها، ممکن است به سه صورت مختلف دیده شوند؛ کتابها

(چسبان)، کتاب‌ها (با نیم فاصله)، کتاب ها (با فاصله). به عنوان یک مثال از یکسان‌سازی، در مورد پسوند «ها» می‌توان شکل‌های «جدا با فاصله» و«جدا بدون فاصله» را نادیده‌گرفت و با مشاهده‌ی هر یک از این دو، آن را به شکل چسبان در آورد. در [۳۴]، یک مجموعه تبدیل‌های نسبتا جامع و موثر برای یکسان-سازی متون فارسی ارائه شده است. تجربه نشان می‌دهد که با استفاده از تبدیل‌های زیر، می‌توان پیکره‌ی متنی را به خوبی یکسان‌سازی کرد:

- تبدیل نویسه‌های «ی» و «ک» عربی به نوع فارسی آن.
- تبدیل نویسه‌های «ؤ» به «و»،«ئ» به «ی» و«أ» به «ا».
- تبدیل نویسه‌های «ۀ» به «ه» در آخر واژه‌ها.
- حذف «ی» از آخر واژه‌هایی مانند «خانه‌ی».
- حذف پسوندهای«تر» و «ترین» به آخر واژه‌ها.
- حذف فاصله بعد از پیشوند«بر» در واژه‌هایی مانند «برمی‌گردد».
- حذف شناسه‌ی «ء» در آخر بعضی واژه‌ها مانند «شهداء».
- حذف پیشوندهای «می»، «نمی»، «بی»، «درمی»، «برمی» از ابتدای واژه‌ها.
- حذف پسوندهای «ها»، «های»، «هایی»،«هایم»،«هایت»، «هایش»، «هایمان»، «هایشان»، «هایتان» از انتهای واژه‌ها.
- چسباندن پیشوند«به» به واژه‌هایی مانند«به‌ندرت» به ابتدای واژه‌ها.
- چسباندن پیشوند«هم» به واژه‌هایی مانند«هم‌چنین» به ابتدای واژه‌ها.

۲-۳-۱-۴- ریشه‌یابی

هدف از ریشه‌یابی حذف اضافات از کلمه و رسیدن به ریشه اصلی کلمه است به طوریکه تمام کلمات هم‌ریشه در تمام متون، دارای یک نمایش منحصربفرد به صورت ریشه‌ی اصلی کلمه باشند. این گام یکی

از مهم‌ترین و موثرترین مراحل پردازش لغوی است که در مراحل بعدی مثل انتخاب واژه‌های کلیدی، به عنوان ویژگی‌های اصلی متن، نقش بسیار تاثیرگذاری را ایفا می‌کند. روش‌های ریشه‌یابی را می‌توان به دو نوع عمده تقسیم نمود [۳۵]:

- **شیوه‌های مبتنی بر شبکه‌های واژه:** در این شیوه‌ها، از ارتباط واژه‌ها در یک شبکه معنایی استفاده می‌شود. واژه‌های هم‌ریشه در یک خوشه قرار می‌گیرند. این شبکه‌ها ساختاری گراف مانند دارند. این دسته از روش‌ها نیازمند بهره‌گیری گسترده از دانش انسانی هستند، به همین دلیل ایجاد و نگهداری آن‌ها بسیار پرهزینه است. هنوز الگوریتم‌هایی که بتواند به صورت خودکار این شبکه‌ها را به گونه‌ای قابل اطمینان ایجاد کنند، توسعه نیافته است. به دلیل کارایی بالای این دسته از روش‌ها، به طور گسترده در سامانه‌های داده‌کاوی و متن‌کاوی استفاده می‌شود.

- **شیوه‌های مبتنی بر ریخت‌شناسی**: رویکرد دوم استفاده از ریخت‌شناسی برای ریشه‌یابی است. این دسته از روش‌ها با بررسی ساخت واژه‌ها، ریشه‌ی کلمه را پیدا می‌کنند [۳۶]. این روش نیازی به معلوم بودن ریشه‌ی واژه‌ها ندارد و تنها نیازمند آگاهی از قواعد ساخت‌واژی است. از آنجا که در خلاصه-سازی بیشتر ریشه‌یابی تصریفی مورد نظر است و قواعد تصریفی اندک هستند، این دسته از روش‌ها به سادگی قابل پیاده‌سازی هستند.

به طور کلی از مشکلات اصلی ریشه‌یابی می‌توان به بیش‌ریشه‌یابی (تولید ریشه‌هایی که هیچ معنایی در زبان ندارند) و کم‌ریشه‌یابی (عدم امکان تولید ریشه برای حالات استثناء) اشاره کرد [۱۴]. اما سه روش معمول ریشه‌یابی که در بسیاری کاربردها مورد استفاده قرار می‌گیرند به شرح زیر است:

۱. ریشه‌یاب جدولی

این روش ریشه‌یابی از لحاظ پیچیدگی پیاده‌سازی، ساده‌ترین روش ممکن، اما بسیار زمانبر است. در این جداول هر کلمه به همراه ریشه‌اش نگهداری شده و برای یافتن ریشه‌ی یک کلمه باید جدول را برای

یافتن کلمه‌ی مورد نظر جستجو کرده سپس ریشه‌ی متناظر با آن را به دست آورد. در واقع این روش یک عمل جستجو است که منجر به ریشه‌یابی می‌گردد. این روش بیشترین دقت را نسبت به دیگر روشها دارا است؛ اما همانگونه که گفتیم سربار زمانی بالایی دارد.

۲. ریشه‌یاب آماری

پایه اصلی این روش یک مجموعه از کلمات با ساخت‌های متفاوت است و عملکرد خوب این روش وابسته به کامل بودن این مجموعه است. در این دسته از ریشه‌یاب‌ها، با روش‌های آماری، وندهایی که در کلمه‌ها تکرار شده‌اند، شناسایی می‌گردد، از مزیت‌های این روش عدم وابستگی آنها به زبانی خاص است. باید توجه نمود که وجود کلمات نادرست در مجموعه مرجعِ این روش‌ها، بر کارایی این ریشه‌یاب‌ها اثر منفی می‌گذارد [۳۷].

۳. ریشه‌یابی به کمک روش Porter

ریشه‌یاب پُرتر[1] [۳۸]، از معمول‌ترین شیوه‌های ریشه‌یابی در زبان انگلیسی است که بر مبنای زبان-شناسی و دسته‌بندی کلمه‌ها به کمک واج‌ها و هجاها عمل می‌نماید. پایه و اساس روش‌های ریشه‌یابی براساس قاعده‌های زبان نیز همین روش پرتر است. در طراحی این ریشه‌یاب‌ها، روال اصلی کار این است که یک ماشین پذیرنده متناهی تولید می‌گردد که معمولا برخی قواعد و حالت‌های خاص در زمان طراحی از قلم افتاده و در انتها باید به آن اضافه گردد، یعنی به مرور زمان و در زمان‌های مقتضی کد برنامه باید تغییر کند. از این رو روند نگهداری و گسترش ریشه‌یاب بسیار پرهزینه می‌گردد.

[1] Porter

۲-۳-۱-۵- خلاصه سازی

منظور از خلاصه‌سازی، دریافت یک متن و تولید یا استخراج یک متن دیگر از آن متن است، به گونه‌ای که متن به دست‌آمده از متن اصلی کوتاه‌تر باشد، نکات اصلی و مهم آن را دربرداشته باشد و بین جملات آن پیوستگی وجود داشته باشد. خلاصه‌سازی باعث کاهش فضای ویژگی‌ها می‌شود.

به طور کلی، دو نوع اصلی برای خلاصه وجود دارد: خلاصه‌ی گزینشی و چکیده [۳۹].

- **خلاصه‌ی گزینشی**، با توجه به معیارهای آرمای، شهودی[1] و یا ترکیبی از این دو تهیه می‌شود. از از آنجا که در تولید این دسته از خلاصه‌ها، جملات متن تغییرات نحوی و معنایی ندارند، می‌توان آن را نوعی گزینش جملات قلمداد کرد. در واقع اگر متن خلاصه با انتخاب جملاتی از متن اصلی به دست آید، نوع خلاصه‌سازی، «استخراجی» یا «گزینشی» است. در حال حاضر خلاصه‌های استخراجی به دلیل آسان و ارزان بودن عمومیت بیش‌تری دارند.
- **خلاصه‌ی چکیده**، تفسیری از متن اولیه است. درتولید خلاصه‌ی چکیده، مفاهیم جملات متن اصلی به شکل کوتاه‌تر بازنویسی می‌شود. به عنوان مثال، جمله «اوسیب، انگور وگیلاس‌ها را خورد» را می‌توان به صورت «او میوه‌ها را خورد» نوشت. به عبارت دیگر اگر خلاصه متن پس از فهم مطالب موجود در متن اصلی تولید شود، خلاصه‌سازی از نوع «چکیده» است [۴۰].

۲-۳-۱-۶- ساده‌سازی

ساده‌سازی از طرق مختلف و در کاربردهای متفاوتی همچون مراجع [۴۱-۴۳]، بر روی متون انجام می‌گیرد. حذف کلمات هم‌معنی و تعمیم کلمات از جمله رویکردهای متداول در زمینه‌ی ساده‌سازی هستند که منجر به همگون‌سازی مستندات شده و سبب می‌شود که مستنداتی که دارای مضامین مشابهی هستند، دارای کلمات کلیدی تقریبا یکسانی باشند و در نتیجه کارایی و دقت سیستم دسته‌بندی

[1] Heuristic

خودکار متون را افزایش دهند. به عنوان مثال در یک مجموعه از اسناد، در دسته‌ی «ورزش»، اسنادی در زمینه‌ی شنا و اسنادی نیز در زمینه بسکتبال وجود دارند. تعمیم کلمات موجود در این دو دسته از اسناد، آنها را به سمت یک مجموعه از کلمات ریشه یعنی ورزش سوق می‌دهد. باید توجه نمود که برای اعمال روش‌های ساده‌سازی به طرق ذکرشده، به یک دانش پیش‌زمینه از لغات در زبان فارسی نیاز است.

۲-۳-۲- انتخاب ویژگی از متن یا شاخص‌بندی متون

یک متن نمی‌تواند به صورت مستقیم توسط یک دسته‌بند یا یک الگوریتم دسته‌بندی تفسیر شود، بلکه باید با استفاده از یک فرآیند شاخص‌بندی، متن را با استفاده از کلمات کلیدی یا ویژگی‌های استخراج شده از کلمات کلیدی نشان داد. به کلمات مهم هر متن که بیان کننده‌ی مفهوم اصلی متن هستند، **کلمه کلیدی** گفته می‌شود. به عبارت دیگر کلمات کلیدی خلاصه‌ای کوتاه از یک متن هستند. استخراج کلمات کلیدی یک گام مهم و تاثیرگزار در بازیابی اطلاعات، صفحات وب، دسته‌بندی متون، خلاصه‌سازی متون و غیره می‌باشد. این مهم کمک می‌کند تا یکنواختی و یک شکلی لازم برای متون مجموعه آموزشی، آزمایشی و اعتبارسنجی فراهم شود. انتخاب شاخص یا ویژگی برای متن، به مسائل مختلفی بستگی داد، همانند:

- واحدهای معنایی (مساله واژگان معنایی[1])
- قوانین معنایی طبیعی برای ترکیب این واحدها (مساله ترکیب معنایی[2])

معمولا در دسته‌بندی متون مشکل دوم در نظر گرفته نمی‌شود (مشابه آنچه در بازیابی اطلاعات وجود دارد) و یک متن d_j، با برداری از وزن عبارت‌هایش نشان داده می‌شود. به عبارت دیگر $d_j=<W_{1j},W_{2j},\ldots, W_{|T|j}>$ به طوریکه T مجموعه عبارت‌هایی است که دست کم یک بار در سرتاسر مجموعه آموزشی آمده

[1] Lexical Semantics
[2] Compositional Semantics

باشند (به آن ویژگی نیز گفته می‌شود) و $0 \leq W_{kj} \leq 1$ باشد. معمولا تفاوت رهیافت‌ها در این زمینه به یکی از دلایل ذیل می‌باشد:

- تفاوت در تعریف چیزی که "عبارت" نامیده می‌شود.
- تفاوت در زمینه محاسبه وزن ویژگی‌ها.

یک انتخاب متداول برای مسئله اول، تعیین هر کلمه به عنوان یک عبارت می‌باشد. به این انتخاب، رهیافت مجموعه کلمات یا کیسه کلمات[1] برای نمایش متون (مستقل از اینکه وزن‌ها دودویی باشند یا نباشند) گفته می‌شود [۴]. محققان از دو رهیافت مختلف با این مساله برخورد کرده‌اند:

- به طور صرفی و نحوی، تعریف عبارت با استفاده از قوانین گرامری زبان
- به طور آماری، در نظر گرفتن یک ترکیب خطی نه به لحاظ قواعد گرامری بلکه بنا به همنشینی آن یا مجموعه یا دنباله‌ای از کلمات که الگوی خاصی را دنبال می‌نماید.

لوئیز(Lewis) در [۴۴] استدلال کرد که استفاده از عبارت‌ها بیشترین ارزش معنایی و کمترین ارزش آماری را دارد. از طرف دیگر استفاده از کلمات کمترین ارزش معنایی و بیشترین ارزش آماری را دارد.

به طور کلی بسته به کاربرد، کل یا بخشی از یک متن می‌تواند به عنوان شاخص‌بندی و ویژگی‌های یک متن مد نظر قرار گیرد. به عنوان مثال هنگامی‌که عنوان یک متن مشخص باشد، می‌توان در متن اصلی اهمیت بیشتری به کلماتی داد که در عنوان نیز حضور دارند. اما هنگامی‌که یک متن قالب خاصی ندارد و به صورت متن تنهاست، تشخیص قسمتی که بیشترین ارتباط را با کل موضوع دارد، یک مساله بغرنج و نا مشخص است.

[1] Bag of Words

۲-۳-۲-۱- روش‌های متداول در انتخاب کلمات کلیدی

در حالت کلی سه روش بسیار معمول برای انتخاب کلمات کلیدی وجود دارد که در بیشتر تحقیقات از آنها استفاده می‌شود [۴] :

- **روش‌های آماری** : در این روش‌ها میزان تکرار یک کلمه در یک متن و یا تعداد متون حاوی یک کلمه اهمیت خاصی دارند و بر اساس یک معیار آماری مثل TFIDF و یا معیارهای ارزش‌گذاری کلمه، کلمات کلیدی متن انتخاب می‌شوند.
- **روش یادگیری ماشینی**: در این شیوه با یک مجموعه متون آموزشی و کلمات کلیدی مشخص برای آن‌ها ، فرآیند انتخاب کلمات کلیدی به عنوان یک مساله دسته‌بندی مدلسازی می‌شود. کلمات بر اساس مشخصه‌هایی که دارا هستند، به کلمات کلیدی و غیر کلیدی دسته‌بندی می‌شوند. احتمالات دسته‌بندی به صورت آماری از مجموعه‌ی آموزشی یاد گرفته می‌شود. این روش‌ها دارای انعطاف‌پذیری بالایی هستند.
- **ترکیب روش‌های تحلیل آماری و زبانشناختی**: علت ترکیب این دو روش این است که اطلاعات آماری بدون پالایش زبان‌شناختی نتایج مفیدی را در اختیار نمی‌دهند. بدون اطلاعات زبانشناختی، کلمات نامطلوبی مانند "از"، "او" و غیره هم به علت تعداد تکرار زیاد انتخاب خواهند شد.

همانطور که گفته شد در روش‌های آماری با استفاده از معیارهای ارزش‌گذاری کلمات و معیارهای وزن‌دهی، به انتخاب کلمات کلیدی پرداخته می‌شود. در بخش بعدی این معیارها معرفی می‌شوند.

۲-۳-۲-۲- طرح های وزن‌دهی به کلمات کلیدی

برای انتخاب کلمات کلیدی مناسب برای هر متن به عنوان ویژگی آن متن، آنچه بسیار مهم است، میزان اهمیت آن ویژگی در آن متن است. میزان اهمیت و تاثیرگزاری یک کلمه در یک متن را به عنوان

وزن آن کلمه تعریف می‌کنیم. چگونگی در نظر گرفتن میزان تاثیرگزاری و اهمیت کلمات و به اصطلاح وزن آنها، به قدری اهمیت دارد که حتی می‌تواند موثرتر از دسته‌بند انتخابی قلمداد گردد [۴۵].

در این بخش به معرفی چند شیوه‌ی مختلف وزن‌دهی که در بازیابی اطلاعات و دسته‌بندی متون بسیار رایج هستند می‌پردازیم:

۲-۳-۲-۲-۱- روش‌های مبتنی بر TF

در این روش‌ها، وزن‌دهی ویژگی‌ها از توزیع آنها در هر یک از متون تبعیت می‌کند.

- روش TF

این روش برای اولین‌بار در مرجع [۴۶] ارائه شد و بدین شکل وزن یک ویژگی را تعیین می‌نمود که اگر ویژگی f_k در متن d_i حضور داشته‌باشد، وزن آن برابر تعداد تکرار آن خواهد بود.

$$W_{ik} = tf(f_k, d_i) = \begin{cases} \#(f_k, d_i) & if\ f_k \in d_i \\ 0 & if\ f_k \notin d_i \end{cases} \quad (۱-۲)$$

در رابطه‌ی ۲-۱ $\#(f_k, d_i)$ تعداد تکرار ویژگی f_k در متن d_i است.

- روش normTF

از آنجائیکه طول متون در مجموعه اسناد یکسان نیست، روش normTF به منظور رفع اثر طول مستند بر روی طرح وزن‌دهی و نگاشت آن در بازه‌ی (۱و۰) از نرمال‌سازی استفاده می‌کند (رابطه ۲-۲).

$$W_{ik} = norm\_tf(f_k, d_i) = {tf(f_k, d_i)}\Big/{\sqrt{\sum_k tf(f_k, d_i)^2}} \quad (۲-۲)$$

- روش logTF

در این روش به منظور رفع تاثیر منفی متفاوت بودن ماهیت ویژگی‌ها و مقادیرشان که منجر به کاهش دقت و کارایی دسته‌بندی می‌گردد، از عملگر لگاریتم استفاده شده‌است. این عملگر اثر ذکرشده را حذف و محدوده مقادیر تخصیص یافته به ویژگی‌ها را یکسان می‌نماید (رابطه ۲-۳).

$$W_{ik} = Logtf(f_k, d_i) = \log\big(tf(f_k, d_i)\big) \quad (۳-۲)$$

- روش ITF

این روش برای اولین بار در مرجع [۴۷] معرفی‌شد که براساس آن وزن هر ویژگی از رابطه ۲-۴ بدست می‌آید.

$$W_{ik} = ITF(f_k, d_i) = 1 - \frac{r}{r + tf(f_k, d_i)} \quad (۴-۲)$$

در معادله‌ی ۲-۴ معمولاً r برابر ۱ درنظر گرفته می‌شود.

- روش Sparck

روش Sparck از تئوری‌های آماری برای وزن‌دهی به ویژگی‌ها استفاده‌کرده و برای اولین‌بار در مرجع [۴۸] ارائه شد.

$$W_{ik} = Sparck(f_k, d_i) = \ tf(f_k, d_i) * (k - \log(p_k)) \quad (۵-۲)$$

در رابطه ۲-۵، k تعداد کل ویژگی‌ها در مجموعه متون و $p_k = \sum_k tf(f_k, d_i)$ است.

۲-۳-۲-۲-۲- روش‌های مبتنی بر IDF

در این دسته از روش‌ها، وزن‌دهی به مشخصه‌ها از توزیع ویژگی f_k در مجموعه متون تبعیت می‌نماید. مبنای اصلی این دسته از روش‌ها بر این اصل استوار است که هر چه تعداد متونی که دارای ویژگی f_k باشند کمتر باشد، این ویژگی برای تمایز بخشیدن به متون، مناسب‌تر خواهدبود.

- روش IDF

این روش برای اولین‌بار در مرجع [۴۹] حوزه‌ی بازیابی اطلاعات مطرح شده‌است و در آن وزن ویژگی f_k از معادله ۲-۶ بدست می‌آید. این طرح وزن‌دهی از متغیر i که نشان‌دهنده شماره‌ی متن است، مستقل است.

$$W_{ik} = idf(f_k, d_i) = \log\left(\frac{|D|}{|D(f_k)|}\right) \quad (۶-۲)$$

در رابطه فوق $|D|$ تعداد کل متون مجموعه و $|D(f_k)|$ تعداد متون از مجموعه اسناد است که ویژگی f_k در آنها حضور دارد. همانطور که مشخص است وزن ویژگی f_k با افزایش $|D(f_k)|$ کاهش می‌یابد و این بدین معنی است که هر چه یک کلمه در متون بیشتری حضور داشته‌باشد، تمایزبخشی و اهمیت آن در دسته‌بندی متون کمتر است.

۲-۳-۲-۲-۳- روش‌های مبتنی بر TFIDF

این روش‌ها ابتدا در حوزه‌ی بازیابی اطلاعات مطرح شده و سپس در دسته‌بندی متون به منظور وزن-دهی به ویژگی‌ها مورد استفاده قرار گرفت.

- روش TFIDF

روش TFIDF متداول‌ترین طرح وزن‌دهی به ویژگی‌ها در زمینه دسته‌بندی و خوشه‌بندی متون است. این روش در واقع حاصل ترکیب روش‌های مبتنی بر TF و IDF است. وزن ویژگی f_k در این طرح از رابطه (۲-۷) بدست می‌آید.

$$W_{ik} = TFIDF(f_k, d_i) = tf(f_k, d_i) * idf(f_k, d_i) \quad (۷-۲)$$

- روش normTFIDF

برای اطمینان از برابری شانس تمام متون با طول‌های متفاوت، بازیابی روش TFIDF در مرجع [۵۰] بصورت نرمال‌شده همانند معادله ۲-۸ ارائه شده‌است.

$$W_{ik} = normTFIDF(f_k, d_i) = {tfidf(f_k, d_i)}\Big/{\sqrt{\sum_k tfidf(f_k, d_i)^2}} \quad (۸-۲)$$

در این بخش به معرفی چند روش وزن‌دهی پرداخته شد. استفاده از یک حد آستانه می‌تواند روش خوبی برای انتخاب و یا کاهش ویژگی‌ها باشد.

۲-۳-۲-۳- معیارهای ارزش‌گذاری ویژگی

در این بخش چند معیار پرکاربرد به عنوان معیاری برای سنجش میزان ارزش یک داده در یک کلاس و یا پایگاه داده معرفی می‌کنیم. در بسیاری از تحقیقات از این معیارها برای انتخاب ویژگی و فیلتر کردن ویژگی‌ها و در نهایت کاهش تعداد ویژگی‌ها استفاده شده است.

۲-۳-۲-۳-۱- آنتروپی[1]

معیاری است که اهمیت داده را در کل کلاس‌ها نشان می‌دهد. اگر مقدار آنتروپی عبارت t_i صفر باشد، یعنی این عبارت فقط در یک کلاس رخ داده است [۲, ۱۲] و احتمالا این عبارت قدرت تفکیک کنندگی خوبی برای دسته‌بندی خواهد داشت. اما اگر این عبارت در کلاس‌های مختلف و با تعداد رخداد مشابه ظاهر شود، مقدار آنتروپی برای آن زیاد خواهد بود و این ویژگی قدرت تفکیک‌کنندگی خوبی نداشته و برای کاهش ویژگی‌ها یک کاندید مناسب است. اگر ${n_i}^k$ ، تعداد رخداد عبارت t_i در کلاس c_k باشد، فرکانس حضور t_i در کلاس c_k ، یعنی ${tf_i}^k$ به صورت رابطه ۲-۹ تعریف می‌شود.

$$ {tf_i}^k = \frac{{n_i}^k}{\sum_{k=1}^{m} {n_i}^k} \qquad (۹-۲) $$

مقدار آنتروپی هم برای عبارت t_i با استفاده از ویژگی مقدار فرکانس حضور عبارت در کلاس به صورت رابطه ۲-۱۰ محاسبه می‌شود.

$$ Entropy(t_i) \quad = -\sum_{k=1}^{r} ({tf_i}^k) * log_2({tf_i}^k) \qquad (۱۰-۲) $$

[1] Entropy

۲-۳-۲-۳-۲- بهره اطلاعات[1]

بهره‌ی اطلاعات که به اختصار Info-Gain یا IG نامیده می‌شود، یک معیار برای ارزش‌گذاری ویژگی-هاست. این معیار برای هر ویژگی نسبت به کلاس‌های مختلف مقداردهی می‌شود، نه نسبت به متون. در کارهای [۶, ۷, ۱۲] از این معیار استفاده شده است. برای کلمه‌ی t_i به عنوان یک ویژگی و کلاس c_k ، بهره اطلاعات $IG(t_i,c_k)$، با فرمول ۲-۱۱ تعریف می‌شود.

$$IG(t_i, c_k) = d(t_i, c_k) \log(\frac{\mathrm{d}(t_i\ , c_k)}{\mathrm{d}(t_i)\mathrm{d}(c_k)}) + d(t'_i, c_k) \log(\frac{\mathrm{d}(t'_i\ , c_k)}{\mathrm{d}(t'_i)\mathrm{d}(c_k)}) \qquad (۲-۱۱)$$

$d(t_i, c_k)$: تعداد متون از کلاس c_k که کلمه‌ی t_i در آنها رخ داده است.

$d(t'_i, c_k)$: تعداد متون از کلاس c_k که کلمه‌ی t_i در آنها رخ نداده است.

$d(t_i, c'_k)$: تعداد متون از تمام کلاس‌ها به جز کلاس c_k که کلمه‌ی t_i در آنها رخ داده است.

$d(t'_i, c'_k)$: تعداد متون از تمام کلاس‌ها به جز کلاس c_k که کلمه‌ی t_i در آنها رخ نداده است.

$d(t_i)$: تعداد متون از کل پایگاه داده که کلمه‌ی t_i در آنها رخ داده است.

$d(t'_i)$: تعداد متون از کل پایگاه داده که کلمه‌ی t_i در آنها رخ نداده است.

$d(c_k)$: تعداد متون کلاس c_k است.

$d(c'_k)$: تعداد متون همه‌ی کلاس‌ها به غیر از کلاس c_k است.

۲-۳-۲-۳-۳- معیار آماری مربع چی[2] یا χ^2

χ^2 یا میزان اثربخشی، درجه وابستگی و مشارکت میان یک عبارت و دسته را تعیین می‌نماید. این کاربرد را می‌توان برپایه این فرضیه دانست که عبارتی که تعداد تکرار بیشتری در یک دسته دارد، می‌تواند ویژگی متمایزکننده‌ی مناسبی برای آن دسته باشد. بنابراین ویژگی‌هایی که دارای مقدار χ^2 کمتری هستند، از اهمیت پایین‌تری برخوردار بوده و در نتیجه حذف آنان تاثیری بر دسته‌بندی نخواهد داشت و با

[1] Information Gain
[2] Chi Square

حذف آنان، فضای ویژگی کاهش می‌یابد، زمانی که χ^2 مقدار صفر دارد یعنی این ویژگی و کلاس کاملا مستقل از هم هستند[۶, ۱۲, ۵۱]. رابطه‌ی ۲-۱۲ میزان اثربخشی عبارت t_i را برای کلاس c_k، محاسبه می‌کند.

$$\chi^2(t_i, c_k) = \frac{\mathrm{N}.[\mathrm{d}(t_i\,, c_k).\,\mathrm{d}(t'_i\,, c'_k) - \mathrm{d}(t'_i\,, c_k).\,\mathrm{d}(t_i\,, c'_k)]}{d(t_i).\,d(t'_i) \;-\; d(c_k).\,d(\,c'_k)}) \qquad (۱۲-۲)$$

متد χ^2 یک متد نظارت‌شده چندمتغیره است که انتخاب عبارات موثر در تفکیک‌پذیری را نه تنها براساس تعداد تکرار آنها در هر دسته، بلکه با توجه به وابستگی عبارات با یکدیگر و نیز وابستگی عبارات با دسته‌ها انجام می‌دهد.

۲-۳-۲-۳-۴- معیار Dominance

این معیار نیز اهمیت یک عبارت را در یک کلاس نشان می‌دهد و نسبت تعداد متون کلاس K که کلمه t_i در آن رخ داده است را به کل متونی از همه‌ی کلاس‌ها که کلمه t_i در آنها رخ داده است، بیان می‌کند و مقدار آن بین صفر و یک است. اگر مقدار Dominance برای یک عبارت نسبت به یک کلاس "۱" باشد، یعنی آن عبارت فقط در آن کلاس رخ داده و در کلاس دیگری رخ نداده است [۷]. فرمول ۲-۱۳ مقدار این معیار را برای عبارت t_i را نسبت به کلاس c_k، محاسبه می‌کند.

$$Dominance(t_i, c_k) = \frac{\mathrm{d}(t_i\,, c_k)}{d(t_i\,, c_k) + \mathrm{d}(t_i\,, c'_k)}) \qquad (۱۳-۲)$$

در بخش قبل در مورد انتخاب کلمات کلیدی بحث شد. در بخش بعدی به معرفی روش‌های استخراج ویژگی می‌پردازیم. **استخراج ویژگی**، فرآیند تشکیل ویژگی‌های جدید از ویژگی‌های انتخابی اولیه است.

۲-۳-۲-۴- روش‌های موجود در استخراج ویژگی

تحقیقات در زمینه‌ی استخراج خودکار ویژگی‌ها جهت دسته‌بندی متون به قبل از سال ۱۹۹۰ برمی-گردد [۷]. در این بخش به متداول‌ترین روش‌ها اشاره می‌کنیم:

۲-۳-۲-۴-۱- عبارات نحوی [1]

اولین روش‌های استخراج ویژگی‌ها به آغاز دهه‌ی ۹۰ میلادی بازمی‌گردد، زمانی‌که برخی محققان از عبارات نحوی به عنوان ویژگی‌ها، برای دسته‌بندی متون استفاده کردند [۵۲-۵۴]. عبارات نحوی به گروهی از کلمات در معنای دستوری، از قبیل عبارات اسمی، عبارات فعلی و عبارات مصدری و غیره گفته می‌شود که با استفاده از فرآیندهای تجزیه، مثل تکنیک‌های پردازش زبان طبیعی به دست می‌آیند. بیشتر کارهای انجام شده، نشان داده‌اند که استفاده از عبارات گرامری بهتر از کیسه‌ی کلمات در دسته‌بندها عمل می‌کند. لوئیز در [۵۳] دلایل موفقیت دسته‌بندی را بر اساس چهار ویژگی مطلوب نشان داد. این چهار ویژگی عبارتند از :

- فقدان افزونگی در شاخص‌بندی عبارات
- توزیع نسبتا یکنواخت ارزش‌ها در شاخص‌بندی
- فقدان ابهام ناشی از شاخص‌بندی
- نویز پایین در ارزش کلمات شاخص‌بندی

لوئیز آزمایشاتی انجام داد که نشان داد عبارات نحوی، سعی در ایجاد چهارمین ویژگی دارند و نه سایر ویژگی‌ها. به دلیل این‌که اولا، چون فرکانس عبارات نحوی در میان کلاس‌های مختلف، متنوع است و ثانیا، ابهام ناشی از ایجاد مدل که نتیجه‌ی کاربرد عبارات مترادف به عنوان معیارهای هم‌ارزی معنایی در میان عبارات نحوی، که از لحاظ لغوی متفاوتند، بسیار قابل توجه است. همچنین با در نظر گرفتن هر کلمه‌ی تنها از یک عبارت T کلمه‌ای، که S مفهوم مترادف دارد، S^T عبارت نحوی با معنای مترادف دارد.

[1] Syntactic Phrases

لوئیز نتیجه گرفت که فرکانس پایین عبارات متمایز معنایی و ابعاد بالای فضای مترادف‌ها، ظرفیت بالقوه‌ی عبارات نحوی را به عنوان یک ویژگی تفکیک کننده کاهش می‌دهد. در [۵۵] نیز، نتیجه‌ی قابل قبولی در دسته‌بندی با استفاده از عبارات نحوی بدست نیامد. علیرغم این نتایج، نویسندگان [۵۶] یک تحقیق مشابه با لوئیز با تغییراتی در جزئیات پیاده سازی، برای انتخاب ویژگی‌های بهتر انجام دادند. برای این کار از الگوریتم‌های Rule-Based استفاده کردند. فرض آنان این بود که برخی قوانین، عبارت نحوی متمایز کننده‌تری در نظر می‌گیرند و در نتیجه دسته‌بندهای بهتری تولید می‌کنند. با وجود این، با انجام آزمایشات نه تنها با استفاده از بهترین عبارات نحوی متمایز کننده، نتایج بهتری نسبت به کاربرد کیسه‌ی لغات بدست نیامد، بلکه کیفیت دسته‌بندی هم کاهش پیدا کرد. نویسندگان مقاله با توجه به نتایج به دست آمده، نتیجه گرفتند که ادامه‌ی کار بر روی عبارات نحوی بی نتیجه خواهد بود.

۲-۳-۲-۴-۲- n-grams کلمات[1]

نویسندگان در [۵۷]، کاربرد n-grams کلمات (دنباله‌ای از n کلمه‌ی مجاور)، با اندازه‌های متفاوت را به عنوان ویژگی در دسته‌بندی متون بررسی کردند. آنها نشان دادند که n-grams های سه کلمه‌ای (ونه بیشتر از سه کلمه) نتیجه‌ای بهتر از کیسه‌ی لغات دارند. کار [۵۸] نشان داد که n-grams کلمات، نه تنها بهتر از کیسه‌ی لغات عمل نمی‌کند، بلکه در مواردی بدتر از آن نتیجه می‌دهد. مقاله [۵۹] نتایج خوبی را در کاربرد n-grams های دو کلمه‌ای گزارش کرد. اما تمرکز بروی کلاس‌هایی بود که تنها ۲,۸٪ از کل متون را تشکیل می‌داد. حتی خود نویسنده هم در پایان اظهار می‌دارد که نتایج حاصله به دلیل وسعت کم داده‌ها، خیلی قابل نتیجه‌گیری نیست. در [۶۰] از یک ریشه‌یاب برای در نظر گرفتن n-gramهای متفاوت مورفولوژیکی، اما با معنی مشابه استفاده شد. در کنار این n-gramsها از ویژگی‌های منفرد تفکیک کننده نیز استفاده شد. نتایج نشان داد که این روش کارایی لازم را ندارد. [۶۱] نیز از ترکیب n-gramها و

[1] Word n-grams

ویژگی‌های منفرد استفاده کرد و فقط ۲٪ از n-gramهای تفکیک کننده برای این کار استفاده شدند نتایجی بهتر از کاربرد کیسه‌ی لغات ارائه شد. در [۶۲] نیز n-gramها برای دسته‌بندی ایمیل‌ها به کار گرفته شد.

به طور خلاصه مطالعه‌ی کارهای مرتبط با استفاده از n-gramها، نشان داد که موفقیت‌ها در شرایط و یا داده‌های خاص بوده و یک موفقیت کلی بدست نیامده است.

۲-۳-۲-۴-۳-ویژگی‌های همرخداد غیر مجاور[1] (C-Features)

از ایده‌ی n-gramها استفاده شده و با حذف محدودیت مجاورت کلمات و فاصله‌ی آنها از هم، ویژگی موثری بنام ویژگی‌های همرخداد غیر مجاور معرفی شد. منظور از غیرمجاور الزاما غیر مجاور بودن نیست بلکه تاکید این است که مانند n-gramها اجباری در نزدیکی و مجاورت کلمات تشکیل دهنده‌ی ویژگی وجود ندارد. ترتیب و فاصله‌ی کلمات یک ویژگی همرخداد محدودیتی نخواهد داشت. این ویژگی‌ها بیشتر در حوزه‌ی بازیابی اطلاعات متنی به کار رفته‌اند و در حوزه‌ی دسته‌بندی کمتر به کار گرفته شده‌اند.

در [۸] رابطه‌ی میان ویژگی‌های منفرد، با در نظرگرفتن کلمات مجزا در متون مشابه بررسی شد. فرض مسئله این بود که نویسندگان تمایل دارند که ایده‌های مرتبط را با استفاده از چندین ویژگی منفرد در عبارت‌های مختلف در متون مشابه تکرار کنند. در این کار از چکیده‌ی متون علمی که معمولا عبارات تکراری و مشخصی دارند استفاده شده است.

نویسنده در [۱۰]، در دسته‌بندی چند کلاسه‌ی متون، از ویژگی‌های همرخداد پرتکرار برای توسعه‌ی متون استفاده کرده و به حذف ویژگی‌هایی که فرکانس حضور آنها در چندین کلاس بالاست، پرداخته است. در این کار نویسنده از تکنیک‌های هرس برای کاهش کلمات کاندید در ساخت ویژگی‌های همرخداد

[1] Non-adjacent Co-occurrences Features

در مقابله با تاثیر حجم نمایی فضای جستجوی کلمات استفاده کرده است، اما جزئیاتی در مورد پیاده-سازی آن ارائه نکرده است.

در [۱۱] نیز کار مشابهی انجام شده و وزن مرتبط با فرکانس حضور برای ویژگی‌ها در نظر گرفته شده است. در این مقاله از ویژگی‌های همرخداد در استخراج قواعد انجمنی برای دسته‌بندهای انجمنی[1] استفاده شده است. در [۹]، نویسنده نشان داد که ویژگی‌های با فرکانس متوسط بهترین و تفکیک کننده‌ترین ویژگی‌ها هستند.

مقاله [۷] نیز نشان داد ویژگی‌های کم تکرار نیز می‌توانند مهم و تفکیک کننده باشند. کارهای قبلی از ویژگی‌های همرخداد با هر تعداد کلمات مختلف استفاده کرده بودند، اما مقاله‌ی مذکور از ویژگی‌های دو کلمه‌ای استفاده کرده است. در ابتدا با استفاده از معیار Info-Gain برخی ویژگی‌های منفرد را حذف کرده، سپس با معرفی معیار Dominance و در نظر گرفتن یک حد آستانه برای آن، به حذف ویژگی‌های بیشتری می‌پردازد. پس از آن با استفاده از ویژگی‌های باقیمانده به تشکیل ویژگی‌های همرخداد پرداخته است. برای توسعه‌ی متون نیز از ویژگی‌های همرخداد انتخابی بر اساس معیار Dominance استفاده کرده است. سپس از دسته‌بندهای KNN، Naïve Baiysian و SVM برای دسته‌بندی متون چهار پایگاه داده 20 Newsgroups، Reuters، OHSUMED و ACM استفاده کرده است.

۲-۳-۳-کاهش ابعاد داده

ابعاد بزرگ فضای ویژگی‌ها در دسته‌بندی متون (یعنی مقدار بزرگ |T|) معمولا دردسرساز می‌باشد. در حقیقت، بزرگ شدن فضای ویژگی‌ها باعث پیچیدگی بیشتر (صرف هزینه زمانی و حافظه‌ی بیشتر) شده و از طرفی عدم وابستگی بین اطلاعات کمتر می‌گردد. داده‌هایی که به هم وابسته‌اند، عموما ارزش دسته-بندی ندارند. بیشتر دسته‌بندها در این موارد به بیش یادگیری برمی‌خورند. بنابراین عموما قبل از دسته-

[1] Associative Classifier

بندی، داده‌ها از فرآیند کاهش ابعاد داده می‌گذرند. روش‌های کاهش بعد داده‌ها در دو حوزه عمومی و محلی(داخل کلاسی) قابل بحث است.

بنا بر آنچه گفته‌شد، مساله کاهش بعد داده‌ها یکی از زمینه‌های اصلی تحقیق در حوزه‌ی متون است. عمده این تحقیقات بر دو محور اصلی استوار است [۴]:

- **انتخاب ویژگی:** در این روش یک مجموعه از عبارات بر مبنای تئوری اطلاعات یا ویژگی‌های آماری از متون انتخاب می‌شود. این رهیافت را "فیلترکردن" می‌نامند چراکه عبارات بی‌ربط از عبارات استخراج‌شده فیلتر می‌شوند [۶۳].
- **استخراج ویژگی:** در این رویکرد عبارات از طریق تابع تبدیل خاصی به یک فضای ویژگی جدید منتقل می‌شوند. این روشها با ایجاد عبارات ترکیبی جدید از مجموعه اصلی، یک فضای ویژگی جدید ساخته و سعی می‌کنند تا با انجام جایگزینی کلمات با مفاهیم‌شان، بعد فضای عبارت را کاهش دهند. به طوری که ممکن است ویژگی‌های نهایی کاملا متفاوت از ترم‌های اولیه باشند.

از شیوه‌های متداول در کاهش ویژگی‌ها، روش‌های PCA [۱] [۶۴, ۶۵] ، LDA [۲] [۶۶] و ICA [۳] [۶۷] را می‌توان نام برد که در دامنه‌های متفاوت کاری و بر روی داده‌های مختلف به خوبی عمل می‌کند. این روش‌ها در دسته‌بندی متون نیز به کار گرفته شده اند.

۲-۳-۴- الگوریتم‌های دسته‌بندی

الگوریتم‌هایی که بصورت خودکار ویژگی‌های دسته‌ها را از متون نمونه یاد می‌گیرند، الگوریتم‌های با ناظر نامیده می‌شوند. در فاز آموزش این الگوریتم‌ها، هر متن تعدادی ویژگی خواهد داشت که در حوزه

[1] Principal Component Analysis
[2] Linear Discriminant Analysis
[3] Independent Component Analysis

بازیابی اطلاعات به آن عبارت[1] می‌گویند و منجر به ساخت یک مدل دسته‌بند می‌گردد. در این کار منظور از ویژگی، همان کلمات کلیدی استخراج‌شده و یا ویژگی‌های همرخداد از متون است.

الگوریتم‌های دسته‌بند باناظر را می‌توان به سه دسته کلی تقسیم نمود[۱]:

۱. **دسته‌بندهای مبتنی بر قاعده**[2]: این دسته‌بندها با استنباط یک مجموعه از قواعد از اسناد که از پیش دسته‌بندی شده‌اند، یادگیری را انجام می‌دهند. الگوریتم ریپر[3] از جمله دسته‌بندهای مبتنی بر قاعده است. قواعد تصمیم ممکن است بصورت درخت تصمیم مانند C4.5 در مرجع [۶۸] باشند.

۲. **دسته‌بندهای خطی**[4]: در این دسته‌بندها، برای هر دسته یک نمایه محاسبه می‌شود که در واقع برداری از اوزان، براساس فرکانس و احتمالات حضور ویژگی خاص است. برای هر دسته و هر سند، امتیاز براساس ویژگی‌های دسته و سند محاسبه می‌شود. از دسته‌بندهای این دسته می‌توان به دسته‌بند بیز اشاره نمود که براساس تخمین شرایط احتمالی عمل می‌کند [۶۹]. ماشین‌های بردار پشتیبان [۱۳]، دسته‌بند خطی بهینه را با استفاده از انتقال به فضای ویژگی بدست می‌آورند. این گروه از دسته‌بندها همچنین شامل الگوریتم‌های یادگیری اکتشافی از هوش مصنوعی مانند پرسپترون می‌باشند که در آن اوزان از راهی پیچیده‌تر بدست می‌آیند [۷۰].

۳. **دسته‌بندهای مبتنی بر نمونه**[5] : این دسته‌بندها یک سند جدید را با پیدا کردن K تا از اسناد نزدیکتر به آن در مجموعه آموزشی، دسته‌بندی می‌کنند و با رأی‌گیری، آن را به دسته‌ی نزدیکتر تخصیص می‌دهند [۷۱] مثل KNN.

[1] Term
[2] Rule Based Classifiers
[3] Ripper
[4] Linear Classifiers
[5] Example Based Classifier

۲-۴-مقالات موجود در دسته‌بندی متون

تحقیقات بسیاری در زمینه دسته‌بندی متون انجام شده است که برخی از آنها را به طور خلاصه می-آوریم. برای مطالعه‌ی بهتر این بخش، مقالات را به دو دسته مقالات فارسی و غیر فارسی تقسیم می‌کنیم.

۲-۴-۱-مقالات مبتنی بر زبان‌های غیر فارسی

در سال ۲۰۱۰ جیانگ[1] و همکاران در [۷۲]، ادعا کردند روش مبتنی بر گرافی که ارائه داده‌اند، از روش مجموعه کلمات در دسته‌بندهای مختلف کارایی بهتری دارد. در این روش یک الگوریتم کاوش گراف وزن‌دار، زیر گراف‌ها را به صورت مکرر استخراج می‌کند. سپس مهم‌ترین زیرگراف‌ها به عنوان ویژگی برای هر متن انتخاب می‌شوند.

کوئی[2] و همکاران [۷۳]، پس از انتخاب ویژگی‌های هر کلاس به صورت محلی، با استفاده از معیارهای TF، DF و IDF، با اعمال روش CHI بر روی ویژگی‌های محلی، یک ویژگی کلی به دست آورده که نتایج خوبی را به بار آورده است.

همین‌طور هو[3] و همکاران [۶۷]، از تحلیل اجزای مستقل (ICA) برای انتخاب ویژگی‌ها استفاده کردند. آنها این روش را با برخی روش‌های سنتی دیگر مثل CHI، DF، IG و MI ترکیب کردند و بهترین نتیجه را از ترکیب ICA با CHI به دست آوردند.

چنگ[4] و همکاران در [۷۴]، روشی بنام TEERS، یک فرمول جدید برای اهمیت ویژگی‌ها ارائه کرده و بر مبنای قابلیت دسته‌بندی، ویژگی‌های شرطی ارائه کردند. پس از کاربرد فرمول فوق از تحلیل میزان همبستگی ویژگی‌ها برای کاهش موثر ویژگی‌ها استفاده کردند. آنها در گام اول نیز Wordnet را برای رفع مشکل کلمات هم‌معنی و کاهش ابعاد داده به کار بردند.

[1] Jiang
[2] Li-Quing Qui
[3] M.Hu
[4] Y.Cheng

پارک[1] نیز در [۷۵]، نشان داد که در دسته‌بندی متون چند برچسبه، خروجی‌های روش‌های خطی چندگانه می‌تواند برای ساختن ویژگی‌های جدید استفاده شود و از آن برای نمایش متن در ابعاد کمتر استفاده نمود. برای این کار از خروجی‌های SVM و LSR[2] در کاهش بعد استفاده کرده و در نهایت از ترکیب آنها ویژگی‌های جدیدی ارائه کرد.

ژو[3] و همکاران [۷۶]، نشان دادند که TFIDF سنتی تنها به خصوصیت تکرار یا بسامد توجه کرده است و توجهی به مکان قرار گرفتن آن‌ها نکرده است. لذا با در نظر گرفتن خصوصیاتی همچون: بسامد، موقعیت و روابط متقابل، یک شیوه‌ی جدید وزن‌دهی بنام TF-IDF-Rel را بر مبنای توسعه کلمات موضوعی ارائه کردند.

در سال ۲۰۰۷، گویال[4] و همکاران [۷۷]، با ترکیب بیز ساده و شبکه عصبی پرسپترون، بر مشکل توزیع بد داده در مجموعه آموزشی فائق آمده و نتایج خوبی را گزارش کرده‌اند.

در همین سال، چن[5] و همکاران [۷۸]، از پرسپترون افزایشی برای دسته‌بندی متن استفاده کردند. مدل ارائه شده دارای این قابلیت است که با افزایش تعداد کلاس‌ها مشکلی در آموزش شبکه ایجاد نشوند.

در سال ۲۰۰۸، وانگ[6] و همکاران [۶۶]، برای دسته‌بندی متون، ابتدا بردار ویژگی هر متن را که دارای ابعاد زیادی است، با استفاده از الگوریتم LDA کاهش داده سپس با استفاده از الگوریتم SVM به طبق-بندی داده‌ها پرداخته است.

در همین سال فنگ و وارن[1] [۷۹]، با ترکیب الگوریتم ژنتیک و شبیه سازی تبرید و نیز با استفاده از نوعی الگوریتم فیدبک بنام Rochio، سیستمی را ارائه کردند که با تولید قوانینی سعی در دسته‌بندی متون دارد.

[1] C.H.Park
[2] Least Squre Regression
[3] Zhou
[4] Goyal
[5] Chen
[6] Zigiang Wang

سو[۲] و همکاران [۸۰]، طی مقاله‌ای یک روش جدید بنام AC-Closed معرفی می‌کردند که پس از استخراج یک سری قوانین از دادگان آموزشی، به طبقه‌بندی متون می‌پردازد.

شی[۳] و همکاران [۸۱]، ابتدا از استفاده از شبکه عصبی مصنوعی ISOMAP، به کاهش ابعاد ویژگی-های دادگان آموزشی پرداختند و سپس با استفاده از SVM تک کلاسه به طبقه‌بندی چند کلاسه‌ی داده-ها پرداختند. آن‌ها ادعا کرده‌اند که کارایی الگوریتم، بدون تغییر در میزان درستی آن افزایش یافته است.

ژو[۴] [۸۲]، روشی ارائه کرد که از مجموعه‌های سخت برای استخراج و کاهش ابعاد استفاده نمود. حاصل این استخراج ویژگی به یک شبکه عصبی FCM داده شد که در آن روابط غیر مهم به روابط همبستگی تغییر داده شده‌اند.

وانگ[۵] و همکاران [۸۳]، برای دسته‌بندی متن، ابتدا از مجموعه‌های سخت برای کاهش ابعاد فضای آموزشی استفاده کردند. سپس با استفاده از الگوریتم ژنتیک به انتخاب ویژگی پرداختند. عمل دسته‌بندی را یک SVM انجام می‌دهد که پارامترهای آن توسط الگوریتم ژنتیک بهینه انتخاب شده‌اند.

یانگ[۶] و همکاران [۸۴]، یک چارچوب کلی برای الگوریتم‌های دسته‌بندی کننده‌ی متن برای استفاده از آنتولوژی پیاده‌سازی کرده و ادعا کردند کارایی خوبی هم دارد.

دینگ[۷] و همکاران [۸۵]، با تغییراتی در الگوریتم بیز ساده، شرط عدم همبستگی داده‌ها با یکدیگر را در آن کمرنگ کرده‌اند. با این تغییر، الگوریتم در مورد داده‌های دارای همبستگی نیز خوب عمل می‌کند.

---

[1] ZHU Zhen-fang Lu Run
[2] Zhitong Su
[3] Lukui Shi
[4] Yu
[5] Wang
[6] Yang
[7] Wang Ding

ژو[1] و همکار [۸۶]، طی مقاله‌ای از مدل تخصیص پنهان دیریکله برای کاهش ابعاد و نیز برای دسته-بندی استفاده کردند. آزمایش‌ها نشان داده که روش ارائه شدخ از بیز ساده و مشابه SVM عمل کرده است.

چن[2] و همکاران [۸۷]، برای دسته‌بندی متون، از مجمعی از دسته‌بندها استفاده کردند. این مجمع شامل سه دسته‌بند SVM، KKN و Centroid است. در این روش داده‌ها به دو کلاس تقسیم می‌شوند. داده‌هایی که به آسانی کلاسشان مشخص می‌شود .داده‌هایی که تعیین کلاسشان سخت‌تر است. برای گروه اول یک دسته‌بند تصمیم می‌گیرد و برچسب آن را مشخص می‌کند اما برای گروه دوم از رای اکثریت استفاده می‌شود.

یوان[3] و همکاران [۸۸]، برای دسته‌بندی متن از ترکیب SVM و KNN استفاده کردند. به این صورت که ابتدا با SVM مرزها را مشخص کرده و سپس در داخل مرزها با استفاده از KNN دسته‌بندی را انجام می‌دهند.

لیو[4] و همکاران [۸۹]، برای دسته بندی متون از ترکیب درخت تصمیم دودویی با SVM بهره بردند. همچنین برای کاهش فضای حالت و بالابردن دقت و سرعت از خوشه‌بندی استفاده کرده‌اند.

چن[5] و همکاران [۹۰]، با ارائه‌ی دو معیار ارزش‌گذاری ویژگی بنام‌های MOR[6] و CDM[7] به استخراج و کاهش ویژگی‌ها پرداخته و از دسته بند بیزین ساده برای دسته‌بندی استفاده کردند. آنها نشان دادند که این معیارهای انتخاب ویژگی، از سایر روش‌ها بهتر عمل می‌کند.

[1] Zhou
[2] Qingxuan Chen
[3] Pingpeng Yuan
[4] Lizhen Liu
[5] Jingnian Chen
[6] Multi-class Odds Ratio
[7] Class Discriminating Measure

وانگ[1] و همکاران [۹۱]، با استفاده از مفاهیم و ارتباط کلمات و مترادفات و روابط انجمنی در ویکی‌پدیا به توسعه‌ی ویژگی‌ها پرداخته و نشان داده که این روش در مقایسه با سایر روش‌های بازنمایی متن قوی‌تر است.

سان[2] و همکاران [۹۲]، یک مطالعه‌ی مقایسه‌ای بر روی روش‌هایی که با داده‌های نامتوازن در کلاس‌ها کار می‌کنند مثل resampling، instance weighting و ... ، انجام دادند و با استفاده از دسته‌بند SVM به بررسی ده روش موجود در زمینه‌ی دسته‌بندی نامتوازن[3] متون پرداختند.

اسریرام[4] و همکاران در [۹۳]، بدلیل کوتاه بودن متون در سرویس‌هایی همچون توئیتر، با استفاده از یک مجموعه از کلمات از پیش تعریف شده در دامنه‌های خاص، به دسته‌بندی متون کوتاه در توئیتر پرداخته، تا مشکل خام بودن متون مورد استفاده کاربران را در این سرویس‌ها از بین ببرد.

جیانگ[5] و همکاران [۹۴]، از گراف‌ها در دسته‌بندی متون استفاده کرده‌اند. آنها با استفاده از یک الگوریتم کاوش گراف وزن‌دار[6]، به جستجوی زیرگراف‌ها به عنوان ویژگی‌های متون پرداخته‌اند. این روش با استفاده از چند دسته‌بند مختلف نتایج خوبی را نشان داده است و نشان داده شده است که در مقابل افزایش داده‌ها و بزرگ شدن گراف‌ها نیز خوب عمل می‌کند.

شینگ هوا [7] و همکاران [۹۵]، با ترکیبی از دو دسته‌بند بیزین و دسته‌بندهای انجمنی ضمن رفع نقصان هریک از آنها به صورت منفرد، دقت بالاتری را در دسته‌بندی ارائه می‌دهند. آنها متون آموزشی را با دسته‌بند بیزین دسته‌بندی کرده و حد آستانه‌های اطمینان را برای کلاس‌های متفاوت با استفاده از داده-های دسته‌بندی شده با بیزین، برای انتخاب قوانین انجمنی تنظیم کرده‌اند. سپس نتایج دسته‌بندی را با

[1] Pu Wang
[2] Sun
[3] Imbalanced Text Classification
[4] Sriram
[5] Jiang
[6] Weighted Graph Mining Algorithm
[7] Shing Hwa

استفاده از این قوانین انجمنی کلاس‌ها بهینه کرده‌اند. نتایج نشان داده است که ترکیب دو دسته‌بند، بهتر از هریک از آنها نتیجه می‌دهد.

لی[1] و همکاران [۹۶]، نشان می‌دهند که داه‌های آموزشی منفی اثر بسیار بدی در دسته‌بندی دارند. ابتدا داده‌های آموزشی مثبت و منفی با فرض طور توزیع یکنواخت، به دو دسته مثبت و منفی دسته‌بندی شده‌اند. در نهایت نشان داده شده است که داده‌های مثبت توزیع یکنواخت دارند، اما این یکنواختی توزیع در داده‌های منفی وجود ندارد. همچنین نشان می‌دهند که اثر منفی و مخرب داه‌دهای منفی، در تنظیم مقادیر و حدود آستانه در معیارهای مختلف است.

خان[2] و همکاران در [۲]، یک مطالعه‌ی مروری بر روش‌های موجود در دسته بندی متون و استخراج ویژگی‌ها انجام داده‌اند. در پایان نشان داده‌اند که پرکاربردترین روش‌های استخراج ویژگی استفاده از Chi Square و InfoGain هستند، گرچه روش‌های دیگر استخراج ویژگی که نتایج بهتری را نیز نشان داده‌اند معرفی شده و نیاز به توسعه‌ی بیشتر دارند. همچنین پرکاربردترین دسته‌بندها، KNN، NB و SVM معرفی شده‌اند که روش‌های ترکیبی معمولا نتایج بهتری را نشان داده‌اند.

ژانگ[3] و همکاران [۹۷]، نیز یک مطالعه‌ی مروری در بازنمایی متن بر روی روش‌های TFIDF، LSI و multi-word داشته‌اند. نتایج نشان داده است که LSI در دسته‌بندی متون انگلیسی و چینی با دقت بیشتری عمل می‌کند. همچنین در بازیابی متون انگلیسی نیز بهتر عمل کرده است.

جیانگ[4] و همکاران [۹۸]، برای کاهش ویژگی‌ها، با استفاده از یک الگوریتم خودسازی بر مبنای شباهت فازی[1]، به خوشه‌بندی ویژگی‌ها پرداخته‌اند. کلمه‌های شبیه به هم در خوشه‌های یکسان قرار

[1] Li
[2] Khan
[3] Zhang
[4] Jiang

گرفته‌اند. هر خوشه با یک تابع عضویت و یک میانگین و انحراف مشخص شده است. در نهایت ویژگی استخراج شده متناظر با هر خوشه، ترکیب وزنی کلمات هر خوشه می‌باشد. نتایج نشان داده است که این الگوریتم سریعتر و با کارایی بیشتر نسبت به سایر روش‌ها عمل می‌کند.

هارون[۲] و همکاران [۹۹]، با استفاده از روش شاخص‌گذاری معنایی پنهان، یک روش دومرحله‌ی برای استخراج ویژگی‌ها ارائه دادند. در مرحله‌ی اول با ارائه‌ی یک روش کاهش ویژگی، به انتخاب ویژگی‌ها پرداخته، سپس یک فضای معنایی جدید بین کلمات با استفاده از روش شاخص‌گذاری معنایی پنهان ایجاد کرده است. این روش بر روی کاربردهایی که با فیلتر کردن اسپم‌ها سر و کار دارند، بهتر از سایر روش‌ها نتیجه داده است.

وان[۳] و همکاران [۱۰۰]، ترکیبی از دسته‌بند KNN و SVM را برای دسته‌بندی متن با عنوان SVM-NN ارائه می‌دهد. این دسته‌بند ترکیبی برای به حداقل رساندن تاثیر پارامتر K در دسته‌بند KNN ارائه شده است. در مرحله‌ی آموزش، SVM به کار گرفته شده است تا نمونه‌های آموزشی هر دسته به بردارهای پشتیبان هر کلاس کاهش پیدا کنند. تصمیم‌گیری دسته‌بندی بر اساس دسته‌ای است که بردارهای پشتیبان آن کمترین فاصله را به داده‌ی تست دارند.

کورسات[۴] و همکاران [۱۰۱]، برای کاهش ویژگی‌ها، از یک روش جدید انتخاب ویژگی احتمالاتی، بر اساس فیلتر کردن، بنام DFS [۵] برای دسته‌بندی متون ارائه می‌دهند. این روش با روش‌های معروفی همچون Information Gain و Chi Square وغیره مقایسه شده است.

---

[1] Fuzzy Similarity-Based Self-Constructing Algorithm
[2] Harun
[3] Wan
[4] Kursat
[5] Distinguishing Feature Selector

یانگ[1] و همکاران [۱۰۲]، الگوریتم CMFS را به عنوان یک روش نوین در انتخاب و کاهش ویژگی‌ها ارائه کردند. این الگوریتم به طور جامع میزان تاثیر یک کلمه را در داخل و خارج یک کلاس اندازه‌گیری می‌کند. این الگوریتم بر روی داده‌های چند مجموعه با استفاده از SVM و بیزین ساده آزمایش شده است. نتایج دسته‌بندی با شش روش معروف و شناخته شده مثل IG، DF،CHI و غیره مقایسه شده است و با استفاده از هر دو دسته‌بند، از همه‌ی روش‌ها نتیجه‌ی بهتری داشته است.

جیانگ[2] و همکاران در [۱۰۳]، برای افزایش کارایی الگوریتم دسته‌بندیKNN ، یک نسخه‌ی بهبود یافته از این الگوریتم را ارائه داده است. این الگوریتم ترکیبی از یک الگوریتم خوشه‌بندی یک مرحله‌ای محدود و دسته‌بند KNN است. در آزمایشات بر روی داده‌های مختلف، این دسته‌بند از سه دسته‌بند معروف KNN، SVM و NB نتایج بهتری داشته است.

صدیق[3] و همکار در [۱۰۴]، پس از پیش‌پردازش و انتخاب ویژگی‌های تفکیک‌کننده‌ی مناسب با استفاده از تئوری Rough Set به دسته‌بندی متون می‌پردازد. بنابراین این مدل از یک زوج مفاهیم دقیق بر اساس تئوری موجود، بنام تقریب‌های بالا و پایین[4] برای دسته‌بندی متون آزمایشی به یک دسته یا بیشتر و حتی زیر دسته‌ها، با دقت بسیار خوب استفاده می‌کند.

سلامت[5] و همکاران در [۱۰۵] ، روش WPCM [6] را برای دسته‌بندی متون خبری وب ارائه می‌دهد. در این کار از ویژگی‌های CPBF [7] و دسته‌بند شبکه‌ی عصبی برای دسته‌بندی استفاده می‌کند و برای کاهش ویژگی‌ها نیز از روشPCA استفاده می‌کند. آزمایشات انجام شده بر روی متون خبری ورزشی وب، با دقت خوبی گزارش شده است.

[1] Yang
[2] Jiang
[3] Sadiq
[4] Lower and Upper Approximations
[5] Selamat
[6] Web Page Classification Method
[7] Class Profile-Based Features

بیلسوس[1] و همکاران در [۱۰۶]، یک عامل هوشمند طراحی کرده است که اخبار روزانه را برای کاربران گردآوری می‌کند. این سیستم بر مبنای بازخورد کاربر، اخبار مورد علاقه‌ی کاربر را جمع‌آوری می‌کند. تمرکز این کار بروی اجزای مدل کردن کاربر سیستم است. این سیستم علاقه‌های کوتاه مدت و بلند مدت کاربر را نیز در نظر می‌گیرد. برای این کار از KNN برای دسته‌بندی علاقه‌های کوتاه مدت و از Naïve Bayesian برای دسته‌بندی علاقه‌مندی بلند مدت کاربر استفاده می‌کند. متد وزن‌دهی بکار رفته نیز TFIDF برای ویژگی‌های استخراجی است. نتایج ارائه شده در این کار رضایت‌بخش نبوده است.

ژائو لیو[2] و همکار در [۱۰۷]، توانسته است با بکارگیری ترکیبی از کلمات منفرد و کلمات همرخداد و نیز ترکیب دسته‌بند SOM و KNN به نتایج بسیار خوبی در دسته‌بندی متون دست پیدا کند. نویسنده ادعا کرده است که این روش با کاهش بعد بالا راه حل مناسبی برای مجموعه داده‌های بزرگ است.

پکار[3] و همکاران در [۱۰۸]، با استفاده از ویژگی‌های همرخداد به دسته‌بندی کلمات بر اساس معنای آن-ها پرداختند. در این تحقیق از دسته‌بندهای بیزین و KNN چندین روش وزن‌دهی مثل Information Gain، Odds Ratio, Gain Ratio وMutual Information باهم مقایسه شده‌اند. نشان داده شده است که ویژگی‌های کلاسی بسیار موثر تر از ویژگی‌های منفرد عمل می‌کنند.

۲-۴-۲-مقالات مبتنی بر زبان فارسی

تحقیقات انجام شده در دسته‌بندی متون فارسی به گستردگی سایر زبان‌ها نیست.

در سال ۱۳۸۶ خانم عربی و همکاران [۱۰۹]، در مقاله خود به استخراج کلمات کلیدی و استفاده از آن برای دسته‌بندی متون فارسی پرداختند. آن‌ها برای این کار از ترکیب روش‌های Wordnet و Porter

[1] Billsus
[2] Xiao Luo
[3] Pekar

استفاده کرده‌اند. روش آن‌ها برای استخراج کلمات کلیدی استفاده از ساختار درهم ریزی Trie است و دادگان مورد استفاده در این مقاله، دادگان همشهری بوده است.

در مقاله دیگری که در همین سال و توسط همین گروه صورت گرفت[۱۱۰]، از الگوریتم کاوشگر مورچه[1] برای استخراج قوانین دسته‌بندی از میان داده‌ها استفاده شده است.

کار دیگری که در همین سال انجام شده است مربوط به احسان بصیری و همکاران [۱۱۱] است. آن‌ها در مقاله‌ای که ارائه کردند، دو الگوریتم KNN و FKNN را با دو روش استخراج ویژگی IG و DF آزمودند، آنها نشان دادند که بهترین ترکیب استفاده از روش FKNN و IG است. با این ترکیب آنها به میانگین دقت ۸۰٪ دست پیدا کرده‌اند.

در سال ۱۳۸۷، مسلم محمدی و همکاران در [۱۱۲]، از شبکه عصبی CC4 بهره جسته‌اند. CC4، یک نوع از الگوریتم‌های دسته‌بند کناره‌ای برای شبکه عصبی سه لایه رو به جلو می‌باشد. این شبکه برای دسته‌بندی اسناد حجیم فارسی مناسب است. هممچنین اگر مقیاس بین فایل‌های متنی نزدیک به هم باشد، این شبکه نتایج بهتری ارائه می‌دهد. در این روش در مرحله پیش‌پردازش کلمات عمومی حذف و سایر کلمات ریشه‌یاب می‌شوند. سپس ویژگی‌های آماری برای کلمات مختلف محاسبه شده، با استفاده از یک روش مبتنی بر فراوانی کلمه، اسناد با اندازه‌های مختلف به یک فضای K بعدی با اندازه‌ی ثابت نگاشت می‌شوند. دقت این دسته‌بند روی دادگان ایسنا حدود ۹۰٪ گزارش شده است.

در همین سال، ایوب باقری و همکاران در [۱۱۳]، با استفاده از بیز ساده و کاهش فضای حالت با روش TFV[2] دسته‌بندی قابل قبولی ارائه دادند.

سید مهدی نقیبی و همکاران [۱۱۴]، نیز در همین سال برای دسته‌بندی متون از شبکه عصبی استفاده کرده‌اند. به این صورت که به تعداد کلاس‌های از پیش تعیین شده شبکه MLP باینری وجود

[1] Ant_miming
[2] Term Frequency Variance

دارد. متن ورودی به هر یک از این شبکه‌ها وارد می‌شود. خروجی این شبکه‌ها به یک MLP ترکیب کننده‌ی نتایج رفته و به این ترتیب برچسب متن ورودی مشخص می‌شود. آنها هم‌چنین با پیشنهاد روش بسامد سند موضوعی ادعا کرده‌اند که تا ۹۹٪ منجر به کاهش فضای حالت شده‌اند.

همچنین در مقاله دیگری که بینا و همکاران [۱۱۵]، ارائه دادند، روش‌های tri-gram و quad-gram با معیارهای اندازه‌گیری فاصله منهتن[۱] ، اندازه‌ی دایس[۲] و ضرب نقطه‌ای، به همراه روش یادگیری KNN مورد بررسی قرار گرفته و بهترین نتیجه از ترکیب quad-gram و ضرب نقطه‌ای به دست آمد.

مقصودی و همکار [۱۱۶]، برای افزایش دقت دسته‌بندی از گنجواژه‌ی معنایی بهره گرفتند. آنها با استفاده ازمعیارهای ICF [۳] و Uni [۴] ، ویژگی‌ها را انتخاب کرده و با گسترش آنها با استفاده از گنجواژه و وزن‌دهی TFIDF و نیز کاربرد دسته‌بند SVM به دسته‌بندی متون پرداختند. آنها آزمایشات خود را در تعداد کلاس‌های مختلف و با تعداد متون مختلف انجام دادند و به نتایج پایداری در این آزمایشات دست یافتند.

سرخی و همکار [۱۱۷]، از روش وزن دهی ویژگی مبتنی بر اطلاعات کلاس درحوزه طبقه بندی مستندات TFCRF برای انتخاب ویژگیها استفاده کردند. برای استخراج ویژگیهای متون فارسی نیز ازروش تجزیه و تحلیل مولفه های اصلی PCA استفاده کردند. نتایج به دست آمده نشان داده است که با درنظر گرفتن روشهای دسته‌بندی KNN و Bayesian درروش پیشنهادی بهبود قابل توجهی در طبقه بندی متون فارسی و کاهش مدت زمان تست با ویژگیهای استخراج شده به دست خواهد آمد.

حسن پور و همکاران در [۱۱۸]، با استفاده از روش PCA و الگوریتم ژنتیک، به استخراج و کاهش ویژگی‌ها پرداخته‌اند. آنها روش وزن‌دهی TFCRF را با استفاده از فاکتورهای ارتباط مثبت و منفی، به کار

[1] Manhattan Distance
[2] Dice Measure
[3] Inverted Conformity Frequency
[4] Uniformity

برده‌اند. فاکتور ارتباط مثبت، نسبت تعداد متون یک دسته که حاوی ویژگی مورد نظر هستند را به کل متون آن دسته نشان می‌دهد. فاکتور ارتباط منفی نیز ، نسبت تعداد متون سایر دسته‌ها که حاوی ویژگی مورد نظر هستند را به کل متون سایر دسته‌ها نشان می‌دهد. آنها با دو دسته‌بند KNN و بیزین به دسته-بندی متون پرداختند ونسبت به کارهای قبلی نتایج بهتری گرفتند.

طاهری‌نیا [۱۱۹] ، به بررسی یکی از روش های یادگیری هوشمند به نام یادگیری نیمه نظارت شده در دسته بندی متون فارسی پرداخت. بسیاری از روش های یادگیری هوشمندانه مانند یادگیری نظارت شده، فقط بر روی داده های آموزشی برچسب دار تکیه می کنند، در شرایطی که بدست آوردن این داده های آموزشی دارای برچسب بسیار پرهزینه است. حال آنکه حجم زیادی از داده های بدون برچسب به سرعت زیاد و با هزینه ی کم در دسترس هستند. در مقابل روش هایی مانند روش یادگیری بدون نظارت فقط بر روی داده های بدون برچسب تکیه می کنند. وی به بررسی روش یادگیری نیمه نظارت شده که ما بین روش های یادگیری نظارت شده و یادگیری بدون نظارت قرار دارد و از ترکیبی از مثال های آموزشی برچسب دار و بدون برچسب برای یادگیری استفاده می کند پرداخته و از این تکنیک برای دسته بندی متون فارسی استفاده کرده ونتایج نسبتا خوبی گرفت.

زمانی و همکاران [۱۲۰]، از روش آنالیز معنایی پنهان احتمالاتی و از مبانی احتمالات و قوانین بیزین ، برای عملیات دسته بندی متون فارسی پایگاه داده همشهری و با استفاده از روش فاصله یابی اقلیدسی در فضای ماتریس های کاهش بعد یافته استفاده کرده است. همچنین روش TFIDF را جهت انتخاب و وزن-دهی کلمات کلیدی موجود در مجموعه متون مورد استفاده، به کار گرفته است. در مرحله هرس، علاوه بر کلمات دستوری مانند حروف اضافه و ربط، کلمات کلیدی زاید نیز به صورت دستی حذف شده‌اند. نتایج حاصل شده در این مقاله نشان می‌دهد که روش بکار رفته در بهبود عملکرد سیستم نقش موثری دارد.

برفامی و همکار [۱۲۱]، ازترکیب دوشبکه عصبی چندلایه پرسپترون MLP دردسته بندی مستندات نیمه ساختیافته XML برروی پایگاه داده روزنامه همشهری استفاده کرده اند. روش TFCRF نیز برای وزن دهی به مستندات انتخاب شده سپس به کمک روش تجزیه و تحلیل مولفه های اصلی PCA به ارزش دهی ویژگیها پرداخته و با الگوریتم KNN تعدادی ازآنها انتخاب و به عنوان ورودی یکی ازشبکه های عصبی استفاده شده‌اند. درمرحله بعد ازویژگیهای ارزش دهی شده به کمک الگوریتم GA تعدادی ویژگی انتخاب شده و به عنوان ورودی یکی دیگر ازشبکه های عصبی استفاده میشود سپس نتایج خروجی این دوشبکه به کمک جمع جبری بایکدیگر ترکیب شده در نهایت مورد تست و ارزیابی قرارگرفته است.

عابدی‌نیا و همکاران در [۱۲۲]، دسته‌بندی لایه ای اطلاعات به عنوان روش اصلی مطرح کرده است و نیز ازمدل فضای برداری برای دسته‌بندی متون فارسی در وب استفاده کرده است. سپس روش جدیدی برای کاوش درمتون فارسی براساس تحلیل معنایی کلمات فارسی مطرح می‌کند که شامل سه بخش استخراج کلمات تحلیل ترکیبات جملات و تحلیل های معنایی جملات است روش پیشنهادی این امکان را به کاربران فارسی زبان می‌دهد که بادقت و سرعت بیشتری متن موردنظر خود را دروب بیابند.

صفیان وهمکاران در [۱۲۳]، با بیان این مسئله که روش سنتی KNN پیچیدگی محاسباتی بالایی دارد و این نقیصه از کاربردی بودن KNN می کاهد، خوشه بندی را یک راه حل برای این مشکل می‌داند. به این روش که تعداد نمونه های یادگیری در دسته بندی را کاهش داده و به این طریق از پیچیدگی زمانی محاسبات می کاهد. در این مقاله، یکی از روشهای ساده مبتنی برای خوشه بندی الگوریتم KNN که قبلا پیشنهاد شده است، انتخاب شده است. این روش بوسیله خوشه بندی پیچیدگی زمانی را تا حد زیادی کاهش می دهد ولی مقداری از کارایی دسته بندی می کاهد. لذا یک مقدار وزنی به مرکز خوشه ها نسبت داده شده و همچنین مقدار K در الگوریتم KNN کاهش داده شده است . آزمایش‌ها برروی مجموعه داده Reuter-21578 نشان داده که این روش می تواند کارایی دسته بندی را افزایش دهد.

حسن‌پور و همکاران [۵۱]، از مشخصه معنایی و توجه به بافت معنایی متون، استفاده کرده‌اند. در رویکرد پیشنهادی، ابتدا مفاهیم توسط هستان شناسی لغوی فارس نت استخراج شده و پس از اعمال فرایند رفع ابهام، با توجه به بافت متن، اوزان مفهوم مرتبط در حین فرایند وزن دهی به وزن عبارات کلیدی اضافه می شود. این مقاله، رویکردی جدید در حین استفاده از روش اولین مفهوم به منظور رفع ابهام از مفاهیم استخراج شده، معرفی می کند که کارایی این روش را بهبود می دهد. معیار $\chi 2$ نیز برای انتخاب ویژگی و روش وزن دهی ویژگی نرمال شده TFIDF در وزن دهی به کار گرفته شده است. نتایج نشان داده است که اعمال دانش معنایی حاصل از هستان شناسی لغوی در فرایند طبقه بندی متون یک گام موثر در بهبود کارایی دسته بند است.

۵-۲ نتیجه‌گیری

در این فصل ضمن تعریف مفاهیم پایه در دسته‌بندی متون، مراحل مختلف سیستم دسته‌بندی متون معرفی شد. انواع روش‌های استخراج و کاهش ویژگی و نیز روش‌های وزن‌دهی به ویژگی‌ها که در تحقیقات مختلف به کار رفته‌اند مطرح گردید. متداول‌ترین روش‌های دسته‌بندی متون نیز معرفی بیان شد. در پایان نیز مروری بر مقالات فارسی و سایر زبان‌ها در حوزه‌ی دسته‌بندی متون انجام شد.

فصل سوم

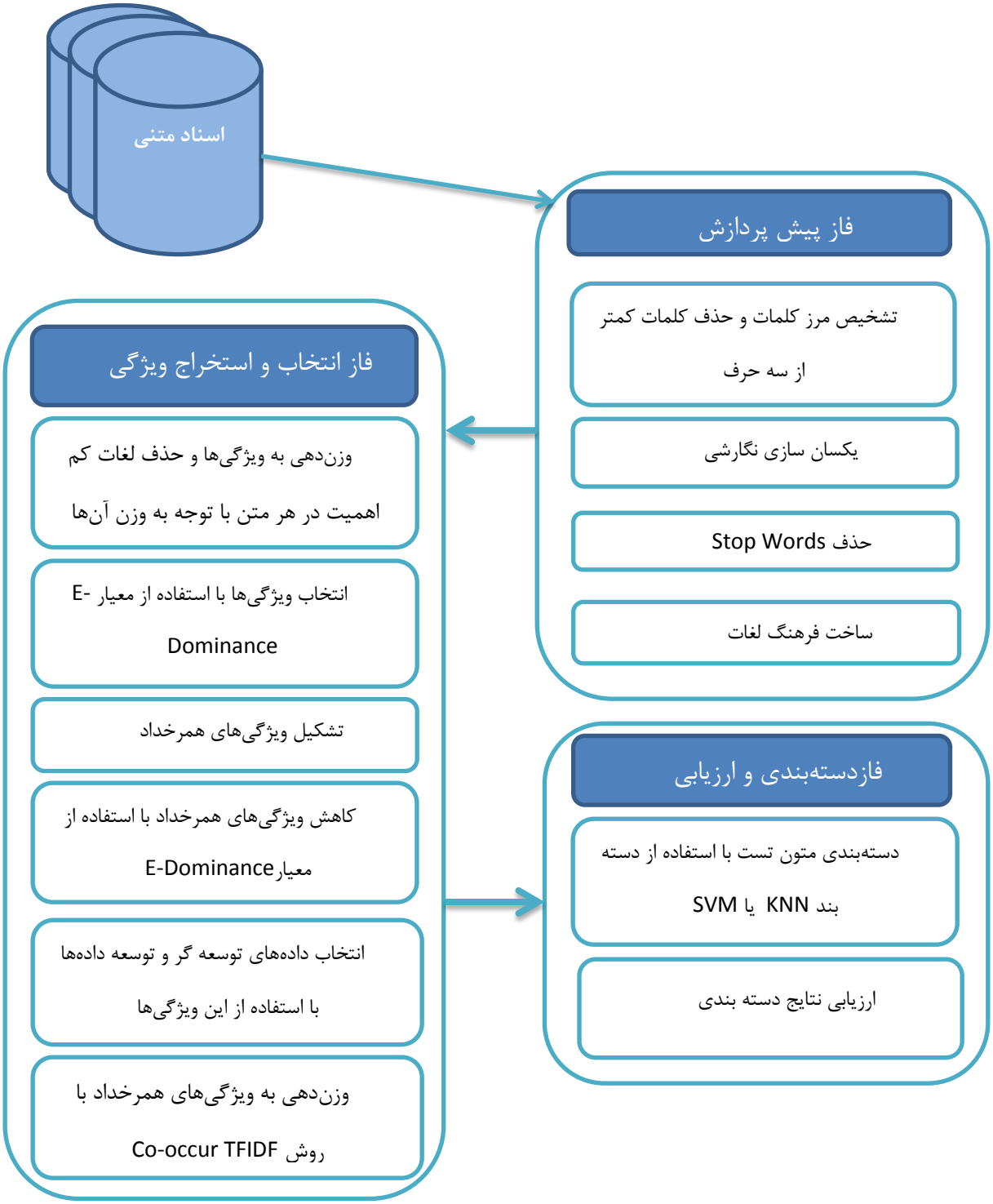
معرفی روش پیشنهادی

۳-۱- مقدمه

در فصل قبل، معماری کلی سیستم دسته‌بندی متون و مراحل آن، تشریح شد و به طور ضمنی با بررسی کارهای گذشته به محدودیت‌ها و نکات ضعف روش‌های موجود، اشاره گردید. در این پایان‌نامه، با ارائه یک معیار اهمیت ویژگی و نیز ارائه یک طرح توسعه یافته از طرح‌های موجود در وزن‌دهی برای ویژگی‌های همرخداد، سعی شده است با مشکلات روش‌های قبلی، یعنی تعداد ویژگی‌های زیاد و فقدان یک روش‌وزن‌دهی کارآمد برای ویژگی‌های همرخداد مقابله شود. نتایج حاصل از اعمال این روش‌ها، نزدیک شدن به هدف مورد نظر را نشان داده است. در این فصل، سیستم دسته‌بند متون فارسی ارائه و تشریح می‌شود، سپس در فصل بعدی نتایج حاصل از ارزیابی آزمایشات انجام شده ارائه می‌گردد. از این‌رو در این فصل به طراحی چارچوب و پیاده‌سازی این سیستم پرداخته می‌شود.

۳-۲- معماری رویکرد پیشنهادی

در این بخش ابتدا یک بلوک دیاگرام از معماری پیشنهادی برای سیستم دسته‌بندی متون ارائه داده، سپس به طور مفصل مراحل مختلف آن تشریح می‌شود. شکل ۳-۱، معماری پیشنهادی را تاحدودی با جزئیات لازم نشان می‌دهد، اما جزئیات بیشتر در قسمت‌های بعدی به صورت مفصل‌تر بیان شده است. معماری پیشنهادی شامل چهار فاز اصلی و هر فاز شامل بخش‌های مختلفی هستند. در اکثر روش‌های دسته‌بندی متون، این مراحل به صورت کلی مشترک بوده و آنچه باعث متفاوت و مجزا بودن کارهای مختلف می‌گردد، جزئیات کار و شیوه‌ی متفاوت در هر مرحله است. بخش‌های مختلف معماری پیشنهادی در ذیل آمده است.



شکل(۳-۱) نمودار کلی معماری پیشنهادی

۳-۲-۱- اجزای تشکیل دهنده‌ی معماری روش پیشنهادی

- فاز انتخاب متون از پایگاه داده
- فاز پیش پردازش

i. تشخیص مرز کلمات و حذف کلمات کمتر از سه حرف

ii. یکسان سازی نگارشی

iii. حذف Stop Words

iv. ساخت فرهنگ لغات

- فاز انتخاب و استخراج ویژگی

i. وزن‌دهی به ویژگی‌ها و حذف لغات کم اهمیت در هر متن با توجه به وزن آنها

ii. انتخاب ویژگی‌ها با استفاده از معیار پیشنهادی E-Dominance

iii. تشکیل ویژگی‌های همرخداد

iv. کاهش ویژگی‌های همرخداد با استفاده از معیار پیشنهادی E-Dominance

v. انتخاب داده‌های توسعه‌گر و توسعه متون با استفاده از این ویژگی‌ها

vi. وزن‌دهی به ویژگی‌های همرخداد با استفاده از روش Co-occur TFIDF

- فاز دسته‌بندی و ارزیابی

i. دسته‌بندی متون تست با استفاده از دسته‌بندهای KNN یا SVM

ii. ارزیابی نتایج دسته بندی

حال پس از ارائه‌ی معماری روش پیشنهادی ، توضیح مفصل‌تری از مراحل مختلف آن بیان می‌گردد.

۳-۲-۲- انتخاب متون از پایگاه داده

پایگاه داده همشهری [۱۲۴]، به دلیل مطابقت با استانداردهای TREC [۱۲۵]، در سال‌های اخیر پرکاربردترین پایگاه داده در حوزه‌ی تحقیقات زبان فارسی بوده است. این پایگاه داده توسط گروه تحقیقاتی دانشگاه تهران ارائه شده است. نسخه‌ی یک این مجموعه نمونه‌ای است که در همایش‌های CLEF در سال‌های ۲۰۰۸ و ۲۰۰۹ برای ارزیابی سامانه‌های بازیابی اطلاعات تک‌منظوره[1] مورد استفاده قرار گرفته‌است. نسخه‌ی دو این پیکره، آخرین نسخه‌ی مجموعه بوده و نسبت به نسخه‌ی یک بزرگتر و جامع‌تر است.

در این تحقیق برای انجام آزمایشات، از پایگاه داده‌ی همشهری۲ استفاده شده است. این نسخه شامل ۳۱۸ هزار سند در قالب فایل‌های XML می‌باشد. داده‌های انتخابی در این پایان‌نامه مطابق جدول ۳-۱ است.

جدول (۳-۱) داده‌های انتخابی از پایگاه داده همشهری ۲

پایگاه داده	دسته‌ها	متون آموزش	متون تست
همشهری ۲	۵ دسته (ورزشی، اقتصادی،سیاسی، علمی ، فرهنگ و هنر)	۱۰۰ متن از هر دسته	۵۰ متن از هر دسته

شکل ۳-۲ یک نمونه از متون این پایگاه داده را که به دسته‌ی "اقتصادی" تعلق دارد، نشان می‌دهد.

[1] Ad Hoc

۳-۲-۳- پیش‌پردازش‌ها

همانطور که گفتیم متون انتخابی باید بعد از پیش‌پردازش‌هایی وارد مرحله انتخاب ویژگی شوند. پیش-پردازش‌های انجام شده بر روی متون ، جهت آماده‌سازی داده‌ها برای دسته‌بندی به صورت ذیل است:

۳-۲-۳-۱-تشخیص مرز کلمات و حذف کلمات کمتر از سه حرف

در این مرحله، با استفاده از کاراکترهای "."،"؟"،"? "،":"،"!"، فاصله، نیم‌فاصله و ... مرز شروع و پایان یک کلمه را شناسایی کرده و با شمارش کاراکترهای موجود در این دو مرز، کلمات کمتر از سه حرف حذف می‌شوند. شکل ۳-۳ ، نمونه‌ای از یک متن انتخابی را پس از اولین مرحله از پیش‌پردازش نشان می‌دهد.

تازه ترین بررسی های آماری نشان می دهد، در روزهای اول اردیبهشت ماه، ارزش سهام بانک ها و لیزینگ ها تحت تأثیر خبر احتمال تثبیت نرخ سود تسهیلات با صعودی جالب توجه روبه رو شده است. در روزهای اخیر ارزش سهام ۵ بانک و لیزینگ فعال در بورس تهران بین ۱/۱ تا ۷/۱ درصد رشد کرد، در این مدت بالاترین سود ناشی از رشد قیمت نصیب سهامداران لیزینگ ایران شده است. ارزش سهام این شرکت لیزینگ در پنج روز اول اردیبهشت ۷/۱ درصد افزایش یافته است. در این میان سودآورترین سهام در بین بانک ها، سهام بانک کارآفرین بوده، قیمت سهام این بانک در روزهای گذشته ۴ درصد رشد کرده است. براین اساس رایان سایپا با رشد قیمتی معادل ۳/۹ درصد، بانک پارسیان با صعودی برابر ۱/۹ درصد و بانک اقتصاد نوین با افزایش قیمتی برابر ۱/۱ درصد ارکستر صعود سهام بانک ها را تکمیل کرده اند.

شکل (۳-۲) نمونه‌ای از یک متن در گروه اقتصادی از پایگاه داده همشهری

تازه ترین بررسی های آماری نشان دهد روزهای اول اردیبهشت ماه ارزش سهام بانک لیزینگ تحت تأثیر خبر احتمال تثبیت نرخ سود تسهیلات صعودی جالب توجه روبه شده است روزهای اخیر ارزش سهام بانک لیزینگ فعال بورس تهران بین درصد رشد کرد این مدت بالاترین سود ناشی رشد قیمت نصیب سهامداران لیزینگ ایران شده است ارزش سهام این شرکت لیزینگ پنج روز اول اردیبهشت درصد افزایش یافته است این میان سودآورترین سهام بین بانک سهام بانک کارآفرین بوده قیمت سهام این بانک روزهای گذشته درصد رشد کرده است این اساس رایان سایپا رشد قیمتی معادل درصد بانک پارسیان صعودی برابر درصد بانک اقتصاد نوین افزایش قیمتی برابر درصد ارکستر صعود سهام بانک تکمیل کرده

شکل (۳-۳) نمونه‌ای از متن با اولین پیش‌پردازش

۳-۲-۳-۲- یکسان سازی نگارشی

در این مرحله حروف مشابه را دسته بندی کرده و با استفاده از کد اسکی آنها، یک حرف را به جای حروف مشابه در تمامی کلمات جایگزین کرده تا یک کلمه با دو حرف مشابه اما غیر یکسان نداشته باشیم؛ مثل «ؤ» به «و»،«ئ» به «ی» و«أ» به «ا» و تبدیل نویسه‌های «ی» و «ک» عربی به نوع فارسی آن. همچنین برخی کلمات که در برخی متون با املای فاصله دار و در برخی متون با املای بدون فاصله حضور داشتند، همگی به شکل بدون فاصله درآمده‌اند.

۳-۲-۳-۳- حذف Stop Words

پس از تشخیص یک کلمه، باید بررسی کنیم که آیا کلمه‌ی مورد نظر جزو کلمات بی اهمیت یا اصطلاحا Stop Words هست یا نه؟ این مرحله کمک شایانی به تشکیل فرهنگ لغت اولیه می‌کند. کلمات ایست نه تنها کمکی به دسته بندی نکرده و نتایج دسته‌بندی را تضعیف می‌کنند، بلکه به سبب بزرگ شدن فرهنگ لغت یا همان بردار ویژگی اولیه، حجم پردازش‌ها و نیز جستجوها را افزایش داده و باعث افزایش قابل توجه زمان فرآیند دسته‌بندی می‌شوند. برای زبان فارسی، هیچ منبع جامعی از Stop Words ارائه نشده است. در کارهای انجام شده در زبان فارسی فقط اشاره شده است که کلمات ایست از فرهنگ لغت تشکیل شده به عنوان بردار ویژگی،حذف شده‌اند و تنها نمونه‌هایی از آنها ارائه شده است. ما در این پایان نامه، با استفاده از معیار DF مجموعه‌ای ۱۰۸۰ کلمه‌ای از کلمات ایست ارائه داده‌ایم که عدم وجود این کلمات در متون، نه تنها تاثیر منفی در نتایج دسته‌بندی ندارد، بلکه زمان دسته‌بندی را نیز به طور محسوسی کاهش می‌دهد. نمونه‌ای از کلمات ایست استخراجی از متون در جدول ۳-۲ آمده است. شکل ۳-۴ نیز متن انتخابی را پس از حذف کلمات ایست نشان می‌دهد. کلماتی که پس از حذف کلمات ایست از لیست لغات حذف می‌شوند به عنوان فرهنگ لغات و یا بردار ویژگی استفاده می‌شوند.

جدول (۳-۲) نمونه‌هایی از کلمات ایست استخراجی

اگرچه	گرچه	زیرا	بود	کمی
این	اند	یک	دیگر	هر
ما	آن	آنها	و	وقتی
نیز	برای	یا	را	بلکه
باید	وی	اما	ممکن	اندک

تازه بررسی آماری نشان روزهای اردیبهشت ماه ارزش سهام بانک لیزینگ تأثیر خبر احتمال تثبیت نرخ سود تسهیلات صعودی جالب توجه روزهای ارزش سهام بانک لیزینگ فعال بورس تهران درصد رشد مدت بالاترین سود ناشی رشد قیمت نصیب سهامداران لیزینگ ایران ارزش سهام شرکت لیزینگ پنج روز اردیبهشت درصد افزایش یافته سودآورترین سهام بانک سهام بانک کارآفرین قیمت سهام بانک روزهای گذشته درصد رشد اساس رایان سایپا رشد قیمتی معادل درصد بانک پارسیان صعودی برابر درصد بانک اقتصاد نوین افزایش قیمتی برابر درصد ارکستر صعود سهام بانک تکمیل

شکل (۳-۴) نمونه‌ای از متن با Stop Word های حذف شده

۳-۲-۴- وزن‌دهی به ویژگی‌ها و حذف لغات کم اهمیت با توجه به وزن آنها

پس از خواندن کلمات متون و حذف کلمات ایست و کلمات کمتر از سه حرف ، حدود ۱۶۸۹۳ کلمه منحصر بفرد در بردار اولیه ویژگی‌ها قرار گرفتند. این تعداد کلمه به عنوان ویژگی، در مراحل بعدی پردازش و دسته‌بندی، بسیار حجیم و زمانبر هستند. بنابراین در مرحله اول از کاهش ویژگی‌ها، با روش TFIDF ویژگی‌ها را وزن‌دهی کرده و با استفاده از سعی و خطا و در نظر گرفتن مقدار آستانه ۵ برای این معیار، در هر متن، لغاتی که دارای وزن کمتری از آستانه‌ی موردنظر هستند، از لیست ویژگی‌های هر کلاس حذف می‌شوند. نهایتا بردار ویژگی اولیه به بردار ویژگی با ۱۶۴۳۰ کلمه کاهش پیدا کرده است.

۳-۲-۵-انتخاب ویژگی‌های کلاسی با معیار E-Dominance

در مرجع [۷] معیار Dominance معرفی شده است. این معیار، میزان اهمیت یک کلمه در هر کلاس را نسبت به کل داده‌ها نشان می‌دهد ؛ به طوری‌که اگر یک کلمه فقط در متون یک کلاس اتفاق بیفتد، آنگاه مقدار Dominance برای آن کلمه در آن کلاس یک خواهد بود. در رابطه‌ی ۳-۱ ، $df(f_i, c_j)$، تعداد متون کلاس c_j است که ویژگی f_i در آنها رخ داده است. یکی از ویژگی‌های Dominance این است که مجموع آن برای یک عبارت f_i در کلاس‌های متفاوت یک است (رابطه‌ی ۳-۲). در رابطه‌ی ۳-۲، k تعداد کلاس‌هاست.

$$\text{Dominance}(f_i, c_j) = \frac{df(f_i, c_j)}{\sum_{j=1}^{k} df(f_i, c_j)} \qquad ۱-۳$$

$$\sum_{j=1}^{k} \text{Dominance}\,(fi, cj) = 1 \qquad ۲-۳$$

مرجع مذکور، از این معیار برای کاهش ویژگی‌ها استفاده کرده است؛ یعنی انتخاب یک حد آستانه برای معیار فوق، کلماتی که مقادیری کمتر از آن آستانه دارند، از لیست ویژگی‌ها حذف می‌شوند. بر اساس این معیار، اگر یک کلمه ، فقط و فقط در یک متن خبری حضور داشته باشد و در هیچ متن دیگری از متون پایگاه داده رخ نداده باشد، مقدار Dominance برای این کلمه، در کلاس شامل متن "۱" خواهد بود و به عنوان یک ویژگی با اهمیت از لیست ویژگی‌ها حذف نخواهد شد. فرض کنیم این کلمه فقط در دو یا سه متن از همان کلاس رخ داده باشد، باز هم شرایط مشابهی داریم و مقدار Dominance برای این کلمه در آن کلاس"۱" و در سایر کلاس‌ها صفر خواهد بود. سوالی که در اینجا مطرح است این است که آیا کلمه‌ای که تنها در یک متن از متون آموزشی رخ داده است، می‌تواند به عنوان یک ویژگی مهم از آن متن

یا آن کلاس قلمداد گردد؟ آیا این ویژگی خاصیت تفکیک کنندگی خوبی دارد؟ یا احتمال اینکه این کلمه در متون آزمایشی رخ داده و این ویژگی باعث شباهت متن مذکور به متن آزمایشی جدید شود چقدر خواهد بود؟ با بررسی فرهنگ لغت تولید شده از متون آموزشی و اعمال ویژگی Dominance برای کاهش تعداد کلمه‌ها در این فرهنگ لغت یا همان بردار ویژگی‌ها، مشاهده می‌شود بسیاری از کلماتی که در برخی کلاس‌ها، مقدار Dominance مساوی با "۱" و یا نزدیک به "۱" و بیشتر از آستانه‌ی مورد نظر دارند، نه تنها نقش مثبتی در دسته‌بندی ایفا نکرده، بلکه امکان کاهش دقت نیز وجود دارد؛ چرا که احتمال رخداد این کلمات در متون مشابه و هم‌دسته از متون آزمایشی، بسیار کم بوده و در محاسبه میزان شباهت آنها مثلا در دسته‌بند KNN که معمولا از فاصله‌ها ی اقلیدسی و کسینوسی استفاده می-کند، تاثیر منفی می‌گذارد. علاوه بر تاثیر منفی در محاسبه‌ی میزان شباهت و کاهش کارایی، ضعف دیگر این معیار در کاهش ویژگی‌ها این است که پس از اعمال این معیار، باز هم کلمات بی‌اهمیتی در بردار ویژگی حضور داشته و حجم و زمان پردازش‌ها را تحت‌الشعاع قرار می‌دهند. با وجود این ضعف‌ها، مزیتی که این ویژگی دارد این است که نشان می‌دهد اگر این ویژگی در یک کلاس مقداری نزدیک به "۱" داشته باشد، می‌توان تا حدود زیادی مطمئن بود، که این ویژگی متعلق به این کلاس است و در سایر کلاس‌ها کمتر اتفاق می‌افتد در نتیجه این ویژگی قدرت تفکیک‌کنندگی خوبی خواهد داشت.

۳-۲-۵ -۱- بهبود معیار Dominance

در بخش قبل گفتیم یکی از نقاط ضعف معیار Dominance، انتخاب برخی از ویژگی‌هایی است که به دلیل رخداد کم در متون یک کلاس، با وجود داشتن قدرت تفکیک‌کنندگی بالا، فقط اندازه بردار ویژگی را افزایش می‌دهند، چرا که این ویژگی‌ها به ندرت اتفاق افتاده و کاربرد موثری در دسته‌بندی ندارند. اما حالت ایده‌آل برای یک ویژگی قدرتمند و تفکیک کننده از یک کلاس این است که اولا آن ویژگی فقط در متون همان کلاس رخ دهد و ثانیا اینکه در همه و یا تعداد موثری از متون آن کلاس رخ دهد، نه تعداد

متون اندک. یعنی اگر یک ویژگی در متون زیادی از یک کلاس رخ دهد، می‌توان با اطمینان گفت آن ویژگی متعلق به این کلاس است. در حالتی که در یک متن و یا تعداد متون اندک رخ دهد، آن ویژگی، ویژگی کلاس نبوده و ویژگی آن متن محسوب می‌گردد. با توجه به مطالب مذکور چگونه یک ویژگی قدرتمند از یک کلاس پیدا کنیم که در آن کلاس جامعیت داشته باشد؟ با توجه به رابطه ۳-۱، مقدار Dominance برای ویژگی f_i در کلاس c_j برابر است با تعداد متونی از کلاس c_j که ویژگی در آنها رخ داده باشد به تعداد متونی از همه‌ی کلاس‌ها (k تعداد کلاس‌ها) که ویژگی f_i در آنها رخ داده باشد. برای مثال اگر یک کلمه فقط در یک متن از یک کلاس رخ داده و در هیچ متن دیگری رخ ندهد ، مقدار صورت و مخرج کسر در رابطه‌ی ۳-۱ برابر بوده و مقدار Dominance برابر با یک خواهد بود. اما همانطور که گفته شد احتمال حضور این ویژگی تفکیک کننده در متون تست نیز خیلی کم خواهد بود.

در این پایان‌نامه از ویژگی مثبت معیار Dominance ، که غالب بودن یک کلاس بر سایر کلاس‌ها در ویژگی است، استفاده شده است. از طرف دیگر سعی شده است نقطه‌ی ضعف این معیار -که برای حضور کم تعداد یک ویژگی در متون یک کلاس، مقدار بالا تخصیص می‌دهد- بر طرف شود. این کار را با اضافه کردن یک ضریب به معیار Dominance انجام شده است.

$$\alpha_{ij} = \frac{df(f_i, c_j)}{d(c_j)} \qquad ۳-۳$$

در رابطه ۳-۳، صورت کسر نشان‌دهنده‌ی تعداد اسناد کلاس c_j است که ویژگی f_i در آن‌ها رخ داده است، و $d(c_j)$ ، تعداد کل اسناد کلاس c_jاست. در نظر بگیرید، اگر $d(c_j) = df(f_i, c_j)$ ،ضریب α_{ij} مساوی با یک خواهد شد زیرا تعداد متون کلاس c_j که ویژگی f_i در آنها رخ دهد، برابر با تعداد کل اسناد کلاس c_j می‌باشد. حالا اگر ضریب α_{ij} را که از این به بعد ضریب جامعیت Dominance [1] می نامیم در

[1] Dominance Comprehensiveness Factor

عبارت اصلی Dominance ضرب کنیم رابطه (۳-۴)، مقدار E-Dominance [1] را خواهیم داشت (رابطه‌ی ۳-۵). عبارت کلی یا E-Dominance، زمانی برابر با ۱ خواهد بود که هم Dominance مقدار یک داشته باشد و هم ضریب α.

$$E-\text{Dominance}(f_i,c_j)=\frac{df(f_i,c_j)}{d(c_j)}*\frac{df(f_i,c_j)}{\sum_{j=1}^{k}df(f_i,c_j)} \qquad ۴-۳$$

$$E-\text{Dominance}(f_i,c_j)=\alpha_{ij}*\text{Dominance}(f_i,c_j) \qquad ۵-۳$$

حال اگر یک برای یک ویژگی مقدار E-Dominance ، در یک کلاس خاص ، یک باشد، معنایش این است که اولا این کلمه فقط در همان کلاس رخ داده است و در کلاس‌های دیگر حتی یک بار هم در هیچ متنی ظاهر نشده است، ثانیا این کلمه در تمامی متون همان کلاس حضور داشته است. با این تفسیر از این معیار، می‌توان مطمئن شد که این ویژگی برای این کلاس یک ویژگی قدرتمند محسوب می‌شود. در فصل بعدی که آزمایشات و نتایج آن‌ها بیان می‌شود، این ادعا با ارزیابی نتایج نشان داده خواهد شد. همان‌طور که گفته شد این معیار زمانی‌که برای یک کلاس مقدار یک داشته باشد، یک ویژگی منحصر بفرد از یک کلاس خواهد بود، اما مقدار یک، یک حالت ایده آل است که خیلی کم اتفاق می‌افتد. برای بدست آوردن مقدار مناسب برای این معیار از روش سعی و خطا استفاده شده و یک حد آستانه برای این ویژگی تعیین شده است. اگر در کلاسی مقدار این ویژگی بیشتر از آن آستانه‌ی انتخابی بود، آن ویژگی به عنوان یکی از ویژگی‌ها که در بردار نهایی ویژگی قرار می‌گیرد، انتخاب خواهد شد. در این تحقیق در مرحله انتخاب ویژگی از معیار E-Dominance برای انتخاب ویژگی‌های مناسب برای مراحل بعدی استفاده شده است. حد آستانه برای این معیار با استفاده از آزمون و خطا، ۰،۰۴ در نظر گرفته شد. چند

[1] Extended Dominance

نمونه از ویژگی‌های استخراجی برای هر کلاس با استفاده از معیار E-Dominance در جدول ۳-۳ آمده است.

جدول (۳-۳) نمونه‌هایی از ویژگی‌های استخراج شده با معیار E-Dominance برای کلاس‌های مختلف

اقتصادی	سیاسی	ورزشی	فرهنگ و هنر	علمی
بورس	نمایندگان	رقابت	سینما	آموزش
رقابت	انتخابات	امتیاز	هنرمندان	کتاب
سهام	خارجی	ورزش	آموزش	یادگیری

۳-۲-۶-تشکیل زوج ویژگی‌های همرخداد به عنوان ویژگی‌های متن

در مرجع [۷]، پس از حذف ویژگی‌های بی‌اهمیت و انتخاب واژه‌های مهم با استفاده از معیار Dominace، واژه‌های همرخداد به عنوان ویژگی‌های نهایی داده‌ها، تشکیل شده و در دسته‌بندی مورد استفاده قرار گرفته‌اند. بردار ویژگی حاصل از تشکیل ویژگی‌های همرخداد بسیار بزرگ و قابل ملاحظه است. این حجم بالا از ویژگی‌ها، در این حقیق به عنوان چالشی بزرگ مورد مطالعه قرار گرفته و سعی شده است، با بهبود ویژگی Dominance -که در بخش ۳-۲-۵-۱ توضیح داده شد – و انتخاب ویژگی‌های مناسب و کاراتر جهت تشکیل ویژگی‌های همرخداد، کاهش چشمگیری در اندازه‌ی این بردار ویژگی ایجاد شود. پس از انتخاب ویژگی‌های مناسب و تفکیک کننده با استفاده از معیار پیشنهادی، به بررسی همرخدادی این ویژگی‌ها در متون مورد مطالعه پرداخته شده است. این ویژگی‌های همرخداد، با تشکیل زوج کلمات، که هر کلمه یکی از کلمات انتخابی در مرحله‌ی قبل است، ایجاد می‌شوند. در بررسی اولیه تعداد این ویژگی‌های همرخداد بسیار زیاد خواهد بود. اگر N، تعداد ویژگی‌های منفرد باشد، $\frac{N(N-1)}{2}$ ویژگی همرخداد تشکیل می‌شود. این تعداد ویژگی برای کل متون، بسیار زیاد خواهد بود. اما در زمان تشکیل این ویژگی‌ها، دیده می‌شود که بسیاری از این زوج ویژگی‌ها در هیچ‌یک از متون رخ نداده

و حذف خواهند شد. اما با این وجود، باز هم تعداد این ویژگی‌ها زیاد بوده و در مراحل بعدی باید فرآیند کاهش ویژگی دنبال شود. در جدول ۳-۴، نمونه‌هایی از ویژگی‌های همرخداد تشکیل شده آمده است.

جدول (۳-۴) نمونه‌هایی از ویژگی‌های همرخداد تشکیل شده

بورس-سهام	نمایندگان-قانون	رقابت-امتیاز	سینما-فیلم	آموزشی-علمی
سرمایه-سود	سیاست-خارجی	فوتبال-گل	هنرمند-موسیقی	الکترونیکی-اطلاعات
رقابت-سود	وزیر- وزارتخانه	ورزشگاه-تماشاگر	آموزشی-فرهنگی	اینترنت-سایت
وزیر-اینترنت	سایت-فوتبال	فرهنگی-وزیر	خارجی-اطلاعات	هنرمند-برگزاری
آموزشی-بورس	برگزاری-رقابت	گل-رقابت	سود-سینما	سیاست-سرمایه

۳-۲-۷-کاهش ویژگی‌های همرخداد با استفاده از معیار پیشنهادی E-Dominance

پس از تشکیل ویژگی‌های همرخداد و توسعه‌ی آنها با استفاده از ویژگی‌های همرخداد توسعه‌گر در مرحله‌ی بعدی، کاهش ویژگی‌ها انجام می‌شود. برای این کار، مثل مرحله‌ی اول کاهش ویژگی‌ها، که برای پیدا کردن عبارات کاندید برای ساخت ویژگی‌های همرخداد انجام شد، در این مرحله نیز از معیار E-Dominance استفاده شده و تعداد ویژگی‌ها کاهش داده می‌شود. در این مرحله حد آستانه برای معیار E-Dominance، مقدار ۰,۰۳۵ با استفاده از آزمون و خطا به دست آمده است.

۳-۲-۸- انتخاب ویژگی‌های توسعه‌گر[1] و توسعه داده‌ها

یکی از نقاط قوت مرجع [۷] توسعه و غنی‌سازی داده‌های یک کلاس با ویژگی‌های مهم آن کلاس است. غنی سازی داده‌ها به این صورت انجام شده است که در فرآیند آموزش پس از یافتن بهترین واژه-های همرخداد به عنوان ویژگی‌های اصلی، در این میان، ویژگی‌های همرخدادی که در تعداد متون بیشتری از یک حد آستانه در یک کلاس خاص تکرار شده‌اند و یا بر اساس یک معیار ویژه، مهم تلقی می-شوند، به عنوان داده‌ی توسعه‌گر یا همان داده‌ای که برای غنی‌سازی داده‌ها به کار خواهد رفت، انتخاب

[1] Developer Features

می‌شوند. به طور مثال اگر دو کلمه‌ی "بورس" و "سرمایه" به عنوان ویژگی همرخداد توسعه‌گر در نظر گرفته شوند و جزء ویژگی‌های با اهمیت کلاس "اقتصادی " محسوب شوند، در فرآیند توسعه وغنی سازی داده‌ها، این زوج ویژگی یعنی " بورس و سرمایه " را به تمام داده‌های کلاس اقتصادی در داده‌های آموزشی اضافه می‌کنیم. با این کار این ویژگی در این کلاس نسبت به سایر کلاس‌ها بارزتر و در میان داده‌های این کلاس عمومی تر و پرتکرار تر خواهد شد. در سایر کلاس‌ها هم در شرایطی این ویژگی به داده اضافه خواهد شد، که یکی از ویژگی‌های این زوج همرخداد، در داده‌ی مورد نظر رخ داده باشد، در این صورت این ویژگی تنها به همین داده از کلاس غیر اقتصادی اضافه می‌گردد. در فرآیند غنی سازی داده‌ها در داده‌های آزمایشی، بدلیل عدم آگاهی از کلاس داده‌ی مورد نظر، از شیوه‌ی دوم استفاده کرده و در صورت وجود هر تک واژه از زوج واژه‌های همرخداد، این ویژگی به داده‌ی مورد نظر اضافه می‌شود. نمونه‌ایی از ویژگی‌های توسعه‌گر در جدول ۳-۵ آمده است. در این پایان‌نامه نیز از تکنیک توسعه متون استفاده شده است.

جدول (۳-۵) نمونه‌ای از ویژگی‌های توسعه‌گر انتخابی از هر کلاس

اقتصادی	سیاسی	ورزشی	فرهنگ و هنر	علمی
بورس- سهام	نمایندگان-قانون	رقابت-امتیاز	سینما- فیلم	آموزشی-علمی

۳-۲-۹-وزن‌دهی ویژگی‌های همرخداد با روش وزن‌دهی Co-Occur TFIDF

علاوه بر کاهش حجم بردار ویژگی‌ها، هدف دیگری که در بهبود کاربرد واژه‌های همرخداد دنبال شده است، ارائه‌ی یک **طرح وزن‌دهی برای ویژگی‌های همرخداد** است. در کارهای مختلفی از ویژگی‌های همرخداد به عنوان ویژگی‌های تفکیک کننده و اصلی برای دسته‌بندی استفاده شده است، اما هیچ گاه یک طرح وزن‌دهی برای این ویژگی‌ها معرفی نشده است. در مرجع [۷] نیز از وزن‌دهی باینری یعنی "صفر" برای عدم حضور ویژگی و "یک" برای حضور ویژگی استفاده شده است. این یک شیوه‌ی پایه برای وزن-

دهی است. اما همانطور که می‌دانیم، دلیل اصلی در ارائه‌ی طرح‌های وزن‌دهی که برای ویژگی‌های ساده معرفی شد، این است که میزان تکرار یا فرکانس حضور یک ویژگی خیلی با اهمیت تر از صرفا حضور یا عدم حضور یک ویژگی قلمداد می‌شود، یعنی داده‌ی با حضور بیشتر باید سهم بیشتری در دسته‌بندی ایفا کند. لذا در ویژگی‌های همرخداد نیز پا را از اهمیت حضور یا عدم حضور فراتر گذاشته و به ارائه‌ی یک طرح وزن‌دهی برای دخالت دادن میزان حضور ویژگی‌های همرخداد در دسته‌بندی می‌پردازیم. در کارهای سابق که ویژگی‌های مورد استفاده برای دسته‌بندی کلمات یا عبارات متن انتخاب شده‌اند، یکی از رایج-ترین شیوه‌های استاندارد وزن‌دهی بنام TFIDF جهت وزن‌دهی و تعیین میزان اهمیت کلمات در هر متن به کار گرفته شده است. در این معیار (رابطه ۳-۶)، آنچه وزن یک کلمه را به عنوان ویژگی آن متن تعیین می‌کند ، اولا تعداد رخداد کلمه یا ویژگی f_k در متن d_i ، یعنی مقدار $tf(f_k, d_i)$ و ثانیا تعداد متونی که ویژگی f_k در آنها حضور دارد یعنی $|D(f_k)|$ ، می‌باشد (رابطه۳-۷). در این رابطه $|D|$، نیز تعداد کل متون است.

$$TFIDF(f_k, d_i) = tf(f_k, d_i) * idf(f_k, d_i) \qquad ۶-۳$$

$$idf(f_k, d_i) = \log\left(\frac{|D|}{|D(f_k)|}\right) \qquad ۷-۳$$

از این معیار استفاده کرده و آن را برای ویژگی‌های همرخداد بسط می‌دهیم. اگر ویژگی $f_{k1,k2}$ شامل دو مولفه یعنی f_{k1} و f_{k2} باشد - که این دو مولفه ، مولفه‌های ویژگی همرخداد هستند- و $|D(f_{k1}, f_{k2})|$ تعداد متونی باشند که پس از عملیات غنی سازی داده‌ها، شامل هر دو ویژگی f_{k1} و f_{k2} باشند آنگاه وزن ویژگی همرخداد $f_{k1,k2}$ با استفاده از رابطه‌ی ۳-۸ محاسبه می‌شود:

$$\text{Co} - \text{Occur TF}IDF(f_{k1,k2}, d_i) = Max(tf(f_{k1}, d_i), tf(f_{k2}, d_i)) * idf(f_{k1,k2}, d_i) \qquad ۸-۳$$

$$idf(f_{k1,k2}, d_i) = \log\left(\frac{|D|}{|D(f_{k1}\ , f_{k2}\ )|}\right) \qquad ۹-۳$$

$Max(tf(f_{k1}, d_i), tf(f_{k2}, d_i))$: مقدار ماکسیمم برای تعداد حضور مولفه اول و دوم ویژگی است. دلیل انتخاب مقدار ماکسیمم این است که در مرحله‌ی توسعه و غنی سازی داده‌ها، به ازای وجود هر مولفه از ویژگی، مولفه‌ی بعدی نیز اضافه شده است. به دلیل همرخداد بودن این ویژگی‌ها، منطقی است که باید به ازای هر بار حضور مولفه‌ی پرتکرارتر، مولفه‌ی دیگر نیز باید حضور داشته باشد، بنابراین این مولفه اضافه می‌شود. تعداد اضافه شدن مولفه دوم، به تعداد اختلاف تعداد مولفه‌های اول و دوم است تا جائی‌که تعداد آنها برابر شود. بنابراین در نهایت تعداد رخداد هر مولفه، به تعداد مولفه‌ی بیشترخواهد بود.

$|D(f_{k1}\ , f_{k2}\ )|$: تعداد متونی است که شامل هر دو مولفه‌ی ویژگی مورد نظر، یعنی f_{k1} و f_{k2} هستند.

۳-۲-۱۰-دسته‌بندی و ارزیابی نتایج

در بخش ۲-۳-۴ به انواع دسته‌بندها در حوزه‌ی دسته‌بندی متون اشاره شد. در این پایان‌نامه از دسته بند SVM و KNN برای دسته‌بندی استفاده شده است. مقدار K در دسته‌بند KNN، با سعی و خطا، ۷ در نظر گرفته شده و فاصله‌ی کسینوسی نیز به عنوان معیار شباهت در این دسته‌بند به کارگرفته شده است. پس از دسته‌بندی داده‌ها، آنچه که می‌تواند درستی و کارایی روش پیشنهادی را مورد تایید قرار دهد، یک ارزیابی درست از نتایج به دست آمده است. در حوزه‌ی بازیابی و دسته‌بندی متون، معیارهای دقت، صحت، فراخوانی و F_1 متداول‌ترین معیارهایی هستند که به کار برده می‌شوند. این معیارها در فصل بعدی به طور کامل توضیح داده می‌شوند.

۳-۳-نتیجه‌گیری

در این فصل، یک معماری کلی از رویکرد پیشنهادی ارائه شد. سپس به توضیح کاملی از هر گام از این رویکرد پرداخته شد. معیار E-Dominace به عنوان معیاری برای انتخاب و کاهش ویژگی‌ها معرفی گردید. پس از تشکیل ویژگی‌های همرخداد و کاهش آنها با معیار مذکور، نحوه انتخاب و بکارگیری ویژگی‌های توسعه‌گر نیز بیان شد. سپس با توسعه روش وزن‌دهی TFIDF پایه برای ویژگی‌های همرخداد، روش وزن-دهی Co-occur TFIDF نیز معرفی شد. نتایج حاصل از اعمال این الگوریتم‌ها بر روی داده‌ها، در فصل بعدی ارائه می‌شود.

فصل چهارم

نتایج تجربی،

تحلیل و ارزیابی

۴-۱- مقدمه

در فصل سوم، سیستم دسته‌بندی متون پیشنهادی معرفی شد. در این فصل ابتدا به معرفی معیارهای ارزیابی مورد استفاده در این پایان‌نامه پرداخته می‌شود. سپس نتایج حاصل از اعمال روش پیشنهادی بر روی داده‌ها، ارائه می‌گردد. روش پیشنهادی در ۵ حالت مختلف آزمایش شده است. در پایان نیز یک ارزیابی کامل از نتایج و مقایسه آن با روش‌های موجود ارائه شده و یک نتیجه‌گیری کلی از روش پیشنهادی ارائه می‌شود.

۴-۲- معرفی معیارهای ارزیابی

معیارهای ارزیابی برای نشان دادن کارایی الگوریتم و نیز مقایسه روش پیشنهادی با سایر روش‌ها، به کار می‌روند. معیارهای Accuracy، Precision، Recall وF-measure پرکاربردترین معیارها در دسته‌بندی و به ویژه در زمینه‌ی دسته‌بندی متن هستند [۱۲۶]. برای معرفی معیارهای ارزیابی، نیاز به معرفی مفاهیم و پارامترهایی داریم که در این معیارها به کار گرفته شده‌اند. این مفاهیم در جدول ۴-۱ آمده‌اند:

جدول (۴-۱) پارامترهای مورد نیاز اندازه‌گیری معیارهای کارایی

	نسبت داده نشده به دسته C_i	نسبت داده شده به دسته C_i
متعلق به دسته C_i	FN_i	TP_i
متعلق به دسته‌ای غیر از C_i	TN_i	FP_i

پس از تعریف مفاهیم اصلی، معیارهای ارزیابی الگوریتم به شرح ذیل می‌باشد :

Accuracy: نسبت نمونه‌های درست دسته‌بندی شده را به کل نمونه‌ها نشان می‌دهد (رابطه ۴-۱):

$$Accuracy: \quad A_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \quad \text{(دقت)} \qquad ۱-۴$$

Precision: این معیار، نسبت تعداد متون درست نسبت داده شده را به کل متون نسبت داده شده به همان دسته نشان می‌دهد (رابطه ۴-۲). مقدار معیار صحت، درجه پایداری[1] الگوریتم دسته‌بند را نشان می‌دهند.

$$Precision: \quad P_i = \frac{TP_i}{TP_i + FP_i} \quad (\text{صحت}) \qquad ۴\text{-}۲$$

Recall: معیار بازخوانی نسبت تعداد متونی را که به درستی در یک دسته قرارگرفته‌اند به تعداد کل متونی که بایستی در آن دسته قرار می‌گرفتند، نشان می‌دهد (رابطه ۴-۳). این معیار، درجه‌ی تمامیت[2] الگوریتم دسته‌بند را نمایش می‌دهد.

$$Recall: \quad R_i = \frac{TP_i}{TP_i + FN_i} \ (\text{بازخوانی}) \qquad ۴\text{-}۳$$

$F_{\beta\ Measure}$: معیاری است که از ترکیب دو معیار $Precision$ و $Recall$ بدست می‌آید، چرا که معیارهای صحت و فراخوانی به تنهایی برای اندازه‌گیری کارایی الگوریتم دسته‌بندی مؤثر نبوده و استفاده-ی مجزای آنها ممکن است نتایج نادرستی از سیستم را ارائه دهد. از این‌رو لازم است که این معیارها به گونه‌ای با یکدیگر ترکیب شوند. این معیار میزان تاثیر معیار میزان صحت و معیار میزان فراخوانی را در ارزیابی دسته‌بند در نظر می‌گیرد و از رابطه‌ی ۴-۴ بدست می‌آید.

$$F_{\beta\ Measure} = \frac{(\beta^2 + 1) * Recall * Precision}{(\beta^2)(Recall + Precision)} \qquad ۴\text{-}۴$$

[1] Soundness
[2] Completness

معمولا مقدار β در حوزه بازیابی و دسته‌بندی متون ۱ در نظر گرفته می‌شود، مقدار ۱ برای β به معنی تاثیری برابر از $Precision$ و $Recall$ خواهدبود. در رابطه‌ی ۴-۵ معیار F_1 تعریف شده‌است.

$$F_{1_{Measure}} = \frac{2 * Recall * Precision}{(Recall + Precision)} \qquad ۵-۴$$

همانطور که گفته شد، بر اساس پارامترهای تعریف شده در فرمول این معیارها، معیارهای معرفی‌شده، به طور مجزا برای هر دسته یا کلاس محاسبه می‌شوند. برای ارزیابی نتایج رویکرد پیشنهادی به صورت کلی باید از مقادیر میانگین هر معیار در تمام دسته‌ها استفاده شود. فرمول میانگین این معیارها در روابط۴-۶ تا ۴-۱۰ آمده است.

$$Mean_{F1} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{F_1(i)}}{\mathrm{m}} \qquad ۶-۴$$

$$Mean_{Precision} = \frac{\sum_{i=1}^{m} Precision(i)}{m} \qquad ۷-۴$$

$$Mean_{Precision} = \frac{\sum_{i=1}^{m} Precision(i)}{m} \qquad ۸-۴$$

$$Mean_{Recall} = \frac{\sum_{i=1}^{m} Recall(i)}{m} \qquad ۹-۴$$

$$Mean_{Accuracy} = \frac{\sum_{i=1}^{m} Accuracy(i)}{m} \qquad ۱۰-۴$$

معیارهای فوق، معیارهایی هستند که در بیشتر تحقیقات در حوزه متون به کار رفته‌اند، لذا معیار مناسبی برای مقایسه‌ی رویکرد پیشنهادی با سایر رویکردها نیز هستند. در این پایان‌نامه نیز معیارهای نامبرده برای ارزیابی کارایی روش پیشنهادی به کار رفته‌اند.

۴-۳- نتایج دسته‌بندی متون

۴-۳-۱- نتایج بکارگیری معیار E-Dominance در دسته‌بندی متون

در بخش ۴-۲، چهار معیار ارزیابی برای بررسی و ارزیابی نتایج حاصل از اعمال روش پیشنهادی بر روی داده‌های انتخابی از پایگاه داده ، معرفی شد. جدول ۴-۳ مقادیر این معیارها را به تفکیک کلاس‌ها و نیز به صورت میانگین، برای حالتی از دسته‌بندی نشان می‌دهد که معیار E_Dominance بر روی ویژگی-های اولیه که همان کلمات منفرد در متون هستند, اعمال شده است. با اعمال این معیار بسیاری از ویژگی‌ها فیلتر شده و با استفاده از ویژگی‌های باقیمانده از این مرحله، بردار ویژگی تشکیل داده شده است. تعداد ویژگی‌ها در این دسته‌بندی، ۱۰۴۱ عبارت انتخابی با معیار E_Dominance بوده و دسته‌بند KNN نیز در این دسته‌بندی استفاده شده است. جدول ۴-۲ تعداد ویژگی‌های انتخابی را به تفکیک هر دسته و نسبت به کل ویژگی‌ها نشان می‌دهد. تعداد مجموع ویژگی‌ها در تمام کلاس‌ها از تعداد ویژگی‌های کل، به دلیل همپوشانی کلاس‌ها بیشتر است.

نرخ تشخیص سیستم که برای مقایسه‌ی روش پیشنهادی با سایر روش‌ها استفاده می‌شود، مقدار میانگین برای معیار Recall و F1 است، که در جدول ۴-۳ مقدار ۰,۹۱۲ را دارا هستند.

جدول (۴-۲) تعداد ویژگی‌های انتخابی از هر دسته

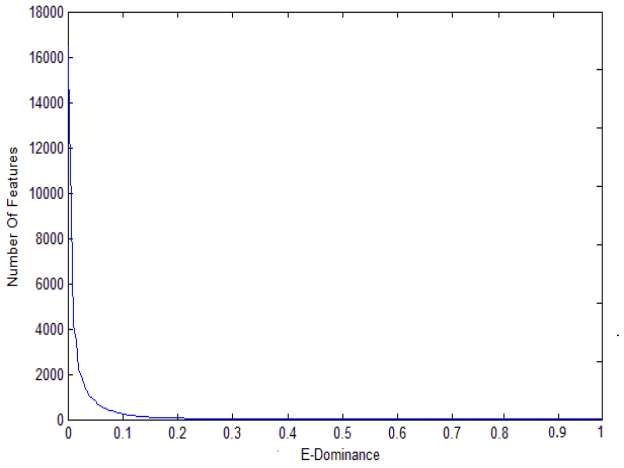
	اقتصادی	سیاسی	ورزشی	فرهنگ و هنر	علمی	کل
تعداد ویژگی‌ها	۲۸۰	۲۲۳	۲۱۳	۱۴۷	۲۱۶	۱۰۷۹
نسبت تعداد ویژگی انتخابی به کل ویژگی‌ها	٪۲۷	٪۲۱	٪۲۰	٪۱۴	٪۲۱	-

جدول (۴-۳) نتایج دسته‌بندی داده‌ها با معیارهای ارزیابی

دسته	*Recall*	*Precision*	F_1	*Accuracy*
اقتصادی	۰,۹۴	۰,۸۷۰	۰,۹۰۴	۰,۹۶
سیاسی	۰,۹۴	۰,۸۸۷	۰,۹۱۳	۰,۹۶۴
ورزشی	۰,۹۰	۰,۸۸۲	۰,۸۹۱	۰,۹۵۶
فرهنگ و هنر	۰,۹۲	۰,۹۳۹	۰,۹۲۹	۰,۹۷۲
علمی	۰,۸۶	۱	۰,۹۲۵	۰,۹۷۲
میانگین	۰,۹۱۲	۰,۹۱۶	۰,۹۱۲	۰,۹۶۵

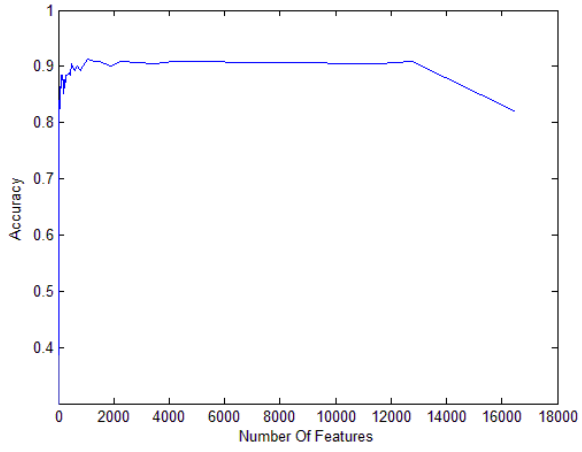
در روش پیشنهادی، ویژگی E-Dominance بعنوان یک معیار اهمیت ویژگی، برای انتخاب ویژگی‌های مناسب معرفی شد. با محاسبه‌ی مقدار این معیار برای ویژگی‌های اولیه و با استفاده از یک حد آستانه برای آن، ویژگی‌ها انتخاب شده‌اند. شکل ۴-۱، بیان کننده تاثیر مقدار حد آستانه این معیار در تعداد ویژگی‌های انتخاب شده است. هر چه مقدار یا حد آستانه انتخابی کمتر باشد، تعداد ویژگی‌های بیشتری انتخاب می‌شوند. مقدار ۰,۰۴ برای این معیار، تعداد ۱۰۴۱ ویژگی را با بیشترین دقت به دست آمده، انتخاب می‌کند.

همان‌طور که گفته شد، انتخاب مقدار آستانه برای E-Dominance، در تعداد ویژگی‌های انتخابی موثر است. مسلما تعداد ویژگی‌های انتخاب شده نیز بر روی دقت سیستم، تاثیرگذار هستند.

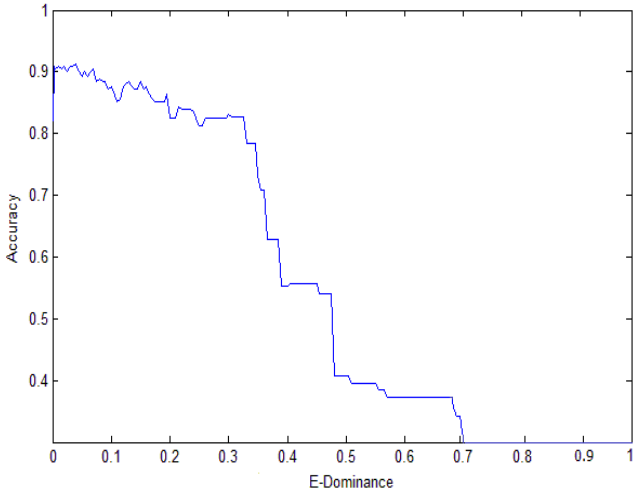


شکل (۴-۱) تاثیر مقدار معیار E-Dominance در تعداد ویژگی‌های انتخابی

شکل ۴-۲ تاثیر تعداد ویژگی‌های انتخابی را بر روی دقت دسته‌بندی نشان می‌دهد. با انتخاب ۱۰۴۱ ویژگی از ویژگی‌های موجود، دقت ۹۱,۲٪ بدست آمده است. این تعداد ویژگی و این دقت دسته‌بندی با انتخاب حدآستانه ۰,۰۴ برای E-Dominance بدست آمده است. بدون استفاده از آستانه و با کل ویژگی‌ها، دقت ۸۲٪ بدست آمد. در هر مرحله از انتخاب آستانه E-Dominance ، مقدار ۰,۰۰۵ به مقدار قبلی اضافه شده است. در آستانه ۰,۰۰۵ دقت سیستم از ۸۲٪ اولیه به ۸۹,۲٪ رسیده و تعداد ویژگی‌ها نیز به ۱۲۷۳۸ عدد رسید. شکل ۴-۳ نیز تاثیر مقدار آستانه‌ی انتخابی برای E-Dominance را در دقت دسته‌بندی نشان می‌دهد.



شکل (۴-۲) تاثیر تعداد ویژگی‌های انتخابی در دقت دسته‌بندی



شکل (۴-۳) تاثیر مقدار معیار E-Dominance در دقت دسته‌بندی

۴-۳-۱-۱-مقایسه معیار E-Dominance با سایر معیارهای اهمیت ویژگی

در جدول ۴-۴، معیار E-Dominance با چند معیار مختلف که پرکاربردترین معیارهای انتخاب ویژگی مورد استفاده در حوزه‌ی متن‌کاوی هستند و در بخش فصل دوم معرفی شده‌اند، مقایسه شده‌است. این مقایسه تاثیر هر یک از معیارها را بر تعداد ویژگی‌های انتخابی و دقت دسته‌بندی نشان می‌دهد. تعداد ویژگی‌های هر روش، با توجه با بهترین دقت دسته‌بندی آورده شده است.

همانطور که در جدول ۴-۴ می‌بینید، بهترین میزان دقت را معیار Dominance و سپس روش پیشنهادی یعنی معیار E-Dominance با اختلاف ناچیز نشان می‌دهند. البته اختلاف دقت دسته‌بندی در میان این روش‌ها قابل ملاحظه نیست. هدف اصلی نیز در اینجا مقایسه‌ی میزان دقت دسته‌بندی نیست. آنچه در این مقایسه مهم است، تعداد ویژگی‌های انتخابی در هر روش است. در مورد تعداد ویژگی‌های انتخابی ، معیار E-Dominance که معیار پیشنهادی است، اختلاف قابل ملاحظه‌ای با سایر معیارها دارد. این معیار توانسته است تعداد ویژگی‌های اولیه که ۱۶۴۳۰ ویژگی بوده‌اند را به کمتر از ۶٪ کاهش دهد و میزان کاهش ۹۳,۶۶٪ را داشته باشد. پس از آن نیز معیار DF، با میزان کاهش ۸۷,۰۱٪ و معیار Chi-

Square با ۸۱,۸۴٪، درکاهش ویژگی‌ها بهترین عملکرد را داشته‌اند. معیار Dominance نیز که بهترین دقت را داراست، ویژگی‌ها را به حدود ۶۰٪ ویژگی‌های اولیه رسانده است. دو معیار TFIDF و Info-Gain نیز در کاهش ویژگی‌ها ضعیف تر عمل کرده‌اند و نتایج نزدیک به هم دارند. معیار Entropy نیز بدترین عملکرد را در کاهش ویژگی‌ها داشته و توانسته است، تنها حدود ۱٪ از داده ها را کاهش دهد.

جدول (۴-۴) مقایسه میزان دقت دسته‌بندی در کاهش ویژگی‌ها با معیار E-dominance و سایر معیارها

معیار	تعداد ویژگی‌ها	Recall	کاهش در ویژگی‌ها
-	۱۶۴۳۰	۰,۸۲	بدون کاهش
DF	۲۱۳۳	۰,۹۰	٪۸۷,۰۱
TFIDF	۹۲۲۱	۰,۸۸۸	٪۴۳,۸۸
Entropy	۱۴۵۸۶	۰,۸۸۸	٪۰,۹۹
Info-Gain	۸۲۶۷	۰,۹۰	٪۴۹,۶۸
Chi-Square	۲۹۴۸	۰,۹۰	٪۸۱,۸۴
Dominance	۹۸۶۶	۰,۹۱۶	٪۳۹,۹۵
E-Dominance (معیار پیشنهادی)	۱۰۴۴	۰,۹۱۲	٪۹۳,۶۶

۴-۳-۲- نتایج بکارگیری ویژگی‌های همرخداد دردسته‌بندی متون

۴-۳-۲-۱- بکارگیری ویژگی‌های همرخداد با دو روش وزن‌دهی و دو دسته‌بند

در بخش ۴-۳-۱ ، نتایج دسته‌بندی با استفاده از ویژگی‌های ساده که همان کلمات منفرد حاصل از فیلتر شدن با معیار E-Dominance هستند، نشان داده شد. ویژگی‌های همرخداد به کار رفته در دسته-بندی، زوج کلماتی هستند که در مرحله‌ی قبل با استفاده از معیار E-Dominance کاندید شده‌اند. پس از

تشکیل عبارات همرخداد از ویژگی‌های کاندید شده و شکل‌گیری ۹۹۴۳۹ زوج عبارت، معیار E-Dominance برای ویژگی‌های همرخداد نیز محاسبه شده و با استفاده از حد آستانه‌ی ۰,۰۳۵ برای این معیار، ویژگی‌ها به ۲۱۵۰ ویژگی همرخداد کاهش داده شده، سپس داده‌ها وارد مرحله‌ی دسته‌بندی شده‌اند. جدول ۴-۵ تعداد ویژگی‌های انتخابی از هر دسته را نشان می‌دهد. جمع تعداد ویژگی‌های همه‌ی دسته‌ها به دلیل همپوشانی برخی کلاس‌ها و انتخاب برخی از ویژگی‌ها در بیش از یک کلاس ، از کل ویژگی‌های انتخابی بیشتراست. سهم دسته‌ی اقتصادی در انتخاب ویژگی‌های همرخداد، بیشتر از سایر دسته‌هاست.

جدول (۴-۵) تعداد ویژگی‌های همرخداد انتخابی از هر دسته

	اقتصادی	سیاسی	ورزشی	فرهنگ و هنر	علمی
تعداد ویژگی انتخابی	۵۰۸	۴۲۳	۴۷۰	۳۹۸	۴۲۰
نسبت تعداد ویژگی-های هر دسته به کل ویژگی‌ها	٪۲۴	٪۲۰	٪۲۱	٪۱۹	٪۲۰

جدول ۴-۶، نتایج دسته‌بندی متون را با استفاده از ویژگی‌های همرخداد، با وزن‌دهی باینری و با استفاده از دسته‌بند KNN نشان می‌دهد. جدول ۴-۷، نتایج دسته‌بندی متون را با استفاده از ویژگی‌های همرخداد، با وزن‌دهی باینری و با استفاده از دسته‌بند SVM نشان می‌دهد. با مقایسه‌ی نتایج در جداول ۴-۶ و ۴-۷، دسته‌بند SVM عملکرد بهتری از دسته‌بند KNN داشته است. میانگین Recall به دست آمده برای KNN، ۰,۹۴ است و میانگین Recall به دست آمده برای SVM، ۰,۹۵۲۰ است. از میان ۲۵۰ متن آزمایشی، دسته‌بند SVM، ۳ متن آزمایشی را بیشتر از دسته‌بند KNN، درست دسته‌بندی کرده است.

جدول (۴-۶) نتایج دسته‌بندی داده‌ها با ویژگی‌های همرخداد با وزن‌دهی باینری و دسته‌بند KNN

دسته	Recall	*Precision*	F_1	*Accuracy*
اقتصادی	۰,۹۸	۰,۹۲۵	۰,۹۵۱	۰,۹۸
سیاسی	۰,۹۴	۰,۹۲۲	۰,۹۳۵	۰,۹۷۲
ورزشی	۰,۹۴	۰,۹۴	۰,۹۴	۰,۹۷۶
فرهنگ و هنر	۰,۹۶	۰,۹۴۱۱	۰,۹۵۰	۰,۹۸
علمی	۰,۸۸	۰,۹۷۸	۰,۹۲۶	۰,۹۷۲
میانگین	۰,۹۴	۰,۹۴۱	۰,۹۴	۰,۹۷۶

جدول (۴-۷) نتایج دسته‌بندی داده‌ها با ویژگی‌های همرخداد با وزن‌دهی باینری و دسته‌بند SVM

دسته	Recall	*Precision*	F_1	*Accuracy*
اقتصادی	۱	۰,۹۰۹۰	۰,۹۵۲۳	۰,۹۸
سیاسی	۰,۹۶	۰,۹۶	۰,۹۶	۰,۹۸۴
ورزشی	۰,۹۴	۰,۹۵۹	۰,۹۴۹	۰,۹۸
فرهنگ و هنر	۰,۹۶	۰,۹۴۱	۰,۹۵۰	۰,۹۸
علمی	۰,۹۰	۱	۰,۹۴۷	۰,۹۸
میانگین	۰,۹۵۰	۰,۹۵۴	۰,۹۵۲	۰,۹۸۱

جداول ۴-۸ و ۴-۹ نتایج دسته‌بندی متون را با استفاده از ویژگی‌های همرخداد و روش وزن‌دهی Co-occur TFIDF، با دسته‌بندهای KNN و SVM نشان می‌دهد. در اینجا نیز دسته‌بند SVM بهتر از KNN عمل کرده است.

جدول (۴-۸) نتایج دسته‌بندی داده‌ها با ویژگی‌های همرخداد با وزن‌دهی Co-occur TFIDF و دسته‌بند KNN

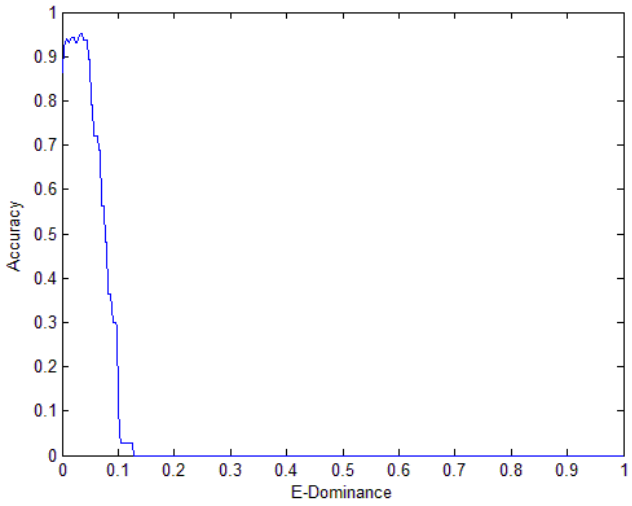
Accuracy	F_1	*Precision*	Recall	دسته
۰,۹۹۲	۰,۹۸۰	۰,۹۶۲	۱	اقتصادی
۰,۹۸۴	۰,۹۶۱	۰,۹۴۲	۰,۹۸۰	سیاسی
۰,۹۸۴	۰,۹۶۰	۰,۹۶	۰,۹۶۰	ورزشی
۰,۹۹۲	۰,۹۸۰	۰,۹۶۲	۱	فرهنگ و هنر
۰,۹۷۶	۰,۹۳۶	۱	۰,۸۸	علمی
۰,۹۸۶	۰,۹۶۳	۰,۹۶۵	۰,۹۶,۴	

جدول (۴-۹) نتایج دسته‌بندی داده‌ها با ویژگی‌های همرخداد با وزن‌دهی Co-occur TFIDF و دسته‌بند SVM

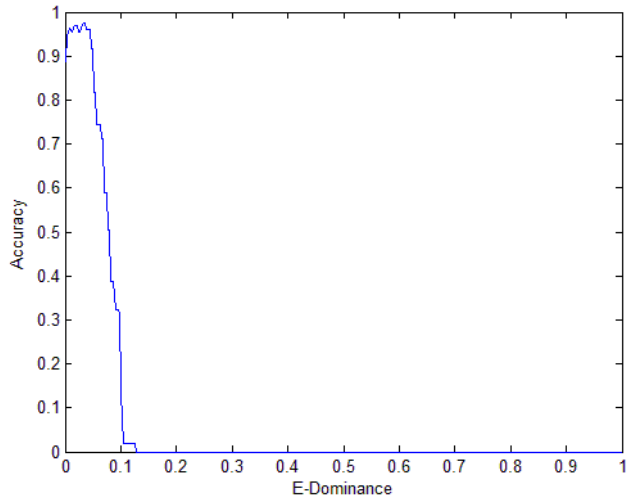
Accuracy	F_1	*Precision*	Recall	دسته
۰,۹۸	۰,۹۷۱	۰,۹۴۳	۱	اقتصادی
۰,۹۹۶	۰,۹۷۰	۰,۹۶۱	۰,۹۸۰	سیاسی
۰,۹۹۲	۰,۹۸۰	۰,۹۸۰	۰,۹۸۰	ورزشی
۰,۹۸۸	۰,۹۹۰	۰,۹۸۰	۱	فرهنگ و هنر
۰,۹۸۸	۰,۹۴۷	۱	۰,۹۰	علمی
۰,۹۸۹	۰,۹۷۲	۰,۹۷۳	۰,۹۷۲	میانگین

شکل۴-۴ تاثیر مقدار معیار E-Dominance را در دقت دسته‌بندی سیستم با استفاده از ویژگی‌های همرخداد و دسته‌بند SVM، نشان می‌دهد. در شکل ۴-۴-الف ویژگی‌ها به صورت باینری وزن‌دهی شده-

اند و در شکل ۴-۴-ب ویژگی‌ها با روش Co-occur TFIDF وزن‌دهی شده‌اند. حد آستانه انتخابی برای این معیار، مقدار ۰,۰۳۵ است که در این حد آستانه، سیستم، بیشترین دقت را داشته است.



الف

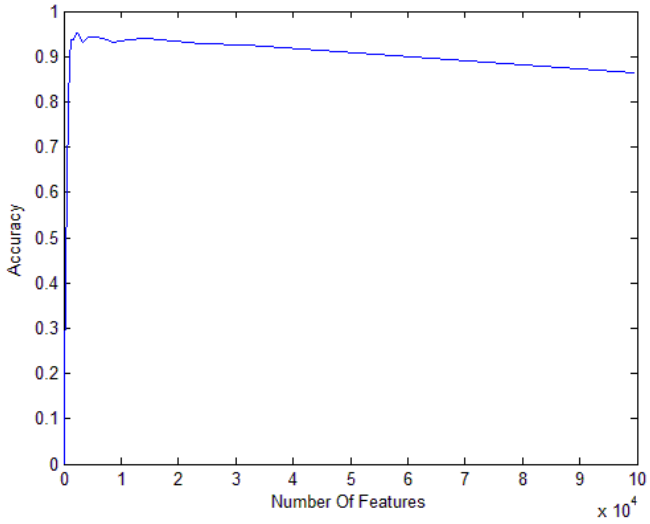


ب

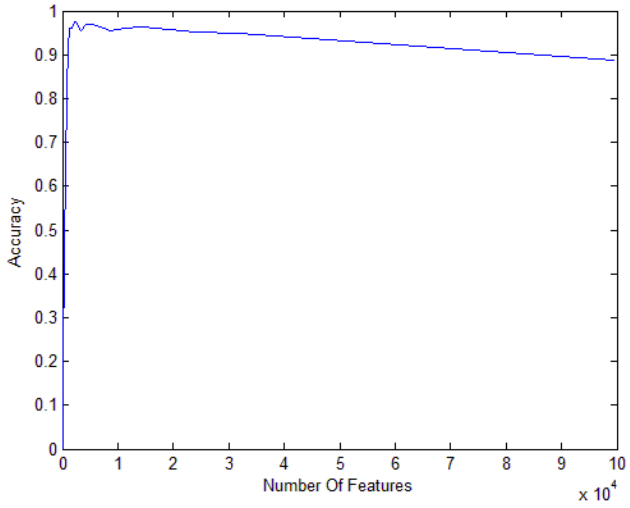
شکل (۴-۴) تاثیر حد آستانه معیار E-Dominance در دقت دسته‌بندی با ویژگی‌های همرخداد با دسته‌بند SVM
(شکل الف مربوط به وزن‌دهی باینری و شکل ب مربوط به وزن‌دهی Co-occur TFIDF)

شکل۴-۵ ، نیز تاثیر تعداد ویژگی‌های انتخابی همرخداد را در دقت دسته‌بندی سیستم با استفاده از دسته بند SVM نشان می‌دهد. در شکل ۴-۵-الف ویژگی‌ها به صورت باینری وزن‌دهی شده‌اند و در شکل

۴-۵-ب ویژگی‌ها با روش Co-occur TFIDF وزن‌دهی شده‌اند. سیستم با تعداد ۲۱۵۰ ویژگی همرخداد و وزن‌دهی باینری، توانسته است بهترین دقت را داشته باشد.



الف



ب

شکل (۴-۵) تاثیر تعداد ویژگی‌های انتخابی همرخداد در دقت دسته‌بندی با دسته‌بند SVM
(شکل الف مربوط به وزن‌دهی باینری و شکل ب مربوط به وزن‌دهی Co-occur TFIDF)

۴-۳-۲-۲- بررسی تاثیر تعداد ویژگی‌های توسعه‌گر در دقت دسته‌بندی

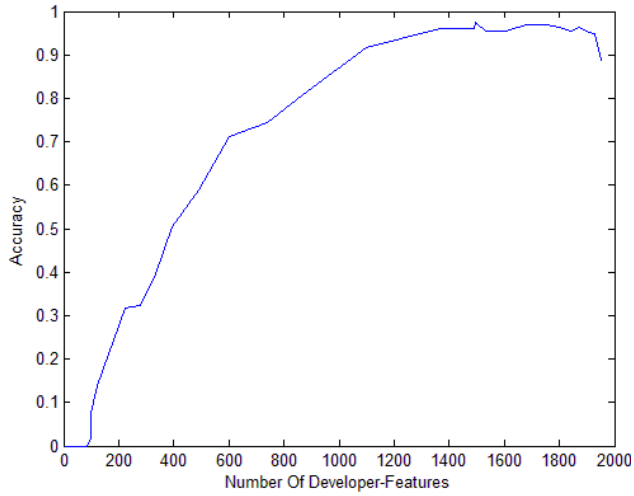
در حالتی که تعداد ویژگی‌های انتخابی بهترین دقت را درسیستم نشان می‌دهد، ۲۱۵۰ ویژگی همرخداد وجود دارد. تعداد ویژگی‌های توسعه‌گر نیز، ۱۴۹۵ ویژگی هستند جدول ۴-۱۰ تعداد ویژگی‌های توسعه‌گر انتخابی به تفکیک هر دسته را نشان می‌دهد. جمع تعداد ویژگی‌های همه‌ی دسته‌ها از کل ویژگی‌های توسعه‌گر بیشتر است. این امر ناشی از همپوشانی برخی کلاس‌ها و انتخاب برخی از ویژگی‌های توسعه‌گر در بیش از یک کلاس است. همچنین این جدول نسبت تعداد ویژگی‌های توسعه‌گر انتخابی از هر دسته را نسبت به کل ویژگی‌های انتخابی آن دسته و نسبت به کل ویژگی‌های توسعه‌گر نشان می‌دهد. نسبت‌ها نشان می‌دهد که دسته‌ی اقتصادی بیشترین و دسته‌ی فرهنگ و هنر کمترین نقش را در تعداد ویژگی-های توسعه‌گر دارند.

جدول (۴-۱۰) ویژگی‌های توسعه‌گر انتخابی از هر دسته

	اقتصادی	سیاسی	ورزشی	فرهنگ و هنر	علمی
تعداد ویژگی توسعه‌گر انتخابی	۴۰۳	۳۲۱	۳۰۲	۱۲۷	۲۰۵
نسبت تعداد ویژگی‌های توسعه‌گر دسته به کل ویژگی‌های دسته	٪۷۹	٪۷۵	٪۶۴	٪۳۱	٪۴۸
نسبت تعداد ویژگی‌های توسعه‌گر دسته به کل ویژگی‌های توسعه‌گر	٪۳۴	٪۲۸	٪۲۰	٪۸	٪۱۴

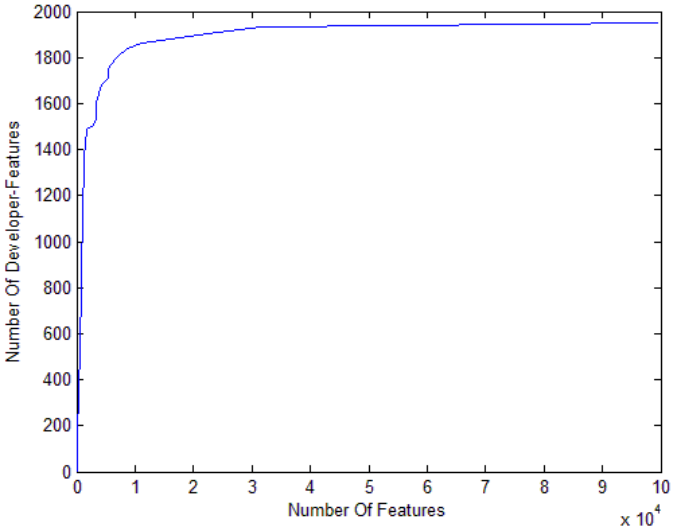
شکل ۴-۶ تاثیر تعداد ویژگی‌های توسعه‌گر انتخابی در دقت سیستم دسته بندی در حالتی که ویژگی‌ها با استفاده از معیار E-Dominace انتخاب شده و ویژگی‌های همرخداد را تشکیل داده‌اند و با استفاده از ویژگی‌های توسعه‌گر توسعه داده شده‌اند و پس از توسعه ا، ویژگی‌ها با استفاده از روش وزن‌دهی Co-

occur TFIDF وزن‌دهی شده و در نهایت با SVM، دسته‌بندی شده‌اند، را نشان می‌دهد. همانطور که می‌بینید در جائیکه تعداد ویژگی‌های توسعه‌گر ۱۴۹۵ بوده است، سیستم بهترین دقت را داشته است.



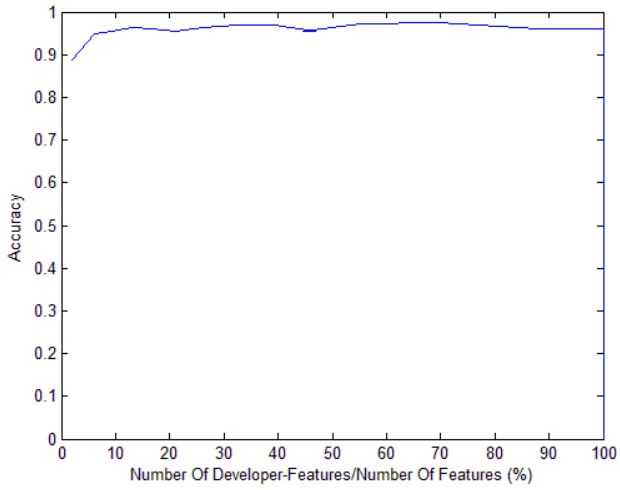
شکل (۴-۶) تاثیر تعداد ویژگی‌های توسعه‌گر در دقت دسته‌بندی

شکل ۴-۷ تعداد ویژگی‌های توسعه‌گر انتخابی را نسبت به کل ویژگی‌های همرخداد انتخابی، در روند کاهش ویژگی‌ها نشان می‌دهد. در حالتی که تعداد کل ویژگی‌ها ۹۹۴۳۸ بوده است، تعداد ویژگی‌های توسعه‌گر، ۱۹۵۰ عدد بوده است. در بهترین حالت سیستم که تعداد ویژگی‌های کل به ۲۱۵۰ ویژگی کاهش یافته است، تعداد ویژگی‌های توسعه‌گر ۱۴۹۵ عدد بوده و ۷۰٪ از ویژگی‌های کل بوده است. در روند کاهش ویژگی‌ها با معیار E-Dominance ، از زمانیکه تعداد ویژگی‌های کل به ۱۳۶۵ عدد رسید، تعداد ویژگی‌های توسعه‌گر نیز با تعداد ویژگی‌های انتخابی برابر بوده است. در شکل۴-۷ بخش اول نمودار، یک خط راست بوده و بعد از آن به صورت منحنی درآمده است.



شکل (۴-۷)تعداد ویژگی‌های توسعه‌گر نسبت به تعداد کل ویژگی‌ها

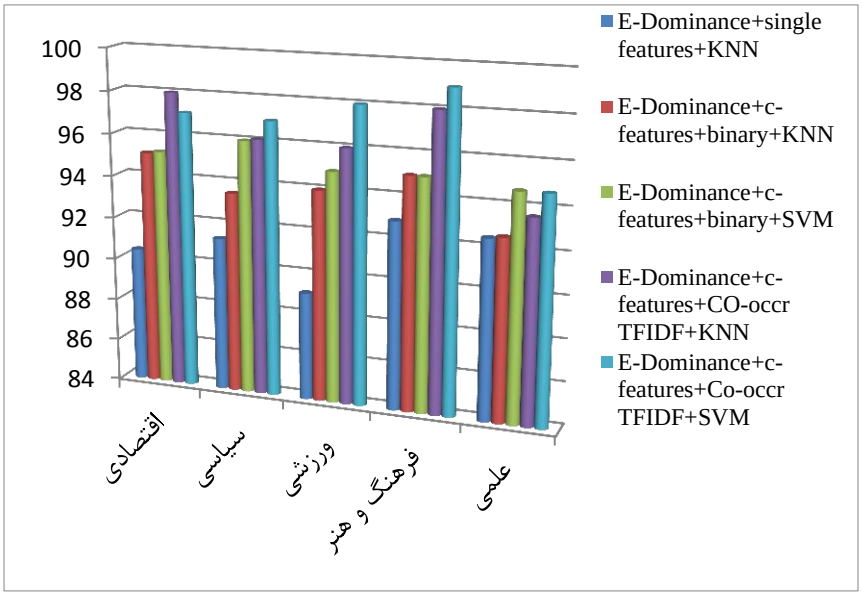
شکل ۴-۸ نیز تاثیر نسبت تعداد ویژگی‌های توسعه‌گر را به تعداد کل ویژگی‌ها در روند کاهش ویژگی‌ها، برای دقت دسته‌بندی نشان می‌دهد. در شکل می‌بینید که وقتی نسبت تعداد ویژگی‌های توسعه‌گر را به تعداد کل ویژگی‌ها، ۷۰٪ بوده است، سیستم بهترین دقت را ارائه داده است.



شکل (۴-۸) تاثیر نسبت تعداد ویژگی‌های توسعه‌گر به تعداد کل ویژگی‌ها در دقت دسته‌بندی

۴-۳-۳- مقایسه نتایج روش پیشنهادی در ۵ حالت مختلف در ۵ کلاس

روش پیشنهادی در ۵ حالت مختلف، پیاده سازی و نتایج آن در بخش‌های قبل ارائه گردید. شکل ۴-۹، معیار F1 را در ۵ حالت مختلف و در ۵ کلاس مختلف نشان می‌دهد. تقریبا در همه‌ی کلاس‌ها ، معیار F1، با به‌کارگیری ویژگی‌های همرخداد انتخابی با معیارE-Dominance و وزن‌دهی به روش Co-occur TFIDF و دسته‌بند SVM، بیشترین مقدار را داشته است.



شکل(۴-۹) مقایسه‌ی معیار F1 برای روش‌های مختلف در ۵ کلاس

۴-۴- مقایسه روش‌پیشنهادی با چند روش موجود

در جدول ۴-۱۱، روش پیشنهادی در حالت‌های مختلف با استفاده از دو روش وزن‌دهی باینری و Co-occur TFIDF و دو دسته‌بند KNN و SVM و با استفاده از ویژگی‌های منفرد و همرخداد آورده شده است. در این جدول نتایج مربوط به مراجع [۵۱, ۱۱۶, ۱۱۸] نیز که برای مقایسه با روش پیشنهادی انتخاب شده‌اند، آورده شده است. مراجع نامبرده به این دلیل انتخاب شده‌اند که در زبان فارسی آزمایشات خود را انجام داده و از پایگاه داده‌ی همشهری استفاده کرده‌اند. همچنین تعداد متون آموزشی و آزمایش و

تعداد کلاس‌ها و نیز موضوعات کلاس‌ها، در این مقالات مشابه این پایان‌نامه است. مرجع [۷] نیز به عنوان مرجع اصلی این پایان نامه نیز برای مقایسه انتخاب شده است. برای مقایسه روش پیشنهادی با مرجع [۷]، روش مورد نظر بر روی متون انتخابی این پایان نامه پیاده‌سازی و تست شده است. مرجع اصلی نتایج خود را بر روی متون زبان انگلیسی ارائه داده است. در فصل دوم روش‌های استفاده شده در این مراجع آورده شده است. همان طور که نتایج در جدول ۴-۱۱ نشان می‌دهد، روش پیشنهادی در حالتی که ویژگی‌ها با استفاده از معیار E-Dominace انتخاب شده و ویژگی‌های همرخداد را تشکیل داده‌اند و نیز با استفاده از روش Co-occur TFIDF وزن‌دهی شده و در نهایت با دسته‌بند SVM، دسته‌بندی شده‌اند، بهترین نتیجه را در میان نتایج ارائه داده است.

نکته قابل توجه در نتایج روش پیشنهادی این است که مقادیر معیارهای F1 و Recall در بیشتر حالات یکسان و یا بسیار نزدیک به هم بوده است. دلیل نزدیکی این مقادیر، نزدیک بودن مقادیر Recall و Precision است که تشکیل دهنده‌ی معیار F1 هستند.

۴-۵- نتیجه‌گیری

در این فصل پس از معرفی معیارهای ارزیابی، نتایج آزمایشات ارائه گردید. معیارE-Dominance با چند معیار ارزش‌گذاری داده دیگر مقایسه شده وکارایی بیشتر را نسبت به دیگر معیارها نشان داد. آزمایشات انجام شده در ۵ حالت مختلف و با استفاده از دسته‌بندهای KNN و SVM انجام شد. دو روش وزن‌دهی باینری و Co-occr TFIDF نیز برای ویژگی‌ها همرخداد به کار رفته اند. در پایان، در حالتی که ویژگی‌ها با استفاده از معیار E-Dominace انتخاب شده و ویژگی‌های همرخداد را تشکیل داده‌اند و و۷۰٪ از ویژگی‌ها به عنوان ویژگی‌های توسعه‌گر، به توسعه وغنی سازی داده‌ها پرداخته‌اند و نیز با استفاده از روش Co-occur TFIDF وزن‌دهی شده و در نهایت با دسته‌بند SVM، دسته‌بندی شده‌اند، بهترین نتیجه را در میان نتایج ارائه داده است.

جدول(۴-۱۱) مقایسه روش پیشنهادی با چند روش دیگر

روش	دسته‌بند	ویژگی‌ها	معیار کاهش ویژگی	وزن‌دهی	تعداد ویژگی‌ها	Recall	F1	توضیح
مرجع [۵۱]	NN	کیسه لغات	Chi-square	TFIDF	۳۱۹۰	۹۴.۳۹	۹۳,۳۰	کاربرد فارس‌نت
مرجع[۱۱۶]	SVM	کیسه لغات	ICF+UNI	TFIDF	–	۸۳	۸۶	کاربرد گنجواژه معنایی
مرجع [۱۱۸]	1-NN	کیسه لغات	PCA + Genetic Algorithm	TFCRF	-	۸۷	۸۹	–
مرجع [۷]	KNN	زوج کلمات همرخداد	Dominance	باینری	۷۰۰۷	۹۳,۲	۹۲,۶۱	کاربرد ویژگی‌های توسعه‌گر
مرجع [۷]	SVM	زوج کلمات همرخداد	Dominance	باینری	۷۰۰۸	۹۴,۸	۹۴	کاربرد ویژگی‌های توسعه‌گر
روش پیشنهادی	KNN	کیسه لغات	E-Dominance	TFIDF	۱۰۴۱	۹۱,۲	۹۱,۲۳	–
روش پیشنهادی	KNN	زوج کلمات همرخداد	E-Dominance	باینری	۲۱۵۰	۹۴	۹۴	کاربرد ویژگی‌های توسعه‌گر
روش پیشنهادی	SVM	زوج کلمات همرخداد	E-Dominance	باینری	۲۱۵۰	۹۵	۹۵,۵	کاربرد ویژگی‌های توسعه‌گر
روش پیشنهادی	KNN	زوج کلمات همرخداد	E-Dominance	Co-occur TFIDF	۲۱۵۰	۹۶,۴۰	۹۶,۴	کاربرد ویژگی‌های توسعه‌گر
روش پیشنهادی	SVM	زوج کلمات همرخداد	E-Dominance	Co-Occur-TFIDF	۲۱۵۰	۹۷,۲	۹۷,۲	کاربرد ویژگی‌های توسعه‌گر

فصل پنجم

نتیجه گیری و کارهای آینده

۵-۱-خلاصه تحقیق

انتخاب ویژگی‌ها در متون و نحوه‌ی وزن‌دهی به ویژگی‌های انتخاب شده، از مهم‌ترین مسائل دسته-بندی خودکار متون هستند. در این تحقیق یک معیار ارزش‌گذاری ویژگی، بنام E-Dominance معرفی شد. این معیار با سایر معیارهای ارزش‌گذاری و وزن‌دهی متداول در حوزه‌ی دسته‌بندی متون مقایسه شد و نشان داده شد که با استفاده از معیار فوق می‌توان به دقتی همسان و یا بیشتر از روش‌های متداول با تعداد بسیار کمتری ویژگی، دست پیدا کرد. کاهش ویژگیها با استفاده از این معیار، حدود ۹۴٪ بوده است. هدف از معرفی این معیار، انتخاب و کاهش ویژگی‌ها برای تشکیل و استخراج ویژگی‌های همرخداد است. با انتخاب کلمات با اهمیت در هر دسته با استفاده از معیار E-Dominance، ویژگی‌های همرخداد تشکیل شد. ویژگی‌های همرخداد هیچ‌گاه در زبان فارسی به عنوان ویژگی‌های متون برای دسته‌بندی استفاده نشده‌اند. در سایر زبان‌ها نیز بندرت به عنوان ویژگی در نظر گرفته شده‌اند. کاربرد اصلی ویژگی‌های همرخداد در بازیابی اطلاعات است. از ویژگی‌های همرخداد تشکیل شده، تعداد زیادی بر اساس حد آستانه معیار فوق، از بردار ویژگی‌ها حذف شدند سپس ویژگی‌هایی که تعداد رخداد بیشتری در یک دسته داشته‌اند، به عنوان ویژگی-های توسعه‌گر، برای توسعه‌ی داده‌های آموزش و آزمایش انتخاب می‌شوند. توسعه‌ی داده‌ها باعث کم شدن شباهت‌های بین کلاسی و زیاد شدن شباهت داخل کلاسی می‌شود. توسعه‌ی داده‌ها در داده‌های آزمایشی نیز به گونه‌ای اعمال می‌شود که بدون توجه به کلاس داده، بتوان داده‌های غنی-تری را برای دسته‌بندی به سیستم ارائه کرد. پس از توسعه‌ی داده‌ها از روش وزن‌دهی Co-occur TFIDF که یک روش توسعه یافته از TFIDF معمول است، برای ویژگی‌های همرخداد به کار گرفته شده است. در کارهای گذشته فقط از روش وزن‌دهی باینری برای ویژگی‌های همرخداد استفاده شده است. پس از استخراج ویژگی و وزن‌دهی به آنها، فاز بعدی دسته‌بندی متون است. در

این کار، از دسته‌بندهای KNN و SVM برای دسته‌بندی استفاده شده است. در بیشتر تحقیقات انجام شده، معمولا SVM از دقت بالاتری نسبت به سایر دسته‌بندها برخوردار بوده است. در روش پیشنهادی نیز، این دسته‌بند نتایج بهتری را ارائه داده است. نتایج آزمایشات در مقایسه با سایر کارها نشان داده است که روش پیشنهادی در حالتی که ویژگی‌ها با استفاده از معیار E-Dominace انتخاب شده و ویژگی‌های همرخداد را تشکیل داده‌اند و نیز با استفاده از روش Co-occur TFIDF وزن‌دهی شده و در نهایت با دسته‌بند SVM، دسته‌بندی شده‌اند، بهترین نتیجه را ارائه داده است. برای نشان دادن کارایی روش پیشنهادی در مقایسه با سایر تحقیقات، از معیارهای Recall و F1 استفاده شده است. این دو معیار در همه‌ی حالت‌ها در روش پیشنهادی، مقادیر بیشتری نسبت به سایر روش‌ها داشته و بهبود دسته‌بندی را نشان دادند.

۵-۲-پیشنهاد برای کارهای آینده

گرچه این کار تحقیقاتی پایان یافته است، اما مسلما دارای نقاط ضعف و کمبودهایی است که در آینده برطرف خواهد شد. از جمله پیشنهادات برای ائامه‌ی کار حاضر عبارتند از:

- طراحی یک ریشه‌یاب موثر در زبان فارسی و کاربرد آن در انتخاب ویژگی‌ها
- توسعه‌ی کلمات ایست استخراجی در زبان فارسی به لیست کامل و استانداردی که مورد استفاده همگان باشد.
- ارائه‌ی یک معیار موثر در انتخاب ویژگی‌های توسعه‌گر برای توسعه‌ی داده‌ها، در این پایان نامه از حد آستانه DF برای این کار بر اساس آزمون و خطا استفاده شده است.
- توسعه داده‌ها با استفاده از ویژگی‌های توسعه‌گر در چند سطح، به نحوی که در هر مرحله ویژگی‌های توسعه‌گر جدید استخراج و مورد استفاده در توسعه داده‌ها قرار گیرند.

مراجع

[۱] مدنی. ص. س. ، " دسته بندی اسناد فارسی به کمک هستان شناسی فارس نت" ، پایان نامه کارشناسی ارشد، دانشکده کامپیوتر و فن آوری اطلاعات، دانشگاه صنعتی شاهرود، بهمن ۱۳۹۱.

[٤] اصدقی. ف.، "طبقه بندی اسناد متنی با استفاده از سیستم ایمنی مصنوعی" ، پایان نامه کارشناسی ارشد، دانشکده کامپیوتر و فن آوری اطلاعات، دانشگاه علم و صنعت ایران ، اسفند ۱۳۸۸.

[۲۸] دستجردی ا. ، زنجانی م. ، "روش جدید خوشه بندی مستندات متنی الکترونیکی فارسی به کمک واژ-هستان شناسی فارس نت"، اولین کنفرانس ملی دانش پژوهان کامپیوتر و فناوری اطلاعات، ۱۳۹۰.

[۲۹] ایران پور م. ، "بررسی مشکلات تعیین حدود جمله وکلمه"، سمینار کارشناسی ارشد، دانشگاه علم وصنعت ایران، ۱۳۸۶.

[۳۰] داداش میری پ. ، "تشخیص انتهای کلمات وایجاد فاصله میان کلمات"، پایان‌نامه کارشناسی، دانشگاه علم وصنعت ایران، ۱۳۸۰.

[۳۱] بیجن خان م. ،"تشخیص کسره ی اضافه"، طرح تحقیقاتی، پژوهشگاه فرهنگ، هنر و ارتباطات، تهران، ۱۳۸۴.

[۳٤] حافظی م. ، ثامتی ح. ، منصوری ن. ، بحرانی، موثق ، " ارائه یک مدل دستوری برای بهبود دقت سیستم های بازشناسی گفتار پیوسته ی فارسی"، دومین گارکاه پژوهشی زبان فارسی و رایانه، ص. ۸۰تا ۹۱، تهران ، ۱۳۸۵.

[۳٥] حسامی‌فرد ر. ، قاسم خانی غ.ر. ، "طراحی یک الگوریتم ریشه یابی برای زبان‌فارسی"، یازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، ص. ۵۱۵ تا۵۱۹، تهران، ۱۳۸۴. .

[۳٦] مشکی م. ، "خلاصه سازی گزینشی چند سندی متون فارسی"، پایان‌نامه ارشد، دانشکده کامپیوتر،دانشگاه علم و صنعت ، ۱۳۸۸.

[۳۷] ابوالحسنی ح. ، اسماعیلی س. ، محمدی م. ، "یک خلاصه ساز آماری برای زبان فارسی"، یازدهمین کنفرانس بین المللی انجمن کامپیوتر ایران، ۱۳۸۴.

[٤۰] آنالویی م. ، مشکی م.، "خلاصه سازی چند سندی متون فارسی با استفاده از یک روش مبتنی بر خوشه بندی" ،اولین کنفرانس ملی مهندسی نرم افزارایران، تهران، ۱۳۸۷.

[۵۱] حسن پور ح. ، مدنی ص. س. ، "بهبود دقت سیستم دسته بندی خودکار اسناد فارسی به کمک هستان شناسی فارس نت،" مجله علمی پژوهشی رایانش نرم و فن آوری اطلاعات، جلد ۳، شماره ۱، ۱۳۹۳ .

[٦٥] حسن پور ح. ، قنبری سرخی ع. ، پارسی ا. ، "استخراج بهترین ویژگی از متون فارسی با استفاده از تجزیه و تحلیل مولفه های اصلی با کمک میانگین یادآوری و الگوریتم ژنتیک"، نخستین کنفرانس بین المللی پردازش خط و زبان فارسی، ۱۳۹۱.

[۱۰۹] عربی نرئی س. ، وحیدی اصل م. ، مینایی بیگدلی ب. ، "استخراج کلمات کلیدی جهت طبقه بندی متون فارسی"، اولین کنفرانس داده کاوی ایران، ۱۳۸۶.

[۱۱۰] عربی نرئی س. ، وحیدی اصل م. ، مینایی بیگدلی ب. ، "رویکردی جدید جهت طبقه-بندی مستندات فارسی با استفاده از الگوریتم کاوشگر مورچه"، هفتمین همایش زبان شناسی ایران، ۱۳۸۶.

[۱۱۱] احسان بصیری ا. ، نعمتی ش.، قاسم آقایی ن. ، "مقایسه طبقه بندی متون فارسی با استفاده از الگوریتم *KNN* و *FKNN*و انتخاب ویژگی ها بر اساس بهره ی اطلاعات و فرکانس سند"، سیزدهمین کنفرانس سالانه انجمن کامپیوتر ایران، ۱۳۸۶.

[۱۱۲] محمدی م.، مینایی بیگدلی ب.، آنالویی م. ، "استفاده از شبکه های عصبی *CC4* برای طبقه بندی اسناد فارسی"، دومین کنفرانس داده کاوی ایران، ۱۳۸۷.

[۱۱۳] احمدزاده،، ا.ب.ح.ف.ف.م.ح.س.م.، ، "طبقه بندی متون خبری فارسی با استفاده از الگوریتم *Naïve Bayes*"، دومین کنفرانس داده کاوی ایران، ۱۳۸۷.

[۱۱٤] محمدی م. ، مینایی بیگدلی ب.، آنالویی م. ، "طبقه بندی اسناد متنی با استفاده از شبکه های عصبی پرسپترون چند لایه "، دومین کنفرانس داده کاوی ایران، ۱۳۸۷.

[۱۱٦] مقصودی ن. ، همایون پور م. ، "ارائه روشی جدید در طبقه بندی متون فارسی با استفاده از دانش معنایی"، پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران، ۱۳۸۸.

[۱۱۷] قنبری سرخی ع. ، ابراهیمی ف. ،" بهبود عملکرد طبقه بندی متون فارسی با استفاده از تجزیه و تحلیل مولفه های اصلی با کمک معیار میانگین یادآوری و دقت "، چهاردهمین کنفرانس دانشجویی مهندسی برق ایران، ۱۳۹۰.

[۱۱۸] حسن پور ح. ، قنبری سرخی ع. ، پارسی ا.، "استخراج بهترین ویژگی از متون فارسی با استفاده از تجزیه و تحلیل مولفه های اصلی با کمک میانگین یادآوری و الگوریتم ژنتیک"، نخستین کنفرانس بین المللی پردازش خط و زبان فارسی، ۱۳۹۱.

[۱۱۹] طاهری نیا م. ، "دسته بندی متون فارسی با استفاده از یادگیری نیمه نظارت شده"، چهارمین کنفرانس مهندسی برق و الکترونیک ایران، ۱۳۹۱.

[۱۲۰] زمانی م. ، هدایت ر. ، صادق زاده م. ، "دسته بندی متون فارسی با استفاده از روش آنالیز معنایی پنهان احتمالاتی"، اولین همایش ملی کاربرد سیستم های هوشمند (محاسبات نرم) در علوم و صنایع، ۱۳۹۲.

[۱۲۱] برفامی م. ، فاطری س. ،"استفاده ازترکیب شبکه های عصبی جهت دسته بندی متون فارسی مبتنی برالگوریتم های *GA,KNN,PCA* جهت کاهش ویژگی"، اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات، ۱۳۹۲.

[۱۲۲] عابدی نیا م. ، الله دادی ل. ، شیخی ف. ، "کاوش متون فارسی دروب بااستفاده ازتحلیل معنایی وروش طبقه بندی چندلایه ای اطلاعات"، اولین همایش ملی رویکردهای نوین در مهندسی کامپیوتر و بازیابی اطلاعات، ۱۳۹۲.

[۱۲۳] صفیان ز. ، بلداجی ، نادرری‌دهکردی م. ، رستگار ح. ، "دسته بندی متون به کمک الگوریتم *KNN* بهبود یافته مبتنی بر خوشه های وزن دار"، اولین کنفرانس ملی نوآوری در مهندسی کامپیوتر و فناوری اطلاعات، ۱۳۹۲.

[2] A.Khan, B.B., L.Hong Lee, K.khan, *"A Review of Machine Learning Algorithms for Text-Documents Classification" , Journal of Advances in Information Technology, VOL. 1, NO. 1, February 2010*.

[3] B.Zhang, *"Learning Features for Text Classication" , Thesis of Doctor of Philosophy, University of Washington, 2013*.

[5] F.Sebastiani, *"Machine learning in automated text categorization", ACM Computing Surveys, Vol. 34, pp. 1-47 ,2002*.

[6] M.YUAN, Y.X.O., Z. XIONG, *"A Text Categorization Method using Extended Vector Space Model by Frequent Term Sets" , Journal of Advances in Information Science and Engineering, 29, 99-114 ,2013*.

[7] F. Figueiredo, L.R., T. Couto, T. Salles, M. A. Goncalves, W. MeiraJr, *"Word co-occurrence features for text classification", Information Systems, 36, pp.843–858, 2011*.

[8] J.Feng, H.L., *"Sat-mod:moderate itemset fittest for text classification", in:WWW'05:Special Interest Tracksand Posters of the 14th International Conference on World Wide Web,NewYork, USA, pp.1054–1055, 2005*.

[9] J.H.H.Cheng, X.Y., C.Hsu, *"Discriminative frequent pattern analysis for effective classification", in: ICDE '07: Proceedings of International Conference on Data Engineering, Istanbul, Turkey, pp.716–725, 2007*.

[10] O.R.Zaiane, M.L.A., *"Classifying text documents by associating terms with text categories, in: CRPITS '02 :Proceedings of The 13th Australasian Conference on Database Technologies, Australian Computer Society,Inc. , Darling hurst, Australia, pp.215–222, 2002*.

[11] R.Rak, W.S., O.R.Zaiane, M.L.Antonie, *"Considering re-occurring features in associative classifiers", in: PAKDD '05: 9th Pacific Asia Conference on Knowledge Discovery and Data Mining, Hanoi, Vietnam, pp.240–24, 2005.*

[12] C. Largeron, C.M., M. Gery, *"Entropy based feature selection for text categorization", ACM Symposium on Applied Computing, TaiChung : Taiwan, Province Of China, 2011*.

[13] Joachims, T., *"Text categorization with support vector machines: Learning with many relevant features", In Proceedings of ECML '98, 10th European Conference on Machine Learning, No.1398, pp137-142, Chemnitz, DE, Springer Verlag, Heidelberg, DE, 1998*.

[14] T.Gaustad, G.B., R. Groningen, *"Accurate Stemming of Dutch for Text Classification," Language and Computers, 2002*.

[15] K.Tzeras, S.H., *"Automatic indexing based on Bayesian inference networks," presented at the Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, Pittsburgh, Pennsylvania, United States, 1993*.

[16] N. J. Belkin , W.B.C., , *"Information filtering and information retrieval: two sides of the same coin? " , Commun. ACM, vol. 35, pp. 29-38, 1992*.

[17] P. J. Hayes, P.M.A., I. B. Nirenburg, L. M. Schmandt, *"TCS: a shell for content-based text categorization," presented at the Proceedings of the sixth conference on Artificial intelligence applications, Santa Barbara, California, United States, 1990*.

[18] E. D. Liddy, W.P., and E. S. Yu, *"Text categorization for multiple users based on semantic features from a machine-readable dictionary," ACM Trans. Inf. Syst., vol. 12, pp. 278-295, 1994*.

[19] G.Escudero, L.M., G.Rigau, *"Boosting Applied to Word Sense Disambiguation," in Machine Learning: ECML 2000. vol. 1810, R. López de Mántaras and E. Plaza, Eds., ed: Springer Berlin Heidelberg, pp. 129-141, 2000.*

[20] Gale, W.A., Kenneth W. Church, and David Yarowsky, *"A method for disambiguating word senses in a large corpus", Computers and the Humanities26, no. 5 ,PP. 415-439, 1992*.

[21] Fürnkranz, J., *"Exploiting Structural Information for Text Classification on the WWW," presented at the Proceedings of the Third International Symposium on Advances in Intelligent Data Analysis, 1999*.

[22] S.Chakrabarti, B.D., P .Indyk, *"Enhanced hypertext categorization using hyperlinks," presented at the Proceedings of the 1998 ACM SIGMOD international conference on Management of data, Seattle, Washington, United States, 1998*.

[23] S.Eyheramendy, A.G., *W.-H. Ju, D. D. Lewis, and D. Madigan, "Sparse Bayesian Classifiers for Text Categorization," Joint Statistical Meeting in San Francisco, 2003*

.

[24] A.Dasgupta, P.D., B.Harb, V.Josifovski, M.W.Mahoney,, *"Feature selection methods for text classification," presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA, 2007*.

[25] H.Kim, P.H., H.Park, *"Dimension Reduction in Text Classification with Support Vector Machines," J. Mach. Learn. Res., vol. 6 ,pp. 37-53, 2005*.

[26] M.Hassel, N.M., *"FarsiSum: a Persian text summarizer," presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, Switzerland, 2004*.

[27] Megerdoomian, R.Z.K., *"Processing Persian Text: Tokenization in the Shiraz Project," Memoranda in Computer and Cognitive Science, vol. MCCS-00-322, April 2000*.

[32] M.R.Davarpanah, M.S., M. Aramideh, *"Farsi lexical analysis and stop word list," Library Hi Tech, vol. 27 pp. 435-449, 2009*.

[33] K.Taghva, R.B., M.Sadeh, " *A List of Farsi Stopwords", ISRI Technical Report, 2003*.

[38] K. Taghva, R.B., and M. Sadeh, , *"A stemming algorithm for the Farsi language", in Information Technology: Coding and Computing. ITCC 2005. International Conference on, pp. 158-162 Vol. 1, 2005.*

[39] Hovy, E., *"Parsimonious and Profligate Approaches to the Question of Discourse Structure Relations", in Workshop on NLG, 1990*.

[41] C.Bouras, V.T., *"W-kmeans: clustering news articles using wordNet," presented at the Proceedings of the 14th international conference on Knowledge-based and intelligent information and engineering systems: Part III, Cardiff, UK, 2010*.

[42] C.Yi-Hsing , H.H.-Y., *"An Automatic Document Classifier System based on Naive Bayes Classifier and Ontology," International Conference on in Machine Learning and Cybernetics, pp. 3144-3149., 2008.*

[43] Z.Karimi , M.S., *"Automatic Text Summarization Systems Farsi," in 12th International Conference of Computer Society of Iran, 2006*.

[44] D.D.Lewi, *"Evaluating and optimizing autonomous text classification systems", In Proceeding of SIGIR '95 ,18th ACM International Conference on Research and Development in Information Retrieval. pp. 246-254, 1995*.

[45] D.Chai, *"A Trainable Web Page Document Summarizer," Natural Language Processing,CS224N/Ling237, 2000*.

[46] H.P.Luhn, *"A statistical approach to mechanized encoding and searching of literary information," IBM J. Res. Dev., vol. 1, pp. 309-317, 1957*.

[47] L.E. , K.J., *"Text Categorization with Support Vector Machines. How to Represent Texts in Input Space," Machine Learning, vol. 46, pp. 423-444 , 2002*.

[48] Jones, K.S., *"Indexing Term Weighting," Information Storage and Retrieval, vol. 9, pp. 619-633, 1973*.

[49] G.Salton, C.S.Y., *"On the specification of term values in automatic indexing," Journal of Documentation, vol. 29, pp. 351-372, 1973*.

[50] G.Salton, J.A., A.Singhal,, *"Automatic text decomposition and structuring," Information Processing & Management, vol. 32, pp. 127-138, 1996*.

[52] C.Apte , F.D., S.M.Weiss, *"Automated learning of decision rules for text categorization", ACM Transactions on Information Systems 12 (3) , pp.233–251, 1994*.

[53] D.D.Lewis, *"Representation and learning in information retrieval", Ph.D. Thesis, Amherst, MA, USA, 1992*.

[54] D.D.Lewis, *"An evaluation of phrasal and clustered representations on a text categorization task", in: SIGIR '92: Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, NY, USA, pp. 37–50, 1992*.

[55] S.Dumais, J.P., D.Heckerman, M.Sahami, *"Inductive learning algorithms and representations for text categorization", in: CIKM '98: Proceedings of the 7th International Conference on Informa- tion and Knowledge Management, ACM, New York, NY, USA, pp. 148–155, 1998*.

[56] S.Scott, S.M., *“Feature engineering for text classification”, in: ICML ’99: Proceedings of the 16th International Conference on Machine Learning, Morgan Kaufmann Publishers, Inc., San Fran- cisco, CA, USA, pp. 379–388, 1999*.

[57]] D.Mladenic , M.G., *”Word sequences as features in text learning”, in:Proceedings of ERK-98, the 7th Electrotechnical and Computer Science Conference, Ljubljana,SL,pp.145–148, 1998.*

[58] urnkran, J.F., *” A study using n-gram features for text categorization”, Technical Report,OEFAI-TR-9830,Austrian Institute for Artificial Intelligence, 1998*.

[59] G.Forman, *”Feature selection for text classification”,in:Computational Methods of Feature Selection”,Chapman and Hall/CRC, pp. 257–276, 2007*.

[60]. M.F.Caropreso, S.M., F.Sebastiani,Alearner, *“independent evaluation of the usefulness of statistical phrases for automated text categorization”,pp.78–102,2001*.

[61] C.M.Tan, Y.F.W., C.D.Lee” ,*The use of bigrams to enhance text categorization”, Information Processing and Management38(4) 529–546, 2002*.

[62] E.Crawford, I.K., J.Patrick, *”Phrases and feature selectionin e-mail classification”,in:P.Bruza,A.Moffat,A.Turpin(Eds.),ADCS, University of Melbourne,Department of Computer Science, pp. 59–62,2004*.

[63] " Y.Yang, J.O.Pedersen., "*A Comparative Study on Feature Selection in Text Categorization," presented at the Proceedings of the Fourteenth International Conference on Machine Learning, 1997.*

[64] Selamat, A., Hidekazu Yanagimoto, and Sigeru Omatu, *"Web news classification using neural networks based on PCA." In SICE 2002. Proceedings of the 41st SICE Annual Conference, vol. 4, pp. 2389-2394. IEEE, 2002*.

[66] W.Z. , X.Q., *"Text categorization based on LDA and SVM." InComputer Science and Software Engineering, 2008 International Conference on, vol. 1, pp. 674-677. IEEE, 2008*.

[67] Hu, M., Shijun Wang, Anhui Wang, and Lei Wang, *"Feature Extraction Based on the Independent Component Analysis for Text Classification." InFuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 2, pp. 296-300. IEEE, 2008*.

[68] W.W.Cohen , Y.S., *"A simple, fast, and effective rule learner," presented at the Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, Orlando, Florida, United States, 1999*.

[69] D.D.Lewis, *"Naive (Bayes) at forty: The independence assumption in information retrieval," MACHINE LEARNING: ECML-98 Lecture Notes in Computer Science, 1998*

[70] I.Dagan, Y.K., D.Roth, , *"Mistake-driven learning in text categorization," The Second Conference on Empirical Methods in Natural Language Processing, pp. 55-63, 1997*

[71] Elkan, C" ,.*Nearest neighbor classification." University of California–San Diego, 2007*.

[72] Jiang, C., Frans Coenen, Robert Sanderson, and Michele Zito, *"Text classification using graph mining-based feature extraction." Knowledge-Based Systems 23, no. 4 : 302-308 ,2010*.

[73] Qiu, L.-Q., Ru-Yi Zhao, Gang Zhou, and Sheng-Wei Yi, *"An extensive empirical study of feature selection for text categorization." In Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on, pp. 312-315. IEEE, 2008*.

[74] Cheng, Y., Ruiling Zhang, Xiufeng Wang, and Qiushuang Chen, *"Text Feature Extraction Based on Rough Set." In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 2, pp. 310-314. IEEE, 2008*.

[75] Park ,C.H., "Dimension reduction using least squares regression in multi-labeled text categorization." In Computer and Information Technology, 2008. CIT 2008. 8th IEEE International Conference on, pp. 71-76. IEEE, 2008.

[76] Sh.Zhou, T.W.L., J.Hu, A.Zhou, , *"Fast Text Classification : A Training-Corpus Pruning Based Approach", in Eighth International conference on Database Systems for Advanced Applications (DASFAA '03), IEEE, 2003*.

[77] G.R.D., *"Knowledge Based Neural Network for Text classification", in IEEE International conference on Granular Computing, IEEE, pp. 542-547, 2003*.

[78] Z.Chen, L.H.a.Y.L.M., *"Incremental Learning for Text", In International Joint conference on Neural Networks, IEEE, Orlando, USA, pp. 1-4, 2007*. .

[79] Zhen-fang, Z., L. Pei-yu, Lu Ran, *"Research of text classification technology based on genetic annealing algorithm." In Computational Intelligence and Design, 2008. ISCID'08. International Symposium on, vol. 1, pp. 265-269. IEEE, 2008*.

[80] Meng, Z.S., Wei Song Dan, Jinhong Li, *"a new associative classifier for text categorization." Intelligent System and Knowledge Engineering, 2008. ISKE 2008. 3rd International Conference on, 2008*.

[81] L.Shi, J.Z., E.Liu, P.He, *“Text classification Based on Nonlinear Dimensionality Reduction Techniques and Support Vector Machines”, in Third International conference on Natural Computation, IEEE, pp. 1-4, 2007*.

[82] Zhou, X., Huaxiang Zhan, *"An algorithm of text categorization based on similar rough set and fuzzy cognitive map." In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 3, pp. 127-131. IEEE, 2008*.

[83] Wang, X., Zhen Hua, Rujiang Bai, *"A hybrid text classification model based on rough sets and genetic algorithms." In Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on, pp. 971-977. IEEE, 2008*.

[84] Yang, Y., Seán Slattery, and Rayid Ghani, *"A study of approaches to hypertext categorization." Journal of Intelligent Information Systems 18, no. 2-3 (2002): 219-241*.

[85] Ding, W., Songnian Yu, Qianfeng Wang, Jiaqi Yu, and Qiang Guo. , *"A Novel Naive Bayesian Text Classifier." In Information Processing (ISIP), 2008 International Symposiums on, pp. 78-82. IEEE ,2008*.

[86] Zhou, S., Kan Li, and Yushu Liu, *"Text Classification Based on a Novel Bayesian Hierarchical Model." In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 2, pp. 218-221. IEEE, 2008*.

[87] Chen, Q., Dequan Zheng, Tiejun Zhao, and Sheng Li, *"A Fusion of Multiple Classifiers Approach Based on Reliability function for Text Categorization." In Fuzzy Systems and Knowledge Discovery, 2008. FSKD'08. Fifth International Conference on, vol. 2, pp. 338-342. IEEE ,2008*.

[88] Yuan, P., Yuqin Chen, Hai Jin, and Li Huang, *"MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification." In Semantic Computing and Systems, 2008. WSCS'08. IEEE International Workshop on, pp. 133-140. IEEE, 2008*.

[89] Wang, L.L.Z.L.A ,.and Haijun He, *"the research of decision support vector machine in web information classification." Computer Supported Cooperative Work , 2008*

[90] Chen, J., Houkuan Huang, Shengfeng Tian, and Youli Qu, *"Feature selection for text classification with Naïve Bayes." Expert Systems with Applications 36, no. 3: 5432-5435, 2009*.

[91] Wang, P., Jian Hu, Hua-Jun Zeng, and Zheng Chen, *"Using Wikipedia knowledge to improve text classification." Knowledge and Information Systems19, no. 3: 265-281, 2009*.

[92] Sun, A ,.Ee-Peng Lim, and Ying Liu, *"On strategies for imbalanced text classification using SVM: A comparative study." Decision Support Systems 48, no. 1: 191-201, 2009*.

[93] Sriram, B., Dave Fuhry, Engin Demir, Hakan Ferhatosmanoglu, and Murat Demirbas, . *"Short text classification in twitter to improve information filtering." InProceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval, pp. 841-842. ACM, 2010*.

[94] Jiang, C., Frans Coenen, Robert Sanderson, and Michele Zito, *"Text classification using graph mining-based feature extraction." Knowledge-Based Systems 23, no. 4, PP. 302-308, 2010*.

[95] Lu, S.-H., Ding-An Chiang, Huan-Chao Keh, and Hui-Hua Huang, *"Chinese text classification by the Naïve Bayes Classifier and the associative classifier with multiple confidence threshold values." Knowledge-based systems 23, no. 6 , PP. 598-604, .2010*.

[96] Li, X.-L., Bing Liu, and See-Kiong Ng, *"Negative training data can be harmful to text classification." In Proceedings of the 2010 conference on empirical methods in natural language processing, pp. 218-228. Association for Computational Linguistics, 2010*.

[97] Zhang, W., Taketoshi Yoshida, and Xijin Tang, *"A comparative study of TF* IDF, LSI and multi-words for text classification." Expert Systems with Applications 38, no. 3,pp 2758-2765, 2011*.

[98] Jiang, J.-Y., Ren-Jia Liou, and Shie-Jue Lee, *"A fuzzy self-constructing feature clustering algorithm for text classification." Knowledge and Data Engineering, IEEE Transactions on 23, no. 3 , pp.: 335-349, 2011*.

[99] Uğuz, H., *"A two-stage feature selection method for text categorization by using information gain, principal component analysis and genetic algorithm."Knowledge-Based Systems 24, no. 7, pp. 1024-1032, 2011*.

[100] Wan, C.H., Lam Hong Lee, Rajprasad Rajkumar, and Dino Isa, *"A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and support vector machine." Expert Systems with Applications 39, no. 15 , pp. 11880-11*, *2012.*

[101] Uysal, A.K., and Serkan Gunal, *"A novel probabilistic feature selection method for text classification." Knowledge-Based Systems 36, pp. 226-235, 2012*.

[102] Yang, J., Yuanning Liu, Xiaodong Zhu, Zhen Liu, and Xiaoxu Zhang, *"A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization." Information Processing & Management 48, no. 4 , pp. 741-754, 2012*.

[103] Jiang, S., Guansong Pang, Meiling Wu, and Limin Kuang, *"An improved K-nearest-neighbor algorithm for text categorization." Expert Systems with Applications 39, no. 1 , pp. 1503-1509, 2012*.

[104] Helvetic, D.S., T. Ahmed, and Sura Mahmood Abdullah., *"Hybrid Intelligent Techniques for Text Categorization." International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol 2 , pp. 23-40, 2014*.

[105] Selamat, A., Hidekazu Yanagimoto, and Sigeru Omatu, *"Web news classification using neural networks based on PCA." In SICE 2002. Proceedings of the 41st SICE Annual Conference, vol. 4, pp. 2389-2394. IEEE, 2002*.

[106] Billsus, D., and Michael J. Pazzani, *" A hybrid user model for news story classification". Springer Vienna, 1999*.

[107] [Xiao Luo Zincir-Heywood, A.N., *"combining word based and word co-occurrence based sequence analysis for text categorization." Machine Learning 2004*.

[108] Pekar, V., Michael Krkoska, and Steffen Staab, *"Feature weighting for co-occurrence-based classification of words." In Proceedings of the 20th international conference on Computational Linguistics, p. 799. Association for Computational Linguistics, 2004*.

[115] Bina, B., Mohamad Hasan Ahmadi, and Maseud Rahgozar, *"Farsi Text Classification Using N-Grams and Knn Algorithm A Comparative Study." InDMIN, pp. 385-390, 2008*.

[124] A. AleAhmad, H.A., E. Darrudi, M. Rahgozar, and F. Oroumchian, *"Hamshahri: A standard Persian text collection", [Online]. Available: http://ece.ut.ac.ir/dbrg/hamshahri/index.html*.

[125] "http://trec.nist.gov."

[126] Y.Yang, *"An Evaluation of Statistical Approaches to Text Categorization," Inf. Retr., vol. 1, pp. 69-90, 1999*.

Abstract

Due to increasing growth of electronic texts, such as news texts, the problem of developing an efficient classifier becomes an important issue in many text-related applications such as news websites. The most important issues in the domain of text classification are efficient feature selection and feature extraction. In previous works, different feature validation criteria have been proposed and used.

In this thesis, a feature validation criterion called E-Dominance is proposed. Using this criterion causes a significant reduction of number of selected features. Features used in this categorization, are called co-occurrence features wich have not been used in classification of Persian texts yet.

In Englisg text classification, researches which use co-occurrence features, usually use binary wheighening method. In this thesis is proposed a wheighening method for co-occurrence features which is called Co-occur TFIDF. IN this study, developer features are used to develop texts, Therefor the problem of class overlapping is solved partially.

Experiments show a significant improvement in the efficiency and accuracy of classification algorithms using E-Dominance criterion and co-occurrence features using Co-occur TFIDF weightening method.

Keywords: Persian text classification, feature selection, feature extraction, E-Dominance criterion, co-occurrence features, developer features, Co-occur TFIDF weightening method.

Shahrood University of Technology
Faculty of Computer Engineering

Persian News Classification Using Artificial Intelligence

Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science (M.Sc.)

Zahra Robati

Supervisor
Dr. Morteza Zahedi

Associate Supervisor
Marzieh Rahimi

Date: September 2014