تاریخ دریافت: ۹۷/۰۶/۲۱
تاریخ پذیرش: ۹۷/۱۲/۰۱

مروری بر روش‌ها و پژوهش‌های دسته‌بندی خودکار متون فارسی

هادی ویسی*
استادیار دانشکده علوم و فنون نوین، دانشگاه تهران
پست الکترونیکی: h.veisi@ut.ac.ir

پویان پارسافرد
کارشناسی ارشد پردیس بین‌المللی، دانشگاه تهران
پست الکترونیکی: pooyanparsafard@ut.ac.ir

چکیده

دسته‌بندی اسناد متنی یا تشخیص عنوان به فرآیند شناسایی خودکار موضوع یک سند متنی (مانند هنری، ورزشی، سیاسی، علمی و ...) گفته می‌شود که در کاربردهای مختلف پردازش زبان طبیعی مانند بازیابی اطلاعات و تحلیل متون مورد استفاده است. یک سامانه دسته‌بندی‌کننده خودکار متون، مشابه اغلب سامانه‌های بازشناسی الگو، از دو گام مهم استخراج ویژگی و دسته‌بندی تشکیل شده است. در این مقاله، مروری بر روش‌های رایج برای استخراج ویژگی و دسته‌بندی در این سامانه‌ها صورت گرفته و پژوهش‌هایی که در این حوزه برای زبان فارسی انجام شده است، مرور شده‌اند. همچنین، تحلیلی از نقاط قوت و ضعف روش‌های موجود و مقایسه کارهای صورت گرفته با همدیگر ارائه شده است.

واژه‌های کلیدی: دسته‌بندی متون فارسی؛ پردازش زبان طبیعی؛ مرور روش‌ها؛ استخراج ویژگی و دسته‌بندی

مقدمه

دسته‌بندی اسناد[1] که به آن تشخیص عنوان[2] نیز گفته می‌شود، بیانگر انتخاب برچسب یا موضوع برای اسناد متنی است. دسته‌بندی اسناد امروزه جزو موضوع‌های مهم و حائز اهمیت در پردازش زبان طبیعی[3] محسوب می‌شود، از آن جهت که نقش ویژه‌ای را در کنترل و اداره حجم روز افزون داده و اطلاعات متنی ایفا می‌کند [۱]. اسناد فاقد این نوع دسته‌بندی به نوعی دچار ابهام مفهوم اولیه در حجم بالای متون خواهند شد و سیستم را وادار به بررسی و خواندن تک‌تک این اسناد فاقد برچسب، توسط نیروی انسانی خواهد کرد. پتانسیل مصرف شده توسط نیروی انسانی به خدمت گرفته شده را می‌توان با الگوریتم‌ها و روش‌های دسته‌بندی اسناد جایگزین کرد تا بتوان از نیروی انسانی به شکل کاراتر و بهینه‌تر در دیگر عرصه‌های پردازش داده استفاده کرد. در راستای این عمل، زمان مفید سیستم و بازدهی آن نیز به طور چشم‌گیری افزایش خواهد یافت. دسته‌بندی اسناد همچنان که حجم اطلاعات متنی را کاهش می‌دهد، مدیریت و جهت‌دهی اسناد در راستای پردازش آن‌ها را آسان‌تر و بهینه‌تر می‌سازد تا در ادامه، بازیابی اسناد نیز با سهولت و سرعت بیشتری انجام بگیرد.

در عصر فناوری ارتباطات که بخش بزرگی از آن را

* نویسندهٔ مسئول

1- Document Classification
2- Topic Identification
3- Natural Language Processing (NLP)

فضای داده‌های متنی و نوشـــتاری تشکیل می‌دهد، فرآیند تولید اسناد با سهولت و سرعت بالا انجام می‌گیرد و همین امر باعث ایجاد روزافزون نمونه‌های فاقد برچســـب از این نوع داده‌ها شده است. این نوع داده‌ها که تحت عنوان متون دسته‌بندی نشده[4] از آن‌ها یاد می‌شود، عمل بررسی و تحلیل محتوای سند را دشوارتر می‌کند. از این رو با توجه به رشد سریع این نوع داده، دسته‌بندی اسناد روز به روز از اهمیت بالاتری برخوردار خواهد شد. توسعه و گسترش اینترنت که در پی آن داده‌های دیجیتال از جمله متن استخراج می‌شوند بســـتر پـــردازش داده را برای فناوری‌هـــای جدیدتر فراهم می‌کند. در مورد مســـئله کلان داده[5]، که همواره موضوعی بحرانی و جدّی در زمینه داده اســـت، یکـــی از راهکارهای ابتدایـــی متداول کـــه در مقدمه هر نوع پـــردازش داده خام صورت می‌گیرد، دسته‌بندی اسناد است.

کاربردهای بســـیار متنوعی برای دســـته‌بندی اســـناد وجود دارد [۱، ۲]. برای مثال می‌توان شناســـایی هرزنامهٔ الکترونیکی، جویشـــگرها، تحلیل محتوای دیجیتال، تعیین اعتبـــار، مشـــخص نمـــودن گروه‌هایی از مشـــتری‌ها که خصوصیات و علایق مشترکی دارند، تشخیص میزان تاثیر داروها و موثر بودن را نام برد.

مراحل و بخش‌های (اصلی) انجام یک فرآیند دسته‌بندی معمولاً بدین ترتیب است [۱]:

- پیش‌پـــردازش: پیش‌پـــردازش اولیـــن مرحلـــه فرآیند دســـته‌بندی به حســـاب می‌آید. عملیاتی همچون حذف ایست‌واژه‌ها[6]، ریشـــه‌یابی[7] کلمات، و رفتارهایی از این قبیل در ایـــن مرحله صورت می‌گیـــرد. از آن‌جایی که اعمال پیش‌پردازش در اکثر مواقع نتایج دســـته‌بندی را بهبود می‌بخشد، این مرحله از اهمیت بالایی برخوردار است به طوری که تحقیقات زیادی در زمینه تاثیرگذاری عملیات پیش‌پرداش بر کیفیت دســـته‌بندی، انجام گرفته است.
- اســـتخراج ویژگی: بعد از پیش پـــردازش، برای اجرای فعالیت‌های آتی نیازمند اســـتخراج عبارت‌های پردازش هســـتیم. در این مقاله، منظور از عبـــارت، همان عبارت پردازش[8] اســـت. هر ســـند حاوی متونی می‌باشـــد که می‌تـــوان آن را در عبارت‌هـــای بزرگ‌تـــر به‌صـــورت پاراگراف یا جمله دید، اما نهایتا هر جمله خود متشـــکل از عبارت‌های کوچک‌تری چون کلمه می‌باشـــد، پس اگر کلمـــه را کوچک‌ترین عبارت پـــردازش در نظر بگیریم، عبارت یـــا عبارت‌های مـــورد پردازش همـــان کلمه یا کلمات اســـتخراجی هســـتند. هر عبارت استخراجی یک ویژگی محســـوب می‌شود، اســـتخراج این ویژگی‌ها ما را بـــا مجموعه‌ای از آن‌ها مواجـــه می‌کند، به طوری که اگـــر بخواهیم بردار را نماینده ســـند در نظر بگیریم هر سند با برداری از مجموعه ویژگی‌های آن سند برابری می‌کند. بنابراین اسناد را می‌توان مجموعه‌ای متشکل از بردارهای ویژگی فرض کرد.
- انتخـــاب ویژگی[9]: این مرحله که متناســـب با نیاز روش دســـته‌بندی، به‌صـــورت اختیاری انجـــام می‌گیرد، عمل انتخاب مجموعه ویژگی‌های مفید از مجموعه ویژگی‌های اولیه را انجام می‌دهد که طی این روند ویژگی‌هایی را که همبستگی کمتری با بقیه ویژگی‌ها دارند حذف می‌نماید و با انتخاب زیرمجموعه‌های کاراتر از ویژگی، ســـرعت دســـته‌بندی افزایش پیدا می کند. این عمل کاهش بعد یا ابعاد، که در راستای حل مسئله افزایش ابعاد[10] (چندبعدی) انجام می‌گیرد منجر به کاهش فضای ویژگی و همچنین فضای مورد نیاز برای پردازش داده‌ها می‌گردد.
- دســـته‌بندی: در پایان، فرآیند دسته‌بندی با انتخاب یک دسته‌بند مناسب تحت یادگیری انجام می‌گیرد. یادگیری در دســـته‌بندی اســـناد متنی، در اکثر حالات به این دو صورت انجام می‌گیرد: یادگیری با نظارت[11] و یادگیری بدون نظارت[12]، که در دســـته‌بندی با نظارت سیستم یا

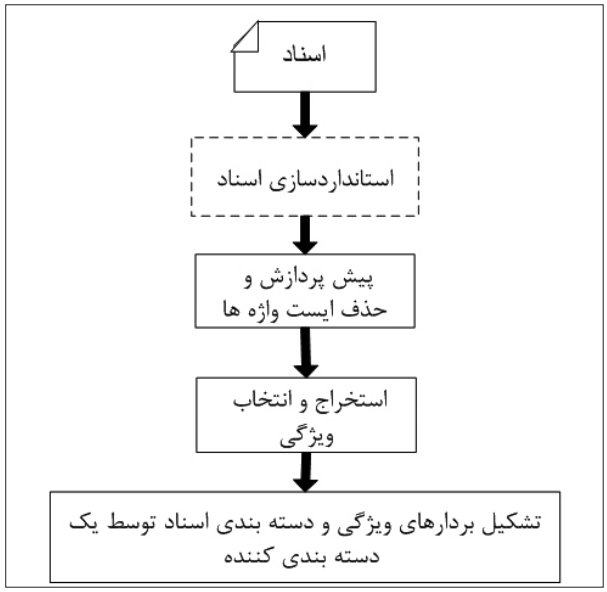
4- Uncategorized
5- Big Data
6- Stopwords
7- Stemming
8- Term
9- Feature Selection
10- High Dimensionality
11- Supervised
12- Unsupervised

دســـته‌بندی‌کننده، تحت داده‌هایی آموزش داده می‌شود که دارای برچســـب‌های از قبل تعیین شـــده باشـــند. به بیانی دیگر اســـناد به همراه برچســـب یا دســـته تعلق گرفته به آن، به منظور آموزش[13] سیستم مورد استفاده قـــرار می‌گیرند. اما در یادگیری بـــدون نظارت، معمولاً الگوریتم‌های خوشه‌بندی[14] هستند که نقش مهمی را ایفا می‌کنند. دســـته دیگری از روش‌هـــا نیز وجود دارند که به آن‌ها نیمه‌نظارتی[15] گفته می‌شـــود و معمولا استفاده همزمان از روش‌های بانظارت و بدون نظارت است.

در این مقاله، روش‌های رایج برای اســـتخراج ویژگی و دسته‌بندی در سامانه‌های دسته‌بندی متون مرور شده است و پژوهش‌هایی که در این حوزه برای زبان فارســـی انجام شده بررسی شده‌اند. برای این منظور، ساختار ارائه مطالب در این مقاله بدین شکل می‌باشد که در بخش دوم، به معرفی مراحـــل و گام‌های یـــک فرآیند دســـته‌بندی پرداخته که در راستای آن انواع دسته‌بندی‌کننده‌ها و روش‌های دسته‌بندی عنوان شـــده و ســـپس، روش‌های گوناگون انتخاب ویژگی شـــرح و توضیح داده خواهد شـــد. همچنین در ادامه همین بخش، انواع معیارهای ارزیابی و روش‌های مقایسه عملکرد فرآیندهای دسته‌بندی همراه با معادله و نحوه محاسبه آن‌ها بیان خواهد شد. در بخش سوم، مروری جامع و موشکافانه به روی کارهای انجام شده و پژوهش‌های گزارش شده در زمینه دسته‌بندی اسناد فارســـی صورت گرفته است که با مطالعه نتایج گردآوری شـــده حاصل از بررسی‌های انجام گرفته و همچنین یک دید کلی و مشـــاهده نتایج در یک نگاه بتوان در بخش چهارم، به نتیجه‌گیری و یک جمع‌بندی مفید، که هدف و ایده نهایی نوشتن این مقاله است، رسید.

۲- دسته‌بندی اسناد متنی

معماری و مراحل انجام یک فرآیند دســـته‌بندی اسناد، عناوین اصلی این بخش را تشکیل می‌دهند. در کنار توضیح



شکل ۱: مراحل کلی یک فرآیند دسته‌بندی اسناد

مراحل ســـعی شـــده اســـت که مهم‌ترین و برجسته‌ترین روش‌های انتخاب ویژگی، انواع روش‌های نمایش اســـناد و دســـته‌بندی‌کننده‌های مورد استفاده شده در دسته‌بندی اســـناد فارســـی معرفی شـــود و توصیفاتی در راســـتای چگونگی روند هـــر کدام ارائه شـــود. همچنین، معیارهای ارزیابی متداول دسته‌بندی به وضوح تشریح خواهند شد.

۲-۱ معماری کلی روش‌های دسته‌بندی اسناد

در فرآیند دســـته‌بندی اسناد، در اولین گام اسناد مورد پردازش تحت یک سری عملیات استانداردسازی یک‌دست می‌شـــوند. در گام بعدی پس از اعمال مرحله پیش‌پردازش و حذف ایســـت‌واژه‌ها، ویژگی‌های مـــورد نظر انتخاب و استخراج می‌شوند تا در نهایت با به نمایش درآوردن اسناد به شکل بردارهایی از ویژگی، دسته‌بندی اسناد توسط یک دسته‌بندی‌کننده مطلوب صورت گیرد. روند و مراحل کلی فرآیند دسته‌بندی اسناد در شکل ۱ قابل مشاهده می‌باشد.

اسناد دریافتی بعد از نرمال کردن و استاندارد شدن، پیش‌پردازش شـــده و ســـپس ویژگی‌های مناسب از سند استخراج شده و پس از آن کار دسته‌بندی انجام می‌شود. هرکدام از این مراحل در ادامه به‌صورت دقیق‌تری تشریح شده است.

13- Train
14- Clustering
15- Semi-Supervised

۲-۲ استانداردسازی اسناد و پیش‌پردازش

متون پیش از ورود به اولین مرحلهٔ پردازش (یعنی پیش‌پردازش) نیازمند یک سری اصلاحات اولیه می‌باشند [۱]. به بیان دیگر، اسناد قبل از این‌که وارد هرگونه پردازش و محاسبات اصلی شوند، بررسی می‌شوند که آیا استانداردها و معیارهای ورود به پردازش متن رعایت شده است یا خیر؟ بدین مفهوم که هر سند ممکن است بسته به منبع[۱۶] آن، ویژگی‌های نوشتاری یا متنی خاصی را دارا باشد، مانند یک سند خبری از یک خبرگزاری که بسته به ضوابط و شیوه نگارش خبرگزاری، از یک سری نویسه‌های ویژه و شکل مانند، در متن استفاده شده است. استانداردسازی متن، کمک بزرگی را به پردازش‌های متنی آتی می‌کند از آن جهت که استخراج اطلاعات ارزشمند از داده خام، هزینه و بار محاسباتی را به نحو مطلوبی کاهش داده و در نتیجه سرعت کار بر روی داده‌ها نیز افزایش می‌یابد. منظور از استانداردسازی، شامل موارد مختلفی است از جمله:

- یکدست کردن نحوه نوشتن کلماتی که دارای صورت نوشتاری مختلف دارد، مانند «مسئول و مسوول»، «میشود، می شود و می‌شود» یا «کتابها و کتاب ها و کتاب‌ها»
- یکسان کردن کدهای نویسه‌ها (کاراکترها) مانند «ک فارسی و عربی» و «ی فارسی و عربی»
- حذف نویسه‌های اضافی مانند بولت
- اصلاح فاصله‌گذاری علائم سجاوندی

ایست‌واژه‌ها به دسته‌ای از کلمات اطلاق می‌گردد که قرارگیری آن‌ها در فرآیند دسته‌بندی اسناد، بی‌اهمیت و بی‌ارزش می‌باشد، مثل حروف اضافه. حذف این عناصر در کنار جایگزینی کلمات با کلمه ریشه معادل آن‌ها (ریشه‌یابی)، فرآیند پیش‌پردازش را تشکیل می‌دهد.

۲-۳ استخراج و انتخاب ویژگی از اسناد متنی

در پردازش زبان طبیعی به‌صورت عمومی و در دسته‌بندی اسناد به‌صورت مشخص، متن اسناد برای پردازش باید به مدل فضای برداری (VSM[۱۷]) برده شود و یا اصطلاحاً از آن ویژگی استخراج شود. ویژگی‌های استخراج شده از متن را می‌توان شامل دو دسته کلی سبد کلمات[۱۸] و جاسازی کلمات[۱۹] دانست. در نمایش سبد کلمات، خود کلمات و یا نمایشی عددی از آن‌ها (مثلا به‌صورت دودویی و یا بر اساس شمارش تعداد آن‌ها) به‌عنوان بردار ویژگی متن در نظر گرفته می‌شود. این رویکرد به دلیل عدم در نظر گرفتن وابستگی بین کلمات و همچنین دارا بودن ابعاد بالای بردارهای ویژگی دارای ضعف هستند. در رویکرد جاسازی کلمات برای استخراج ویژگی، از یادگیری و برداری کردن هر کلمه یا هر سند استفاده می‌شود. برخی از روش‌های این دو گروه از استخراج ویژگی به‌صورت زیر هستند:

- روش‌های مبتنی بر سبد کلمات [۳]
  - ♦ بردارهای دودویی
  - ♦ فراوانی عبارت[۲۰] (TF)
  - ♦ فراوانی عبارت – معکوس فراوانی سند[۲۱] (TF-IDF)
- روش‌های مبتنی بر جاسازی کلمات
  - ♦ تحلیل معنایی پنهان[۲۲] (LSA) [٤]
  - ♦ تحلیل آماری معنایی پنهان[۲۳] (PLSA) [٥]
  - ♦ بردار کلمات مبتنی بر شبکه عصبی[۲٤] [٦، ۷]

در روش‌های مبتنی بر سبد کلمات، برای به نمایش در آوردن هر سند، تعداد n کلمه (معمولا بر اساس فراوانی تکرار در یک پیکره مانند داده‌های آموزش) را انتخاب کرده و به ازای هر سند یک وزن به آن کلمه داده می‌شود تا این بردار nبعدی نمایشی از آن سند باشد. این موضوع در رابطه (۱) برای سند d_k نشان داده شده است.

$$d_k = (w_{1.k}, w_{2.k}, \dots, w_{i.k}, \dots, w_{n.k}) \qquad (۱)$$

در این رابطه $w_{i.k}$ بیانگر کلمه iام در سند d_k است و یا

16- Source

17- Vector Space Model
18- Bag of Words
19- Word Embedding
20- Term Frequency
21- Term Frequency-Inverse Document Frequency
22- Latent Semantic Analysis
23- Probabilistic Latent Semantic Analysis
24- Neural Network based Word Vectors

در صورت استفاده از Nتایی[25]ها، بیانگر Nتایی iام است.

در نمایش برداری دودویی، مقادیر $w_{i.k}$ یک (در صورت حضور کلمه iام در سند d_k) و یا صفر (در صورت عدم حضور کلمه iام در سند d_k) است. در فراوانی عبارت، مقدار $w_{i.k}$ برابر با تعداد تکرار کلمه iام در سند d_k است که معمولا از مقدار نرمال شده آن (روی جمع تکرار کلمات و یا روی مقدار بیشنه تکرارها مانند رابطه (۲) استفاده می‌شود.

(۲)

$$tf_{i,k} = \frac{freq_{i,k}}{max_l freq_{l,k}} \quad \rightarrow \quad tf_{i,k} = freq(w_{i.k})$$

که در این معادله، $tf_{i,k}$ نمایانگر فراوانی عبارت iام در سند k و $max_l freq_{l,k}$ فراوانی پرتکرارترین عبارت در سند kام می‌باشد.

در روش فراوانی عبارت – معکوس فراوانی سند (TF-IDF) از ترکیب دو مقدار فراوانی عبارت (معادله ۲) و معکوس فراوانی سند (معادله ۳) به مقدار وزن حاصل شده در رابطه ٤ می‌رسیم [۳]. معادله ۳ بیان می‌کند idf_i معکوس (نرمال شده با لگاریتم) فراوانی تعداد اسنادی است که عبارت iام در آن رخ داده است، N تعداد همه سندها n_i و تعداد سندهایی که شامل امین عبارت هستند. ایده IDF این هست که اگر عبارتی (کلمه‌ای) در همه سندها تکرار شده باشد، آن عبارت (کلمه) کمکی به ما در تمایز بین دسته‌های مختلف نمی‌کند و عبارت‌هایی (کلماتی) مفیدتر هستند که فقط در تعداد محدودی سند تکرار شده باشند. به‌عنوان مثال کلمه «تورینگ» در متن‌های علمی و فناوری و کلمه «رونالدو» در متن‌های ورزشی بیشتر از متن‌های دیگر رخ می‌دهند و از این‌رو، این دو کلمه به ترتیب برای تمایز دسته‌های علمی و ورزشی مناسب‌تر هستند.

$$idf_i = \log \frac{N}{n_i} \qquad (۳)$$

$$TF - IDF_{i,k} = tf_{i,k} \times idf_i = \frac{freq_{i,k}}{max_l freq_{l,k}} \times \log \frac{N}{n_i} \qquad (٤)$$

روش‌های دیگری مانند آنتروپی نرمال شده برای استخراج ویژگی از متن وجود دارند که دارای منطقی مشابه TF-IDF هستند.

در دسته دوم روش‌ها، سعی می‌شود هر کلمه یا هر سند به نحوی به یک بردار عددی تبدیل شود که هم تعداد ویژگی‌ها کمتر باشند و هم بردارها حاوی اطلاعاتی دیگر (معنایی، نحوی و ...) در مورد آن کلمه یا سند باشند به نحوی که کلمات مشابه (مثلا از نظر معنایی) دارای بردارهای نزدیک به هم باشند. به‌عنوان مثال بردار کلمه «خانه» با بردار کلمه «منزل» نزدیک به همدیگر باشند.

در میان روش‌های جاسازی کلمات برای استخراج ویژگی، روش تحلیل معنایی پنهان (LSA) [٤] که آن را نمایه‌سازی معنایی پنهان (LSI[26]) می‌نامند به‌عنوان یکی از روش‌های جبری–آماری برای تبدیل کلمات و اسناد به بردار ویژگی مورد استفاده قرار می‌گیرد. این روش سعی در یافتن معنی پنهان کلمات (عنوان یا موضوع سند) در سندها دارد. در این روش، فرض بر این است که کلماتی که به طور همزمان در یک سند با موضوع مشخص رخ می‌دهند، از نظر معنایی به هم مرتبط هستند و سندهایی که دارای موضوع مشابهی هستند، حاوی کلمات مشابهی هستند. در این روش، در یک سند، کلمات قابل مشاهده هستند ولی عنوان/موضوع آن پنهان (Latent) است. این روش با بهره‌گیری از تجزیه مقادیر تکین (SVD[27])، ماتریس عبارت–سند ساخته شده از فراوانی تکرار عبارت‌ها در اسناد را تجزیه می‌کند و با نگاشت بردار فراوانی سند یا عبارت به زیرفضاهای کوچک‌تر، بردار سند یا عبارت را پیدا می‌کند. اگر A ماتریس عبارت–سند دادگان موجود باشد، تجزیه SVD آن به‌صورت رابطه (۵) خواهد بود که در آن T ماتریس بردارهای کلمات، D ماتریس بردارهای مستندات و S یک ماتریس قطری حاوی مقادیر ویژه است که جهت کاهش بعد استفاده می‌شود.

$$A = TSD^T \qquad (۵)$$

با انتخاب k مقدار از بیشترین مقادیر ویژه (از ماتریس)

25- N-gram

26- Latent Semantic Indexing (LSI)
27- Singular Value Decomposition

قطری این ماتریس اســتخراج شــده و ابعاد ماتریس A با توجه به رابطه ٦ کاهش می‌یابد.

$$A \approx A_k = T_k S_k D_k^T \quad (٦)$$

در واقــع، روش LSA متــون و کلمــات را به یک فضای معنایی مشــترک تصویر می‌کند و آن را می‌توان یک روش کاهش ابعاد بردارهای ویژگی دانست که از ایده‌ای مشابه به تحلیل اجزای اصلی[28] [۸] بهره می‌برد. با این روش، بردارهای حاصل (در حدود صد ویژگی) از بردارهای سبد کلمات (در حدود چند هزار ویژگی) حاصل می‌شود که بردارهای عبارات حاوی اطلاعات معنایی پنهان آن عبارت‌ها هست.

نســخه احتمالــی ایــن روش PLSA [٥] نــام دارد که متغیرهای پنهــان (عنوان/موضوع) را بــا متغیرهای قابل مشــاهده (ســندها و کلمات) مرتبط می‌کند. مشکل اصلی اســتفاده از روش‌هایــی همچــون LSA این اســت که این روش‌ها بدون نظارت هستند یعنی نسبت به توزیع دسته‌ها کور هســتند و آن را در نظر نمی‌گیرند. بنابراین ویژگی‌ای که با روش LSA انتخاب می‌شود، لزوما در راستای توزیع دسته اســنادی که بهترین جداکننده هستند، قرار نمی‌گیرد [۱]. استفاده از روش‌های پیش‌برنده[29] در کنار ویژگی‌های مفهومی تهیه شده از روش PLSA بدون نظارت باعث بهبود دقت این نوع دسته‌بندی شده‌اند [۹].

تحلیل معنایی پنهان احتمالاتی تجزیه را توسط یک مدل آماری به نام مدل وجه[30] آغاز می‌کند. مشاهدات به‌صورت (w, d) از کلمات و مستندات موجود هستند. احتمال رخداد کلمه در سند با استفاده از رابطه ۷ محاسبه می‌گردد که در آن c موضوعات موجود بوده و احتمالات شــرطی توسط الگوریتم درست‌نمایی بیشینه[31] محاسبه می‌شوند [٥].

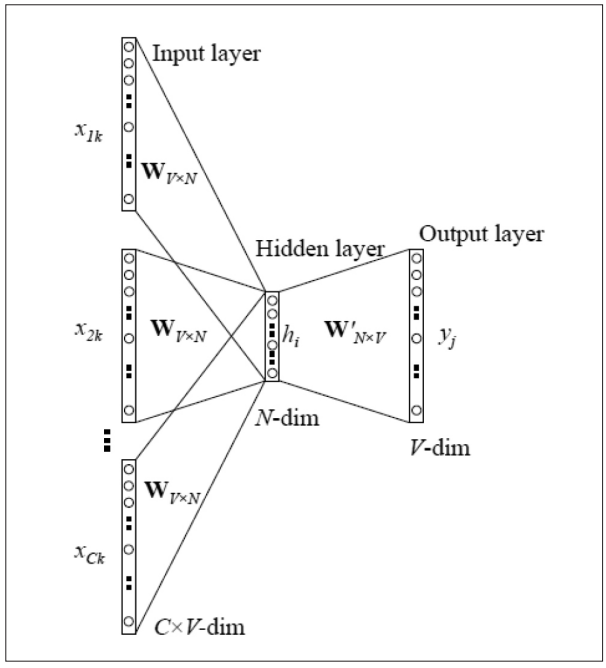
$$P(w,d) = \sum_c P(c)P(d|c)\,P(w|c) = P(d)\sum_c P(c|d)\,P(w|c) \quad (٧)$$

روش بردار کلمات، یکــی از روش‌های نوین بازنمایی متون برای تبدیل کلمات به بردار ویژگی با استفاده از توان یادگیری شبکه‌های عصبی مصنوعی[32] است که موفقیت آن در مدل‌ســازی مختلف پردازش زبــان طبیعی، آن‌ها را به روشی اصلی تبدیل کرده است. در این رویکرد، با استفاده از پیکره‌هــای متنی[33] برای هر کلمه یک بــردار (در حدود یک‌صد تا سی‌صد ویژگی) به‌صورت بی‌نظارت یاد گرفته می‌شود که به دلیل استفاده از اطلاعات بافت[34]، بردارهای حاصل دارای اطلاعات معنایی و نحوی هستند [۱۰]. منظور از اطلاعات بافت بــرای یک کلمه، اطلاعات کناری آن کلمه در متــن مانند کلمات اطراف آن اســت. هرچند روش‌های مختلفــی برای ایجاد بردار کلمات وجــود دارد اما از میان آن‌ها سبد کلمه پیوسته (CBOW)[35] و پرش-چندتایی[36] [۱۱] رایج‌تر بــوده و در ابزارهای متن‌بازی مانند Word2Vec[37] در دسترس هستند.

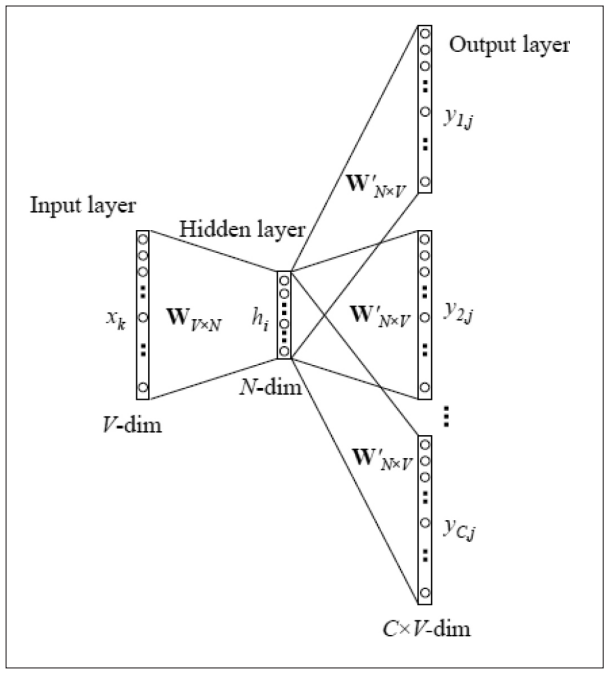
مــدل ســبد پیوســته کلمــات از بافــت متنــی [w(t-2),w(t-1),w(t+1),w(t+2,] ... ...]، بــرای پیش‌بینی کلمه w(t) اســتفاده میشــود. این مدل از مدل زبانی شبکه عصبی پیش‌خور گرفته شده است اما در آن لایه غیر خطی پنهان حذف شده است. ورودی شبکه، بردارهای کدگذاری شده کلمات موجود در بافت با کدگذاری یکی فعال[38] هستند. وزن بین لایه ورودی و لایه خروجی به‌صورت ماتریس W با ابعاد N*V اســت که در آن V تعــداد واژگان و N ابعاد (تعداد ویژگی) هر کلمه اســت. با اســتفاده از این ماتریس وزنــدار میتوان برای هر کلمه موجود در واژگان یک وزن محاسبه کرد و هر ســطر ماتریس W بازنمایی N بعدی از بردارهای کلماتی اســت که در ورودی دریافت شــده‌اند. ســپس در لایه پنهان میانگین بردارهای کلمات محاســبه میشود. از لایه پنهان به لایه خروجی نیز ماتریس وزن‌دار W' با ابعــاد N*V وجود دارد. بــردار حاصل از خروجی این ماتریس وزنــدار به‌عنوان توزیع احتمالاتی کلمه هدف

28- Principal Component Analysis
29- Boosting
30- Aspect Model
31- Maximum Liklihood

32- Artificial Neural Network
33- Text Corpus
34- Context Information
35- Continuous Bag of Words
36- Skip-gram
37- https://code.google.com/p/word2vec/
38- One-Hot Eencoding



شکل ۲: معماری شبکه عصبی سبد پیوسته کلمات [۱۲]



شکل ۳: معماری شبکه عصبی پرش چندتایی [۱۲]

برگردانده می‌شود. شکل ۲ معماری مدل سبد پیوسته کلمات را نشان می‌دهد.

روش پرش چندتایی مشابه با مدل سبد پیوسته کلمات است اما هدف آن تعیین پارامترهایی است که احتمال کلمه جاری را در بافت متن به حداکثر برساند. معماری این مدل در شکل ۳ آورده شده است که در آن با استفاده از کلمهٔ ورودی w(t)، بافت متن [w(t-2),w(t-1),w(t+1),w(t+2,] ... ...] را پیش‌بینی می‌شود. به طور دقیقتر، از کلمه جاری به‌عنوان ورودی یک دسته‌بند خطی –لگاریتمی با لایه بازتاب پیوسته برای پیش‌بینی کلمات بافت متن در یک اندازه پنجره قبل و بعد از کلمه جاری بهره برده شود. ورودی شبکه، بردار کد گذاری شده کلمه ورودی است و از لایه ورودی به لایه پنهان ماتریس وزندار W مرتبط با بردار کلمه ورودی وجود دارد.

روش‌های مرتبط دیگری جهت ایجاد بازنمایی برداری کلمات با رویکرد جاسازی کلمات وجود دارد که از آن‌ها می‌توان به روش‌های مبتنی بر خوشه‌بندی مانند خوشه‌بندی براون[۳۹][۱۳]، روش‌های مبتنی بر شبکه عصبی برای ایجاد مدل زبانی[۴۰][۱۴]، روش تخصیص پنهان دیریکله (LDA)[۴۱] [۱۵]، روش مدل‌سازی عنوان تنک (STC)[۴۲] [۱۶] و روش بردار جهانی[۴۳] [۱۷] اشاره کرد.

39- Brown Clustering
40- Neural Network Language Modeling

۲-۳-۱ انتخاب ویژگی

در حالتی که از رویکرد سبد کلمات به‌عنوان ویژگی استفاده می‌کنیم، با ابعاد بالای بردارهای ویژگی روبه‌رو هستیم که با افزایش بردارهای ویژگی مشکل افزایش ابعاد (چندبعدی) بردارهای ویژگی پدیدار می‌گردد که حل این مسئله خود به بهبود کیفیت دسته‌بندی اسناد نیز کمک شایانی خواهد نمود. راه حل مورد استفاده برای این مشکل استفاده از روش‌های انتخاب ویژگی است. در کل روش‌های انتخاب ویژگی به دو دسته تقسیم‌بندی می‌شوند، روش‌های پوشه‌ای[۴۴] و روش‌های تصفیه‌ای (فیلتری)[۴۵]، که روش‌های پوشه‌ای از الگوریتم‌های فراگیری اکتشافی[۴۶] استفاده می‌کنند و با تخمین کارایی سیستم نتیجه‌گیری می‌کنند که آیا مجموعه ویژگی‌های انتخاب شده سودمند هستند یا خیر. روش‌های پوشه‌ای جواب‌های بهتری را برای مسائل مقیاس کوچک ارائه می‌دهند اگر چه

41-Latent Dirichlet Allocation (LDA)
42- Sparse Topical Model (STC)
43- Global Vector For Word Representaion
44- Wrapper Methods
45- Filter Methods
46- Heuristic Learning Algorithms

زمان محاسباتی بالایی را می‌طلبند تا آن‌جا که در مسائلی با تعداد ویژگی‌های بالا تقریبا ناکارآمد به نظر می‌آیند. در مقابل روش‌های تصفیه‌ای زمان کمتری را بابت مسائل با مقیاس‌های بزرگ‌تر مصـــرف می‌کنند پس می‌توانند نقش مفیدتری را نســـبت به روش‌های پوشه‌ای، در دسته‌بندی اسناد بازی کنند. روش‌های تصفیه‌ای به جای به کارگیری الگوریتم‌های فراگیری اکتشافی، از قدرت تمایز ویژگی‌ها و دیگر اندازه‌گیری‌های آماری استفاده می‌کنند.

همان‌گونه که اشـــاره شـــد، روش‌هـــای متنوعی برای انتخاب ویژگی وجود دارند که هر کدام بسته به نوع روابط و ضوابط مدون شـــده، اقدام به انتخاب ویژگی می‌نمایند. در زیر به چند نمونه از بارزترین آن‌ها اشاره شده است:

- خی‌دو آماری[47] (χ^2)
- بهره اطلاعات[48] (IG)
- اطلاعات متقابل[49] (MI)
- فراوانی سند[50] (DF)
- اهمیت عبارت[51] (TS)

در فرایند انتخـــاب ویژگی به روش خی‌دو آماری (χ^2)، یک ویژگی بر اســـاس میزان همبســـتگی‌اش با یک دســـته یا طبقه انتخاب می‌شـــود [۱۸]. نحوه محاســـبه این معیار به‌صورت رابطه ۸ تعریف می‌شود [۱۹]:

(۸)

$$\chi^2(t_i, c_j) = \frac{N \times (a_{ij}d_{ij} - b_{ij}c_{ij})^2}{(a_{ij} + b_{ij}) \times (a_{ij} + c_{ij}) \times (b_{ij} + d_{ij}) \times (c_{ij} + d_{ij})}$$

که در معادله ۸، N معرف تعداد اســـناد اســـت و سایر متغیرها به‌صورت زیر هستند:

a_{ij} تعداد سندهایی است که حاوی ویژگی t_i و متعلق به دسته c_j می‌باشند،

b_{ij} تعداد اســـنادی اســـت که حاوی ویژگی t_i نیست اما متعلق به دسته c_j می‌باشند،

c_{ij} تعداد سندهایی اســـت که حاوی ویژگی t_i هست اما متعلق به دسته c_j نیست،

d_{ij} تعداد اسنادی است که نه حاوی ویژگی t_i است، و نه متعلق به دسته c_j .

انتخـــاب ویژگی بـــه روش بهره اطلاعـــات (IG)، تعداد بیت‌های اطلاعاتی را می‌شـــمرد که از آن‌ها برای پیش‌بینی دسته در زمان حضور یا عدم حضور یک ویژگی در سند، استفاده می‌کند [۲۰]. رابطه آن به‌صورت زیر است:

$$IG(t_i) = -\sum_{j=1}^{C} P(c_j) \times \log P(c_j) + P(t_i) \times \sum_{j=1}^{C} P(c_j|t_i) \times \quad (۹)$$
$$\times \log P(c_j|t_i) + P(\bar{t}_i) \times \sum_{j=1}^{C} P(c_j|\bar{t}_i) \times \log P(c_j|\bar{t}_i)$$

با توجه به رابطه ۹، C نمایانگر تعداد دســـته‌های سند و $P(c_j)$ احتمال رخداد دســـته c_j، $P(t_i)$ احتمال رخداد ویژگی t_i، $P(\bar{t}_i)$ احتمال رخ ندادن ویژگی احتمال شرطی رخداد دســـته c_j، وقتی که ویژگی t_i هم رخ دهد، $P(c_j|\bar{t}_i)$ احتمال شـــرطی رخداد دسته c_j، وقتی که ویژگی t_i هم رخ ندهد، است.

روش اطلاعات متقابل (رابطه ۱۰)، یک استاندارد همبستگی بین ویژگی و دســـته است و می‌تواند برای تشخیص ارتباط ویژگی‌ها با یک دســـته خاص بـــه کار رود. اندازه اطلاعات متقابل با استفاده از نظریهٔ اطلاعات حاصل شده که آن‌ها را بین دسته‌ها و ویژگی‌ها مدل‌سازی می‌کند و باید متذکر شد کـــه این روش، ضعفی را همراه دارد که آن، تمایل به انتخاب ویژگی‌های تنک[52] است که $P(t_i)$ پایینی دارند [۱۹].

$$MI(t_i) = \sum_{j=1}^{C} P(c_j) \times \log \frac{P(t_i|c_j)}{P(t_i)} \quad (۱۰)$$

در رابطـــه ۱۰، $P(t_i)$ احتمال رخداد ویژگی t_i، $P(t_i|c_j)$ احتمال شـــرطی رخداد ویژگی t_i وقتی که دسته c_j هم رخ دهد و $P(c_j)$ احتمال رخداد دسته c_j است.

در انتخاب ویژگی به روش اهمیت عبارت (TS)، اهمیت یک عبارت t با یک دسته c [۲۱] این‌گونه تعریف می‌شود:

(۱۱)

$$TS(t,c) = \begin{cases} \dfrac{\log(\max\{P(t), P(c)\})}{1 - \log(\min\{P(t), P(c)\})}, & if\ P(t,c) = 0 \\ \dfrac{\log(\max\{P(t), P(c)\}) - \log P(t,c)}{1 - \log(\min\{P(t), P(c)\})}, & Otherwise \end{cases}$$

47- Chi Square Statistic
48- Information Gain
49- Mutual Information
50- Document Frequency
51- Term Significance

52- Sparse

که در اینجا، P(t) بیانگر احتمال این که یک سند حاوی عبارت t باشد، P(t) احتمال این که یک سند متعلق به دسته c باشـــد و P(t, c) احتمال این که یک سند متعلق به دسته c باشد و حاوی عبارت t باشد، است.

برای اندازه‌گیری ارتباط یک عبارت در فضای سراسری ویژگی، همه امتیازات معین دسته در نظر گرفته می‌شوند، با فرض این که تعداد دســـته‌ها باشد، وزن یک عبارت در همه دسته‌ها این‌گونه محاسبه می‌شود (رابطه ۱۲).

$$TS_{max}(t) = \max_{i=1,\dots,m} \{TS(t, c_i)\} \qquad (۱۲)$$

همـــه ویژگی‌هـــا، با توجـــه بـــه وزن اهمیـــت عبارت TS_{max} (t) به ترتیب کاهشی رتبه‌بندی خواهند شد و سپس به تعداد از پیش تعیین شده، عبارت‌هایی که بالاترین وزن‌ها را دارند برای فرآیند دسته‌بندی انتخاب می‌شوند. الگوریتم زیر فرآیند انتخاب ویژگی اهمیت عبارت را برای دسته‌بندی بیان می‌کند و همچنین این الگوریتم، دارای ملاحظاتی است که در [۲۱]، بیان شده‌اند.

۲-٤ روش‌هـای دسـته‌بندی بـرای اسـتفاده در دسته‌بندی اسناد متنی

به‌صـــورت کلی همه دســـته‌بندی کننده‌های موجود را می‌توان برای کاربرد دسته‌بندی متن هم به کار برد که هر کدام به تنهایـــی پارامترهای قابل تنظیم مختص به خود را دارند که تغییر در هر پارامتر الگوریتم، ممکن اســـت نتایج دســـته‌بندی را کاملا متفاوت ســـازد. گاهی اوقات چندین روش دسته‌بندی (دسته‌بندی‌کننده‌ها) به روی یک مجموعه داده مشترک نتایج خوبی را ارائه می‌دهند، در این موقعیت زمان مورد نیاز برای آموزش سیستم است که تعیین کننده عملکرد بهتر اســـت. تا به‌حال تلاش‌های بسیاری در زمینه دسته‌بندی اسناد انجام گرفته است که دسته‌بندی‌کننده‌های متفاوت با درصدهای موفقیت متفاوت گزارش شده است، از جمله برجسته‌ترین این دســـته‌بندی‌کننده‌ها می‌توان به موارد زیر اشاره کرد:

- بیز ساده[۵۳]
- K نزدیک‌ترین همسایه (KNN[۵۴])
- ماشین بردار پشتیبان (SVM[۵۵])
- شبکه عصبی مصنوعی (ANN)
- مدل مخفی مارکوف[۵۶] (HMM)
- درخت تصمیم[۵۷]
- Nتایی آماری[۵۸]
- قوانین فازی[۵۹]

مبنای دســـته‌بندی به روش بیز ســـاده، احتمال‌هاست و احتمال حضور کلمات در یک ســـند، تعیین کننده دســـته آن سند هســـتند. توجه دسته‌بندی به روش K نزدیک‌ترین همســـایه به اسناد آموزش اســـت و روش آن بدین ترتیب اســـت که ابتدا شـــباهت اســـناد آموزش با ســـند ورودی محاســـبه می‌شـــود، ســـپس به تعداد K تا ســـند از اسناد آموزش که به ســـند ورودی شبیه‌تر هســـتند، همسایگان ســـند مورد نظر محسوب می‌شـــوند. روش ماشین بردار پشتیبان یکی از موفق‌ترین روش‌های دسته‌بندی است که تا به حال برای دســـته‌بندی متون گزارش شـــده است و با پیدا کردن بهترین مرز تصمیم‌گیری بین دسته‌های مختلف (با شروع از دســـته‌بندی دو دســـته‌ای) از روی مجموعه داده آمـــوزش، یـــک اَبَرصفحـــه[۶۰] بین این هر دو دســـته به‌دســـت می‌آورد. در این روش، سعی می‌شود فاصله بین صفحات مرزی دو دسته و اَبَرصفحه دسته‌بندی بیشینه[۶۱] باشد. شـــبکه عصبی مصنوعی یک مدل ریاضی است که ســـاختار و پردازش‌های مغز انسان را شبیه‌سازی می‌کند و به‌صورت جلورو[۶۲] یا بازگشـــتی[۶۳] استفاده می‌شود. یک شبکه عصبی معمولا شـــامل لایه ورودی (که ویژگی‌های ســـند یا کلمات را دریافت می‌کند)، لایه خروجی (که بیانگر دســـته‌ها یا موضوعات مختلف اســـت) و تعدادی لایه‌های مخفی اســـت [۲۲]. از شـــبکه‌های موفق مورد استفاده در

53- Naive Bayes

54- K-Nearest Neighbors
55- Support Vector Machines
56- Hidden Markov Model
57- Decision Tree
58- Statistic N-Gram
59- Fuzzy Rules
60- Hyperplane
61- Maximum
62- Forward
63- Recurrent

دسته‌بندی متون می‌توان به شبکه جلورو پرسپترون چند لایه (MLP)[64] و شبکه بازگشتی حافظه کوتاه مدت ماندگار (LSTM)[65] اشاره کرد. روش مدل مخفی مارکوف به‌عنوان روشی موفق در مدل‌سازی داده‌های متوالی مانند گفتار، می‌تواند با مدل‌سازی احتمالاتی دسته‌ها نسبت به دسته‌بندی متون از روی مشاهدات (کلمات) اقدام کند [۲۳]. تصمیم و غلبه[66]، روند کاری دسته‌بندی به روش درخت تصمیم را تشکیل می‌دهد به طوری که داده‌های آموزش از قبل دسته‌بندی شده توسط عبارت تقسیم‌بندی می‌شوند. انتخاب عبارت را می‌توان بر عهده معیار انتخاب ویژگی بهره اطلاعات، یا معیارهای گوناگون دیگر گذاشت. پس در ادامه مجموعه آموزش به دو زیرمجموعه تقسیم می‌شود، یکی زیرمجموعه‌ای که شامل عبارت می‌شود و دیگری زیرمجموعه‌ای که عبارت را شامل نمی‌شود. فرآیند مذکور را دوباره برای آن دو زیر مجموعه تکرار می‌کنیم. به همین شکل، ادامه اسناد را تقسیم‌بندی می‌نماییم و این روند تا زمانی تکرار می‌شود و ادامه پیدا می‌کند که همه اسناد موجود در یک زیرمجموعه متعلق به یک دسته باشند. گره[67] آخر تبدیل به برگ می‌شود به عبارتی دیگر، فرآیند تقسیم و غلبه را تا زمانی ادامه می‌دهیم که دیگر عبارتی برای تقسیم‌بندی زیرمجموعه‌ها موجود نباشد [۱].

۲-۵ معیارهای ارزیابی سامانه‌های دسته‌بندی اسناد متنی

برای معرفی معیارهای ارزیابی دسته‌بندی اسناد، ابتدا باید پارامترهای مورد استفاده در این معیارها را بشناسیم. این معیارها شامل برچسب‌هایی است که نشان می‌دهد آیا سند مورد آزمون در دسته مورد نظر برچسب‌گذاری شده است یا خیر، و آیا سند مورد آزمون در دسته مورد نظر دسته‌بندی شده است یا خیر. با این توصیفات، می‌توان چهار معیار ارزیابی دسته را این‌گونه تعریف کرد:

- فراخوانی j دسته که همان $Recall(c_j)$ است.
- دقت دسته که همان $Precision(c_j)$ است.
- معیار دسته که همان $F-Measure(c_j)$ است.
- درستی دسته که همان $Accuracy(c_j)$ است.

$$Recall(c_j) = \frac{A_j}{A_j + B_j} \qquad (۱۳)$$

در این معادله، معیار فراخوانی دسته j درصد سندهای درست دسته‌بندی شده، در میان همه سندهایی است که باید به دسته هدف[68] تخصیص می‌یافتند.

$$Precision(c_j) = \frac{A_j}{A_j + C_j} \qquad (۱۴)$$

با توجه به معادله ۱۴، معیار دقت دسته j درصد سندهای درست دسته‌بندی شده، در میان همه سندهایی است که به دسته هدف تخصیص یافته‌اند.

$$F-Measure(c_j) = \frac{2P(c_j) \times R(c_j)}{P(c_j) + R(c_j)} \qquad (۱۵)$$

و در معادله ۱۵، معیار اندازه F دسته j یک میانگین هارمونیک[69] از $Recall(c_j)$ و $Precision(c_j)$، می‌باشد.

$$Accuracy(c_j) = \frac{A_j + D_j}{A_j + B_j + C_j + D_j} \qquad (۱۶)$$

در معادله ۱۶، درستی، بیانگر درصد تعداد سندهای درست دسته‌بندی شده است.

برای تجمیع نتایج روی همه نمونه‌ها، از میانگین‌گیری استفاده می شود که این تجمیع می‌تواند میانگین‌گیری ریز[70] یا میانگین‌گیری درشت[71] باشد. معیارهای ارزیابی ریز که از معادلات ۱۳، ۱۴ و ۱۵ مشتق می‌شوند، این‌گونه تعریف می‌شوند:

$$R_{micro} = \frac{\sum_{j=1}^{c} A_j}{\sum_{j=1}^{c}(A_j + B_j)} \qquad (۱۷)$$

$$P_{micro} = \frac{\sum_{j=1}^{c} A_j}{\sum_{j=1}^{c}(A_j + C_j)} \qquad (۱۸)$$

$$F_{micro} = \frac{2P_{micro} \times R_{micro}}{P_{micro} + R_{micro}} \qquad (۱۹)$$

64- Multi-Layer Perceptron
65- Long Short-Term Memory
66- Decision and Conquer
67- Node
68- Target Category
69- Harmonic Mean
70- Micro Averaging
71- Macro Averaging

در میانگین‌گیری درشت وزن همه دسته‌ها برابر در نظر گرفته می‌شود و به‌صورت رابطه‌های زیر محاسبه می‌شوند.

$$R_{macro} = \frac{1}{C} \times \sum_{j=1}^{c} R(c_j) \quad (۲۰)$$

$$P_{macro} = \frac{1}{C} \times \sum_{j=1}^{c} P(c_j) \quad (۲۱)$$

$$F_{macro} = \frac{2P_{macro} \times R_{macro}}{P_{macro} + R_{macro}} \quad (۲۲)$$

$$Accuracy_{macro} = \frac{1}{C} \times \sum_{j=1}^{c} Accuracy(c_j) \quad (۲۳)$$

۳- مروری بر پژوهش‌های انجام شده در دسته‌بندی اسناد فارسی

توضیحات مراحل و همچنین روند کارها و پژوهش‌های برجسته انجام شده در زمینه دسته‌بندی اسناد فارسی، همراه با گزارش بهترین نتایج به‌دست آمده توسط آن‌ها، در این بخش گردآوری شده‌اند تا با هدف تحلیل و بررسی روش‌ها و الگوریتم‌های استفاده شده، بتوان در انتهای این مقاله یک نتیجه‌گیری مناسب از کارهای انجام شده در زمینه زبان فارسی داشته باشیم.

هر زبان طبیعی، پیچیدگی‌ها و قواعد خاص خود را دارد و زبان فارسی نیز از این قاعده مستثنی نیست. با توجه به ساختار زبان فارسی، ویژگی‌هایی در این زبان وجود دارد که کار دسته‌بندی را (نسبت به زبان انگلیسی) کمی دشوارتر می‌سازند. به‌عنوان مثال، برای اشاره به چند نمونه از این دشواری‌ها می‌توان بحث تفکیک کردن عبارت‌ها (عبارت‌های پردازش) و یا تشخیص فضاهای خالی (واقعی) را مطرح کرد. در زمینه دسته‌بندی اسناد تا به امروز، مطالعات مختلفی به روی اسناد فارسی صورت گرفته است. اما تا آن‌جایی که سعی شده بیشتر تحقیقات و بررسی‌های برجسته‌ای (در محدوده دسته‌بندی اسناد فارسی) که تا الان صورت گرفته را، در این پژوهش با ذکر نتایج آورده‌ایم. آزمایش‌های دسته‌بندی به روی زبان فارسی با سیستم‌های دسته‌بندی متفاوت انجام گرفته است، که با توجه به تنوع حضور دسته‌بندی‌کننده‌های مختلف و همچنین استفاده از انواع گوناگون روش‌های انتخاب ویژگی و ترکیب و تلفیق الگوریتم‌های کاربردی دیگر، نتایج گسترده‌ای به‌دست آمده است. مشابه کار جاری، برخی کارهای مروری مختصر به زبان فارسی وجود دارد [۲۴] که فقط به مرور روش‌های یادگیری بسنده کرده‌اند و کارهای فارسی را بررسی نکرده‌اند.

مجموعه آزمایش‌های دسته‌بندی مرتبط با هر دسته‌بندی‌کننده در بخش مربوطه بررسی شده‌اند. سپس در انتهای فرآیند بررسی همه این آزمایش‌ها به روی اسناد فارسی، نگاهی کلی به تاثیرات روش‌های مختلف مورد استفاده و نحوه استفاده از آن‌ها بر کیفیت دسته‌بندی اسناد فارسی خواهیم داشت.

۳-۱ k نزدیک‌ترین همسایه (KNN)

روش KNN که یکی از دسته‌بندی‌کننده‌های معمول است، درصد بزرگی از آزمون‌های انجام شده در زمینه دسته‌بندی را به خود اختصاص داده است. این روش دسته‌بندی که به تنهایی قادر به ارائه یک دسته‌بندی خوب و مطلوب نیست، معمولاً با روش‌ها و الگوریتم‌های دیگر به کار می‌رود. آزمون KNN به روش‌های مختلف بر روی اسناد فارسی صورت گرفته است. در ادامه به ذکر شرایط، روش‌ها (روش انتخاب ویژگی و ...) و پارامترهای به کار گرفته در هر آزمون می‌پردازیم.

آزمایش انجام گرفته بر روی مجموعه داده فارسی همشهری، دسته‌بندی به روش k نزدیک‌ترین همسایه و انتخاب ویژگی TF-IDF می‌باشد. طی بررسی [۲۵] این روش دسته‌بندی نیز به مانند روش ماشین بردار پشتیبان، که هر دو بر روی یک مجموعه داده مشترک آزمایش شده‌اند، نتایج خوبی را در بر داشته است و بیانگر تاثیر بسزای تعداد ویژگی در روند نتایج است.

روش دیگر استفاده از مقدار kهای متفاوت است.

بررسی نتایج این روش نشان می‌دهد که هر چقدر مقدار k به سمت اعداد پایین‌تری میل کند نتایج بهتری را در پیش خواهیم داشت. به عبارتی دیگر، اجرای الگوریتم به منظور پیش‌بینی دسته نمونه جدید، هر چقدر با تعداد همسایگان کمتری صورت بگیرد نتایج بهتری را در پی خواهد داشت.

فاکتور سوم مورد بررسی معیار فاصله می‌باشد، این معیار با استفاده از روابط شباهت، میزان شباهت یا عدم شباهت نمونه‌ها را نسبت به یک‌دیگر می‌سنجد. بررسی‌ها نشان می‌دهد که معیار فاصله کازین، بهترین عملکرد را در محاسبه میزان شباهت بین نمونه‌های آموزش و آزمون ارائه می‌دهد و بعد از آن، معیار فاصله همبستگی نیز معیار مفیدی برای اندازه‌گیری شباهت مورد نظر است.

روش دسته‌بندی k نزدیک‌ترین همسایه و روش ماشین بردار پشتیبان، با توجه به این که به روی یک مجموعه داده مشترک فارسی، تحت یک تحقیق مشترک [۲۵] مورد آزمون و بررسی قرار گرفتند، هر دو نتایج خوبی را گزارش دادند با این تفاوت که، روش k نزدیک‌ترین همسایه عملکرد بهتری را نسبت به روش ماشین بردار از خود نشان داد، آن چه که به‌صورت مشترک بین این دو الگوریتم تاثیر صعودی از خود به جای می‌گذارد، تعداد ویژگی‌ها است که با افزایش آن‌ها روند صعودی دقت دسته‌بندی را مشاهده خواهیم کرد. انتخاب یک معیار شباهت درست و درخور در روش k نزدیک‌ترین همسایه، عملکرد دسته‌بندی را بهبود خواهد داد.

یک بررسی دیگر دسته‌بندی اسناد فارسی، بر روی مجموعه داده‌ای با نام همشهری انجام گرفته است [۲۶]. در این مطالعه کارایی دسته‌بندی متاثر از استفاده Nتایی حروف برای تبدیل اسناد به بردارهای عددی مورد بررسی قرار گرفته است. همچنین اثر دسته‌بندی‌کننده بهبود یافته k نزدیک‌ترین همسایه نیز مورد مطالعه قرار گرفته است. در دسته‌بندی به روش k نزدیک‌ترین همسایه، توزیع نامتقارن اسناد آموزش در دسته‌های مختلف به روی دسته‌بندی‌کننده تاثیر می‌گذارد. از آن‌جایی که در الگوریتم k نزدیک‌ترین همسایه، دسته برنده دسته‌ای است که بیشترین همسایه را داراست، یک دسته با اسناد آموزش بیشتر نیز شانس بیشتری برای برنده شدن خواهد داشت.

روش Nتایی حروف، یک روش متداول برای نمایش بردارها به شکل عددی است. در این روش، هر سند متنی به تکه‌هایی به طول N از حروف هم‌جوار تقسیم می‌شود. بردار مرتبط با هر سند شامل یک لیست از Nتایی‌های غیر تکراری، حاوی عدد تکرار هر Nتایی است. در این بررسی اثر طول Nتایی‌های متفاوت بین ۲ و ۱۰ بر کارایی دسته‌بندی‌کننده تحلیل می‌شود.

در این مطالعه آزمایش‌های متعددی به منظور ارزیابی عملکرد روش جدید انجام گرفته، دقت دسته‌بندی‌کننده با رویه اعتبارسنجی ۵ تایی اندازه‌گیری شده، به این معنا که متن به پنج بخش منحصر تقسیم می‌شود و الگوریتم برای هر بخش یک بار اجرا می‌شود که هر بار یک بخش به‌عنوان مجموعه آزمون، و مابقی ٤ بخش دیگر به‌عنوان مجموعه آموزش مورد پردازش قرار می‌گیرند. نهایتاً از نتایج ۵ بار اجرا میانگین گرفته می‌شود.

در آزمایش اول، دسته‌بندی‌کننده KNN بهبود یافته بر روی مجموعه داده اعمال شده است، که عملکرد دسته‌بندی‌کننده k نزدیک‌ترین همسایه بهبود یافته در مقایسه با حالت بدون تغییر الگوریتم بهتر بوده است. در آزمایش دوم، از طول Nتایی‌های متفاوت در عملیات پیش پردازش استفاده شده و سپس دسته‌بندی‌کننده نوین و بهینه شده اعمال شده است. در این حالت بهترین عملکرد با طول Nتایی ۸ بوده است. در آزمایش سوم، حذف ایست‌واژه‌ها و کلمات با فراوانی سند کم در دست اقدام قرار گرفته که در این آزمایش عدد مورد نظر برای پارامتر فراوانی سند کم برابر با ۱ است، نتیجه آزمایش نشان می‌دهد که حذف کلمات با فراوانی سند ۱، موجب تاثیر منفی در این روش می‌شود. در آزمایش چهارم، یک مجموعه داده جدید حاوی ۱۰ دسته از موضوعات مختلف جهت مقایسه دو روش

دسته‌بندی اسناد فارسی، یعنی دسته‌بندی با استفاده از الگوریتم k نزدیک‌ترین همسایه بهبود یافته (روش مورد بحث) و دسته‌بندی با استفاده از الگوریتم ماشین بردار پشتیبان با در نظر گرفتن گنجینه واژگان، روش مطرح شده در [۲۷]، به کار رفته است. نتیجه این آزمایش بیانگر نتایج بهتر در مورد روش اول است.

در یک مطالعه دیگر [۲۸]، مجموعه داده همشهری مورد بررسی قرار گرفته که اسناد مورد بررسی شامل ۱۶۰۰ سند آموزش و ۴۰۰ سند آزمون می‌باشد. از آن‌جایی که روش K نزدیک‌ترین همسایه به تنهایی نتایج مطلوبی را ارائه نمی‌دهد، اعمال روش‌ها و الگوریتم‌های دیگر به KNN به منظور بالا بردن دقت و کارایی الگوریتم در دستور اجرا قرار می‌گیرند. در این تحلیل، از شبکه واژگان استفاده شده است تا سندهای قرار گرفته در یک دسته شباهت بیشتری را به هم داشته باشند. با استفاده از این شبکه (واژگان)، فراوانی کلماتی که با هم در ارتباطند از دقت بیشتری برخوردار می‌شود.

در شبکه واژگان، نحوه نمایش بردارهای اسناد به‌صورت مجموعه‌ای از کلمات است. فرآیند آغازین پس از استخراج ویژگی، ریشه‌یابی کلمات است. بعضی از کلمات به اشکال دستوری متفاوتی استفاده می‌شوند اما کماکان یک معنی را دارند. عمل ریشه‌یابی می‌تواند به حل مشکل چندشکلی بودن کمک کند. همچنین کلمات پر تکراری مانند "در" و "بر" مناسب برای تشخیص دسته اسناد نیستند، بنابراین باید حذف شوند. اما بعضی از این کلمات پیشوند کلمات دیگر هستند و می‌توانند در معنای کلمه تاثیرگذار باشند و حتی باعث تغییر در دسته اسناد شوند. بنابراین به جای قرار دادن آن‌ها در لیست ایست‌واژه‌ها و حذف مستقیم آن‌ها، از روش بهره اطلاعات استفاده می‌شود و سپس کلمات با فراوانی بالا و بهره اطلاعات پایین حذف می‌شوند. این عملیات این اطمینان را می‌دهد که کلمات با اثرگذاری کم در دسته‌بندی حذف خواهند شد. شبکه واژگان [۲۹، ۳۰]، یک پایگاه داده لغوی از هر زبان است.

هر کلمه موجود در سند به شبکه واژگان داده می‌شود و فراوانی آن محاسبه می‌شود.

۳-۲ Nتایی آماری

دسته‌بندی به روش Nتایی، بر روی مجموعه داده همشهری صورت گرفته است [۳۱]. سودمندی استفاده از این روش این است که در مقایسه با روش‌های یادگیری دسته‌بندی سنتی در حین انجام دسته‌بندی، ویژگی‌های با فراوانی کم را نادیده نمی‌گیرد. روش مورد نظر یک نوع مدل‌سازی آماری زبان (به‌صورت متنی و نوشتاری) محسوب می‌شود. مدل‌سازی آماری زبان، احتمالات رخدادهای دنباله کلمات را در مجموعه آموزش محاسبه می‌کند، سپس با استخراج دنباله کلمات از مجموعه آزمون بیشترین احتمال را برای دسته مربوط به سند مورد پرسش پیش‌بینی می‌کند [۳۱]. در واقع هدف مدل‌سازی، قرار دادن دنباله‌های کلماتی که به واقع رخ داده‌اند در احتمالات بالا و همچنین قرار دادن دنباله‌هایی از کلمات که هرگز رخ نداده‌اند در احتمالات پایین است.

در واقع، اصل استفاده از مدل زبانی Nتایی به‌عنوان دسته‌بندی‌کننده متن بیان این مطلب است که دسته سند مورد پرسش باید به گونه‌ای تعریف شود که سند پرسش بیشترین شباهت را با مدل دسته تعیین شده داشته باشد. بنابراین یک مدل زبانی مجزا برای هر دسته، آموزش داده می‌شود و دسته‌بندی یک سند جدید با ارزیابی میزان درست‌نمایش تحت هر دسته، و سپس انتخاب دسته صورت می‌گیرد. پارامتر n یک فاکتور کلیدی در مدل‌سازی زبانی Nتایی به حساب می‌آید. n کوچک اطلاعات کافی را برای یک مدل وابستگی‌های کلمات با دقت بالا ارائه نمی‌دهد و همچنین n خیلی بزرگ نیز مشکلات پراکندگی داده را در آموزش سیستم به وجود می‌آورد. یک چند جمله‌ای خالص بیز ساده، نام یک حالت خاص از دسته‌بندی‌کننده (Nتایی) است که در آن n=1 است و در این حالت از هموارکننده افزودن یک[۷۲] استفاده می‌شود.

72- Add-One

منظور از هموارکننده، تکنیک‌هایی هستند که برای تنظیم بیشینه درست‌نمایی تخمین زده شده از میان احتمالات به منظور تولید احتمالات دقیق‌تر، به کار می‌روند. تکنیک‌های هموارسازی نه تنها احتمالات صفر را از بین می‌برند بلکه در تلاشند تا دقت مدل را نیز بهبود ببخشند [۳۲]. چند نمونه از مهم‌ترین روش‌های هموارسازی از این قرار است [۳۳، ۳٤]: هموارکننده افزودن یک برای جلوگیری از احتمالات صفر وانمود می‌کند که رخ دادن هر Nتایی ناچیزتر از آن چیزی است که در واقعیت هست، این روش به هر شمارش، ۱ را اضافه می‌کند. ایده هموار کننده تخفیف قطعی[۷۳] این است که، احتمالات کلمات دیده شده را با تفریق کردن یک مقدار ثابت از تعدادشان می‌کاهد. رفتار هموارکننده عقب‌گرد[۷٤] بدین گونه است که به دنبال کوچک‌تر شدن N تایی‌هاست، به این معنا که در مدل‌سازی زبانی سه‌تایی، تنها از احتمالات سه‌تایی استفاده نمی‌شود بلکه احتمالات دوتایی و یکتایی نیز در نظر گرفته می‌شود، اما در هموارکننده مذکور اگر سه‌تایی پیدا نشد به سراغ دوتایی می‌رود و اگر دوتایی هم پیدا نشد به سراغ یکتایی می‌رود. طبق نتایج حاصل شده، هموارکننده عقب‌گرد بهترین دقت را در مجموعه داده مورد بررسی و در مدل با n=3 داشته است. با توجه به نتایج بررسی شده و در نظر گرفتن n متعادل برای مقابله با مشکلات پراکندگی داده، مدل‌سازی زبانی سه‌تایی بهترین دقت را در دسته‌بندی اسناد فارسی ارائه می‌دهد و حتی این نتایج با اعمال روش هموارسازی عقب‌گرد بهتر نیز خواهند شد.

۳-۳ شبکه عصبی مصنوعی

فرآیند آموزش دسته‌بندی به این روش، بر روی ۱۰۵۰ سند خبری از مجموعه اسناد فارسی همشهری ۲ انجام گرفته است که روش دسته‌بندی یادگیری چندی‌سازی برداری، یک روش دسته‌بندی الگویی است که بر اساس شبکه‌های نگاشت خودسازمان‌ده می‌باشد [۳۵، ۳٦]. در این روش فضای ورودی به چند ناحیه منحصر به فرد تقسیم می‌شود سپس برای هر ناحیه یک بردار ساخته می‌شود که نمایانگر یک دسته خاص تعریف شده است. بردار وزنی برای یک واحد خروجی، معمولا با یک بردار کتاب کد منتسب می‌شود که نمایانگر دسته‌ای است که واحد نمایش می‌دهد. طی فرآیند آموزش، محل قرارگیری واحدهای خروجی با تنظیم کردن وزن‌هایشان از میان یک آموزش با نظارت با مبنای دسته‌بندی‌کننده بیز، صورت می‌گیرد. روش دسته‌بندی جلو رو به روی مجموعه داده فارسی تحت بررسی [۳۷] اعمال شده که در پی آن نتایج خوبی گزارش شده است.

۳-٤ بیز ساده

بیز ساده یکی از روش‌های آماری برای دسته‌بندی به شمار می‌آید، بدین گونه که دسته‌های مختلف، هر کدام به‌عنوان یک فرضیه دارای احتمال در نظر گرفته می‌شوند. استفاده از تکنیک‌های کمکی و روش‌های انتخاب ویژگی مختلف در کنار بیز ساده، قالب آزمون‌های صورت گرفته برای این روش از دسته‌بندی را (در این پژوهش) تشکیل می‌دهند. یکی از این تکنیک‌ها احتساب بردار نماینده است. در این روش از مجموعه داده همشهری شامل ۱٦۰۰۰۰ خبر فارسی بین سال‌های ۱۹۹۷ تا ۲۰۰۲ میلادی استفاده شده است. تاثیر بردار نماینده، بردار متشکل از کلمات مرتبط با هم همراه با درجه ارتباطشان، بر بهبود کیفیت دسته‌بندی به روی اسناد فارسی در این بررسی [۳۸] مورد مطالعه قرار گرفته است. در این فرآیند، بردار نماینده حکم یک پایه و اساس برای ارزیابی بهتر روش بیز ساده را دارد.

تکنیک دیگر استفاده از روش‌های انتخاب ویژگی فراوانی سند (DF)، واریانس فراوانی عبارت (TFV)، اطلاعات متقابل (MI)، اطلاعات متقابل تغییریافته (MMI) است. این بررسی [۳۹] که به روی مجموعه داده‌ای متشکل از ۸۲۹ نظردهی فارسی صورت گرفته، تنها مطالعه‌ای است که با این حجم کم داده (فقط حاوی دو دسته، نظرات مثبت و نظرات منفی) صورت گرفته است. این روش‌ها عملکردهای بهتری نسبت

73- Absolute Discounting
74- Back-Off

به روش MI سنتی دارند. دلیل عملکرد ضعیف MI این است که این روش فقط از اطلاعات بین ویژگی و دسته متناظر با هم (جاری) استفاده می‌کند و با بقیه ویژگی‌ها و دسته‌های دیگر کاری ندارد. وقتی سه روش DF،TFV و MMI را با هم مقایسه می‌کنیم، مشاهده می‌کنیم که MMI در رقابت بر سر دسته منفی، با بهبودهای ۱/۲ و ۱/۷٦ درصدی بر آن دو روش دیگر برتر است اما، وقتی که به سراغ نتایج F دسته مثبت می‌رویم دو روش DF و TFV برتری‌های ۰/۰٤ و ۰/۳ درصدی را نسبت به عملکرد روش MMI دارا هستند. از طرفی مزیت روش MI این است که از تمام اطلاعات وابسته به یک ویژگی استفاده می‌کند و همچنین تمام فاکتورهای بین ویژگی‌ها و دسته‌ها را در نظر می‌گیرد نه بخشی از آن‌ها را. روش MMI علاوه بر دسته‌بندی نظرات فارسی (که حاوی دو دسته بودند) می‌تواند به دیگر حوزه‌های دسته‌بندی اسناد (با دسته‌های بیشتر) نیز کمک کند.

در [۲٤] با محوریت دسته‌بندی متون کوتاه فارسی (متون توئیت‌ها)، از ویژگی TF-IDF و چهار روش بیز ساده، نزدیک‌ترین همسایه، درخت تصمیم و SVM استفاده شده است و نتیجه‌گیری شده است که روش‌های بیز ساده و نزدیک‌ترین همسایه به ترتیب بهترین و ضعیف‌ترین روش برای دسته‌بندی ٦ دسته از متون بوده است.

۳-۵ منطق فازی

دسته‌بندی اسناد فارسی به روش منطق فازی در [٤۰] بررسی شده است که به روی ۹۰۰ سند استخراج شده از وبگاه‌های همشهری، ویکی‌پدیا، دیگر وبگاه‌های فارسی، انجام گرفته است. دو سوم این مجموعه به‌عنوان مجموعه داده آموزش و مابقی تحت عنوان مجموعه داده آزمون مورد استفاده قرار گرفته است. در این روش دسته‌بندی روابط فازی که مبنای دسته‌بندی هستند از نظریهٔ فازی مشتق می‌شوند.

با هدف دسته‌بندی اسناد، با توجه به ویژگی‌های خاص زبان فارسی می‌توان از مدل فازی استفاده کرد. ایده اصلی این روش، از پیوندهای بین دسته‌های از پیش تعریف شده‌شان مشتق می‌شود، که این پیوندها از تابع عضویت فازی به‌دست می‌آیند. مقدار تابع مذکور بین صفر و یک است که مقدار صفر مربوط به حالتی است که عنصر انتخاب شده عضو هیچ دسته‌ای نیست و مقدار یک بیانگر این حالت است که عنصر انتخاب شده کاملا عضو یک دسته است، همچنین مقادیر بین صفر و یک نیز معرف عضویت و درجه شباهت عنصر با دسته مربوطه است [٤۰].

یکی از روش‌های تعیین کننده مقادیر تابع عضویت استفاده از عوامل هوشمند، بر اساس قوانین فازی (If ... Then) است. برای تعریف این عوامل استاندارد خاصی وجود ندارد بلکه عموما اندیشه‌هایی که به ارتباطات مهم می‌پردازند در پیش‌زمینه این عوامل قرار می‌گیرند. عوامل هوشمند اعمال شده در این بررسی، به منظور تعیین میزان درجه عضویت عبارت‌های اسناد آموزش به دسته‌ها استفاده می‌شوند. در پی آن، برای دریافت پیوندهای فازی از میان کلمات، عوامل هوشمند سازوکارهای شناسایی دسته اسناد مورد پرسش را فراهم می‌کنند [٤۰].

۳-۶ ماشین بردار پشتیبان

روش ماشین بردار پشتیبان، یکی از بهترین و محبوب‌ترین روش‌های دسته‌بندی متون به حساب می‌آید. استفاده از این روش در مسائل دسته‌بندی روند جدیدی است که در سالیان اخیر مورد توجه بسیاری قرار گرفته است. همچنین توابع (هسته) گوناگونی که می‌توانند در دل محاسبات این روش قرار بگیرند، نتایج جالبی را در شرایط متفاوت از خود برجای می‌گذارند. یکی از دلایل محبوبیت این روش، انعطاف‌پذیری بالای آن در شرایط متفاوتی از همکاری با دیگر الگوریتم‌هاست. روش انتخاب ویژگی، استفاده از توابع هسته گوناگون، استفاده از گنجینه واژگان و در نظر گرفتن روابط معنایی و .... همگی شرایطی هستند که در آزمون‌های دسته‌بندی اسناد فارسی به روش ماشین بردار پشتیبان پیش رو مورد بررسی قرار گرفته‌اند.

روش ماشـــین بردار پشـــتیبان، به روی مجموعه داده فارســـی همشهری که جزو مجموعه داده‌های معتبر اسناد فارسی محســـوب می‌شـــود، نتایج خوبی را در بر داشته اســـت [۲۵]. از روش انتخاب ویژگی TF-IDF برای انتخاب برترین ویژگی‌های هر دسته استفاده شده است و همچنین از ۷۰ درصـــد مجموعه داده همشـــهری به‌عنوان مجموعه آموزش و مابقی که ۳۰ درصد می‌باشد به‌عنوان مجموعه داده آزمون استفاده گردیده است.

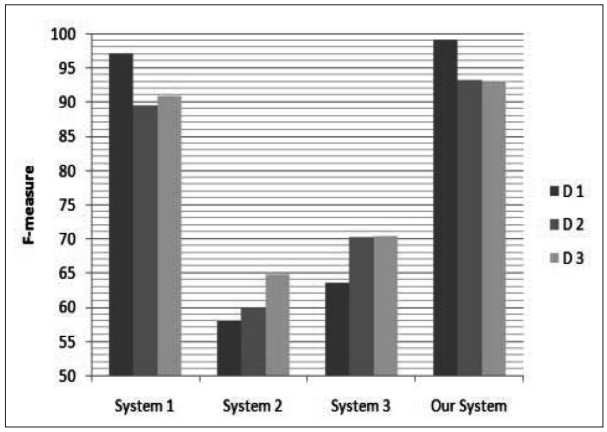
عامل دیگر تاثیرگذار بر کیفیت عملکرد ماشـــین بردار پشـــتیبان تابع هسته می‌باشد [۲۵]. انتخاب یک تابع هسته مناســـب از اهمیت بالایی برخوردار است، زیرا تابع هسته تعیین کننده فضای ویژگی اســـت که نمونه‌های مجموعه داده آموزش دسته‌بندی می‌شوند.

دسته‌بندی اسناد فارسی با استفاده از گنجینه واژگان و روش SVM، به هدف بررسی تاثیر آن در نتایج دسته‌بندی صورت گرفته اســـت [۲۷]. به‌کارگیری و عدم به‌کارگیری گنجینه واژگان، موردهای مورد آزمایش هستند که در پی آن گزارش‌هایی حاکی از بهبود نتایج در شـــرایط حضور گنجینه واژگان ارائه شـــده‌اند. لازم به ذکر است مجموعه داده مورد اســـتفاده متشـــکل از اســـناد موجود در وبگاه ویکی‌پدیا فارسی، آرشـــیو روزنامه همشهری، و نشریات سروش و رشد [٤١] می‌باشند.

گســـترش بـــردار ویژگـــی با اضافـــه کـــردن کلمات اســـتخراجی از گنجینه واژگان، عملی است که هدفش کمک به دســـته‌بندی‌کننده در زمان‌هایی است که مجموعه داده آموزش کامل و مناسب همه دسته‌ها نیست. فرهنگ طایفی [٤٢] نام گنجینه واژگانی اســـت که در این بررسی استفاده می‌شود. با توجه به ساختار گنجینه واژگان، گنجینه برای هر دســـته به هدف یافتـــن کلمات مرتبطی کـــه نمایندگان بهتری برای آن دســـته هستند جســـتجو می‌شود. انتخاب یک کلمه از گنجینه مبتنی بر همان ویژگی‌هایی است که از فاز انتخاب ویژگی اولیه حاصل شـــده‌اند به این ترتیب که، هر کلمه انتخاب شـــده (از فاز انتخـــاب ویژگی) در گنجینه جستجو می‌شـــود، همه کلمات مرتبط (در گنجینه) با کلمه انتخاب شده، نامزد اضافه شدن به بردار ویژگی می‌شوند. ســـرانجام اگر فراوانی کلمه نامزد شده در اسناد متعلق به دسته مورد نظر در مجموعه داده آموزش، از یک آستانه از قبل تعیین شده بزرگ‌تر باشند و کلمه قبلا در بردار ویژگی موجود نباشـــد، به بردار ویژگی اضافه خواهد شـــد. نکته قابل ذکر این اســـت که اگر آستانه مقداری کوچک در نظر گرفته شـــود کلمات غیر مرتبط در بردار ویژگی به تناسب زیاد خواهند شد. همچنین انتخاب یک مقدار آستانه بزرگ، کلمات مرتبط بیشـــتری را به بردار ویژگی اضافه خواهد کرد. مقدار آســـتانه نقش مهمی را در اضافه نشدن کلمات غیر مرتبط به بردار ویژگی ایفا می‌کند.

در بررســـی [٤٣]، اســـناد متشـــکل از ۳ مجموعه داده همشهری نســـخه۲ [٤٤]، پرسیکا [٤٥] و تابناک با برچسب دسته‌های مشـــترک به‌عنوان مجموعه داده آموزش و سه مجموعه داده متفاوت D1، D2 و D3 با ٦ دســـته مشـــترک کـــه به ترتیب حـــاوی ۵۸۵، ۹۷٦ و ۱۲۲۵ ســـند به تصادف انتخاب شـــده می‌باشـــد به‌عنوان مجموعه داده‌های آزمون استفاده شده است. در دســـته‌بندی اسناد فارسی به روش ارتباطـــات معنایی بین کلمات، آن چه که باعث تغییرات مهم و کارایی بهتر نســـبت به روش‌های دیگر آماری می‌شـــود، نحوه وزن‌دهی عبارت‌هاست. روش وزن‌دهی عبارت‌ها بر اساس روابط معنایی بین کلمات برتری است که روش‌های دیگـــر وزن‌دهی از به کارگیـــری آن معذورند و به جای آن از اطلاعات آماری اســـتفاده می‌کنند. این روش ارتباط بین عبارت‌ها را اندازه‌ای از وابستگی در نظر می‌گیرد، همچنین برای تعیین معناهای (ارتباطات معنایی) دســـته‌ها بر اساس قرارگیری عبارت‌ها در (برچسب‌های) دسته‌ها، بردار ویژگی هر دسته را با استفاده از گنجینه واژگان گسترش می‌دهد.

سیستم دسته‌بندی اسناد فارسی مذکور (مورد بررسی) با سیستم‌های طراحی شده دیگر برای اسناد فارسی مورد مقایســـه قرار گرفته‌اند. برای این منظور بهترین نتایج سه سیستم که شـــبیه سیستم مورد بررســـی بوده‌اند، مورد



شکل ۴: مقایسه سیستم‌های دسته‌بندی اسناد فارسی [۴۳]

مطالعه قرار گرفتند. نیاز به توضیح است که سه سیستم با شرایط یکسان و با آزمایش به روی هر سه مجموعه داده D1، D2 و D3 پیاده‌سازی شده‌اند (شکل ٤). این سیستم‌ها عبارتند از:

سیستم ۱: دسته‌بندی اسناد فارسی با روش انتخاب ویژگی DF و استفاده از SVM و گنجینه واژگان توسط [۲۷].

سیستم ۲: دسته‌بندی اسناد فارسی با روش انتخاب ویژگی IG و استفاده از SVM (به جای KNN) و گنجینه واژگان توسط [۲۸].

سیستم ۳: دسته‌بندی اسناد فارسی با نمایش عددی بردارها به روش Nتایی و احتساب ضریب دسته با استفاده از KNN (معیار شباهت دایس) توسط [۲٦].

با این توضیح که سیستم ۱، فقط از گنجینه واژگان به منظور گسترش بردارهای ویژگی و TF-IDF به هدف وزن‌دهی استفاده کرده است و اطلاعاتی مانند ارتباطات معنایی را در نظر نگرفته است، همچنین سیستم‌های ۲ و ۳ از اطلاعات آماری برای وزن‌دهی استفاده کرده‌اند.

در بررسی [٤٦] که به روی مجموعه داده فارسی بی‌جن‌خان [٤٧] انجام شده است، به منظور نمایش برداری کلمات کلیدی اسناد، از روش TF-IDF استفاده شده است. در گام بعدی، به وسیله این بردارها و با استفاده از الگوریتم‌های تحلیل آماری معنایی پنهان، به بردارهای ویژگی نمایش‌دهنده اسناد رسیده است. در ادامه با دسته‌بندی‌کننده‌های گوناگون (k نزدیک‌ترین همسایه، بیز ساده و SVM)، به دسته‌بندی اسناد پرداخته است. این روش با به‌کارگیری مجموعه کلمات اسمی به‌عنوان، کلمات کلیدی و استفاده از دسته‌بندی‌کننده SVM بهترین مقدار F را برای نه دسته، ۸۸ درصد گزارش داده است.

در یکی از کارهای اخیر در [٤٨] از روش‌های مدل‌سازی عنوان برای استخراج ویژگی استفاده شده و روش SVM برای دسته‌بندی متون فارسی روی پیکره بیجن‌خان به کار گرفته شده است که در آن روش‌های مختلف مدل‌سازی عنوان شامل LDA، PLSA و STC با روش TF-IDF روی تعداد مختلف دسته (عنوان) بررسی شده و نشان داده شده است که روش‌های مدل‌سازی عنوان به طور چشمگیری کارایی بالاتری را نتیجه می‌دهند.

۳-۷ مدل مخفی مارکوف (HMM)

این روش مبتنی بر مدل مخفی مارکوف که جزو روش‌های آماری است، به مدل‌سازی توالی دنباله کلمات یک سند می‌پردازد. در بررسی [۲۳]، که به روی مجموعه داده فارسی بی‌جن‌خان انجام شده است، اسناد مورد دسته‌بندی شامل هشت دسته ادبی، مذهبی، اقتصادی، هنری، پزشکی، تاریخی، سیاسی و ورزشی می‌باشند. در این روش، هر سند به‌صورت رشته‌ای از کلمات متوالی در نظر گرفته می‌شود، همچنین از قابلیت این روش در مدل‌سازی داده‌های ترتیبی، بهره گرفته می‌شود. مدل مخفی مارکوف در میان سایر دسته‌بندی‌کننده‌ها (شبکه عصبی، بیز ساده، شبکه بیز، ماشین بردار پشتیبان، درخت تصمیم و k نزدیک‌ترین همسایه)، بیشترین مقدار میانگین دقت (۸۳درصد)، فراخوانی (۷٦ درصد) و F (۷۹ درصد) را به روی هشت دسته، گزارش داده است. پس از مدل مخفی مارکوف، شبکه عصبی است که برترین نتایج را ارائه داده است.

خلاصه و جمع‌بندی

در این بخش سعی شده که نتایج همه آزمایش‌های دسته‌بندی اسناد فارسی که در بخش قبل بیان شدند در جدول ۱ گردآوری شود. در هر سطر از جدول به ترتیب

نوع دسته‌بندی‌کننده، جزئیات دسته‌بندی (روش انتخاب ویژگی، شرایط، پارامترهای الگوریتم، الگوریتم‌های مکمل و کمکی) و بهترین نتیجه حاصل از ارزیابی دسته‌بندی آورده شده است. لازم به ذکر است که هر خط افقی، یک جدا کننده بین دو مجموعه داده متفاوت می‌باشد به این معنا که نتایج قرار گرفته بین دو خط افقی حاصل از اعمال آزمایش‌های دسته‌بندی به روی یک مجموعه داده مشترک است. بنابراین مقایسه نتایج قرار گرفته بین خطوط افقی از اعتبار بیشتری (نسبت به مقایسه به روی کل نتایج جدول) برخوردار است و دلیل بدیهی این امر نیز، یکسان بودن شرایط استفاده از مجموعه داده در آزمایش‌های دسته‌بندی (قرار گرفته بین خطوط افقی) است.

زمانی که طبیعت یک مجموعه داده با فرضیات و تدابیر مورد نظر یک دسته‌بندی تطبیق داده شود، دقت آن دسته‌بندی افزایش پیدا می‌کند. اما عدم تطبیق و ضعف تناسب یک مدل نیز می‌تواند ناشی از انتخاب نوع ویژگی‌های استفاده شده در مدل، انتخاب روش‌های امتیازدهی، انتخاب توابع شباهت و خیلی از عوامل درگیر دیگر باشد. حال با یک دید کلی می‌توان به بحث در مورد دسته‌بندی‌کننده‌ها پرداخت.

دسته‌بندی‌کننده بیز ساده، یک دسته‌بندی‌کننده خیلی ساده است که خیلی خوب روی داده‌های متنی و عددی عمل می‌کند، ساده پیاده‌سازی می‌شود و از نظر بار محاسباتی بسیار ارزان‌تر از هر الگوریتم دسته‌بندی دیگر است. یکی از محدودیت‌های این دسته‌بندی‌کننده این است که زمانی که ویژگی‌ها با هم ارتباط بسیار قوی دارند، ضعیف عمل می‌کند. همچنین با توجه به قواعد دسته‌بندی اسناد، این نوع دسته‌بندی‌کننده ناتوان از شمارش تعداد رخداد کلمات در یک بردار ویژگی است و اما در مقابل، دسته‌بندی‌کننده k نزدیک‌ترین همسایه، مشخصات محلی یک سند را حفظ می‌کند ولی زمان زیادی را بابت دسته‌بندی می‌طلبد و همچنین پیدا کردن مقدار بهینه k همواره یک مشکل جدی است چنان‌که اگر نقاط، توزیع یکنواختی را نداشته باشند این عمل دشوارتر نیز خواهد شد. همان‌گونه که انتخاب یک مقدار k مناسب در کیفیت دسته‌بندی تاثیرگذار است، عملکرد این نوع دسته‌بندی‌کننده به معیار فاصله اعمال شده نیز وابسته است. عموما انتخاب یک k کوچک، در جایی که اسناد اشتراکات (روی‌هم‌افتادگی) زیادی را با هم دارند و به دسته‌های مختلف تعلق دارند بهتر نتیجه می‌دهد چرا که، دسته‌های فشرده‌تری را حاصل می‌نماید. در حالی که مقدار kهای بزرگ‌تر در مقابل نوفه‌ها ایمن‌تر و مقاوم‌تر است و مرزهای متعادل‌تری را بین دسته‌ها قائل می‌گردد. اگرچه دسته‌بندی‌کننده مرکز ثقل ساده و خطی است، اما اغلب مدل آن از عدم تطبیق داده‌ها با فرضیاتش رنج می‌برد. این دسته‌بندی‌کننده همیشه با این فرض اقدام به دسته‌بندی می‌کند که اگر شباهت یک سند با مرکز ثقل یک دسته نسبت به دسته‌های دیگر بیشترین مقدار را داشته باشد، آن سند متعلق به همان دسته مورد نظر (با بیشینه شباهت) است. قطعا این فرضیه در موارد خاص و استثنا همیشه خوب جواب نمی‌دهد. یک دسته‌بندی‌کننده مبتنی بر قانون مانند درخت تصمیم، به راحتی قابل فهم است و پیچیدگی مسئله را می‌کاهد اما زمان زیادی را بابت آموزش اسناد مصرف می‌کند. همچنین با توجه به ساختار درختی شکل آن، در سطوح بالاتر یک زیردرخت اشتباهاتی رخ می‌دهد که گاهی قابل چشم‌پوشی نمی‌باشد. در اداره متغیرهای پیوسته ناتوان است و از مشکلات عدم تناسب نیز رنج می‌برد. دسته‌بندی‌کننده‌های بر پایه و اساس شبکه عصبی، نتایج بسیار خوبی را در دامنه‌های پیچیده ارائه می‌دهند و همچنین قادر به اداره داده‌های پیوسته و گسسته می‌باشد. از معایب آن می‌توان زمان زیاد آموزش و دشوار بودن تفسیر نتایج یادگیری توسط کاربران را نام برد. در سال‌های اخیر استفاده همه‌گیر از روش‌های یادگیری عمیق[75] و کارایی بالاتر آن‌ها لزوم توجه و به کارگیری این روش‌ها را برای دسته‌بندی متون فارسی پررنگ‌تر می‌کند، به ویژه پیش‌بینی می‌شود استفاده از شبکه‌های بازگشتی مانند LSTM در این موضوع منجر به بهبود کارایی این سیستم‌ها شود.

75- Deep Learning

جدول ۱: خلاصه‌ای از پژوهش‌های صورت گرفته در دسته‌بندی اسناد فارسی

نتیجه ارزیابی	جزئیات (روش انتخاب ویژگی، شرایط، پارامترهای الگوریتم، الگوریتم‌های مکمل و کمکی)				دسته‌بندی‌کننده	مرجع
ریز P=97.3	معیار فاصله = Cosine	K=3	تعداد ویژگی=۴۰۰۰	انتخاب ویژگی=TFIDF	KNN	[۲۵]
ریز P=98.8	معیار فاصله = Cosine	K=1	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	KNN	[۲۵]
ریز P=67	معیار فاصله = Euclidean	K=3	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	KNN	[۲۵]
ریز P=45.2	معیار فاصله = Cityblock	K=3	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	KNN	[۲۵]
ریز P=94.3	معیار فاصله = Correlation	K=3	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	KNN	[۲۵]
ریز F=90.3	معیار فاصله = Dice	K=3	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
ریز F=93.9	معیار فاصله = Dice	K=8	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
میانگین F=91.2	-	K=10	احتساب گنجینه واژگان	انتخاب ویژگی=IG	KNN	[۲۸]
ریز F=91	معیار فاصله = Dice	K=8	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
ریز F=89	-	-	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
Etp&Ppx=98	-	N=3	هموارکننده Back-Off	تعداد نمونه در train=9000	Nتایی آماری	[۳۱]
درشت F=80.3	-	α=1	-	-	LVQ 1	[۳۷]
درشت F=84.4	-	α=1	-	-	OLVQ 1	[۳۷]
درشت F=88.4	-	α=1	-	-	LVQ 2.1	[۳۷]
درشت F=88.8	-	α=1	-	-	LVQ 3	[۳۷]
درشت F=89.9	-	α=1	-	-	OLVQ 3	[۳۷]
درشت F=79.8	-	-	-	-	SVM	[۳۷]
درشت F=68.6	-	-	-	-	KNN	[۴۹]
درشت F=57.5	-	-	-	انتخاب ویژگی=MI	بیز ساده	[۳۹]
درشت F=83	-	-	-	انتخاب ویژگی=DF	بیز ساده	[۳۹]
درشت F=83.5	-	-	-	انتخاب ویژگی=TFV	بیز ساده	[۳۹]
درشت F=84	-	-	-	انتخاب ویژگی=MMI	بیز ساده	[۳۹]
درشت F=78.4	-	-	احتساب بردار نماینده	انتخاب ویژگی=MI	بیز ساده	[۳۸]
درشت F=75.3	-	-	-	انتخاب ویژگی=MI	بیز ساده	[۳۸]
درشت F=91	-	-	-	-	قوانین فازی	[۴۰]
ریز F=93.8	تابع هسته=Polynomial	-	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	SVM	[۲۵]
ریز F=82.2	تابع هسته=Linear	-	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	SVM	[۲۵]
ریز F=93.1	تابع هسته=RBF	-	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	SVM	[۲۵]
ریز F=82.8	تابع هسته=MLP	-	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	SVM	[۲۵]
ریز F=83.8	تابع هسته=Quadtaric	-	تعداد ویژگی=۱۰۰۰	انتخاب ویژگی=TFIDF	SVM	[۲۵]

ریز F=90	–	–	–	انتخاب ویژگی=DF	SVM	[۲۷]
ریز F=94	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=99	احتساب وزن معنایی	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=97.1	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=71.3	–	–	–	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=58	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=IG	SVM	[۲۸]
میانگین F=64	معیار فاصله = Dice	K=8	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
میانگین F=93.1	احتساب وزن معنایی	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=89.5	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=78	–	–	–	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=60	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=IG	SVM	[۲۸]
میانگین F=70	معیار فاصله = Dice	K=8	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
میانگین F=92.8	احتساب وزن معنایی	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=90.8	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=78.6	–	–	–	انتخاب ویژگی=DF	SVM	[۲۷]
میانگین F=65	–	–	احتساب گنجینه واژگان	انتخاب ویژگی=IG	SVM	[۲۸]
میانگین F=70	معیار فاصله = Dice	K=8	احتساب ضریب دسته	نمایش عددی بردار=Nتایی	KNN	[۲۶]
میانگین F=88	–	–	–	PLSA & TFIDF	SVM	[۴۶]
میانگین F=79	–	–	States ۹	هر عنوان ۵ سند	HMM	[۲۳]
میانگین F=73	HID-OUT=sigmoid	–	2900 in, 46 hi, 8 ou	MLP	NN	[۲۳]
میانگین F=71	تابع هسته=Polynomial	–	–	–	SVM (SMO1)	[۲۳]
میانگین F=67	معیار فاصله = Euclidean	K=50	–	–	KNN	[۲۳]
درستی=74	–	–	–	TFIDF	بیز ساده	[۵۰]
درستی=67	–	–	–	TFIDF	SVM	[۵۰]
درستی=66	–	C4.5	–	TFIDF	درخت تصمیم	[۵۰]
درستی=46	–	K=3	–	TFIDF	KNN	[۵۰]
درستی=67	–	–	–	PLSA	SVM	[۴۸]
درستی=68	–	–	–	LDA	SVM	[۴۸]
درستی=87	–	–	–	$MedSTC^2$	SVM	[۴۸]

در قیـــاس با بقیه دســـته‌بندی‌کننده‌ها ماشـــین بردار پشـــتیبان، عمل تشخیص مشـــخصات ذاتی داده را بهتر و دقیق‌تر انجام می‌دهد. این روش دسته‌بندی قابلیت یادگیری مســـتقل در فضای ویژگی چندبعدی را داراست. پیچیدگی دســـته‌بندی‌کننده در تنظیـــم پارامترهـــا و تعیین هســـته محاسبات اســـت. نهایتا آن چه را که به طور کلی می‌توان نتیجه گرفت این است که دســـته‌بندی‌کننده‌های مختلف با توجـــه به خصوصیـــات داده به روی مجموعـــه داده‌های متفاوت، عملکردهای متفاوتی را از خود نشـــان می‌دهند و هیچ گاه نمی‌توان به طور قطع، این ادعا را داشت که یکی از آن‌ها در همه موارد برترین دسته‌بندی‌کننده است، هر چند که ماشـــین بردار پشـــتیبان با مدل نمایشی VSM، در اکثر مسائل دسته‌بندی اسناد نتایج خوبی را گزارش می‌دهد.

در مورد روش‌های استخراج ویژگی، می‌توان گفت که در میان روش‌های کلاسیک، TF-IDF به‌صورت کلی کارایی قابل قبولی داشـــته اســـت و روش‌های مبتنی بر جاسازی کلمات و مدل‌ســـازی عنوان به دلیل بهره‌گیری از اطلاعات بیشتر نقش موثری در بهبود کارایی دارند. در میان کارهای مرور شده، تحقیقی که از روش‌های جاسازی کلمات مبتنی بر شـــبکه عصبی که در ســـال‌های اخیر بسیار رایج شده، اســـتفاده کند، یافت نشد که برآورد می‌شود این نمایش از متون، بتواند منجر به بهبود کارایی سیستم‌های دسته‌بندی متون شود.

مراجع

1. Aggarwal, C.C., Zhai, C.X., 2012. A Survey of Text Classification Algorithms. Mining Text Data Book, 163-222.
2. Sebastiani, F., 2002. “Machine learning in automated text categorization”. ACM computing surveys (CSUR), 34(1), pp.1-47.
3. Salton, G., Yang, C.S., 1999. On the Specification of Term Various in Automatic Indexing. Journal of Documentation, 29(2), 351-372.
4. Deerwester, S., Dumais, S., Harshman, H., 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 391.
5. Hofmann, T., 1999. Probabilistic Latent Semantic Indexing. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 50-57.
6. Mikolov, T., Yih, W. T., & Zweig, G., 2013. “Linguistic regularities in continuous space word representations”. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).
7. Pennington, J., Socher, R. and Manning, C., 2014. “Glove: Global vectors for word representation”. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).
8. Jolliffee, I.T., 2002. Principal Component Analysis. Springer.
9. Cai, L., Hofmann, T., 2003. Text categorization by Boosting Automatically Extracted Concepts. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 182-189.
10. Bengio Y., Ducharme R., Vincent P., and Janvin C., 2003, "A neural probabilistic language model," The Journal of Machine Learning Research, vol. 3, pp. 1137-1155.
11. Mikolov T., Chen K., Corrado G., and Dean J., 2013, "Efficient estimation of word representations in vector space," In ICLR.
12. Rong X., 2014, “word2vec Parameter Learning Explained”.
13. Brown P. F., Desouza P. V., Mercer R. L., Pietra V. J. D., and Lai J. C., 1992, "Class-based n-gram models of natural language," Computational linguistics, vol. 18, pp. 467-479.
14. Bengio Y., Schwenk H., Senécal J.-S., Morin F., and Gauvain J.-L., 2006, "Neural probabilistic language models," in Innovations in Machine Learning, pp. 137-186.
15. Blei D. M., Ng A. Y., and Jordan M. I., 2003, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993-1022.
16. Zhu, J. and Xing, E.P., 2012. Sparse topical coding. arXiv preprint arXiv:1202.3778.
17. Pennington J., Socher R., and Manning C. D., 2014, "Glove: Global vectors for word representation," Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014), vol. 12, pp. 1532-1543.
18. Yang, Y., Pedersen, J.O., 1997. A Comparative Study on Feature Selection in Text Categorization. Proceedings of the 14th International Conference on Machine Learning 97, 412-420.
19. Zong, W., Wu, F., Chu, L., Sculli, D., 2015. A Discriminative and Semantic Feature Selection Method for Text Categorization. International Journal of Production Economics, 215-222.
20. Aghdam, M.H., Ghasem-Aghaee, N., Basiri, M.E., 2009. Text Feature Selection Using Ant Colony Optimization. Expert Syst. Appl. 36 (3), 6843-6853.
21. Basu, T., Murthy, C.A., 2012. Effective Text Classification by a Supervised Feature Selection Approach. Data Mining Workshops (ICDMW), IEEE 12th International Conference, 918-925.

22. Chen, J., Pan, H., Ao, Q., 2012. Study a Text Classification Method Based on Neural Network Model. Proceedings of the MSEC International Conference on Multimedia, Software Engineering and Computing, Springer Berlin Heidelberg, 128, 471-475.
23. Gharavi, E., Veisi, H., 2014. A Hidden Markov Model for Persian Text Classification. 3rd National Computational Linguistics Conference.
۲۴. احسان ریحانی آرانی، سیدمحمدرضا لاجوردی. ۱۳۹۵. بررسی روش‌های طبقه‌بندی خودکار اسناد متنی. سومین کنفرانس ملی مهندسی برق و کامپیوتر سیستم‌های توزیع شده و شبکه‌های هوشمند.
25. Farhoodi, M., Yari, A., 2010. Applying Machine Learning Algorithms for Automatic Persian Text Classification. 6th International Conference on Advanced Information Management and Service (IMS), 318-323.
26. Elahimanesh, M.H., Minaei-Bidgoli, B., Malekinezhad, H., 2012. Improving K-Nearest Neighbor Efficacy for FarsiText Classification. The International Conference on Language Resources and Evaluation (LREC), 1618-1621.
27. Maghsoodi, N. and Homayounpoor, M., 2011. Using Thesaurus to Improve Multiclass Text Classification. Part II, LNCS 6609, 244-253.
28. Parchami, M., Akhtar, B., Dezfoulian, M.H., 2012. Persian Text Classification Based on K-NN Using Wordnet. IEA/AIE, LNAI, Springer-Verlag Berlin Heidelberg, 7345, 283-291.
29. Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A., 2009. A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. NAACL Proceedings of Human Language Technologies, Annual Conference of the North American Chapter of the Association for Computational Linguistics, 19-27.
30. Fellbaum, C., 1998. WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, Cambridge.
31. Farhoodi, M., Yari, A., Sayah, A., 2011. N-gram Based Text Classification for Persian Newspaper Corpus. Digital Content, Multimedia Technology and its Applications (IDCTA), IEEE, 55-59.
32. Peng, F., Huang, X., 2007. Machine Learning for Asian Laguage Text Classification. Journal of Documentation, 63(3), 378-397.
33. Ramasundaram, S., Victor, S.P, 2010, Text Categorization by Backpropagation Network. Intrenational Journal of Computer Applications, 8(6), 1-5.
34. Chen, S.F., Goodman, J., 1999. An Emperical Study of Smoothing Techniques for Language Modeling. 13(4), 359-394.
35. Kohonen, T., 1990. Improved versions of Learning Vector Quantization. In proceeding of the National Joint Conference of Neural Networks, San Diego, 545-550.
36. Kohonen, T., 1995. Self-Organization and Associative Memory. Berlin: Springer-Verlag.
37. Pilevar, M.T., Feili, H., Soltani, M., 2009. Classification of Persian Textual Documents Using Learning Vector Quantization. In Proceedings of the International Conference on Natural Language Processing and Knowledge Engineering, IEEE, 1-6.
38. Jafari, A., Hosseinejad, M., Amiri, A., 2011. Improvement in Automatic Classification of Persian Documents by Means of Naïve Bayes and Representative Vector. 1st International eConference on Computer and Knowledge Engineering (ICCKE), IEEE, 226-229.
39. Saraee, M., Bagheri, A., 2013. Feature Selection Methods in Persian Sentiment Analysis. Natural Language Processing and Information Systems, Springer-Verlag Berlin Heidelberg, 7934, 303-308.
40. Yari, A., Abbasi, A., MomenBellah, S., 2010. Presenting a fuzzy relation to classify the Persian Web documents. Intelligent Computing and Intelligent Systems (ICIS), IEEE International Conference, 2, 220-223.
41. Bijankhan, M., 2008. 100 Millions Word Farsi Corpus. Technical Report, Research Center for Intelligent Signal Processing.
42. Fararoy, J., 2008. Farhang-e Teyfi. Tehran: Hermes Press.
43. Parseh, S., Baraani, A., 2014. Improving Persian Document Classification Using Semantic Relations between Words. International Journal of Advanced Studies in Computers, Science and Engineering, ARXIV, 16-22.
44. AleAhmad, A., Amiri, H., Darrudi, E., Rahgozar, M., Oroumchian, F., 2009. Hamshahri: A Standard Persian Text Collection. Knowledge-Based Systems, 22, 382-387.
45. Eghbalzadeh, H., Hosseini, B., Khadivi, S., Khodabakhsh, A., 2012. Persica: A Persian Corpus for Multipurpose Text Mining and Natural Language Processing. Sixth International Symposium on Telecommunications (IST), IEEE, 1207-1214.
۴۶. خاکی اردکانی بصیرا، ۱۳۹۲. به‌کارگیری مدل‌های موضوع پنهان جهت دسته بندی مستندات فارسی و ارائه راهکارهای مناسب جهت بهبود آن. پایان نامه، دانشگاه صنعتی شریف.
47. Bijankhan, M., Sheykhzadegan, J., Bahrani, M., Ghayoomi, M., 2011. Lessons from Building a Persian Written Corpus: Peykare. Language Resources and Evaluation, 45, 143-164.
48. Ahmadi, P., Tabandeh, M. and Gholampour, I., 2016, May. Persian text classification based on topic models. In Electrical Engineering (ICEE), 2016 24th Iranian Conference on (pp. 86-91). IEEE.
49. Bina, B., Ahmadi, M.H., Rahgozar,M.,2008. Farsi Text Classification Using N-gram and Knn Algorithm A Comparative Study. DMIN, 385-390.
۵۰. انامرادنژاد عیسی، حبیبی جعفر. ۱۳۹۶. بررسی کارایی الگوریتم‌های کلاس‌بندی متن بر روی متون کوتاه فارسی. کنفرانس بین‌المللی فناوری اطلاعات، مهندسی کامپیوتر و مخابرات .