دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد هوش مصنوعی

خوشه‌بندی نیمه‌نظارتی متون با استفاده از تعبیه کلمات

نگارنده: سید مجتبی سجّادی

استاد راهنما

دکتر هدی مشایخی

استاد مشاور

دکتر حمید حسن‌پور

تیر ماه ۱۴۰۰

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای **سید مجتبی سجادی** با شماره دانشجویی **۹۷۰۸۹۷٤** رشته **مهندسی کامپیوتر** گرایش **هوش مصنوعی و رباتیکز** تحت عنوان **خوشه‌بندی نیمه‌نظارتی متون با استفاده از تعبیه کلمات** که در تاریخ ۱٤۰۰/۰٦/۱۰ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار شد به شرح ذیل اعلام می گردد:

الف) درجه عالی: نمره ۲۰-۱۹ ☑ ب) درجه خیلی خوب: نمره ۱۸/۹۹ – ۱۸ ☐
ج) درجه خوب: نمره۱۷/۹۹–۱۶ ☐ د) درجه متوسط: نمره ۱۵/۹۹-۱۴ ☐
ه) کمتر از ۱۴ غیر قابل قبول و نیاز به دفاع مجدد دارد ☐
نوع تحقیق: نظری ☐ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استاد راهنمای اول	دکتر هدی مشایخی	استادیار	
۲- استاد راهنمای دوم			
۳- استاد مشاور	دکتر حمید حسن‌پور	استاد	
۴- استاد داور اول	دکتر مرتضی زاهدی	استادیار	
۵- استاد داور دوم	دکتر مریم خدابخش	استادیار	
۶- نماینده تحصیلات تکمیلی	مهندس محسن فرهادی	مربی	

نام و نام خانوادگی رئیس دانشکده: دکتر علیرضا الفی

تاریخ و امضاء و مهر دانشکده:

تقدیم به:

پدر بزرگوار و مادر مهربانم،

دو فرشته ای که

برایم زندگی،بودن و انسان بودن را معنا کردند.

سپاس خداوندی را که بیاراست ارواح ما را به وجود اصل، و بپیراست اشباح ما را به سجود وصل، و در ما پوشید حله زندگی، و بر ما کشید رقم بندگی. گوهر جان در نهاد ما نهاد ضنانتی‌و خلعت ایمان بر سر ما افکند بی منتی. سواد دل ما را با نور شمع معرفت آشنایی داد و در اطباق احداق‌ما به کمال قدرت روشنایی نهاد.

و با تقدیر و تشکر شایسته از اساتید فرهیخته **سرکار خانم دکترهدی مشایخی** و **جناب آقای دکتر حمیدحسن‌پور** که با نکته های دلاویز و گفته های بلند، صحیفه‌های سخن را علم پرور نمودند و همواره راهنما و راه‌گشای نگارنده در اتمام واکمال پایان نامه بوده‌اند.

تعهد نامه

اینجانب **سید مجتبی سجادی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی کامپیوتر-هوش مصنوعی و رباتیکز** دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **خوشه‌بندی نیمه‌نظارتی متون با استفاده از تعبیه کلمات** تحت راهنمائی **دکتر هدی مشایخی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه 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 پایان‌نامه 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 می‌باشد 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 پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، 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 ۲۰ تیر ۱۴۰۰

امضای دانشجو: سید مجتبی سجادی

چکیده

خوشه‌بندی متن، در کاربردهای متنوعی از تحلیل متن مورد استفاده قرار می‌گیرد. در عملیات خوشه-بندی، روشی که برای بازنمایی متن استفاده می‌شود، تاثیر بسزایی در نتایج خواهد داشت. برخی روش-های متداول بازنمایی اسناد مبتنی بر کیسه‌ی کلمات، به تعداد تکرار لغات وابسته می‌باشند و بردارهای سندی با طول زیاد و تُنُک تولید می‌کنند. علاوه بر این، روش‌های مبتنی بر تعبیه‌کلمات و شبکه‌ی عصبی، مانند Doc2Vec از تفسیرپذیری پایین رنج می‌برند که با روی کار آمدن روش‌های مبتنی بر مفهوم، این نواقص تا حد زیادی برطرف شده است. با این وجود روش‌های موجود خوشه‌بندی نیمه‌نظارتی اسناد، از نمایش مفهومی اسناد استفاده نمی‌کنند. در این تحقیق، یک روش خوشه‌بندی نیمه‌نظارتی مبتنی بر مفهوم ارائه می‌گردد که از هر دو نوع داده‌ی برچسب‌دار و بدون برچسب به منظور ایجاد خوشه‌بندی با کیفیت‌تر استفاده می‌کند. اسناد بر اساس مفاهیمِ استخراج شده از مجموعه کلمات تعبیه-شده نمایش داده می‌شوند. این نحوه‌ی بازنمایی، علاوه بر حفظ اطلاعات مجاورتی اسناد، از تفسیرپذیری بالایی نیز برخوردار است. سپس فرآیند خوشه‌بندی نیمه‌نظارتی، ساختار کلی خوشه‌ها را با استفاده از داده‌های بدون برچسب و جایگاه دقیق مراکز خوشه را با بهره‌گیری از داده‌های برچسبدار تعیین می‌کند. ما همچنین نظریه‌ی مفاهیم نیمه‌نظارتی و روش جدید خوشه‌بندی اسناد را بر اساس وزن چنین مفاهیمی پیشنهاد می‌دهیم. نتایج این روش از طریق خوشه‌بندی نیمه‌نظارتی اسناد مورد ارزیابی قرار گرفته است. آزمایشات بر روی دو مجموعه دادگان متنی رویترز و 20-NewsGroup نشان می‌دهد که روش پیشنهادی در جنبه‌ی کیفیت خوشه‌بندی حداقل ۱۰ درصد و در دقت طبقه‌بندی متن، حداقل پنج درصد در مقایسه با سایر روش‌های موجود بهتر عمل می‌کند.

کلمات کلیدی: خوشه‌بندی نیمه‌نظارتی، خوشه‌بندی اسناد، تعبیه‌کلمات، نمایش مبتنی بر مفهوم، داده‌ی برچسب‌دار، طبقه‌بندی متن

لیست مقالات

- Seyed Mojtaba Sadjadi, Hoda Mashayekhi, Hamid Hassanpour, “Semi-supervised Concept-based Document Clustering”, *Information Processing & Management* (**Submitted on Jun 2021)**

- S. M. Sadjadi, H. Mashayekhi, H. Hassanpour, A Two-Level Semi-supervised Clustering Technique for News Articles, International Journal of Engineering, Vol. 34, No. 12, (2021), doi: 10.5829/ije.2021.34.12c.10

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱ : مقدمه

امروزه با گسترش روزافزون شبکه‌ی ارتباطات و اینترنت در دنیا، شبکه‌های اجتماعی، وبسایت‌های خبری، انجمن‌های یادگیری برخط[1] و کتابخانه‌های کتاب‌های الکترونیکی به منبع اصلی و بزرگی از داده‌های متنی تبدیل شده‌اند [1]. مدیریت و تحلیل این حجم عظیم از داده‌های متنی امری انکار ناشدنی در دنیای امروز به شمار می‌رود. به منظور استخراج اطلاعات با معنی از این داده‌های متنی، عملیات داده‌کاوی[2] بر روی آنها اعمال می‌شود. داده‌کاوی متنی به فرآیندی گفته می‌شود که در آن استخراج اطلاعات با اهمیت از متون با استفاده از یادگیری الگوهای آماری توسط مدل یادگیرنده انجام می‌پذیرد.

مهمترین گام در داده‌کاوی متنی خوشه‌بندی است که برای استفاده از آن به طور کلی از روش‌های استخراج اطلاعات، پردازش زبان طبیعی و تکنیک‌های یادگیری ماشین استفاده می‌شود. همچنین به منظور سازماندهی و مدیریت حجم زیادی از داده‌های متنی، از عملیات خوشه‌بندی استفاده می‌شود. خوشه‌بندی متن یا سند، به تکنیکی گفته می‌شود که در آن مجموعه‌ی اسناد به گروه‌های مجزایی تقسیم می‌شوند. اساس تقسیم‌بندی این گروه‌های کوچک یا بزرگ، شباهت بین محتوای اسناد است، بدین صورت که اسناد مشابه را در گروه‌هایی یکسان تقسیم‌بندی می‌کند [2]. خوشه‌بندی بر خلاف دسته‌بندی یا طبقه‌بندی[3] که به عوامل بیرونی برای برچسب‌گذاری مجموعه‌دادگان نیاز دارد، یک روش مهم برای یادگیری بدون نظارت و کشف ساختار ذاتی داده‌های بدون برچسب به‌شمار می‌رود [3]. این تکنیک که منجر به تجزیه و تحلیل جامع داده‌های متنی می‌گردد، دارای کاربردهای وسیعی در زمینه-های گوناگون می‌باشد. از جمله سیستم‌های توصیه‌گر [4] که اخیرا مورد توجه بسیاری از کسب و کارهای برخط قرار گرفته است. این نوع سیستم‌ها الگوریتم‌هایی هستند که توسط مدل یادگیرنده

[1] Online Learning Froum

[2] Data Mining

[3] Classification

دقیق‌ترین و مناسب‌ترین محصولات و سرویس‌های مشابه به نیاز کاربر را براساس رفتار کاربر و یا سلایق مشترک با سایر کاربران پیشنهاد می‌کنند. پیش‌نیاز این الگورریتم‌های توصیه‌کننده فرآیند خوشه‌بندی آیتم‌ها[۴] و سلایق کاربران می‌باشد. در کاربردی دیگر، برای نمونه در زمینه پزشکی، به منظور تشخیص و دسته‌بندی بیماری‌های مختلف [5] و یا پیش‌بینی دارو برای بیماری‌ها [6] از این تکنیک در ابتدایی-ترین مراحل تحلیل و یادگیری ماشین به وفور استفاده می‌شود. از جمله موارد دیگر برای کاربرد خوشه-بندی متون و اسناد می‌توان به تشخیص هرزنامه‌ها در پست الکترونیک و پیام‌های کوتاه الکترونیکی [7]، بازیابی اطلاعات [8]، مدلسازی عناوین[۵] [9]، تجزیه و تحلیل سری‌های زمانی و پردازش سیگنال [10]، طبقه‌بندی احساسات و نظرات کاربران در شبکه‌های اجتماعی [11]، خلاصه‌سازی متون یا اسناد[12] و غیره اشاره کرد.

۱-۱ مولفه‌های خوشه‌بندی اسناد

در اکثر الگوریتم‌های خوشه‌بندی اسناد، سه مرحله‌ی اصلی از قبیل پیش‌پردازش متن[۶]، نمایش (بازنمایی) متن[۷] و خوشه‌بندی وجود دارد. هر یک از این سه مورد نقش بسزایی در عملکرد و کیفیت خوشه‌های نهایی ایجادشده توسط مدل دارند. به همین منظور در ادامه به اختصار به بررسی هر یک از این موارد و روش‌های مختلف آن‌ها می‌پردازیم.

[۴] Items

[۵] Topic Modeling

[۶] Text Pre-processing

[۷] Text-representation | Document-representation

۱-۱-۱ پیش پردازش متن

پس از جمع‌آوری اطلاعات متنی، نیاز به جداسازی این داده‌ها به واحدهای تشکیل‌دهنده و پاکسازی آن‌ها از هرگونه علائم نگارشی، اعداد و نشان‌ها غیرمجاز است که به مفاهیم و محتوای اسناد مرتبط نیستند. به این فرآیند که تبدیل داده‌ی خام به فرمتی قابل فهم و خوانا برای ماشین است، پیش‌پردازش گفته می‌شود. برخی از کلمات و اجزای تشکیل دهنده متن به دلیل فقدان اطلاعات با معنی، در فرآیند یادگیری ماشین و تمیز دادن داده‌ها از یکدیگر نقشی موثری ندارند، از این رو برای حذف این کلمات، با استفاده از روش‌های مختلف که در فصل آینده تشریح خواهند شد، پیش‌پردازشی بر روی متن صورت خواهد گرفت.

۱-۱-۲ نمایش (بازنمایی) متن

نمایش متن که عموما کیفیت خوشه‌بندی حاصله به آن بستگی دارد، مجموعه روش‌هایی را شامل می‌شود که در آنها اسناد متنی به بردارهای ویژگی عددی[۸] تبدیل می‌شوند تا برای الگوریتم‌های خوشه-بندی و یادگیری‌ماشین قابل درک و استفاده باشند. کیفیت نمایش متن از دو جنبه مورد ارزیابی قرار می‌گیرد: کیفیت معنایی[۹] و کیفیت آماری[۱۰] [13].

کیفیت معنایی به این موضوع اشاره دارد که بردار متن تولید شده به چه اندازه تفسیرپذیر بوده و به چه میزان محتوای متن را توصیف می‌کند. یکی از معروف‌ترین روش‌های نمایش متن که به شهودی بودن و تفسیرپذیری بالا مشهور است، روش کیسه‌کلمات[۱۱] [14] می‌باشد که بردار اسناد را بر اساس تکرار

[۸] Numerical Feature Vectors

[۹] Semantic Quality

[۱۰] Statistical Quality

[۱۱] Bag-of-Words

کلمات تشکیل‌دهنده آن متن نمایش می‌دهد. اگرچه این روش نمایش متن قابلیت تفسیرپذیری زیادی دارد، اما به دلیل آنکه تمام کلمات یکتا[۱۲] موجود در متن را به عنوان ویژگی‌های بردار سند درنظر می‌گیرد، از بزرگی ابعاد و طولانی بودن بردار سند رنج می‌برد. همچنین زمانی که حجم مجموعه اسناد زیاد باشد، تعداد لغات یکتا زیاد می‌شود در نتیجه این روش قادر نخواهد بود تا اطلاعات مجاورتی بین اسناد را به خوبی حفظ کند.

کیفیت آماری بررسی می‌کند که چه مقدار بردار سند تولید شده قادر است تا اسناد را در دسته‌های گوناگون به خوبی از یکدیگر تمیز دهد. برخی روش‌ها از جمله شبکه عصبی کانوولوشن[۱۳] [15]، شبکه عصبی بازگشتی[۱۴] [16] و Doc2Vec [17] اسناد را با طول بردار کم نمایش می‌دهند و می‌توانند اطلاعات مجاورتی بین اسناد را به خوبی حفظ کنند. با این وجود، تفسیر نمایش‌های متن بدست آمده پیچیده است، زیرا مقدار هر ویژگی از طریق ساختارهای پیچیده وزن شبکه عصبی محاسبه می‌شود. برای غلبه بر ضعف روش‌های قبلی بازنمایی اسناد، اخیراً رویکرد دیگری ارائه شده است که از مفاهیم[۱۵] برای نمایش اسناد استفاده می‌کند [18]. پس از تعبیه کلمات تشکیل‌دهنده متون به فضای برداری، بردارهای کلمه تولید شده توسط فرآیندی خوشه‌بندی می‌شوند و هر یک از این خوشه‌ها یک مفهوم را شکل می‌دهند. اسناد بر اساس این مفاهیم استخراج شده نمایش داده می‌شوند، بنابراین بردارهای سند ابعاد معقولی دارند و همچنین نسبت به سایر روش‌های نمایش متن قابلیت تفسیر بالاتری خواهند داشت.

[۱۲] Unique

[۱۳] Convolutional Neural Networks (CNN)

[۱۴] Recurrent Neural Networks (RNN)

[۱۵] Concepts

۱-۱-۳ خوشه بندی

از آنجایی که فرآیند برچسب‌گذاری تمامی داده‌های موجود در مجموعه دادگان امری پرهزینه بوده و همچنین استفاده از روش‌های خوشه‌بندی بدون نظارت از دقت مناسبی برخوردار نمی‌باشند، استفاده از تعداد کمی داده دارای برچسب برای بهبود کیفیت خوشه‌بندی مورد توجه قرار گرفته‌است [19]. ایده اصلی در این روش‌ها بدین صورت است که داده‌های بدون برچسب شاکله‌ی کلی خوشه‌ها را مشخص می‌کنند و تعداد کمی از داده‌های دارای برچسب به منظور تنظیم مرکز خوشه‌ها مورد استفاده قرار می‌گیرند. این روش که از هر دو نوع داده‌ی برچسب‌دار و بدون برچسب بهره می‌برد، با نام خوشه‌بندی نیمه‌نظارتی شناخته می‌شود [20]. به طور کلی، الگوریتم‌های خوشه‌بندی نیمه نظارت شده به دو دسته تقسیم می‌شوند: رویکردهای مبتنی بر شباهت و رویکردهای مبتنی بر جستجو. در الگوریتم‌های مبتنی بر شباهت، ملاک تشابه اساسی داده‌ها مطابق با محدودیت ها و برچسب‌های داده‌های نظارت شده سازگار و تنظیم می‌شود. روش‌های پیشنهادی ژانگ و همکاران [21] و دارا و همکاران [22] بر این فرض استوار هستند که یک مدلِ‌مخلوط[۱۶]، جمعیت داده را تولید می‌کند. در این روش‌ها ، داده‌های بدون برچسب، برچسب‌گذاری شده و برای آموزش مدل استفاده می‌شوند. با این حال، مشخص نیست که به چه میزان برچسب‌گذاری مجدد داده‌ها نیاز است و این برچسب‌گذاریِ مجدد چقدر قابل اعتماد است. در روش‌های مبتنی بر جستجو، الگوریتم خوشه‌بندی نیمه نظارت شده با هدف جستجوی خوشه‌ی مناسب، با استفاده از برچسب‌ها یا محدودیت‌های ارائه شده توسط کاربر اصلاح می‌شود. خوشه‌بندی نیمه نظارت شده TESC [23] از یک روش خوشه‌بندی مبتنی بر جستجو برای کاربرد طبقه بندی

[۱۶] Mixture Model

متون استفاده می‌کند. هدف اصلی این روش بهبود کیفیت طبقه‌بندی با استفاده از روش مناسب خوشه-بندی است. این روش از خوشه‌بندی نیمه نظارتی سلسله مراتبی برای شناسایی اجزای[۱۷] تشکیل‌دهنده متن و استفاده از این اجزا برای پیش‌بینی برچسب برای داده های بدون برچسب استفاده می‌کند.

۱-۲ چالش‌های خوشه‌بندی اسناد

۱-۲-۱ مقدار زیاد داده‌های متنی

ما در دوره داده‌های بزرگ و حجیم زندگی می‌کنیم، جایی که مقدار زیادی داده از طریق منابع مختلف تولید می‌شود. تحقیقات انجام شده توسط محققان نشان داده است که اکثر داده های تولید شده امروزه در قالب متن است. گزارش منتشر شده توسط شرکت DataStax [24] ، در مورد انواع مختلف داده‌های تولید شده و ذخیره شده در صنایع مختلف نشان داد که بخش عظیمی از نوع داده تولید شده مربوط به قالب‌های متنی است. جدول (۱-۱) خلاصه‌ای از این گزارش را ارائه می‌دهد. از این‌رو، خوشه‌بندی این مقدار زیاد داده به صورت برچسب‌دار تلاش‌های طاقت فرسایی را می‌طلبد و عملا غیرممکن است. این امر مستلزم توسعه تکنیک های نوآورانه و استراتژی های موثر برای خوشه‌بندی نیمه‌نظارتی متون است.

جدول ۱-۱. تنوع نوع داده‌ها در صنایع مختلف [24]

	فیلم	عکس	صدا	متن
بانک	متوسط	متوسط	کم	زیاد
ساخت و تولید	متوسط	متوسط	کم	زیاد
خرده فروشی	متوسط	کم	کم	زیاد

[۱۷] Components

سلامت	کم	زیاد	کم	زیاد
ارتباطات	زیاد	متوسط	زیاد	زیاد
ساخت و ساز	کم	متوسط	کم	متوسط
دولت	زیاد	متوسط	زیاد	زیاد
آموزش	زیاد	متوسط	زیاد	متوسط

۱-۲-۲ بازنمایی عددی کلمات و اسناد

همانطور که در بخش‌های گذشته اشاره شد، به طور کلی در خوشه‌بندی اسناد متنی، لازم است اسناد متنی را به نمایش عددی تبدیل کنیم. از یک طرح وزنی برای بیان اصطلاحات و کلمات بر اساس تعداد وقوع آن‌ها در سند استفاده می‌شود. ابتدایی‌ترین روش رخداد کلمات[۱۸] است که مبتنی بر تعداد وقوع یک اصطلاح در یک سند می‌باشد. علاوه بر این، «رخداد کلمات – معکوس رخداد سند»[۱۹] یکی دیگر از طرح‌های وزنی است که عمدتا مورد استفاده قرار می‌گیرد. هدف این روش نشان دادن اهمیت هر کلمه در اسناد بر اساس این واقعیت است که کلماتی که غالبا در تعداد کمی از اسناد وجود دارند و در دیگر اسناد نادر هستند، بیشتر با برچسب سند مربوطه نزدیک هستند.

۱-۲-۳ تُنُکی[۲۰] بردار اسناد

تنکی بردار سند یک مسئله حائز اهمیت در خوشه‌بندی اسناد است. بردار سند در مقیاس بزرگ به طور قابل توجهی تُنُک است، که می تواند تأثیر منفی بر زمان پردازش داشته باشد. در مجموعه اسناد، آن

[۱۸] Term Frequency

[۱۹] Term Frequency and Inverse Document Frequency (TF-IDF)

[۲۰] Sparsity

سندی که کلمات کمتری نسبت به کل کلمات مجزا در مجموعه اسناد دارند، بردار سند تُنُک‌تری خواهند داشت. این امر باعث می‌شود که بردار سند تعداد زیادی سلول صفر داشته باشد. بنابراین پرداختن و تجزیه و تحلیل چنین بردار تنکی یک کار چالش برانگیز است.

۱-۲-۴ انتخاب مراکز اولیه

در برخی از روش‌های خوشه‌بندی مانند K-Means، انتخاب مراکز اولیه نقش مهمی در شکل‌گیری نتایج خوشه‌بندی دارد [25]. انتخاب مقادیر تصادفی به عنوان مراکز اولیه هنگام اجرای الگوریتم در دفعات بعدی به نتایج خوشه‌بندی متفاوتی منجر می‌شود [26]. تعیین روش‌های جدید برای انتخاب مراکز اولیه خوشه‌بندی، یک نتیجه خوشه بندی با کیفیت و منحصر به فرد را تضمین می‌کند. اگرچه چندین تلاش برای حل این مشکل انجام شده است اما هنوز هم نیاز به یافتن یک روش استاندارد وجود دارد که برای کاربردهای مختلف قابل استفاده باشد.

۱-۲-۵ معیار شباهت یا عدم تشابه (فاصله)

در فرآیند خوشه‌بندی ، از اندازه‌گیری شباهت و عدم تشابه (فاصله) برای ارزیابی میزان نزدیکی یا تفکیک بین اسناد استفاده می‌شود. اصولاً، از این معیارها برای ارزیابی میزان شباهت دو سند استفاده می‌شود [27]. پایین بودن مقدار اندازه‌گیری‌شده فاصله نشانگر درجه بالایی از نزدیکی است. برخی از معیارهای فاصله که در خوشه‌بندی اسناد بیشتر استفاده می‌شوند عبارتند از: منهتن[۲۱]، اقلیدوسی[۲۲]، کسینوسی[۲۳]،

[۲۱] Manhattan

[۲۲] Euclidean

[۲۳] Cosine

مینکوفسکی[۲۴] و غیره. در معیارهای شباهت، مقدار بین $[1 , 1-]$ یا $[1 , 0]$ است. صفر یا $1-$ نمایش عدم شباهت و عدد یک نمایانگر شباهت مطلق است.

۱-۲-۶ روش های ارزیابی خوشه

ارزیابی خوشه برای اندازه‌گیری کیفیت یک خوشه استفاده می‌شود و در سنجش اینکه نتایج خوشه‌بندی چقدر خوب و دقیق است، مفید می‌باشد. علاوه‌براین، برای درک چگونگی تفکیک و میزان تراکم خوشه‌ها از معیارهای ارزیابی استفاده می‌شود. برخی اندازه‌گیری‌ها مانند اطلاعات متقابل عادی[۲۵]، خطای میانگین مربعات[۲۶]، آنتروپی[۲۷]، دقت[۲۸] و غیره برای ارزیابی خوشه‌های حاصله از فرآیند خوشه‌بندی استفاده می‌شود.

۱-۳ تعریف مسئله و اهداف پایان‌نامه

این پژوهش به چالش‌های موجودِ خوشه‌بندی اسناد از قبیل **مقدار زیاد داده‌های متنی، بازنمایی عددی کلمات و اسناد، تنکی بردار سند** و **نحوه‌ی انتخاب مراکز اولیه** می‌پردازد و برای حل این چالش‌ها روش‌هایی را ارئه می‌دهد. در این پایان‌نامه، ما یک روش خوشه‌بندی نیمه‌نظارتی جدید مبتنی بر مفاهیم را ارائه می‌دهیم که برخلاف برخی از الگوریتم‌های خوشه‌بندی نیمه‌نظارتی رایج [22][21][20]، از داده‌های بدون‌برچسب و یک مجموعه محدود از داده‌های برچسب‌دار به طور همزمان

[۲۴] Minkowski
[۲۵] Normal Mutual Information (NMI)
[۲۶] Mean squared error
[۲۷] Entropy
[۲۸] Accuracy

در فرآیند خوشه‌بندی استفاده می‌کند. با توجه به دانش ما، هیچ یک از روش‌های خوشه‌بندی نیمه‌نظارتی فعلی از نمایش مفهومی اسناد استفاده نمی‌کنند. این بازنمایی (نمایش مفهومی) این مزیت را دارد که برخلاف سایر روش‌های نمایش اسناد که بردارهای سند را مستقیماً براساس تکرار وقوع واژه‌های آن‌ها می‌سازند، اسناد با مفاهیم مشترک را در یک خوشه قرار می‌دهد. این روش می‌تواند ویژگی‌های اساسی و ویژگی‌های متمایز اسناد را با حفظ تفسیرپذیری بردار سند ثبت کند. از دیگر مزایای آن می‌توان به حفظ اطلاعات مجاورتی بین اسناد و اطلاعات معنایی غیرخطی بین کلمات هر سند اشاره نمود و همانطور که در بخش آزمایشات نشان داده شده است، با استخراج نیمه نظارتی مفاهیم موجود در متن و خوشه-بندی سلسله‌مراتبی بردارهای مفهوم ساخته شده، کیفیت خوشه‌بندی را به طور قابل توجهی بهبود می‌بخشد.

پس از تعبیه‌ی کلمات موجود در متون مجموعه دادگان به فضای برداری، با استخراج مفاهیم از مجموعه-ی بردارِ کلمات تولید شده، این مفاهیم به عنوان اجزای تشکیل‌دهنده اسناد شناسایی شده و اسناد بر اساس آن‌ها (مفاهیم مستخرج) نمایش داده می‌شوند. این نمایش اسناد (بردارهای مفهومی اسناد) در فرآیند خوشه‌بندی نیمه‌نظارتی استفاده می‌شوند. روش پیشنهادی می‌تواند به صورت دلخواه از داده‌های دارای برچسب در مرحله استخراج مفاهیم یا مرحله خوشه‌بندی اسناد استفاده کند. اگرچه بازنمایی اسناد بر اساس مفاهیم سابقاً توسط برخی از مطالعات خارج از زمینه خوشه‌بندی نیمه‌نظارتی انجام شده است، در این پایان‌نامه، ما ایده جدید مفاهیم نیمه نظارتی را برای نمایش و بازنمایی اسناد ارائه می‌دهیم. مفاهیم نیمه نظارت شده، مولفه‌های معنایی مجموعه متون و همخوانی آنها با برچسب های کلاس داده‌ها را به دست می‌آورند. مدل پیشنهادی قابل تفسیر است و درک عمیق تر و شفاف‌تری از مجموعه عملیات صورت گرفته برای استدلال را فراهم می کند. علاوه بر این، از آنجایی که اسناد به عنوان مجموعه ای از

مفاهیم برچسب‌دار بیان می‌شوند، درک شهودی از اجزای معنایی تشکیل‌دهنده اسناد و تناظر با برچسب-های متناظر آن‌ها حاصل می‌شود. در مرحله خوشه‌بندی نیمه نظارت شده ، اسنادِ بدون برچسب برای شناسایی ساختار کلی خوشه‌ها و تعداد کمی از اسناد دارای برچسب برای تعیین دقیق‌تر مراکز خوشه استفاده می‌شوند.

ما در بخش‌های پایانی یک ارزیابی جامع از روش پیشنهادی را با استفاده از دو مجموعه دادگان متنی شناخته شده انجام می‌دهیم و روش پیشنهادی را در هر دو جنبه کیفیت خوشه‌بندی و دقت طبقه‌بندی با چندین الگوریتم کلاسیک و جدید مقایسه می‌کنیم. نتایج ارزیابی‌ها نشان می‌دهد که روش پیشنهادی دارای برتری قابل توجهی نسبت به روش‌های گذشته خوشه بندی نیمه‌نظارتی اسناد است.

مشارکت اصلی این پایان‌نامه به شرح زیر است:

• پیشنهاد یک روش جدید خوشه‌بندی نیمه‌نظارتی اسناد بر اساس نمایش مفهومی اسناد.

• ارائه روش‌هایی که شامل استفاده مجموعه محدودی از اسناد برچسب‌دار در فرآیند استخراج مفهوم یا در مرحله خوشه‌بندی اسناد است.

• ارائه رویکرد جدید مفاهیم نیمه‌نظارتی برای نمایش اسناد.

• انجام ارزیابی جامع برای تجزیه و تحلیل روش‌های پیشنهادی بر اساس کاربردهای خوشه‌بندی و طبقه‌بندی اسناد.

۱-۴ ساختار پایان‌نامه

ادامه پایان‌نامه به شرح مقابل می‌باشد: در فصل دوم، مطالعات گذشته مربوط به پیش‌پردازش متن و نمایش سند و خوشه‌بندی نیمه‌نظارتی اسناد مرور می‌شود. روش جدید خوشه بندی نیمه‌نظارتی پیشنهادی ما در فصل سوم ارائه می‌گردد. فصل چهارم به نتایج تجربی، تجزیه و تحلیل دقیق آزمایشات و مقایسه روش پیشنهادی با سایر روش‌ها می‌پردازد. در نهایت، نتیجه‌گیری‌ها و کارهای پیشنهادی آینده در بخش پنجم بیان می‌شوند.

فصل۲ : مروری بر کارهای گذشته

۲-۱ مقدمه

در سال‌های اخیر پژوهشگران بسیاری در زمینه‌های متفاوت خوشه‌بندی اسناد، مطالعاتی را انجام دادند. همانطور که در فصل گذشته گفته شد، علاوه بر الگوریتم خوشه‌بندی، نحوه‌ی نمایش متن نیز بر کیفیت خوشه‌های حاصله نقش بسزایی خواهد داشت. از این رو در این فصل، در ابتدا مروری بر روش‌های مهم پیش‌پردازش متن و نمایش اسناد خواهیم داشت و سپس به الگوریتم‌های خوشه‌بندی نیمه‌نظارتی اسناد خواهیم پرداخت.

۲-۲ پیش پردازش متن

قبل از هرگونه فرآیند خوشه‌بندی یا یادگیری ماشین، عملیات مختلفی برای پاکسازی متون و تبدیل آن‌ها به فرمت قابل درک برای ماشین انجام می‌گردد. روش‌های گوناگونی که برای پیش‌پردازش متون استفاده می‌شوند، به شرح ذیل می‌باشند:

۱. تبدیل کردن تمامی حروف بزرگ[۲۹] موجود در متن به حروف کوچک[۳۰]
۲. تبدیل کردن اعداد به حروف متنی و یا حذف کامل اعداد از متون
۳. حذف کردن علائم نگارشی[۳۱] ، علائم لهجه[۳۲] و علائم تشخیص[۳۳]
۴. بسط یا گسترش اختصارها

[۲۹] Uppercase Letters
[۳۰] Lowercase Letters
[۳۱] Punctuations
[۳۲] Accent Marks
[۳۳] Diacritics

۵. تصحیح کردن خطاهای املایی موجود در متن[۳۴]

۶. حذف فضاهای خالی[۳۵] و یکپارچه کردن متون

۷. حذف کلمات بی‌اثر[۳۶]

۸. کانونی‌سازی داده‌های متنی[۳۷]

۹. جداسازی و قطعه‌بندی متون به کلمه یا کلمات[۳۸]

۱۰. ریشه‌یابی کلمات موجود در متن[۳۹]

۱۱. برچسب‌گذاری نقش دستوری کلمات موجود در متون[۴۰]

۱۲. روش تقطیع[۴۱] یا تجزیه سطحی جملات[۴۲] موجود در متون

۱۳. بازشناسی موجودیت‌های اسم‌دار[۴۳] در داده‌های متنی

۱۴. استخراج عبارات هم‌اتفاق[۴۴] در متون

در اینجا به توضیح برخی از مهمترین پیش‌پردازش‌های متنی خواهیم پرداخت. «جداسازی کلمات» به روشی اطلاق می‌شود که در آن یک متن به واحدهای کوچکتری به نام توکن[۴۵] تجزیه می‌شود. کلمات، علائم نگارشی، اعداد و غیره از مجموعه‌ی واحدهای متن هستند که با نام توکن شناخته می‌شوند. نیاز

---

[۳۴] Spell Correction
[۳۵] Whitespaces
[۳۶] Stopwords Elimination
[۳۷] Text Canonicalization
[۳۸] Tokenization
[۳۹] Stemming
[۴۰] Part Of Speech tagging | POS Tagging
[۴۱] Chunking
[۴۲] Shallow Parsing
[۴۳] Named Entity Recognition
[۴۴] Callocation
[۴۵] Token

است برای نمایش متن به صورت عددی و برداری، واحدهای تشکیل‌دهنده آن شناسایی و جداسازی شوند.

«کلمات بی‌اثر» یا «کلمات توقف» کلماتی در متن را گویند که به کرّات در تمام متون دیده می‌شوند. برای مثال حروف اضافه در زبان انگلیسی و کلمات ربط در زبان فارسی در این مجموعه قرار می‌گیرند. از آنجایی که این کلمات بار معنایی و محتوایی خاصی ندارند و متمایز کننده خوبی برای خوشه‌بندی اسناد به‌شمار نمی‌روند، حذف آنها قبل از خوشه‌بندی و یادگیری توسط ماشین امری ضروری می‌باشد.

«ریشه‌یابی کلمات» به فرآیندی گفته می‌شود که در آن کلمات موجود در متن به شکل ریشه‌ای خود بازمی‌گردند. به عنوان مثال کلمه‌ی ”Processing“ در زبان انگلیسی از ریشه‌ی ”Process“ است. الگوریتم‌های ریشه‌یابی معمولا یکی از دو عملیات بُن‌واژه‌یاب[۴۶] و یا Lemmatization را اجرا می‌کنند. در عملیات بُن‌واژه‌یاب بخش‌های "عطفی[۴۷]" و "مورفولوژیک[۴۸]" از انتهای کلمات حذف می‌شوند اما در عملیات Lemmatization بخش های موفولوژیک و عطفی کلمات حذف نمی‌شوند بلکه از "مجموعه دادگان دانش لغوی[49]" به منظور یافتن ریشه کلمات استفاده می‌شود.

«برچسب‌گذاری نقش دستوری کلمات» در متن، الگوریتم دیگری است که در آن بر اساس محتوای متن و جمله، نقش دستوری (صفت، قید، فعل، فاعل، مفعول و غیره) هر کلمه تشکیل دهنده متن تعیین می‌شود. «تقطیع» روش دیگری از پیش‌پردازش‌های زبان طبیعی است که در آن کلماتی که نقش دستوری آنها در متون توسط «برچسب‌گذاری نقش دستوری» مشخص شده است را به واحدهای زبانی

[۴۶] Stemming
[۴۷] Inflectional
[۴۸] Morphological
[۴۹] Lexical Knowledge Base

مرتبه بالاتر (گروه‌های اسمی، گروه‌های فعلی و غیره) وصل می‌کند. در این صورت است که می‌توان ساختار درختی جملات متون را رسم کرد و نمایش داد.

روش «بازشناسی موجودیت‌های اسم‌دار» الگوریتم‌هایی هستند که موجودیت‌های موجود در متن را به دسته‌های از پیش‌تعیین‌شده‌ای (نام مکان‌ها، نام اشخاص، نام اشیا) نسبت می‌دهند. مزیت این الگوریتم‌ها در این است که مفاهیم موجود در متن به راحتی قابل دریافت می‌باشند و می‌توان در آینده از این مفاهیم برای نمایش اسناد استفاده کرد.

ممکن است در یک مجموعه دادگان از تمام روش‌های ذکر شده استفاده نشود و متناسب با کاربرد و نوع اطلاعات مستخرج مورد نیاز، از یک یا چند روش پیش‌پردازش متن به منظور ایجاد ساختاری مناسب برای تبدیل متن به بردارهای عددی بهره برده شود. پس از اعمال پیش‌پردازش‌های مورد نیاز، توکن‌های بدست آمده برای ایجاد بردارهای بازنمایی متن مورد استفاده قرار می‌گیرند.

۲-۳ نمایش اسناد

۲-۳-۱ کدگذاری One-hot

برای نمایش اسناد و متون به صورت عددی، یکی از ساده‌ترین روش‌ها کدگذاری به صورت بردار One-hot [28] می‌باشد. در این روش، بردار کلمه به صورت برداری از اوزان (مقادیر عددی) نمایش داده می‌شود که هر درایه از این بردار نمایانگر یک کلمه در مجموعه تمام کلمات لغت‌نامه خواهد بود. برای هر کلمه نیز تنها درایه مربوطه در بردار یک خواهد بود و مابقی درایه‌های بردار صفر می‌شوند. برای نمونه فرض کنید لغت‌نامه تنها شامل ۵ کلمه است. همانطور که در شکل (۲-۱) دیده می‌شود، بردار کلمه دارای پنج درایه و در جایگاه کلمه مورد نظر عدد یک و مابقی عدد صفر قرار می‌گیرد.

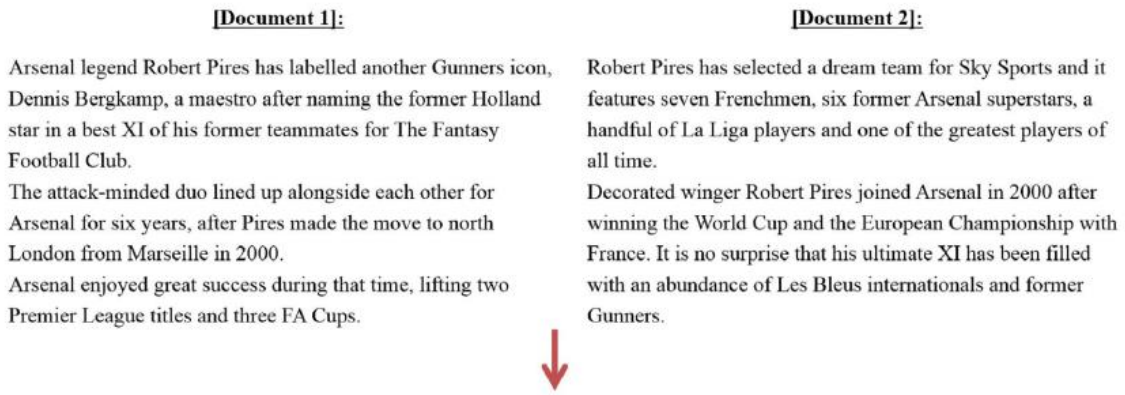
شکل ۲-۱. کدگذاری one-hot برای نمایش کلمه

پس از ایجاد بردار One-hot هر کلمه، می‌توان با جمع کردن بردارهای تشکیل‌دهنده متون به بردار کلی متن دست پیدا کرد که تحت عنوان روشی دیگر در **بخش ۲-۲-۲** تشریح می‌شود. این روش نمایش متن به دلیل آنکه به تعداد کلمات لغت‌نامه وابسته است، طول و پیچیدگی محاسباتی زیادی خواهد داشت و روش مناسبی به نظر نمی‌رسد.

۲-۳-۲ روش کیسه کلمات (Bag-of-Words)

متداول‌ترین روش‌های بازنمایی اسناد اغلب به روش‌های مبتنی بر کلمات تکیه کرده‌اند [29] [30] [31]، که در آن‌ها یک سند اساساً با تعداد وقایع کلمات در یک سند نشان داده می‌شود. برای دهه‌ها ثابت شده است که این روش در حل کاربردهای مختلف متن کاوی موثر است [32] [33] [34]. یکی از مزایای اصلی آن تولید بردارهای سندی قابل تفسیر بصری است، زیرا هر ویژگی بردار سند نشان دهنده وقوع یک کلمه خاص در مجموعه اسناد است. با توجه به این ویژگی‌های قابل فهم، بردارهای سندی که از رویکرد کیسه‌کلمات تولید شده‌اند به‌راحتی قابل تفسیر می‌باشند. اگر دو سند از شکل (۲-۲) به عنوان اسناد مشابه محاسبه شده باشند، می‌توان دلیل تشابه آن‌ها را با مشاهده مستقیم و مقایسه ویژگی‌های هر بردار سند توضیح داد. به عنوان مثال، این دو بردار سند را می‌توان به دلیل این که تعداد تکرار کلمات Arsenal، Legend، Robert و Pires نزدیک هستند، مشابه دانست. در نتیجه، می‌توان

پذیرفت که این دو بردار سند مشابه هستند، زیرا هر دو در مورد یک بازیکن از تیم فوتبال خاص بحث می‌کنند.



	X[1]: Arsenal	X[2]: Legend	X[3]: Robert	X[4]: Pires	…
Document 1	3	1	1	2	…
Document 2	2	0	2	2	…

شکل ۲-۲.بردار اسناد تولید شده توسط رویکرد کیسه کلمات [18]

رویکرد کیسه کلمات، هنگامی که تعداد اسناد ارائه شده بسیار زیاد باشد، می‌تواند مشکل‌ساز باشد. با افزایش تعداد اسناد، تعداد کلمات منحصر به فرد در کل مجموعه اسناد نیز به طور طبیعی افزایش می‌یابد. در نتیجه، بردارهای سند تولید شده نه تنها تنک می‌شوند، بلکه ابعاد آن‌ها نیز بزرگ خواهد بود. هرچه ابعاد و تنکی بردارهای سند افزایش یابد، معیارهای رایج فاصله و شباهت در نمایش مجاورت مناسب بین اسناد بی اثر می‌شوند. بعلاوه، روش کیسه کلمات فرض می‌کند که تمام کلمات موجود در اسناد مستقل هستند. با این حال، انواع مختلف کلمات مانند مترادف‌ها و زیرنام[۵۰]ها معمولا اطلاعات مشابه را در سند توصیف می‌کنند. بنابراین، این فرض استقلال کلمات، تاثیر نامطلوبی در محاسبه مجاورت اسناد دارد. در نتیجه، مدل های پردازش متنِ ساخته شده با رویکرد کیسه کلمات می‌توانند در

[۵۰] Hypernym

اندازه‌گیری فاصله و شباهت مناسب بین بردارهای سند با ابعاد بالا، تُنُک و ناموفق باشند. اگرچه تکنیک-های مختلف کاهش ابعاد [35] وجود دارد، اما این روش‌ها تفسیرپذیری ذاتی رویکرد کیسه کلمات را از دست می‌دهند. برخی از روش‌های بازنمایی متن، مانند تجزیه و تحلیل معنایی نهفته[۵۱] (LSA) [36] و تخصیص پنهان دیریکله[۵۲] (LDA) [37]، ابعاد بردار سند را کاهش می‌دهند. بر اساس تجزیه مقدار منفرد و تجزیه و تحلیل آماری ماتریس کلمات، این روش‌ها ماتریس کلمات را به یک بردار متراکم و با بعد کم تقلیل می‌دهند. اگرچه این روش‌ها به مراتب بهتر از روش کیسه‌کلمات عمل می‌کنند، اما ماتریس را در فضای خطی کاهش می‌دهند و در کشف روابط معنایی غیرخطی میان کلمات ناکام هستند.

۲-۳-۳ روش تعبیه کلمات[۵۳]

تعبیه‌ی کلمه [38] مجموعه‌ای از تکنیک‌های مدل‌سازی زبان در پردازش زبان طبیعی است که در آن می‌توان روابط معنایی کلمات را در یک فضای برداری کم حجم و متراکم ثبت کرد. یکی از این تکنیک‌ها، روش Word2Vec [39] بر این فرض استوار است که کلمات و اصطلاحاتی که در موضوعات مشابه وجود دارند، معانی مشابهی دارند. مدل تعبیه شده Skip-Gram [40] شامل یک شبکه‌ی عصبی دولایه است که برای تبدیل یک مجموعه متن بزرگ خام به یک فضای برداری چند بعدی استفاده می‌شود. همانطور که از نام آن پیداست، Word2Vec هر کلمه را با یک بردار منحصر به فرد، استخراج شده از وزن شبکه عصبی، نشان می‌دهد که میزان شباهت معنایی بین کلمات را حفظ می‌کند. بسیاری از برنامه های دنیای واقعی از تعبیه‌ی کلمه از قبل آموزش دیده استفاده می‌کنند. یکی از مشهورترین آنها روش میانگین Word2Vec است [41]. اگرچه Word2Vec یک روش نمایش کلمه است، اما بدون تغییرات

[۵۱] Latent Semantic Analysis

[۵۲] Latent Dirichlet Allocation

[۵۳] Word Embedding

قابل توجهی می‌تواند به روشی برای بازنمایی اسناد گسترش یابد. بنابراین، ما ابتدا Word2Vec را قبل از بحث در مورد بازنمایی سند مبتنی بر Doc2Vec، تشریح خواهیم کرد.

Word2Vec مبتنی بر فرضیه توزیع شده[۵۴] [36] است، که بیان می‌کند کلماتی که در متون مشابه رخ می‌دهند، تمایل دارند که معنای مشابه داشته باشند [42]. بر اساس این فرض، Word2Vec از یک شبکه‌ی عصبی ساده برای تعبیه و جاسازی کلمات در فضای برداری پیوسته استفاده می‌کند. از طریق آموزش وزن شبکه مدل Word2Vec (Skip-Gram) با استفاده از یک پنجره لغزان با اندازه از پیش-تعریف‌شده، کلمات همسایه‌ی کلمه ورودی را در اندازه خاص پیش‌بینی می‌کند.

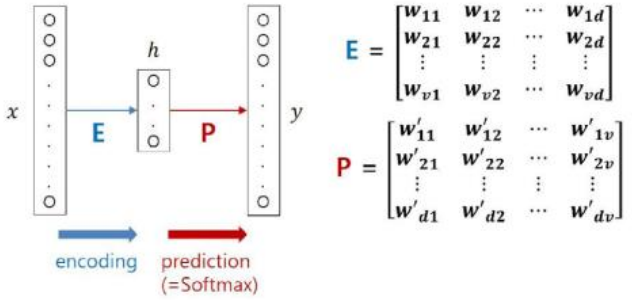
همانطور که در شکل (۲-۳) نشان داده شده است، اندازه لایه ورودی x برابر V است، این مقدار معادل با کل کلمات منحصر به فرد در یک مجموعه اسناد می‌باشد. هر گره[۵۵] از لایه ورودی یک کلمه منحصر به فرد را با رمزگذاری One-hot نشان می‌دهد. از طریق ماتریس E، که اساساً ماتریس وزن هر گره ورودی به هر گره لایه مخفی[۵۶] است، متن کلمه ورودی تعبیه‌شده و توسط لایه مخفی h نمایش داده می‌شود. درنتیجه، تعداد گره های مخفی d بیانگر ابعاد بردارهای کلمه و فضای تعبیه‌شده است. بردار متن تعبیه شده h پس از آن در بردار کلمه متناظر از ماتریس P ضرب می‌شود، که P ماتریس وزن هر گره پنهان به هر گره در لایه خروجی است. از طریق P ، کلمات متن اطراف کلمات ورودی با تابع فعال‌ساز Soft-Max پیش‌بینی می‌شوند، که هدف آن به حداکثر رساندن ضرب خارجی بین بردارهای متن تعبیه شده کلمات ورودی و بردارهای کلمه حاصله است. سپس، مقادیر احتمال پیش‌بینی شده برای هر کلمه به‌وسیله‌ی مقدار هر گره در لایه خروجی y نشان داده می‌شود. همانند لایه ورودی، اندازه لایه خروجی y مجددا V است. با بررسی اینکه آیا کلمات پیش‌بینی شده واقعاً در اطراف کلمات ورودی

[۵۴] Distributed Hypothesis

[۵۵] Node

[۵۶] Hidden Layer

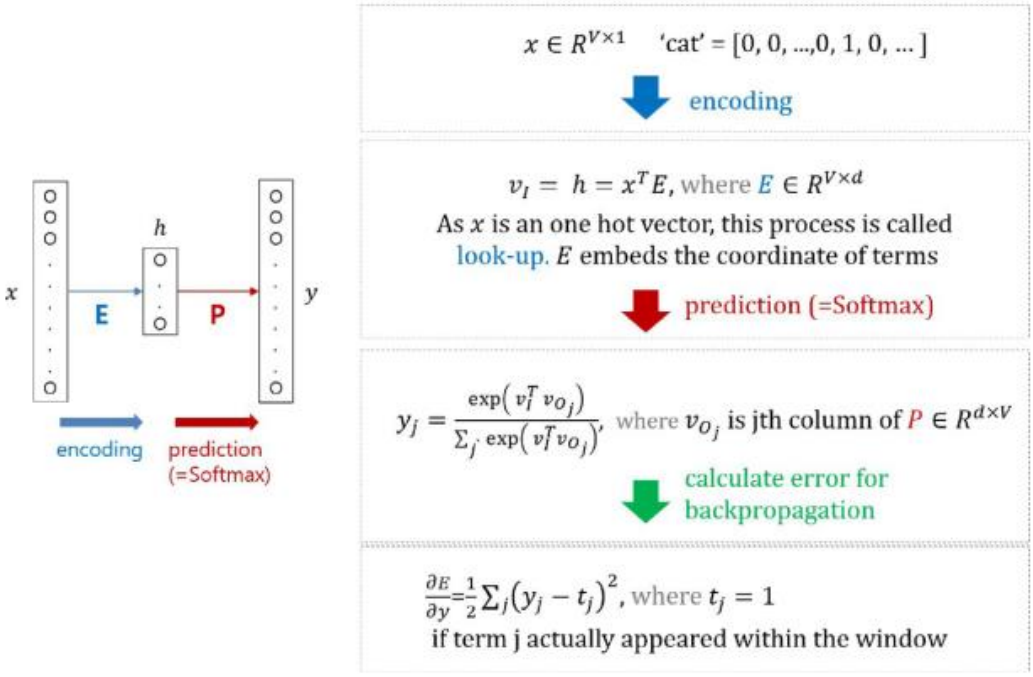
رخ داده اند، دقت پیش‌بینی ارزیابی می‌شود. از طریق رویکرد Back-Propagation، مقادیر (وزن) بردارهای تعبیه‌شده و بردارهای کلمه به روز می‌شوند.



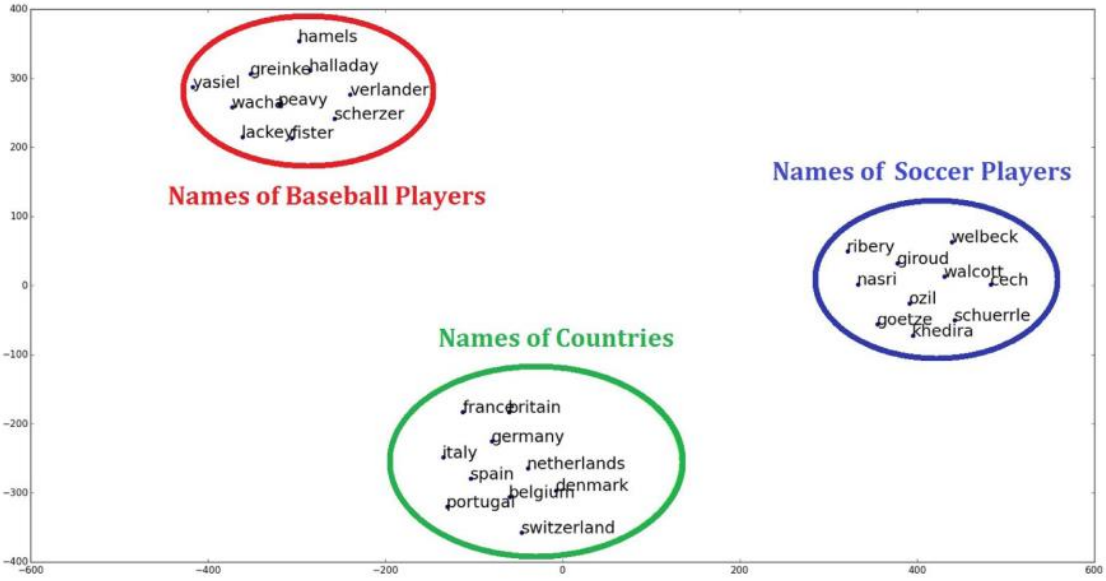
$x \in R^{V\times 1}$	One hot representation of a word. Size of input layer is equal to the number of terms
$h \in R^{d\times 1}$	Encoded context vector
$y \in R^{V\times 1}$	Predicted word prob. Size of output layer is also equal to the number of terms
$E \in R^{V\times d}$	Embedding Matrix
w_{ij}	Coordinate of context vector. Weight between the ith input node and jth hidden node
$P \in R^{d\times V}$	Word Matrix
w'_{ij}	Coordinate of word vector. Weight between the ith hidden node and jth output node

شکل ۲-۳. معماری مدل شبکه‌ی عصبی Word2Vec (Skip-Gram) [18]

توضیحات در مورد آموزش شبکه‌ی عصبی Skip-Gram در شکل (۲-۴) دیده می‌شود. همانطور که کلمات در یک فضای تعبیه‌شده پیوسته نشان داده می‌شوند، می‌توان در این فضا تکنیک‌های متداول یادگیری ماشین و داده کاوی را برای حل کارهای مختلف متن کاوی به کار برد [43] [44] [45].



شکل ۲-۴. آموزش شبکه‌ی عصبی Skip-Gram [18]



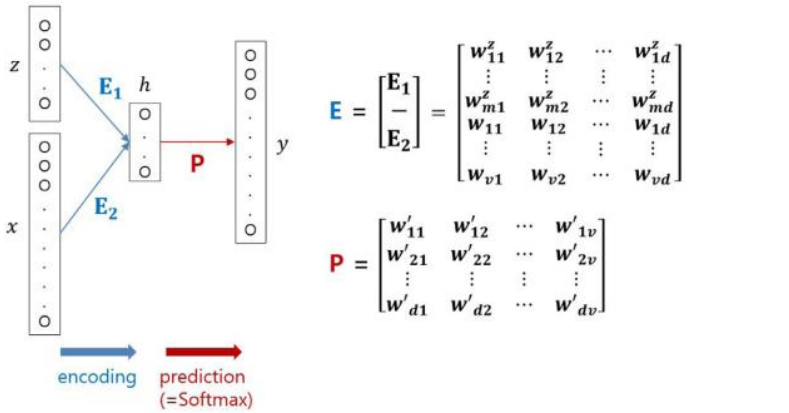
شکل ۲-۵. فضای تعبیه شده با استفاده از t-sne [18]

شکل (۲-۵) نمونه ای از چنین فضای تعبیه‌شده‌ای را نشان می‌دهد که توسط t-sne [46] مصور شده است. در این شکل، کلماتی تعبیه شده است که نشان‌دهنده نام بازیکنان بیس بال، نام بازیکنان فوتبال

و نام کشورها می‌باشد. در حالی که کلمات با معانی مشابه در نزدیکی یکدیگر قرار دارند، کلمات با معانی مختلف در فاصله دور از یکدیگر قرار گرفته‌اند. در مقایسه با روش کیسه کلمات، که در آن بعد و تُنُکی بردار سند می تواند به طور قابل توجهی افزایش یابد، می توان از مدل Word2Vec برای ساخت بردارهای سند متراکم با ابعاد مناسب استفاده کرد.

۲-۳-۴ روش Doc2Vec

معماری و آموزش شبکه عصبی در Doc2Vec با Word2Vec یکسان است. تنها تفاوت در این است که اسناد نیز به عنوان ورودی در شبکه گنجانده شده‌اند. مشابه کلمات در مدل Word2Vec ، اسناد با یک کدگذاری One-hot نشان داده می‌شوند و از طریق یک ماتریس تعبیه شده در یک فضای پیوسته جاسازی می‌شوند. همانطور که در شکل (۲-۶) نشان داده شده است، $E1$ یک ماتریس تعبیه شده برای اسناد را نشان می‌دهد، در حالی که $E2$ یک ماتریس تعبیه را برای کلمات است.



$z \in R^{m \times 1}$ One hot representation of a document. Size of input layer is equal to the number of documents

w^z_{ij} Coordinate of document embedding vector. Weight between the ith input node of z and jth hidden node

شکل ۲-۶. معماری مدل Doc2Vec [18]

قدرت بازنمایی Doc2Vec در کاربردهای خوشه‌بندی و طبقه‌بندی اسناد موثر است، و از روش Word2Vec بهتر عمل می‌کند [47]. حتی وقتی ابعاد بردارهای سند کوچکتر از روش کیسه کلمات انتخاب شده باشند، Doc2Vec به اندازه کافی اطلاعات متنی کلمات و اسناد را در خود جای داده‌است، در نتیجه از مدل‌های مبتنی بر کیسه کلمات بهتر عمل می‌کند. مدل Doc2Vec با وجود قدرت بازنمایی موثر، قادر به ارائه تفسیر بصری در بردارهای سند تولید شده خود نیست. از آنجا که هر بردار سند از طریق یک شبکه‌ی عصبی آموزش دیده است، هر درایه بردار سند فقط قدرت اتصال بین گره‌های ورودی و گره‌های لایه‌ی مخفی را نشان می‌دهد. در نتیجه، درک آنچه که هر ویژگی از بردار سند چه محتوای واقعی از سند را نشان می دهد، دشوار است. بنابراین، اگر یک مدل کاربرد متن مانند طبقه‌بندی سند از این بردارهای سند تولید شده در بازنمایی Doc2Vec آموزش دیده باشد، در ارائه توضیحات شهودی برای منطق عملیاتی موجود در مدل موفق نیست. داشتن بازنمایی خوب از یک سند، هدف نهایی متن کاوی نیست. برای اینکه این روش‌های بازنمایی تأثیرات و پیامدهای معناداری در محیط واقعی کسب و کار داشته باشند، ضروری است که بازنمایی اسناد بتواند درک روشنی از متن را ارائه دهد.

۲-۳-۵ روش‌های مبتنی بر مفهوم

در این پژوهش، اسناد بر اساس مفاهیم ارائه می‌شوند. در ادامه، برخی از مطالعات پیشرفته در این زمینه مرور خواهند شد. کیم و همکاران [18] روش دیگری برای ارائه اسناد به نام Bag-of-Concepts (BoC) پیشنهاد کردند. BoC با خوشه‌بندی کلمات بردار تولید شده توسط مدل Word2Vec مفاهیم اسناد را استخراج می‌کند. سپس بردار سند را با استفاده از فرکانس مفاهیم موجود در اسناد ایجاد می‌کند. روش BoC یک روش بدون نظارت است که بر بازنمایی اسناد متمرکز است. این روش یک راه حل خاص برای کاربردهایی مانند خوشه‌بندی و طبقه‌بندی اسناد ارائه نمی‌دهد.

لی و همکاران [48] طرح جدیدی را برای بازنمایی اسناد مبتنی بر مفهوم معرفی کردند، که بطور خودکار دانش مفهومی مناسبی را از یک پایگاه دانش خارجی بدست می‌آورد و سپس کلمات و عبارات اسناد را با رویکرد احتمالاتی به مفاهیم تبدیل می‌کند. آن‌ها با استفاده از یک پایگاه دانش، درک و تفسیرپذیری بهتری را برای بردارهای سند فراهم می‌کنند تا برای انسان قابل فهم باشد. آن‌ها همچنین با هدف بهبود الگوریتم BoC، مفاهیم با معانی مشابه را خوشه‌بندی کردند تا بتوانند ابهام مفاهیم را کاهش دهند. روش آن‌ها در زمینه طبقه‌بندی اسناد ارزیابی می‌شود.

لو و همکاران [49] طرح مبتنی بر مفهوم جدیدی را برای خوشه‌بندی و مصورسازی اسناد پزشکی ارائه‌دادند. در این پژوهش تعبیه مفهوم[۵۷] از طریق شبکه‌های عصبی برای به دست آوردن ارتباط بین مفاهیم فرا گرفته می‌شود. بردارهای سند که براساس مفاهیم تولید شده‌اند، می‌توانند برای خوشه‌بندی مجموعه‌های اسناد استفاده شوند. در [50]، تجزیه مفهوم برای خوشه‌بندی متن‌های کوتاه استفاده می‌شود. این مطالعه یک روش تجزیه را ارائه می‌دهد که با شناسایی جوامع کلمات معنایی[۵۸] در شبکه همزمانی کلمات وزنی[۵۹] استخراج شده از مجموعه متن کوتاه، بردارهای مفهومی تولید می‌کند. این بردارهای تولید شده در فرآیند خوشه‌بندی متون کوتاه مورد استفاده قرار می‌گیرند.

لازم به ذکر است که مطابق با دانش ما، هیچ یک از روش‌های نمایش متن از رویکرد نیمه‌نظارتی استفاده نمی‌کنند. از این رو ایجاد مدلی جهت استخراج نیمه‌نظارتی مفاهیم با استفاده از تعداد محدودی داده‌ی برچسب‌دار می‌تواند مفید واقع شود.

۲-۴ خوشه‌بندی نیمه‌نظارتی

خوشه‌بندی نیمه‌نظارتی به عنوان جایگزینی برای روش‌های متداول بدون نظارت در جایی که دانش جزئی از برچسب داده‌ها وجود دارد، به کار گرفته می شود تا عملکرد فرآیند خوشه‌بندی بهبود پیدا کند. یک بررسی جامع از برخی الگوریتم‌های خوشه‌بندی نیمه‌نظارت‌شده توسط باسو و همکاران در پژوهشی ارائه شده است [51].

دارا و همکاران [22] از SOM (نقشه خودسازمانده[۶۰]) برای خوشه‌بندی نیمه نظارتی استفاده می‌کنند. ابتدا داده‌های بدون برچسب، برچسب‌دهی می‌شوند و سپس از همه داده‌ها (داده‌های برچسب‌دار اصلی و برچسب‌دهی‌شده) برای آموزش مدل طبقه‌بندی‌کننده استفاده می‌شود. در روش پیشنهادی آن‌ها

[۵۷] Concept Embedding

[۵۸] Semantic Word Communities

[۵۹] Weighted Word Co-occurrence Network

[۶۰] Self-Organizing Map

مشخص نیست که چه میزان به برچسب‌دهی داده‌های بدون برچسب نیاز است. باسو و همکاران روش MCP KMEANS [52] را پیشنهاد دادند، که ترکیبی از دو الگوریتم خوشه‌بندی مبتنی بر جستجو و شباهت است. اگرچه ترکیبی از این دو رویکرد ممکن است عملکرد آن‌ها را بهبود بخشد، اما تابع هدف آن‌ها ممکن است به مینیموم محلی برسد. ترکیبی از الگوریتم های Naive Bayes و EM (NBEM) نیز برای خوشه‌بندی نیمه نظارتی اسناد استفاده شده است [21]. این مدل در روندی تکراری داده‌های بدون برچسب را برچسب‌گذاری می‌کند و از این داده‌های تازه برچسب‌خورده برای آموزش مجدد مدل استفاده می‌کند. این حلقه برچسب‌گذاری مجدد و آموزش مجدد تا زمان همگرایی مدل ادامه دارد. ژانگ و همکاران [23] روشی را با نام TESC برای طبقه‌بندی متن با استفاده از خوشه‌بندی سلسه‌مراتبی نیمه نظارتی توسعه دادند. برخلاف روش‌های مرسوم، در این روش از داده‌های برچسب‌خورده و بدون برچسب همزمان استفاده می‌شود. روش TESC فرض می‌کند که مجموعه داده‌ها از اجزای مختلفی تشکیل شده است و از یک فرآیند خوشه‌بندی برای بدست آوردن مولفه‌ها متن استفاده می‌کند. در ابتدا تمام داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند، سپس الگوریتم خوشه‌بندی سلسه مراتبی نزدیک‌ترین خوشه‌ها را طی یک حلقه بر اساس برچسب آن‌ها ادغام می‌کند. حلقه تا زمانی ادامه می‌یابد که دیگر خوشه‌ی بدون پیمایشی باقی نمانده باشد. اجزای تشکیل‌دهنده متن با این روش استخراج می‌شوند و در فرآیند طبقه‌بندی متن و پیش‌بینی برچسب برای داده های بدون برچسب مورد استفاده قرار می‌گیرند. در این روش اکثر داده‌های آموزش دارای برچسب می‌باشند و تعداد کمی داده‌ی بدون برچسب برای تنظیم خوشه‌ها استفاده می‌شود.

برای بهبود عملکرد خوشه‌بندی اسناد با داده‌های نظارتی، یک روش نیمه نظارت شده برای فاکتورگیری[۶۱] مفاهیم توسط لو و همکاران ارائه شده است [53]. اطلاعات نظارتی به صورت دو مجموعه محدودیت

[۶۱] Factorization

جفتی ارائه می شود. آن‌ها جفت محدودیت‌های[۶۲] مجازات و پاداش را در فاکتورگیری مفهوم لحاظ می‌کنند، که می تواند اطمینان حاصل کند که نقاط داده مربوط به یک خوشه در فضای اصلی هنوز در همان خوشه در فضای تبدیل شده هستند. این پژوهش یک مکانیزم مجازات پویا را ارائه می‌دهد که برای واریانس درون خوشه‌ای به خوبی عمل می‌کند، به عنوان مثال، اجازه می‌دهد تا نقاط داده غیرمشابه با همان برچسب خوشه، دورتر از نقاط مشابه ترسیم شود. به این ترتیب می‌توان ساختارهای خوشه‌بندی معقول‌تری را در فضای نمایش متن یاد بگیرد. در مطالعه دیگری، از خوشه‌بندی متن با استفاده از تولید خودکار محدودیت‌ها برای طبقه‌بندی اسناد استفاده شده است [54]. ساختار ذاتی داده‌ها با استفاده از یک الگوریتم خوشه‌بندی جزئی مورد مطالعه قرار می‌گیرد. الگوریتم خوشه‌بندی این امکان را می‌دهد که مجموعه‌ای از محدودیت‌های must-link/cannot-link ایجاد شود، که می‌تواند در مرحله خوشه-بندی نیمه نظارت شده استفاده شود. سپس، این محدودیت‌ها به عنوان عامل نیمه نظارتی در الگوریتم خوشه‌بندی سلسله‌مراتبی استفاده می‌شوند.

لی و همکاران [55] یک روش خوشه‌بندی نیمه‌نظارتی توزیع شده را پیشنهاد کردند. الگوریتم خوشه-بندی همان روش TESC است، با این تفاوت که خوشه‌بندی توزیع شده و جمع‌آوری نتایج از زیر خوشه‌ها انجام می‌شود. در پژوهش مسعود و همکاران [56] روش جدیدی برای انتخاب جفت محدودیت‌ها از داده‌های بدون برچسب برای خوشه‌بندی نیمه نظارتی اسناد ارائه شده است. مجموعه داده توسط الگوریتم I-nice و بدون هیچ گونه اطلاعات برچسب‌گذاری به خوشه‌ها تقسیم می‌شود. از هر خوشه اولیه، یک گروه داده متراکم انتخاب می‌شود و در این گروه‌های متراکم، داده‌های با اطلاعات بیشتر با استفاده از روش تخمین چگالی محلی شناسایی می‌شوند. داده‌های شناسایی شده برای تشکیل مجموعه-ای از جفت‌های محدودیت تحت عنوان ناظر در خوشه‌بندی نیمه نظارتی استفاده می‌شوند.

[۶۲] Pairwise Constraints

گان و همکاران [57] بیان می‌کنند که دانش قبلی در صورت جمع‌آوری نادرست می تواند عملکرد خوشه‌بندی نیمه نظارتی را کاهش دهد. ایده اصلی روش آن‌ها این است که وقتی برچسب یک نمونه برچسب‌دار ریسکی (نامطمئن) است، پیش‌بینی‌های نمونه برچسب‌دار و نزدیکترین نمونه‌های بدون برچسب همگن باید مشابه باشند. این کار با خوشه‌بندی بدون نظارت و سپس ایجاد یک نمودار محلی برای مدل سازی روابط بین نمونه‌های برچسب‌زده‌شده و نزدیکترین نمونه‌های بدون برچسب انجام می‌شود. به منظور مقایسه و رسیدن به درک جامعی از روش‌های ذکر شده، برخی از روش‌های خوشه-بندی نیمه‌نظارتی متون به اختصار در جدول (۲-۱) درج شده‌اند.

جدول ۲-۱. خلاصه ای از برخی روش های نیمه نظارتی موجود در خوشه بندی اسناد

نویسندگان	سال	شرح مختصر	چالش‌ها
دارا و همکاران [22]	۲۰۰۲	استفاده از نقشه خودسازمانده برای خوشه‌بندی نیمه نظارتی – برچسب-گذاری داده‌های بدون برچسب و آموزش مدل با تمامی داده‌ها.	مشخص نیست که چه میزان به برچسب‌دهی داده‌های بدون برچسب نیاز است.
باسو و همکاران [52]	۲۰۰۳	ترکیبی از دو الگوریتم خوشه‌بندی نیمه‌نظارتی مبتنی بر جستجو و مبتنی بر شباهت.	رسیدن تابع هدف به مینیموم محلی - عدم استفاده همزمان از هر دو نوع داده‌ی برچسب‌دار و بدون برچسب
ژانگ و همکاران [21]	۲۰۱۵	این مدل در روندی تکراری داده‌های بدون برچسب را برچسب‌گذاری می‌کند و از این داده‌های تازه برچسب‌خورده برای آموزش مجدد مدل استفاده می-کند.	زمان اجرای بالا و نیاز به ادامه آموزش تا همگرایی مدل.
ژانگ و همکاران [23]	۲۰۱۵	خوشه‌بندی سلسله مراتبی نیمه نظارتی – استفاده همزمان از هر دو نوع داده‌ی برچسب‌دار و بدون برچسب.	اکثر داده‌های آموزش دارای برچسب هستند و فقط از تعداد محدودی داده‌ی بدون برچسب استفاده می‌کند.
لو و همکاران [53]	۲۰۱۶	اطلاعات نظارتی به صورت دو مجموعه محدودیت جفتی مجازات و پاداش ارائه می شود.	دقت پایین و نامشخص بودن عملکرد در واریانس بین خوشه‌ای.

لی و همکاران [55]	۲۰۱۹	خوشه‌بندی نیمه‌نظارتی توزیع شده - جمع‌آوری نتایج از زیر خوشه‌ها.	پیچیدگی محاسباتی و زمانی بالا- دقت و کیفیت پایین‌تر نسبت به خوشه‌بندی روش‌های مشابه.
مسعود و همکاران [56]	۲۰۱۹	انتخاب جفت محدودیت‌ها از داده‌های بدون برچسب برای خوشه‌بندی نیمه نظارتی اسناد	ایجاد ناظر با اتکا بر داده‌های بدون برچسب - عدم قطعیت جفت محدودیت‌های انتخابی.

در ستون چالش‌های جدول (۲-۱) به محدودیت‌ها و معایب اکثر روش‌های خوشه‌بندی نیمه‌نظارتی پرداخته شده است، اما با توجه به دانش ما ذکر این نکته ضروری است که هیچ یک از روش‌های موجود خوشه‌بندی نیمه‌نظارتی اسناد، از بازنمایی مفهومی و مدلسازی مفهوم برای نمایش اسناد بهره نمی‌برند و در نتیجه این امر را می‌توان به‌عنوان یک چالش اساسی برای تمامی روش‌های ذکر شده بیان کرد.

۲-۵ جمع‌بندی

در این فصل روش‌های بازنمایی اسناد و همینطور رویکردهای سابق خوشه‌بندی نیمه نظارتی اسناد به همراه مزایا و معیاب ذکر شدند. با مطالعه روش‌ها و تحقیقات مرور شده در این فصل، نیاز به روشی جدید برای خوشه‌بندی نیمه‌نظارتی مبتنی بر بازنمایی مفهومی اسناد احساس می‌شود. روش پیشنهادی این تحقیق که در فصل بعدی بسط داده می‌شود، به منظور غلبه بر معایب روش‌های پیشین خوشه‌بندی متون ارائه می‌شود. اگرچه روش پیشنهادی ما همچنین از داده های دارای برچسب و بدون برچسب به صورت همزمان استفاده می‌کند، اما با رویکردهای قبلی مانند TESC متفاوت است. در روش TESC، کمتر از دو درصد داده‌ها بدون برچسب هستند و اکثر داده‌ها برچسب‌دار می‌باشند. در روش پیشنهادی و ارزیابی‌های ما، بخش بزرگی از داده‌ها بدون‌برچسب هستند و ممکن است از تعداد محدودی از داده های دارای برچسب استفاده شود. این تفاوت به طور قابل توجهی هزینه برچسب‌گذاری داده‌ها را در

کاربردهای واقعی کاهش می‌دهد. بعلاوه، بسیاری از روش‌های خوشه‌بندی نیمه نظارتی ذکر شده از مسئله بازنمایی اسناد غافل می‌شوند، که می‌تواند تا حد زیادی بر نتیجه خوشه بندی تأثیر بگذارد. در این تحقیق، به غیر از استفاده از بازنمایی مفهومی اسناد در خوشه‌بندی نیمه‌نظارتی، ما رویکرد مفاهیم برچسب‌دار از طریق استخراج مفاهیم به صورت نیمه نظارتی را پیشنهاد می‌دهیم.

فصل۳ : روش پیشنهادی

۳-۱ مقدمه

این پژوهش یک روش نوآورانه خوشه‌بندی نیمه‌نظارتی برای اسناد مبتنی بر نمایش مفهومی آن‌ها ارائه می‌دهد. به دلیل آنکه در روش پیشنهادی مفاهیم متن شناسایی و استخراج می‌شوند، به راحتی طبقه‌ها و زیرطبقه‌های جزئی متن قابل شناسایی و مورد دسترسی قرار می‌گیرند. این روش از یک خوشه‌بند سلسه‌مراتبی به منظور یافتن گروه‌های سند استفاده می‌کند که کلیات خوشه‌ها را با کمک داده‌های بدون برچسب و مراکز دقیق خوشه‌ها را با داده‌های برچسب دار تنظیم می‌کند. همچنین یکی از بزرگترین مزایای این روش بهره‌گیری از تعداد اندکی داده‌ی برچسب‌دار در مقایسه با تعداد داده‌های بدون برچسب می‌باشد که کاربرد فراوانی را در کار با داده‌های دنیای واقعی فراهم می‌آورد. برای درک بهتر و شهودی روش پیشنهادی، مسئله به صورت مدل ریاضی فرموله شده و توضیحات مبسوط ارائه می‌گردد.

۳-۱-۱ مدل ریاضی مسئله

فرض می‌شود که مجموعه اسناد ورودی $D = \{D^l, D^u\}$ به مجموعه‌ی اسناد برچسب‌دار (D^l) و بدون-برچسب (D^u) تقسیم می‌شود. یک سند برچسب‌دار d یک برچسب متناظر در مجموعه تمام برچسب‌ها (L) دارد ($label(d) \in L$). هر سند از مجموعه ای از کلمات تشکیل شده است. اجتماع تمام کلمات از تمام اسناد با نماد W نشان داده می شود. هدف ما رسیدن به یک مدل خوشه‌بندی از اسناد $C = \{C_1, \dots, C_m\}$ است، که در آن شرایط زیر حاکم باشد:

$$C_i \cap C_j = \emptyset \ (1 \le i \ne j \le m) \quad \text{و} \quad \cup_{1 \le i \le m} C_i = D$$

۳-۲ کلیات روش پیشنهادی

هدف از این پژوهش ارائه‌ی یک الگوریتم خوشه‌بندی نیمه‌نظارتی است، بگونه‌ای که بتوان با استفاده از تعداد کمی داده‌ی برچسب‌دار به خوشه‌های با کیفیتی برای کاربردهای تحلیل متن دست یافت. شکل (۳-۱) فرآیند دقیق روش پیشنهادی را نشان می‌دهد که در سه فاز اصلی توصیف شده است: پیش-پردازش، نمایش اسناد و خوشه بندی. در مرحله‌ی پیش‌پردازش متن از روش‌های متداول برای تبدیل متن به بردارهای عددی حامل اطلاعات با معنی استفاده می‌شود. نمایش اسناد بر اساس مفاهیم استخراج شده از متنِ مجموعه‌ی اسناد می‌باشد. رویکرد خوشه‌بندی پیشنهادی نیمه‌نظارت‌شده است و از اسناد برچسب‌دار جهت بهبود کیفیت خوشه‌بندی بهره می‌برد. براساس اینکه آیا می‌خواهیم از برچسب‌ها در فاز نمایش اسناد یا در فاز خوشه‌بندی اسناد استفاده کنیم، دو روش پیشنهاد می‌شود. دو روش پیشنهادی در این پژوهش در دو رنگ مختلف در شکل (۳-۱) ارائه شده است: SSConE [۶۳] با رنگ قرمز و SSClusE [۶۴] با رنگ آبی. مراحل مشترک این دو با رنگ سیاه مشخص شده است. در روش اول (SSConE)، اسناد برچسب‌دار را می‌توان در مدل‌سازی مفهوم[۶۵] استفاده کرد، که منجر به رویکرد استخراج نیمه نظارت شده مفهوم می‌شود. در روش دوم (SSClusE)، اسناد برچسب‌دار در فاز خوشه‌بندی اسناد استفاده می‌شوند و منجر به رویکرد استخراج نیمه‌نظارتی خوشه‌های اسناد می‌شوند. با این وجود، هر دو روش یک مدل خوشه‌بندی از اسناد را ارائه می‌دهند. جدول (۳-۱) مجموعه نمادهایی را که در توصیف الگوریتم و روش پیشنهادی استفاده شده است، ارائه می‌دهد. در ادامه این فصل، سه فاز اصلی از روش پیشنهادی یعنی پیش‌پردازش، نمایش و خوشه‌بندی اسناد به طور مفصل تشریح می‌شود.

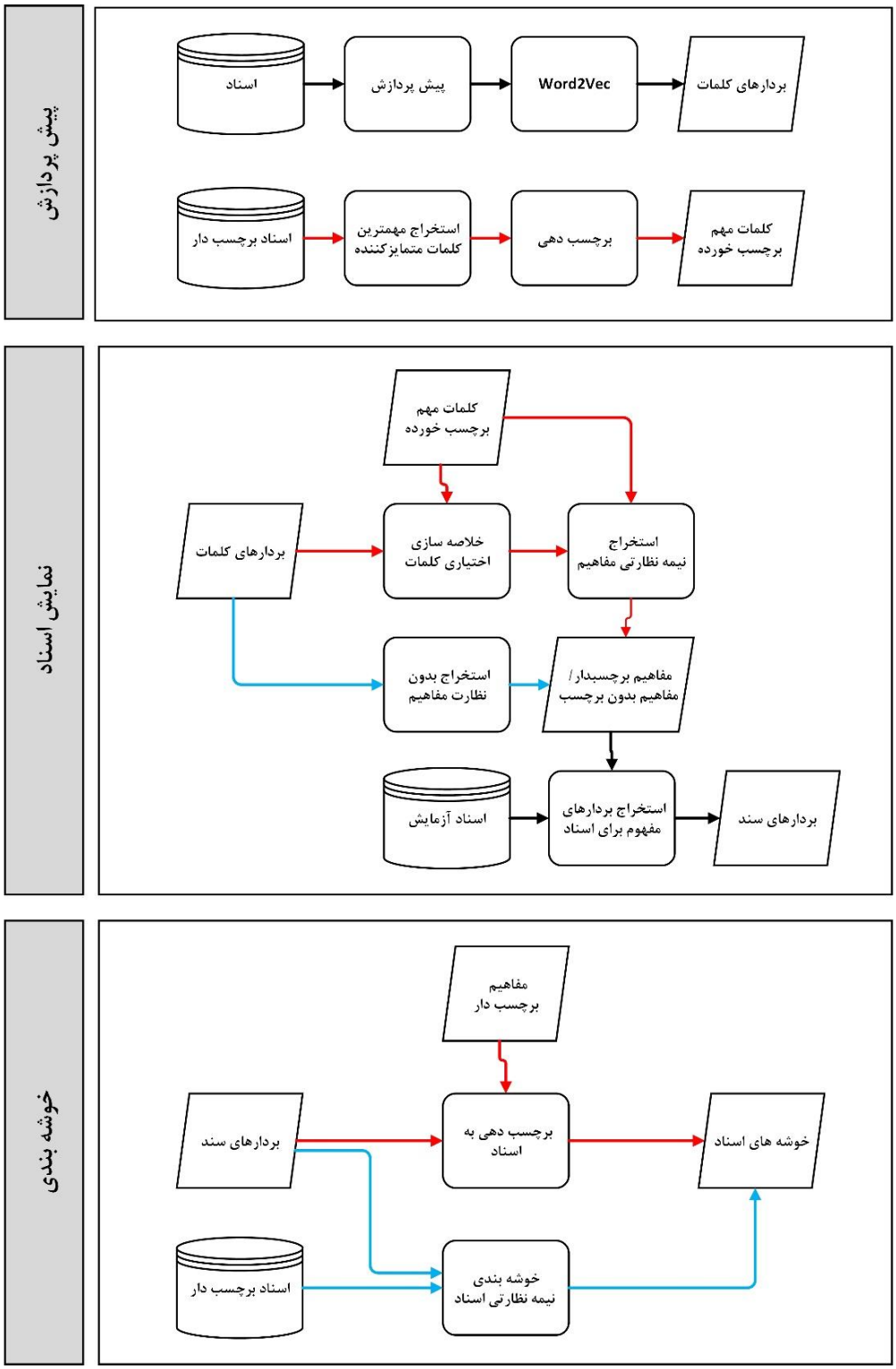
[۶۳] Semi-Supervised Concept Extraction

[۶۴] Semi-Supervised Cluster Extraction

[۶۵] Concept Modeling

جدول ۳-۱. توضیحات نماد های مورد استفاده در روش های پیشنهادی

نماد ها	توضیحات
$d\ ,D$	مجموعه‌ی اسناد، یک سند
$D^u\ ,D^l$	مجموعه‌ی اسناد برچسب‌دار، بدون‌برچسب
$\overrightarrow{d_i^j}\ ,\overrightarrow{d_i}$	بردار مفهوم سند d_i ، j امین جزء از بردار سند
$l\ ,L$	مجموعه‌ی برچسب‌ها، یک برچسب
$w\ ,W$	مجموعه‌ی لغات متون، یک لغت
W^l	یک مجموعه لغات برچسب‌دار
$T_i\ ,T$	مجموعه‌ی مفاهیم، یک مفهوم
$C_i\ ,C$	مجموعه‌ی خوشه‌ها، یک خوشه
n, m, z	تعداد مفاهیم، تعداد خوشه‌ها، تعداد اسناد
$i,j\ ,k$	استفاده شده در شمارش حلقه‌های الگوریتم



شکل ۳-۱. فرایند کلی روش پیشنهادی خوشه‌بندی نیمه‌نظارتی اسناد.

۳-۳ جزئیات فازهای اصلی روش پیشنهادی

۳-۳-۱ فاز پیش‌پردازش

در ابتدا، پیش‌پردازش‌های متداولی از جمله تبدیل حروف بزرگ به حروف کوچک، حذف کامل اعداد و نشانه‌های غیرمجاز، حذف علائم نگارشی و لهجه، حذف فضاهای خالی و حذف کلمات بی‌اثر بر روی متون موجود در مجموعه‌ی اسناد انجام می‌شود. پس از این مراحل، عملیات جداسازی کلمات (توکن‌ها) متن انجام می‌گیرد. نیاز است این کلمات به بردارهای عددی قابل فهم برای ماشین تبدیل شوند. به همین منظور با توجه به مزایای ذکر شده در فصل دوم، از یک مدل تعبیه کلمات Word2Vec [58] برای این امر استفاده می‌کنیم. این مدل تعبیه کلمات با توجه به جایگاه کلمات در متون، یک شبکه‌ی عصبی را آموزش می‌دهد، که وزن‌های این شبکه‌ی عصبی بردارهای کلمات مورد نظر می‌باشند. کلمات بدست آمده از اسناد (توکن‌ها) به عنوان ورودی به این مدل شبکه‌ی عصبی وارد می‌شوند. در نتیجه هر لغت در مجموعه‌ی لغات متون (W)، با یک بردار در فضای تعبیه شده نمایش داده می‌شود. همچنین نکته‌ی مهم این است که تعبیه کلمات، ارتباط معنایی بین کلمات را به خوبی حفظ می‌کند.

با توجه به شکل (۳-۱) اگر اسناد برچسب‌دار، در فاز نمایش اسناد برای مدلسازی مفهومی به صورت نیمه‌نظارتی (SSConE) مورد استفاده قرار بگیرند، یک مجموعه‌ی لغات برچسب‌دار (W^l) مورد نیاز است. به همین منظور، مجموعه‌ی محدودی از کلماتی که قدرت متمایزکنندگی بیشتری نسبت به سایر کلمات دارند، از اسناد برچسب‌دار (D^l) استخراج می‌شوند (W^l) و به صورتی که در ادامه تشریح می‌شود، برچسب‌گذاری خواهند شد. برای برچسب‌گذاری این کلمات، از مدل TF-IDF رابطه (۳-۱) که یک مدل وزن‌دهی به کلمات متن است، استفاده می‌شود. این مدل به هر کلمه در مجموعه‌ی اسنادِ برچسب‌دار،

یک وزن مطابق با رابطه (۳-۱) نسبت می‌دهد. در این رابطه $tf_{w,d}$ تعداد رخداد کلمه w در سند d را نشان می‌دهد و df_w بیانگر تعداد اسناد برچسب‌داری است که شامل کلمه‌ی w می‌باشند.

$$TF - IDF(w,d) = tf_{w,d} \times \log(\frac{|D^l|}{df_w}) \qquad (۱-۳)$$

یک مجموعه‌ی محدود کلمات با بیشترین وزن از بین تمام کلمات اسناد برچسب‌دار انتخاب می‌گردند و مطابق با آنکه در کدام سند رخ داده‌اند، با برچسب آن سند برچسب‌دهی می‌شوند. برای سادگی توضیح الگوریتم، مجموعه‌ی لغات بدون برچسب با W^u نمایش داده می‌شوند که در این صورت $W^l \cup W^u = W$. کلمات برچسب خورده بعداً در فاز نمایش برای استخراج مفاهیم به صورت نیمه نظارت شده استفاده می شوند.

۳-۳-۲ فاز نمایش اسناد

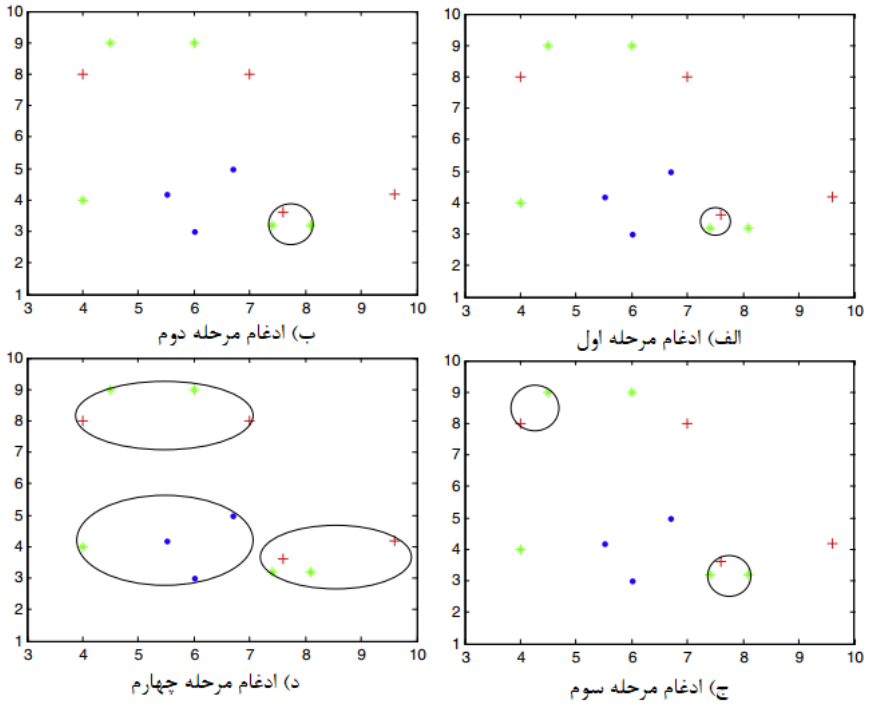
نمایش اسناد بر اساس مفاهیم استخراج شده از مجموعه‌ی اسناد است. در فاز نمایش، ما در ابتدا یک مجموعه‌ی مفاهیم $T = \{T_1, \dots, T_z\}$ را از مجموعه‌ی کلمات W استخراج می‌کنیم. بدین ترتیب که هر مفهوم از مجموعه‌ی منحصربه‌فردی از کلمات تشکیل شده است، یعنی $T_i \cap T_j = \emptyset$ $(1 \le i \ne j \le z)$ و $\cup_{1 \le i \le z} T_i = W$. روش کلی استخراج مفهوم شامل اجرای یک الگوریتم خوشه‌بندی در مجموعه کلمات W برای تقسیم آن به چندین خوشه است که هرکدام نمایانگر یک مفهوم هستند. همانطور که در ادامه توضیح داده شد، مفاهیم ممکن است در یک رویکرد نیمه‌نظارتی یا بدون نظارت استخراج شوند. در رویکرد اول، به مجموعه‌ی کلمات برچسب‌خورده W^l نیاز است. پس از ایجاد مفاهیم، هر سند d با یک بردار مفاهیم $\vec{d}$ نمایش داده می‌شود.

۳-۳-۲-۱ استخراج نیمه‌نظارتی مفاهیم

استخراج نیمه‌نظارتی مفهوم یک الگوریتم خوشه بندی نیمه نظارت شده را در مجموعه کلمات W^l ∪ W^u اجرا می کند. هر خوشه حاصله که به عنوان یک مفهوم در نظر گرفته می‌شود، دارای برچسب بر اساس کلمات اساسی آن خوشه می‌باشد. الگوریتم خوشه‌بندی نیمه‌نظارتی (الگوریتم شماره ۱) از الگوریتم TESC الهام گرفته شده است [23]، و همین روش در ادامه در خوشه بندی نیمه‌نظارتی اسناد در **بخش ۳-۳-۳-۲** استفاده می‌شود. در این رویکرد خوشه‌بندی، داده‌های بدون برچسب برای یافتن ساختار کلی و داده‌های دارای برچسب برای تنظیم مراکز خوشه‌ها مورد استفاده قرار می‌گیرند.

الگوریتم (۱) روال خوشه‌بندی نیمه‌نظارتی را نشان می‌دهد که بر روی مجموعه‌ی داده X با هدف رسیدن به خوشه‌های نهایی P اجرا شده است. برای استفاده از یک علامت ثابت، ما یک برچسب ساختگی U را به داده‌های بدون برچسب اختصاص می‌دهیم تا تمامی داده‌ها دارای برچسب باشند. در ابتدا، هر یک از نقاط داده ورودی x_i ، به عنوان یک خوشه‌ی کاندید S_i با مقدار برچسب یکسان با x_i در نظر گفته می‌شود (خط ۲ تا ۷). حلقه‌ی اصلی الگوریتم حداقل به اندازه‌ی $|X|$ مرتبه اجرا می‌شود تا در مجموعه‌ی خوشه‌های کاندید S حداکثر یک عضو باقی بماند. از بین تمام جفت خوشه‌ی کاندید، دو کاندید نزدیک‌تر S_i و S_j (دو خوشه با کمترین فاصله‌ی کسینوسی بین مراکزشان) انتخاب می‌شوند (خط ۱۱). سپس براساس برچسب کاندیدها، یا یک خوشه جدید ایجاد می‌شود (خط ۱۶ تا۲۰) یا خوشه‌ها حفظ می‌شوند (خط ۱۲ تا ۱۵). اگر هیچ‌کدام از خوشه‌های کاندید برچسب U نداشته باشند و $label(S_i) \neq$ $label(S_j)$ دو خوشه حفظ می‌شوند (عدم ادغام) و به خوشه‌بندی نهایی P منتقل می‌گردند. در حالت دیگر، دو خوشه‌ی کاندید به یک خوشه جدید ادغام می‌شوند S_k، که جایگزین خوشه‌های انتخاب شده در مجموعه‌ی کاندید می‌شود. اگر دو خوشه‌ی کاندید انتخاب شده برچسب یکسانی داشته باشند، برچسب خوشه جدید (S_k) با برچسب آن‌ها یکسان است، یا اگر یکی از خوشه‌های کاندید برچسب U

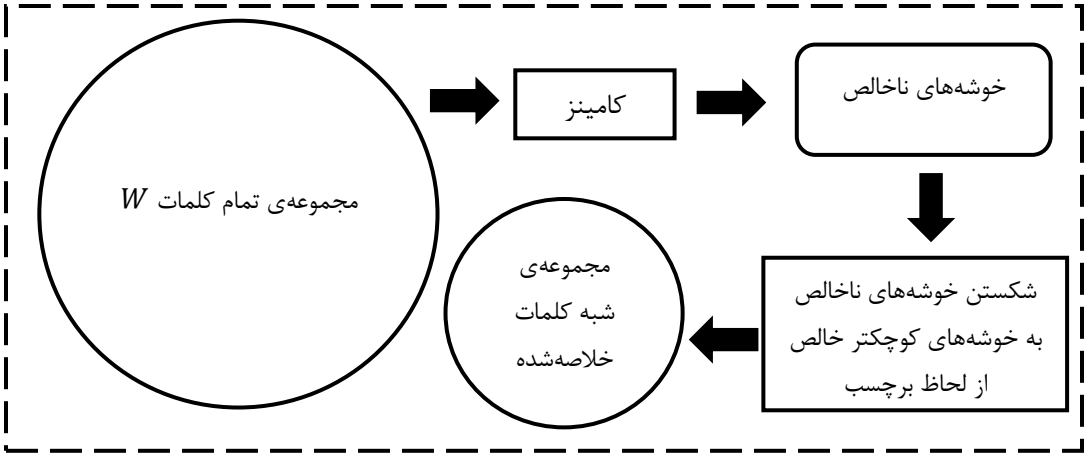
داشته باشد، برچسب خوشه جدید برابر با برچسب خوشه‌ی کاندیدی خواهد بود که برچسبش U نباشد. الگوریتم در هر مرحله با انتخاب جفت خوشه‌های کاندید تکرار می‌شود. خوشه‌های نهایی با کمتر از سه عضو، به عنوان نویز حذف می‌گردند (خط ۲۴). خوشه‌های نهایی برچسب خود را می‌گیرند و مفاهیم برچسب‌دار را نشان می‌دهند. تعداد خوشه ها (m) توسط الگوریتم خوشه‌بندی تعیین می‌شود و به شکل و تعداد داده‌های ورودی بستگی دارد. از این‌رو در این رویکرد تعداد مفاهیم (z) و در نتیجه طول بردارهای سند ($|\vec{d}|$) از پیش تعیین نشده و متغیر است. برای روشن شدن نحوه‌ی کار الگوریتم خوشه‌بندی نیمه‌نظارتی شکل (۳-۲) مراحل اجرا الگوریتم را بر روی ۱۲ نقطه داده نشان می‌دهد. داده‌های آبی از کلاس اول، داده‌های قرمز از کلاس دوم و داده‌های سبز بدون برچسب می‌باشند.



شکل ۳-۲. مثال از الگوریتم شماره یک و مراحل خوشه بندی داده‌ها

در مرحله اول دو نزدیکترین داده شناسایی و در یک خوشه ادغام می‌شوند شکل (۳-۲ الف). در مرحله بعد همان خوشه‌ی ایجاد شده کمترین فاصله را با یک داده بدون برچسب دیگر دارد، پس این سه داده در یک خوشه جدید قرار می‌گیرند شکل (۳-۲ ب). در مرحله سوم با توجه به فواصل بین خوشه‌ها، یک خوشه‌ی جدید دیگر ایجاد می‌شود شکل (۳-۲ ج). در مرحله‌ی پایانی، بعد از تکرارهای زیاد همانطور که در اشکال الف تا ج نشان داده شد، ۱۲ نقطه داده در ۳ خوشه مطابق با شکل (۳-۲ د) جای می‌گیرند.

از آنجا که مجموعه W بزرگ است و پیچیدگی الگوریتم خوشه بندی نیمه‌نظارتی $O(|W|^3)$ است، برای کاهش پیچیدگی زمانی، کلمات می‌توانند به صورت اختیاری قبل از استخراج مفاهیم خلاصه شوند. برای این منظور، خوشه بندی کامینز کروی[۶۶] [59] برای خوشه‌بندی W استفاده می‌شود (بخش خلاصه-سازی اختیاری کلمات در شکل (۳-۱)). پس از خوشه‌بندی، خوشه‌های حاصل بیشتر بررسی می‌شوند تا هر خوشه‌ای که شامل چندین برچسب باشد، به خوشه‌های خالص کوچکتر تک برچسبی شکسته شود. مراکز خوشه مطابق با اعضای آن خوشه محاسبه و برچسب‌دهی می‌شوند و این خوشه‌ها به عنوان شبه کلمه لیبل‌خورده[۶۷] جدید برای ورودی خوشه‌بندی نیمه‌نظارتی بکار گرفته می‌شوند شکل (۳-۳).



شکل ۳-۳. روال کلی بخش خلاصه‌سازی کلمات برای ورود به خوشه‌بند نیمه‌نظارتی

[۶۶] Spherical K-means

[۶۷] Pseudo Labeled Words

الگوریتم ۱. فرآیند خوشه بندی نیمه نظارتی

Input:

X : a set of labeled and unlabeled data

Output:

P : a set of labeled data clusters

SSC_Procedure:

// Initialization
Construct a cluster candidate set $S = \{\}$
For each $x_i \in X$
    Establish a cluster candidate S_i using each x_i and label S_i with l_{x_i} ;
    Set centroid of S_i as x_i ;
    $S \leftarrow S_i$;
End for
// Clustering
$n \leftarrow$ the number of cluster in S ;
While $n > 1$
    Find a cluster candidate pair (S_i, S_j) in S, which has the smallest distance between centroids among all the possible cluster candidate pairs in S ;
    If S_i and S_j are all labeled cluster candidates and have different cluster labels
        $P \leftarrow S_i$ and S_j ;
        Remove S_i and S_j from S ;
        $n \leftarrow n - 2$;
    Else
        Merge S_i and S_j into a new cluster candidate S_k ;
        Remove S_i and S_j from S ;
        Add S_k to S ;
        $n \leftarrow n - 1$;
    End if
End while
// Output
Remove the clusters whose size is smaller than 3 from P ;
Output P ;

۳-۳-۲-۲ استخراج بدون نظارت مفاهیم

اگر از داده‌های برچسب‌دار در استخراج مفاهیم استفاده نشود، الگوریتم خوشه بندی بدون نظارت کامینز کروی با استفاده از فاصله کسینوسی برای خوشه بندی مجموعه کلمات W استفاده می‌شود. خوشه‌های حاصل به عنوان مفاهیمی بدون برچسبِ استخراج شده از کلمات در نظر گرفته می‌شوند. کامینز کروی یک روش خوشه‌بندی بدون نظارت است و برای تعداد خوشه ها در فرآیند خوشه‌بندی به مقدار از پیش تعیین شده‌ای نیاز دارد. به همین دلیل، تعداد مفاهیم استخراج شده توسط این رویکرد، شماری ثابت و از پیش‌تعیین‌شده دارد.

۳-۳-۲-۳ ساخت بردار اسناد

خوشه‌های ایجاد شده توسط خوشه‌بندی نیمه‌نظارتی یا خوشه‌بندی بدون نظارت به عنوان مفاهیم اسناد محاسبه می‌شوند و بردارهای سند از این مفاهیم ساخته می‌شوند. با توجه به کارایی خوشه‌بندی و فضای معنایی آموزش دیده توسط Word2Vec، کلمات با معنای مشابه در یک خوشه گروه‌بندی می‌شوند. در نتیجه، هر کلمه در متون به عنوان عضوی از یک مفهوم در نظر گرفته خواهد شد. از آنجا که هر کلمه ممکن است در متون زیادی تکرار شود، یک تمیزدهنده مناسب برای برنامه های یادگیری ماشین به‌شمار نمی‌رود [60]، بنابراین رابطه فرکانس سند-معکوس فرکانس مفهوم (CF-IDF) رابطه (۳-۲) بر روی بردارهای کلمهِ تولید شده اعمال می‌شود تا اثرات سو کلمات مشترک بین مفاهیم را از بین ببرد.

$$CF - IDF(c_i, d_j, D) = CF(c_i, d_j) \times \log \frac{|D|}{|d \in D \; ; \; c_i \in D|} \qquad (۳-۲)$$
$$where \; (c_i, d_j, D) = (Concept_i, Document_j, Corpus)$$

طول بردار سند ممکن است براساس رویکرد استخراج مفاهیم متفاوت باشد. در استخراج نیمه‌نظارتی مفهوم، تعداد مفاهیم به طور خودکار توسط الگوریتم خوشه‌بندی تعیین می‌شود و به ویژگی‌های داده و ساختار داده‌های دارای برچسب بستگی دارد. در استخراج بدون نظارت مفهوم، تعداد مفاهیم دلخواه است و با تنظیم تعداد خوشه در کامینز کروی، توسط کاربر قابل تعریف است. اثر تغییر تعداد مفاهیم در فصل نتایج بررسی شده است. تغییر تعداد مفاهیم در رویکرد دوم، طول بردارهای سند را تغییر می‌دهد. بنابراین، فضا و پیچیدگی های محاسباتی مرحله خوشه بندی تحت کنترل کاربر می‌باشد.

۳-۳-۳ فاز خوشه‌بندی

تا این فاز بردارهای سند استخراج شده‌اند و باید خوشه بندی سند انجام شود. انتظار می‌رود که دو سند حاوی مفاهیم مشابه دارای بردارهای سند مشابه باشند. با توجه به اینکه آیا مفاهیم به صورت نیمه‌نظارتی یا بدون نظارت استخراج می شوند، روند خوشه بندی اسناد پیگیری می‌شود. در حالت اول، خوشه بندی اسناد با استفاده از مفاهیم برچسب خورده انجام می‌شود، در حالی که در مورد دوم، خوشه‌بندی اسناد به صورت نیمه نظارت شده مورد نیاز است. در زیر دو روش به نوبت تشریح شده‌اند.

۳-۳-۳-۱ خوشه بندی اسناد مبتنی بر مفاهیم برچسب‌دار

هر ورودی یا خانه در بردار سند $\vec{d}$ با یک مفهوم استخراج شده مرتبط است. اگر مفاهیم توسط الگوریتم نیمه‌نظارتی از مجموعه کلمات استخراج شوند، هر مفهوم T_j دارای برچسب مربوطه است. با این وجود با شناسایی موثرترین مفهوم هر سند می‌توان فرآیند خوشه‌بندی اسناد را انجام داد. برچسب این مفهوم به عنوان برچسب خوشه‌ی سند $label(d)$ درنظر گرفته می‌شود. برای تعیین موثرترین مفهوم در بردار سند، راهکارهای متفاوتی قابل استفاده می‌باشد. در این پژوهش ما سه روش مختلف را ارائه و ارزیابی

می‌کنیم که در ادامه و در **بخش ۴-۳** قابل درسترسی است. دقیق‌ترین روش از بین سه روش ارائه شده، در رابطه (۳-۳) فرموله شده است. در این روش برچسبی که بیشترین وزن کل را در بردار CF-IDF از سند d داشته باشد به عنوان برچسب سند اختصاص داده شده است.

$$label(d) = \underset{l}{\operatorname{argmax}} \sum_{label(T_j)=l} \overrightarrow{d^J} \qquad (۳-۳)$$

۳-۳-۳-۲ خوشه‌بندی مبتنی بر مفاهیم بدون برچسب

اگر روند استخراج مفاهیم از مجموعه کلمات بدون نظارت انجام شود، مجموعه مفاهیم برچسب‌هایی با خود ندارند. بنابراین، بردارهای سند از طریق الگوریتم خوشه‌بندی نیمه‌نظارتی که بالاتر در الگوریتم ۱ معرفی شده بود، با کمک اسناد دارای برچسب خوشه‌بندی می‌شوند. ورودی الگوریتم (X) با بردارهای اسناد $\bigcup_{d\in D} \vec{d}$ مقداردهی می‌شود و خروجی (P) مجموعه‌ی خوشه‌های اسناد خواهد بود. همانند قبل، به اسناد بدون برچسب یک برچسب ساختگی U اختصاص داده‌شده است. این الگوریتم خوشه‌هایی از اسناد را تولید می‌کند که هر یک دارای برچسب مناسبی می‌باشند.

۳-۳-۴ خوشه‌بندی داده‌های آزمایش

پس از آنکه خوشه‌بندی بر روی داده‌های آموزش صورت گرفت، داده‌های آزمایش به نزدیک‌ترین خوشه‌ی ممکن انتساب داده می‌شوند. به آن صورت که فاصله‌ی کسینوسی بردار سند داده آزمون با تمامی مراکز خوشه‌ها محاسبه شده و در ماتریسی ذخیره می‌شود. ماتریس به صورت صعودی مرتب می‌شود و سپس خانه اول آن که نشان دهنده خوشه‌ای با کمترین فاصله به داده آزمایش است، به‌عنوان خوشه‌ی آن داده انتخاب می‌شود.

۳-۴ جمع‌بندی

هر دو روش خوشه‌بندی نیمه‌نظارتی اسناد که در این پژوهش معرفی شده‌اند، مزایای متعددی را شامل می‌شوند. روش های کلاسیک بازنمایی متن قادر به حفظ روابط معنایی غیرخطی بین کلمات نیستند. بردارهای سند تولید شده توسط برخی از پیشنهادهای اخیر مانند Doc2Vec بصری و قابل درک و تفسیر نیستند. رویکرد اتخاذ شده در این پایان‌نامه نه تنها روابط معنایی غیرخطی بین کلمات را حفظ می‌کند بلکه به دلیل استخراج مفاهیم در مجموعه اسناد، از تفسیر و شهودیت بالایی نیز برخوردار است. از آنجا که مفاهیم اجزای تشکیل‌دهنده متن را جدا می‌کنند، افزایش تعداد مفاهیم می تواند زیر شاخه‌ها و زیرکلاس‌های بیشتری را برای اسناد کشف کند. از آنجا که الگوریتم به تعداد کمی داده برچسب‌دار احتیاج دارد، سربار و هزینه برچسب‌گذاری اسناد پایین خواهد بود. بر این اساس، در کاربردهایی مانند تجزیه و تحلیل شبکه‌های اجتماعی که مقادیر زیادی از داده‌ها بدون برچسب هستند، این روش می تواند کارایی و نتیجه‌ی خوبی بجای بگذارد. درپایان، منطق فرآیند خوشه‌بندی روشن است و خوشه‌های حاصل از آن قابل تفسیر و توضیح هستند.

فصل۴ : نتایج و ارزیابی

۴-۱ مقدمه

در این فصل جزئیات پیاده‌سازی برای روش پیشنهادی بیان می‌شود. همانطور که در فصل گذشته ذکر شد، روش پیشنهادی شامل سه فاز اصلی پیش‌پردازش، نمایش و خوشه‌بندی اسناد می‌باشد. عملکرد هر فاز در خروجی خوشه‌بندی انجام‌شده موثر است. از این رو باید آزمایشاتی جامع و کامل برای بررسی عوامل موثر در خروجی صورت پذیرد. ابتدا به پایگاه‌های داده‌ی استفاده شده در این پژوهش پرداخته می‌شود، سپس برای مدل ارزیابی، معیارهای ارزیابی، آزمایشات و نتایج آن‌ها توضیحات کاملی ارائه خواهد شد.

آزمایشاتی برای تعیین عملکرد روش پیشنهادی در هر دو جنبه کیفیت خوشه‌بندی و دقت طبقه‌بندی انجام می‌شود. رویکردهای خوشه‌بندی نیمه نظارتی پیشنهادی به عناوین اختصاری SSConE و SSClusE در آزمایشات و نمودارها برچسب‌گذاری شده‌اند. SSConE روشی است که مفاهیم به‌صورت نیمه نظارتی استخراج می‌شوند و از داده‌های برچسب‌دار در مرحله‌ی استخراج مفاهیم استفاده شده است. اما در روش SSClusE اسناد برچسب‌دار در مرحله خوشه‌بندی سند به مدل وارد شده و اسناد در فرآیندی نیمه نظارتی خوشه‌های مناسب خود را پیدا می‌کنند. روش‌های پیشنهادی با خوشه‌بندی کامینز، TESC (BoW) [23]، Doc2Vec [17] ، کیسه مفاهیم (BoC) [18] و Naïve Bayes Expectation-Maximization (NBEM) [21] مقایسه می‌شوند.

کلیات الگوریتم پیشنهادی این پژوهش به صورت زیر است:

۱. پیش‌پردازش‌های مورد نیاز بر روی مجموعه‌ی اسناد
۲. جداکردن کلمات اسناد یا به عبارت دیگر توکنایز کردن کلمات
۳. خوشه‌بندی توکن‌های ایجاد شده و ساخت مفاهیم (نیمه نظارتی یا بدون ناظر)

۴. تولید بردار اسناد با استفاده از مفاهیم استخراج شده

۵. اجرای مراحل ۱ تا ۴ برای اسناد آموزش و آزمایش

۶. خوشه‌بندی اسناد آموزش و ایجاد خوشه‌های اسناد (نیمه نظارتی یا بدون ناظر)

۷. شناسایی خوشه‌ی داده‌ی آزمایش با پیدا کردن خوشه‌ای که بیشترین شباهت را با داده‌ی آزمایش دارد.

۴-۲ پایگاه داده

به منظور نشان دادن عملکرد روش پیشنهادی و کاربرد آن، از دو پایگاه داده‌ی متداول پردازش متن یعنی Reuters-21578 و 20NewsGroup در آزمایشات استفاده می‌شود. در ادامه در مورد هر یک از این پایگاه‌های داده توضیحات بیشتری ارائه می‌شود.

۴-۲-۱ پایگاه داده‌ی Reuters-21578

اسناد موجود در مجموعه Reuters-21578 در سال ۱۹۸۷ در شبکه خبری رویترز منتشر شده است. اسناد توسط پرسنل رویترز و گروه Carnegie، جمع‌آوری و برچسب‌گذاری شده‌اند. در سال ۱۹۹۰، اسناد توسط رویترز برای اهداف تحقیقاتی در اختیار آزمایشگاه بازیابی اطلاعات گروه علوم کامپیوتر و اطلاعات در دانشگاه ماساچوست قرار گرفت. قالب بندی اسناد و تولید پرونده های اطلاعاتی مرتبط در سال ۱۹۹۰ توسط دیوید دی لوئیس و استفان هاردینگ در آزمایشگاه بازیابی اطلاعات انجام شد. در این پایگاه داده که به صورت برخط[۶۸] قابل دسترسی است، ۲۱۵۷۸ خبر از وبسایت رویترز قرار دارد و در ۱۳۵ کلاس

[۶۸] https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection

تقسیم‌بندی می‌شوند. برای مقایسه بهتر، مشابه تحقیق [23] ما زیرمجموعه‌ای از پایگاه داده را مورد استفاده قرار می‌دهیم. در این تحقیق، ۲۱۲۳ سند از Reuters-21578 در چهار دسته‌ی "agriculture" (۵۸۲ سند) ، "crude" (۵۷۸ سند) ، "trade" (۴۸۵ سند) و "interest" (۴۷۸ سند) به طور تصادفی انتخاب می‌شوند. هر سند موجود در این پایگاه داده شامل چندین مورد از جمله تاریخ، موضوع، نویسنده، مکان، متن و غیره است که با تگ‌هایی قبل از متن اصلی مشخص شده‌اند. برای استخراج هر کدام از این اطلاعات نیاز به پیش‌پردازش‌هایی برای تمیز کردن متون می‌باشد. شکل (۴-۱) تصویری از یک سند موجود در این پایگاه داده را نشان می‌دهد.

```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="19420" NEWID="3002">
<DATE> 9-MAR-1987 05:03:09.75</DATE>
<TOPICS><D>money-fx</D><D>interest</D></TOPICS>
<PLACES><D>france</D></PLACES>
<PEOPLE></PEOPLE>
<ORGS></ORGS>
<EXCHANGES></EXCHANGES>
<COMPANIES></COMPANIES>
<UNKNOWN>
&#5;&#5;&#5;RM
&#22;&#22;&#1;f0423&#31;reute
u f BC-BANK-OF-FRANCE-SETS-M   03-09 0112</UNKNOWN>
<TEXT>&#2;
<TITLE>BANK OF FRANCE SETS MONEY MARKET TENDER</TITLE>
<DATELINE>    PARIS, March 9 - </DATELINE><BODY>The Bank of France said it invited offers
of first category paper today for a money market intervention
tender.
    Money market dealers said conditions seemed right for the
Bank to cut its intervention rate at the tender by a quarter
percentage point to 7-3/4 pct from eight, reflecting an easing
in call money rate last week, and the French franc's steadiness
on foreign exchange markets since the February 22 currency
stabilisation accord here by the Group of Five and Canada.
    Intervention rate was last raised to eight pct from 7-1/4
on January 2. Call money today was quoted at 7-11/16 7-3/4 pct.
 REUTER
&#3;</BODY></TEXT>
</REUTERS>
```

شکل ۴-۱. تصویری از نمونه سند موجود در پایگاه داده Reuters-21578

۴-۲-۲ پایگاه داده 20NewsGroup

مجموعه داده‌های 20NewsGroup مجموعه‌ای از تقریباً ۱۸ هزار سند از گروه‌های خبری است که به طور مساوی در ۲۰ گروه مختلف خبری تقسیم شده است. این مجموعه در ابتدا توسط کِن لنگ با هدف

یادگیری فیلترکردن مقالات گردآوری شده است. مجموعه 20NewsGroup به مجموعه داده‌های پرطرفداری برای آزمایش در کاربردهای متنی تکنیک‌های یادگیری ماشین، مانند طبقه‌بندی متن و خوشه‌بندی متن تبدیل شده است. داده‌ها در ۲۰ دسته خبری مختلف سازمان یافته‌اند که هر یک مربوط به یک موضوع متفاوت است. برخی از گروه‌های خبری ارتباط بسیار نزدیک با یکدیگر دارند (به عنوان نمونه comp.sys.ibm.pc.hardware / comp.sys.mac.hardware)، در حالی که برخی دیگر ارتباط زیادی ندارند (به عنوان مثال misc.forsale / soc.religion.christian). این مجموعه داده به دو زیرمجموعه تقسیم می‌شود: یکی برای آموزش و دیگری برای آزمایش (یا برای ارزیابی عملکرد). تقسیم بین مجموعه آموزش و آزمایش براساس اسناد منتشر شده قبل و بعد از یک تاریخ خاص است. در این پژوهش ۲۳۳۹ سند از چهار دسته، یعنی "sci" (۵۹۴ سند) ، "rec" (۵۹۷ سند) ،" comp" (۵۸۴ سند) و "talk" (۵۶۴ سند) به طور تصادفی انتخاب می‌شوند. هر سند دارای مشخصاتی از قبیل موضوع، دسته، تاریخ، سازمان، تعداد خطوط، شناسه، لغات اصلی و غیره است که در ابتدای هر سند قرار داده شده است. به همین دلیل برای جدا کردن این اطلاعات از متن نیاز به پیش‌پردازش می‌باشد.

۳-۴ مدل ارزیابی

در فاز پیش‌پردازش، روش‌های معمول مانند تبدیل حروف بزرگ به کوچک، حذف علائم غیرمجاز، حذف کلمات توقف و غیره انجام می‌شود. برای کاهش پیچیدگی آموزش شبکه‌ی عصبی Word2Vec، کلماتی که در کل متونِ مجموعه داده، کمتر از پنج بار دیده شده‌اند حذف می شوند و به ترتیب ۱۱۲۵۱ و ۲۶۳۴۲ کلمه منحصر به فرد، در مجموعه داده های Reuters-21578 و 20NewsGroup باقی می‌ماند. در بخش خلاصه سازی کلمات در فاز استخراج نیمه نظارتی مفهوم، بردار کلمات به ۲۰۰۰ شبه کلمه برچسب‌دار خلاصه می‌شوند.

در بخش استخراج بدون نظارت مفاهیم، از آنجا که تعداد مفاهیم طول بردارهای سند را تعیین می‌کند، بر عملکرد خروجی سیستم تأثیر می‌گذارد. عملکرد مدل پیشنهادی SSClusE، از نظر دقت طبقه‌بندی، هنگامی که تعداد مفاهیم تغییر می‌کند، در جدول (۴-۱) نشان داده شده است. با توجه به نتایج این جدول، بهترین دقت، زمانی حاصل می‌شود که تعداد مفاهیم بین ۴۰۰ تا ۵۰۰ باشد و پس از آن دقت به طور قابل توجهی تغییر نمی‌کند. به همین دلیل در تمام آزمایشاتِ روش SSClusE تعداد مفاهیم ۴۰۰ انتخاب شده است.

جدول ۴-۱. دقت طبقه بندی اسناد بر روی پایگاه داده Reuters-21578 برای مقادیر متغیر تعداد مفاهیم (Z)

	تعداد مفاهیم					
	۱۰۰	۲۰۰	۳۰۰	۴۰۰	۵۰۰	۶۰۰
روش SSClusE	۵۷,۰۲	۶۴,۹	۷۶,۸	۷۸,۸۲	۷۹,۴۷	۷۹,۳۷
روش BoW	۶۲,۱۶					

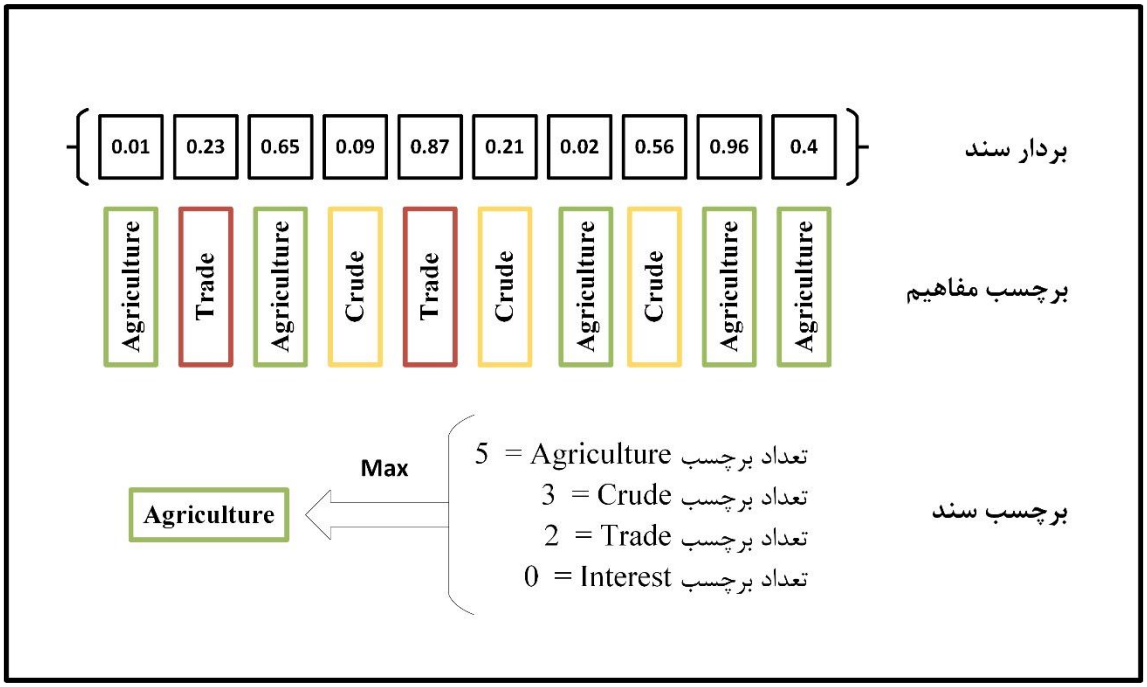
همانطور که در **بخش ۳-۳-۳-۱** توضیح داده شده است، هنگامی که مفاهیم دارای برچسب هستند (روش SSConE)، وظیفه اختصاص یک سند d به یک خوشه، نیازمند یافتن موثرترین برچسب در بردار سند $\vec{d}$ است. سه روش مختلف در این زمینه پیشنهاد و ارزیابی می‌شود که موثرترین برچسب را در بردار سند انتخاب می‌کند. این سه روش عبارتند از:

(۱) برچسب با بیشترین وزن در بردار سند $\vec{d}$:

این روش در واقع بزرگترین درایه در بردار سند $\vec{d}$ را پیدا کرده و برچسب این درایه (هر درایه در بردار سند معرف یه مفهوم می‌باشد) را به عنوان برچسب سند مورد نظر انتساب می‌دهد.

(۲) برچسب با بیشترین فراوانی در بین مفاهیم تشکیل‌دهنده بردار سند $\vec{d}$:

از آنجایی که هر درایه در بردار سند معرف یک مفهوم است و هر مفهوم برچسب مخصوص به خود را دارد، فراوانی این برچسب‌ها می‌تواند معیاری برای تشخیص برچسب سند مورد نظر باشد. به همین علت در این روش برچسب با بیشترین فراوانی در بردار سند به عنوان برچسب سند منتسب می‌گردد. برای درک بهتر، شکل (۴-۲) نحوه‌ی انتساب برچسب را برای بردار سندی فرضی نشان می‌دهد.

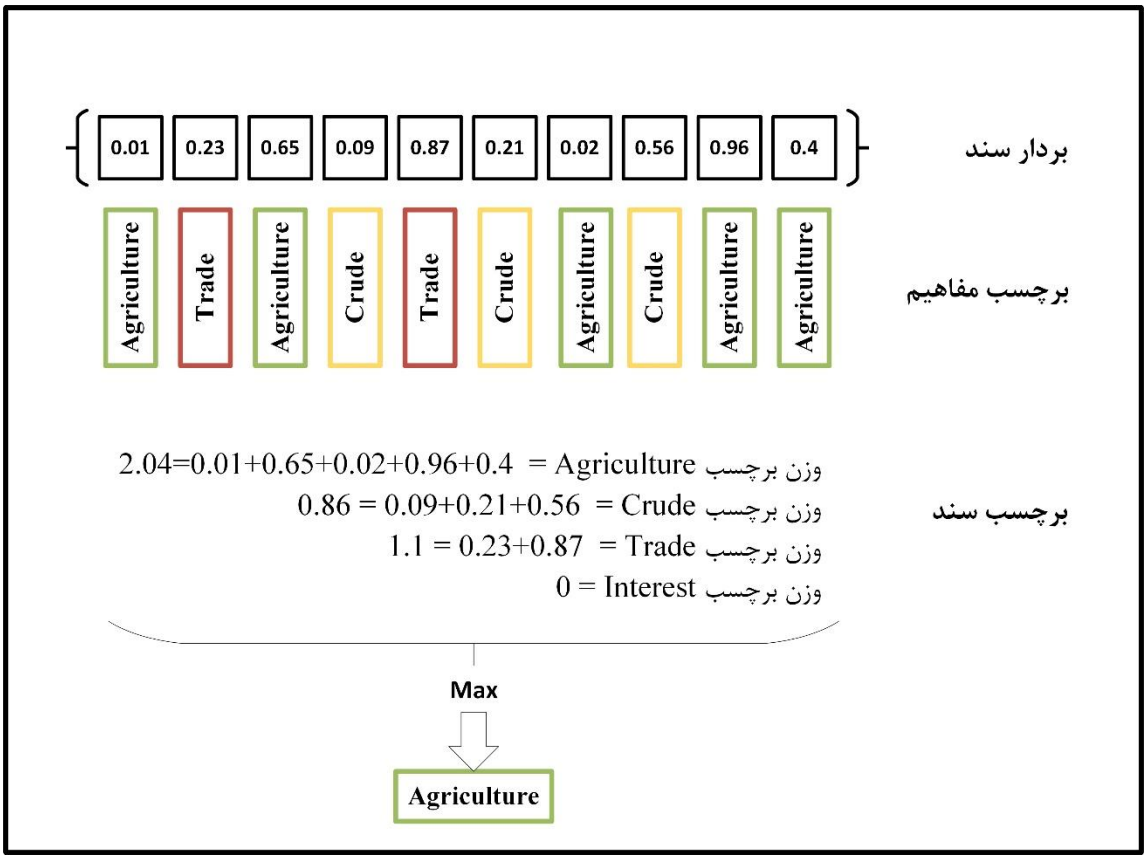


شکل ۴-۲. روش دوم نحوه‌ی تخصیص برچسب به سند در SSConE – مثال فرضی بر روی پایگاه داده Reuters-21578

(۳) برچسب با بزرگترین وزن تجمیعی در بردار سند $\vec{d}$:

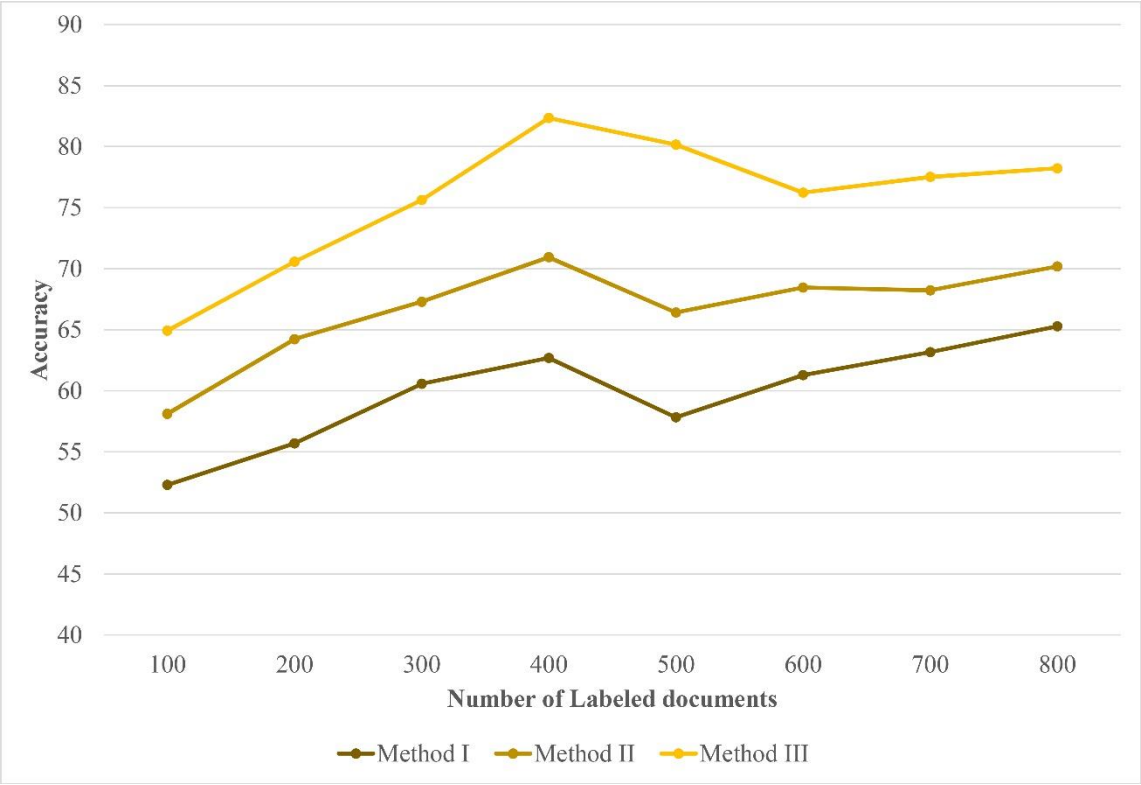
همانطور که در **بخش ۳-۳-۳-۱** رابطه (۳-۳) شرح داده شده است. به عبارت دیگر این روش بیان می‌کند که کدام مفهوم وزن بالاتری در سند مورد نظر دارد. مجموع مقادیر درایه‌ها برای هر برچسب در

بردار سند محاسبه شده و برچسبی که بیشترین وزن را داشته باشد، به عنوان برچسب سند انتخاب می‌شود. شکل (۴-۳) بیانگر این موضوع می‌باشد.



شکل ۴-۳. روش سوم نحوه‌ی تخصیص برچسب به سند در SSConE – مثال فرضی بر روی پایگاه داده Reuters-21578

هر سه روش مورد آزمایش قرار گرفته‌اند و با تغییر تعداد اسناد برچسب‌دار، نتایج دقت طبقه‌بندی ثبت و در شکل (۴-۴) مشاهده می‌شود. همانطور که مشاهده شد، روش سوم بالاترین دقت را در طبقه‌بندی اسناد نشان می‌دهد. در بقیه آزمایشات، این روش برای انتساب خوشه در SSConE استفاده می‌شود.



شکل ۴-۴. دقت طبقه‌بندی اسناد روش پیشنهادی (SSConE) با روش‌های مختلف تخصیص برچسب برای Reuters-21578.

انتخاب نقاط اولیه در الگوریتم کامینز کروی ضروری است و عملکرد خوشه‌بندی را تحت تأثیر قرار می‌دهد. برای حل این مسئله، وقتی کامینز کروی در الگوریتم مورد استفاده قرار می‌گیرد (به عنوان مثال هنگام استخراج شبه کلمات)، الگوریتم چندین بار با نقاط اولیه تصادفی اجرا می‌شود و خوشه-بندی‌ای با حداقل فاصله درون خوشه انتخاب می‌شود.

در پایان ذکر این نکته مهم است که تمامی آزمایشات صورت گرفته، بر روی یک سیستم شخصی با پردازنده **Intel Core i5** و **۶ گیگابایت** حافظه انجام شده است.

۴-۴ معیارهای ارزیابی

۴-۴-۱ معیار خطای میانگین مربعات نرمال‌شده[۶۹] ($NMSE$)

برای مقایسه کیفیت خوشه‌بندی روش پیشنهادی با سایر روش‌ها، خطای معیار میانگین مربعات نرمال-شده به اختصار $NMSE$ که در روابط (۴-۱) و (۴-۲) تعریف شده، استفاده می‌شود. در این روابط، μ مجموعه‌ای از مراکز خوشه، X مجموعه‌ای از نقاط داده و μ_{c_i} مرکز خوشه‌ی داده x_i است. معیار $NMSE$ میانگین فاصله بین نقاط درون خوشه و مرکز خوشه را محاسبه می‌کند و مقادیر کوچکتر نشان دهنده واریانس کمتر در خوشه و به دنبال آن خوشه‌بندی با کیفیت‌تری است.

$$MSE\ (X,\mu) = \frac{1}{N}\sum_{i}(x_i - \mu_{c_i})^2 \qquad (۱\text{-}۴)$$

$$NMSE(X,\mu) = MSE(x,\mu)/MSE(x,0) \qquad (۲\text{-}۴)$$

۴-۴-۲ معیار اطلاعات متقابل عادی[۷۰] (NMI)

معیار اطلاعات متقابل عادی (NMI) یک معیار اعتبارسنجی خوشه‌ای است که کیفیت خوشه‌بندی داده‌ها را با توجه به برچسب‌های از قبل داده شده آن‌ها ارزیابی می کند. NMI ارزیابی می‌کند که الگوریتم خوشه‌بندی به چه صورت می‌تواند برچسب‌های اصلی داده را بازسازی کند [61]. از این معیار زمانی می‌توان استفاده کرد که برچسب داده‌ها موجود باشد. خروجی این معیار عددی در بازه‌ی $[0,1]$ است که مشابهت آماری بین برچسب خوشه‌های تولید شده و برچسب‌های اصلی داده‌ها را نمایان می‌کند.

[۶۹] Normalized Mean Squared Error (NMSE)

[۷۰] Normal Mutual Information

مقدار صفر یک انتساب خوشه ناموفق را نشان می‌دهد درحالی‌که مقادیر نزدیک به یک نشان می‌دهد که خوشه‌بندی می تواند کلاس‌های واقعی داده را دوباره ایجاد کند. معیار NMI نسبت به معیار آنتروپی عملکرد بهتری را در نمایش کیفیت خوشه‌ها از خود نشان می‌دهد. علت این امر این است که معیار آنتروپی به تعداد خوشه‌ها وابسته است و هرچه این تعداد بیشتر باشد، معیار آنتروپی افزایش می‌یابد. اما معیار NMI این‌گونه نیست و با افزایش تعداد خوشه‌ها الزاما افزایش ندارد. رابطه (۴-۳) تعریف ریاضی این معیار را نشان می‌دهد.

$$NMI = \frac{I(C;K)}{(H(C) + H(K))/2} \qquad (۳\text{-}۴)$$

در این رابطه $I(X;Y) = H(X) - H(X|Y)$ اطلاعات متقابل بین متغیرهای تصادفی X و Y، $H(X)$ آنتروپی شانون X و $H(X|Y)$ آنتروپی مشروط X با توجه به Y است.

۴-۴-۳ دقت [۷۱] طبقه‌بندی

دقت طبقه بندی معیاری است که عملکرد یک طبقه‌بندی را با یک مقدار عددی اندازه‌گیری می‌کند. این مقدار نشان می‌دهد که چند نمونه به درستی طبقه‌بندی شده‌اند. با تقسیم تعداد نمونه‌های طبقه-بندی شده صحیح بر تعداد کل نمونه‌ها، مقدار دقت همانطور که در معادله (۴-۴) نشان داده شده بدست می‌آید. در این رابطه $\hat{y}_i$ پیش بینی کلاس برای نمونه l است.

$$Accuracy = \frac{\sum_i 1(\hat{y}_i = y_i)}{|X|} \qquad (۴\text{-}۴)$$

[۷۱] Accuracy

۴-۵ نتایج

این بخش به گزارش، بررسی و تحلیل نتایج آزمایشات انجام شده می‌پردازد. با توجه به صحبت‌های گذشته، عملکرد روش پیشنهادی از دو جنبه مورد ارزیابی قرار گرفته می‌شود. جنبه‌ی کیفیت خوشه-بندی اسناد و جنبه‌ی دقت طبقه‌بندی. جنبه‌ی کیفیت خوشه‌بندی اسناد، شامل ارزیابی کیفیت خوشه-بندیِ روش پیشنهادی با دو معیار $NMSE$ و NMI است که برای مشاهده برتری روش پیشنهادی، آزمایشاتی جهت مقایسه با سایر روش‌ها انجام می‌گیرد. همچنین در جنبه‌ی دقت طبقه‌بندی اسناد، با تغییر پارامترهای مختلف، میعار دقت اندازه‌گیری می‌شود و با سایر روش‌های طبقه‌بندی متن مقایسه صورت می‌پذیرد.

۴-۵-۱ ارزیابی کیفیت خوشه‌بندی

۴-۵-۱-۱ اثر نسبت نگهداری[۷۲]

در آزمایش اول، ما تأثیر تغییر اندازه مجموعه داده‌های آموزشی را ارزیابی می‌کنیم. نسبت نگهداری درصد داده‌های مورد استفاده برای آموزش مدل را نشان می‌دهد. در ابتدا، ۸۰ درصد از داده‌ها برای آموزش استفاده می‌شود، و در نهایت در شکل (۴-۵) در یک روند نزولی به ۲۰ درصد کاهش می‌یابد. یک مجموعه ۲۰۰تایی از اسناد برچسب‌دار در مجموعه آموزش در همه نسبت‌ها حفظ می‌شود به عبارت دیگر مابقی داده‌های آموزشی بدون برچسب می‌باشند. بنابراین ، با کاهش نسبت نگهداری، درصد تعداد داده‌های دارای برچسب به غیر برچسب در مجموعه آموزش افزایش می‌یابد. تغییر در مجموعه داده‌های

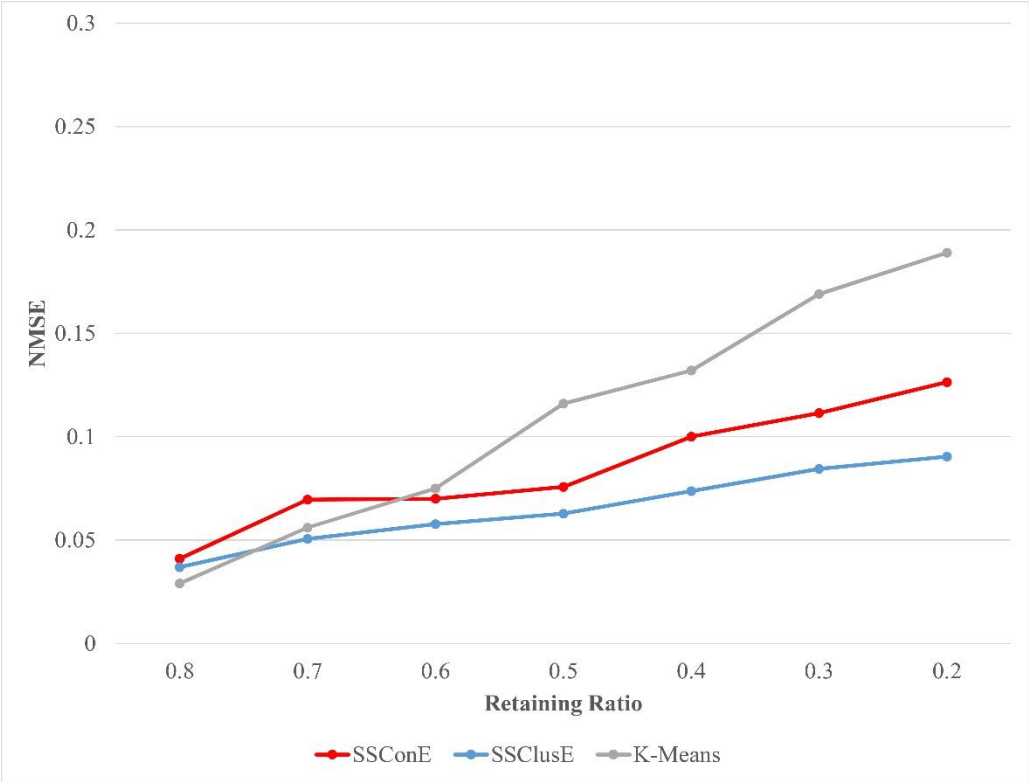
[۷۲] Effect of Retaining Ratio

آموزشی منجر به نتایج مختلف خوشه‌بندی هنگام ساخت مدل می‌شود. بنابراین، این پیکربندی می‌تواند تأثیر نسبت بین تعداد اسناد برچسب‌دار و بدون برچسب را نشان دهد. معیار $NMSE$ بر روی مجموعه‌ی آزمون در شکل (۴-۵) اندازه‌گیری و عملکرد الگوریتم‌های پیشنهادی با K-Means مقایسه می‌شود. تعداد خوشه‌های سند تولید شده توسط روش پیشنهادی SSClusE برای نسبت‌های نگهداری مختلف در جدول (۴-۱) نشان داده شده است. تعداد خوشه های نهایی تولید شده با روش SSConE برابر با تعداد دسته‌های داده است.

جدول ۴-۲ . تعداد خوشه های تولید شده توسط روش SSClusE متناظر با نسبت های نگهداری آزمایش شکل (۴-۷)

	نسبت نگهداری						
	۰,۲	۰,۳	۰,۴	۰,۵	۰,۶	۰,۷	۰,۸
Reuters-21578	۴۹	۶۳	۸۲	۸۹	۹۱	۱۰۸	۱۱۰
20NewsGroup	۵۲	۶۱	۸۶	۱۰۰	۱۱۶	۱۲۱	۱۲۷

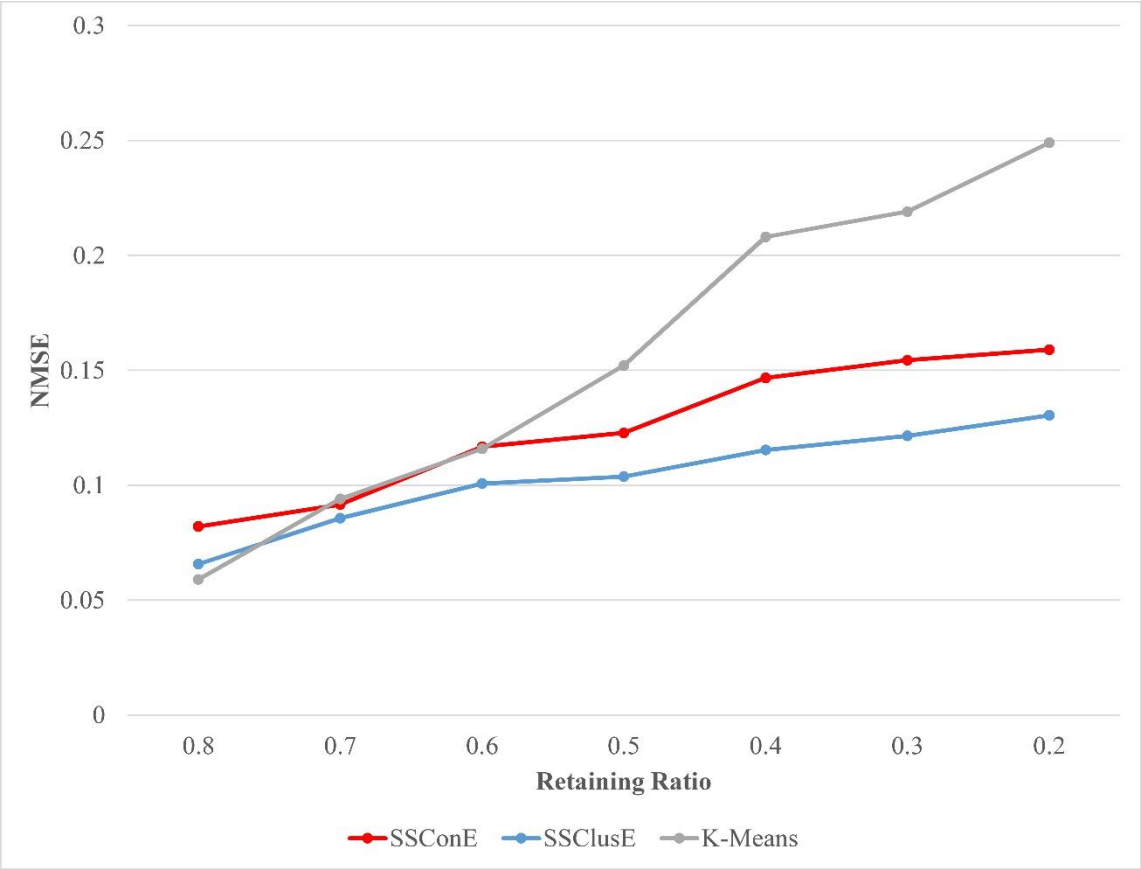
همانطور که انتظار می‌رفت، با نسبت‌های نگهداری بزرگتر و در نتیجه مجموعه‌های آموزشی بزرگتر، خوشه‌های بیشتری توسط الگوریتم تولید می‌شوند و مقادیر $NMSE$ کمتری بدست می‌آیند. کاهش نسبت نگهداری، که به معنای کوچک‌کردن مجموعه آموزش است، خوشه‌های کمتری و مقادیر بزرگتر $NMSE$ تولید می‌کند.



شکل ۴-۵ . مقادیر NMSE از خوشه‌بندی اسناد پایگاه داده Reuters-21578 برای روش‌های پیشنهادی و خوشه‌بندی K-Means با نسبت نگهداری متغیر

نتایج گزارش شده در نمودارهای شکل (۴-۶) نشان می‌دهد که روش پیشنهادی در تمام نسبت‌های نگهداری کمتر از ۰,۷ در هر دو مجموعه‌ی اسناد Reuters-21578 و 20NewsGroup از خوشه‌بندی K-Means بهتر عمل می‌کند. در این نسبت نگهداری (۰,۷)، مقدار $NMSE$ روش K-Means برای گروه Reuters-21578 و 20NewsGroup به ترتیب ۰,۰۵۶ و ۰,۰۹۴ است، در حالی که SSClusE دارای مقادیر ۰,۰۵۰ و ۰,۰۸۵ برای معیار $NMSE$ است. با کاهش اندازه مجموعه‌ی آموزشی، الگوریتم در مقایسه با K-Means بدون نظارت، خوشه‌هایی با کیفیت بالاتری تولید می‌کند، زیرا همانطور که در شکل (۴-۶) دیده می‌شود، هرچه نسبت نگهداری کاهش می‌یابد شیب نمودار برای روش K-Means با

سرعت بیشتری افزایش پیدا می‌کند و این بدین معنی است که افزایش زیادی را برای $NMSE$ تجربه می‌کند. این آزمایش توجیه می‌کند که استفاده از داده‌های دارای‌برچسب برای هدایت خوشه‌بندی بدون نظارت، خوشه‌هایی با فاصله درون خوشه‌ای کمتری شکل می‌دهد. بعلاوه، این نشان می‌دهد که حتی زمانی که از مجموعه کوچکی از داده‌های دارای برچسب استفاده می‌شود، الگوریتم پیشنهادی قادر به تولید مدل خوشه‌بندی بهتری است. در مقایسه با روش K-Means، شیب کمتری برای مقدار $NMSE$ مشاهده شده است که نشانگر برتری روش‌های پیشنهادی در خوشه‌بندی با تعداد کم اسناد برچسب‌دار است. یک مشاهده جالب این است که SSClusE در این آزمایش عملکرد بهتری نسبت به SSConE دارد.

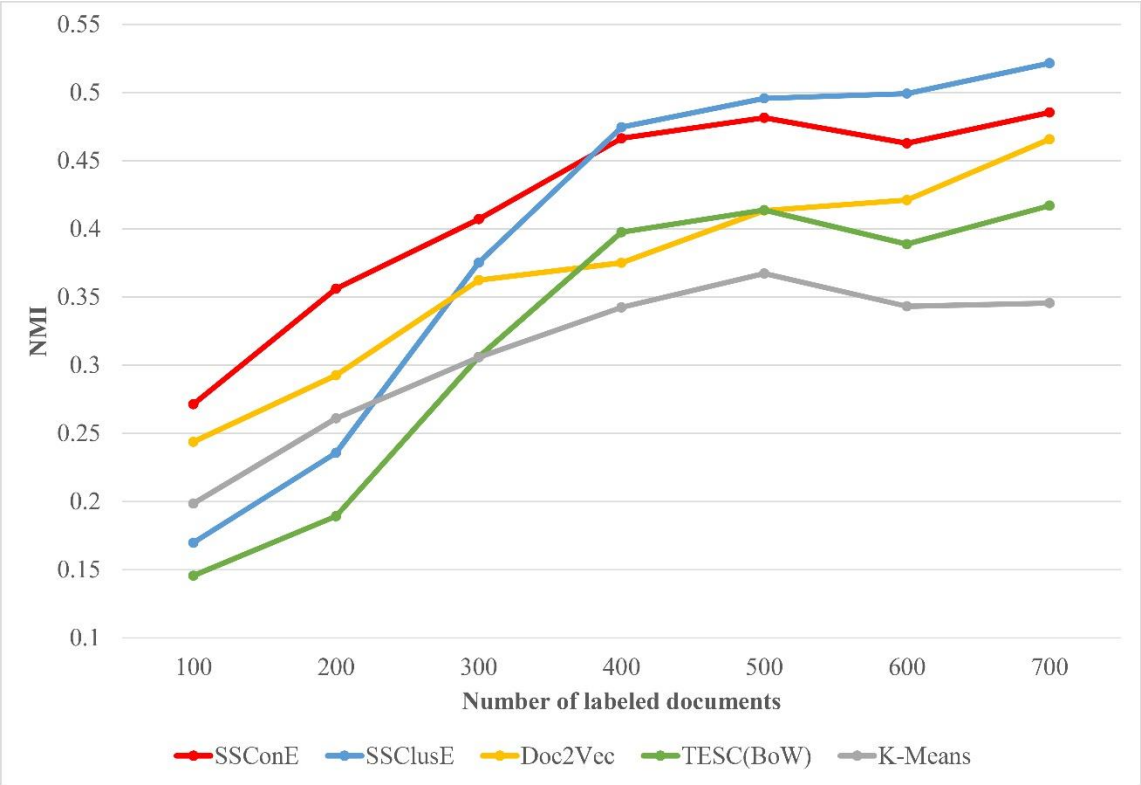


شکل ۴-۶ . مقادیر NMSE از خوشه‌بندی اسناد پایگاه داده 20NewsGroup برای روش‌های پیشنهادی و خوشه‌بندی K-Means با نسبت نگهداری متغیر

۴-۵-۱-۲ اثر تعداد اسناد برچسب‌دار

تعداد اسناد دارای برچسب تأثیر محسوسی بر کیفیت خوشه‌بندی انجام شده دارد. برای مشاهده این اثر، آزمایش دیگری را انجام دادیم که در آن اندازه مجموعه آموزش ثابت می‌ماند، اما اندازه زیر مجموعه داده‌های دارای برچسب در این مجموعه متغیر است. با تعداد متفاوت داده‌های دارای برچسب، مدل‌های خوشه‌بندی گوناگونی ساخته می‌شوند. این پیکربندی می‌تواند عملکرد روش‌های پیشنهادی را هنگام مواجهه با نسبت‌های مختلف داده‌های برچسب‌دار ارزیابی کند. این آزمایش می‌تواند به طور خاص نشان دهد که آیا روش‌های پیشنهادی این پژوهش، هنگام داشتن فقط یک زیرمجموعه کوچک از داده‌های

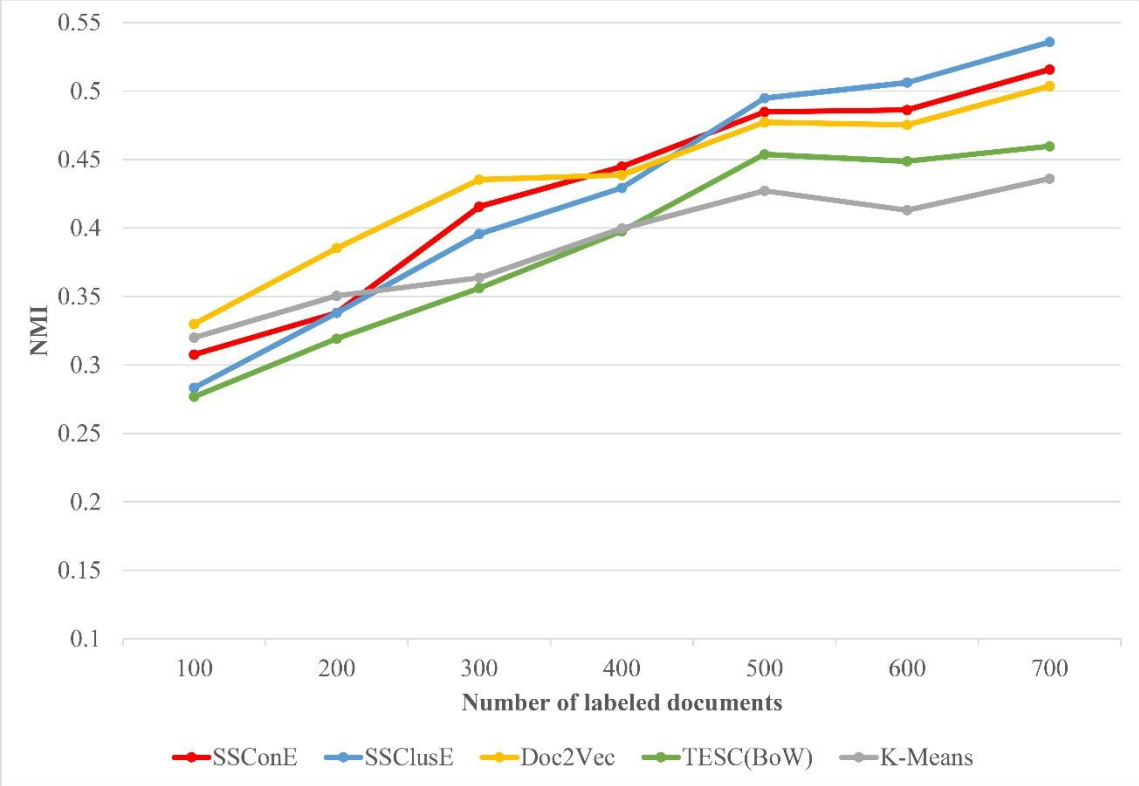
دارای برچسب از لحاظ عملکردی مقاوم هستند و آیا می‌توانند از تعداد زیادی داده دارای برچسب برای بهبود عملکرد خود بهره‌مند شوند.



شکل ۴-۷. مقادیر NMI خوشه بندی اسناد پایگاه داده Reuters-21578 برای روش های پیشنهادی در مقایسه با K-Means، TESC (BoW) و Doc2Vec در تعداد مختلف اسناد دارای برچسب

شکل (۴-۷) مقادیر NMI را نشان می‌دهد که با تغییر تعداد اسناد برچسب‌دار هنگام ثابت نگه‌داشتن تعداد اسناد بدون برچسب بدست آمده است. مطابق شکل (۴-۷)، خوشه‌بندی حاصل‌شده از روش پیشنهادی بر روی پایگاه داده Reuters-21578 از کیفیت بهتری نسبت به سایر روش‌ها برخوردار است. همچنین قابل توجه است که با افزایش تعداد اسناد برچسب‌دار، مقدار NMI و در نتیجه کیفیت خوشه-بندی به شکل قابل محسوسی افزایش می‌یابد. به عنوان مثال، در بدترین حالت، هنگامی که تنها ۵ درصد اسناد دارای برچسب باشند (۱۰۰ سند)، مقدار NMI برای روش SSConE عدد ۰,۲۷۱،

SSClusE عدد ۰,۱۶۹، روش TESC (BoW) عدد ۰,۱۳۰، روش Doc2Vec عدد ۰,۲۴۵ و روش K-Means عدد ۰,۱۹۸ است. در مورد دیگر، وقتی ۲۳ درصد از اسناد دارای برچسب هستند (۴۰۰ سند)، مقدار NMI برای روش SSConE به عدد قابل توجه ۰,۴۶۵ می‌رسد. برای مجموعه‌ی داده اسناد 20NewsGroup در شکل (۴-۸)، با تعداد ۷۰۰ سند دارای برچسب، مقدار NMI عدد ۰,۵۴ است و در بدترین حالت، وقتی ۱۰۰ داده برچسب‌دار وجود دارد، مقدار NMI عدد ۰,۳۱۱ است. در این پایگاه داده، مشابه پایگاه داده Reuters-21578 روش پیشنهادی عملکردی برتر از سایر روش‌ها از خود بجای می-گذارد.



شکل ۴-۸. مقادیر NMI خوشه بندی اسناد پایگاه داده 20NewsGroup برای روش های پیشنهادی در مقایسه با K-Means، TESC (BoW) و Doc2Vec در تعداد مختلف اسناد دارای برچسب.

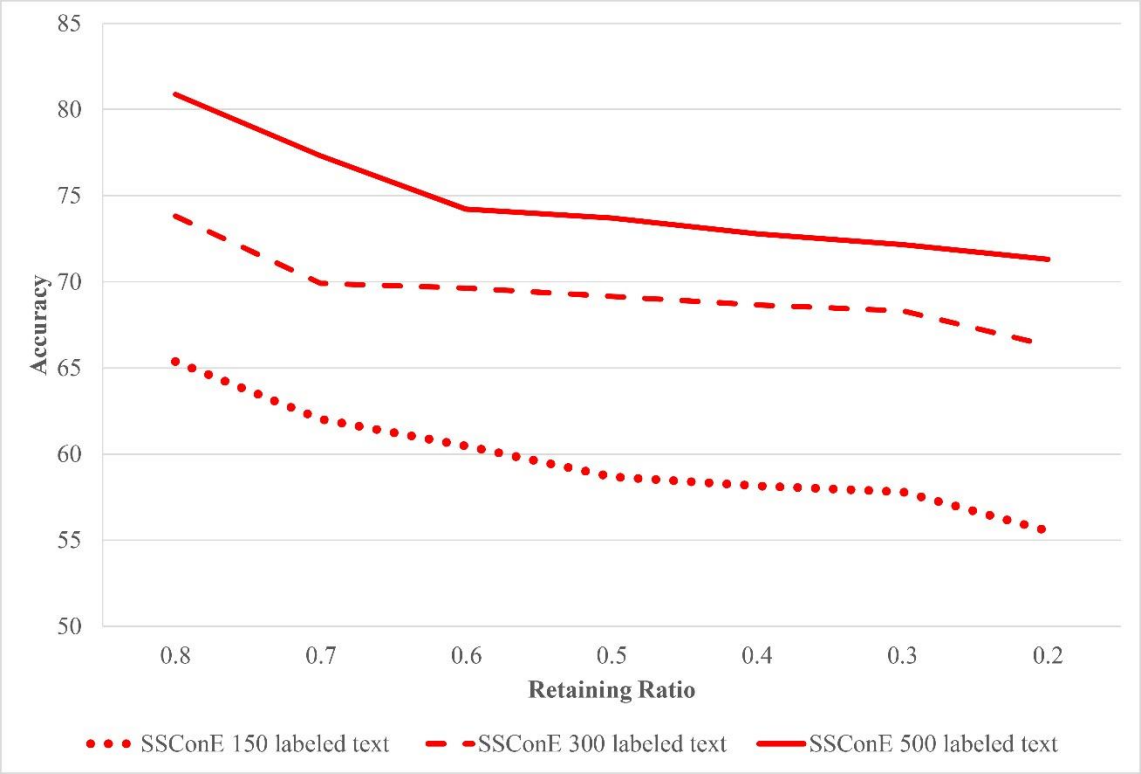
از نمودارهای اشکال (۴-۷) و (۴-۸)، می‌توان دریافت در حالتی‌که تنها مجموعه‌ی کوچکی از داده‌ها دارای برچسب هستند، روش‌های پیشنهادی برتر از همه روش‌های دیگر هستند. با افزایش تعداد اسناد برچسب‌دار، اختلاف بین NMI روش‌های پیشنهادی و سایر روش ها افزایش می‌یابد. این نشان می‌دهد که هرچه داده‌های دارای برچسب بیشتری وجود داشته باشد، کیفیت خوشه‌بندی انجام شده نسبت به سایر روش ها بهتر است. همچنین، تأیید می‌کند که طراحی الگوریتم از مزیت داده‌های برچسب‌دار بیشتر برای کشف خوشه‌های داده بهره می برد. مشاهده جالب دیگر این است که وقتی داده‌های برچسب-دار کمتر در دسترس است، استخراج مفهوم نیمه‌نظارتی (SSConE) در هر دو مجموعه‌ی داده اسناد عملکرد بهتری نسبت به استخراج مفهوم بدون نظارت (SSClusE) دارد. به تدریج با افزایش تعداد داده‌های دارای برچسب، SSClusE در این ارزیابی پیشی می‌گیرد. روش‌های پیشنهادی با استخراج و تجزیه اجزای متن در مقایسه با سایر روش‌ها، گام مهمی در جهت خوشه‌بندی بهتر اسناد برمی‌دارند. با استفاده از داده‌های دارای برچسب، مراکز خوشه بهتر تنظیم می‌شوند و می‌توان خوشه‌بندی قابل اطمینان‌تری را انتظار داشت.

۴-۵-۲ ارزیابی دقت طبقه‌بندی

۴-۵-۲-۱ اثر تعداد اسناد برچسب‌دار

اشکال (۴-۹) و (۴-۱۰) دقت روش پیشنهادی را زمانی که نسبت نگهداری و تعداد داده‌های دارای برچسب تغییر می‌کند، نشان می‌دهند. محور عمودی میانگین دقت‌ها را در طبقه‌بندی اسناد آزمون نشان می‌دهد و محور افقی تعداد اسناد برچسب‌داری را که برای آموزش مدل استفاده می‌شود. نتایج مشخص می‌کند که روش‌های پیشنهادی می‌توانند از تعداد بیشتری اسناد دارای برچسب برای بهبود

عملکرد کار طبقه‌بندی متن بهره مند شوند. علاوه بر این، تعداد اسناد بدون برچسب بر دقت کار طبقه بندی تاثیر می‌گذارد. اگر داده های بدون برچسب زیاد باشد، بایاس[۷۳] هر خوشه افزایش می‌یابد. این بدان معناست که بین مرکز خوشه‌ها و تراکم جمعیتی داده‌ها اختلاف بوجود خواهد آمد. اما اگر تعداد داده‌های بدون برچسب کم باشد، واریانس هر خوشه کاهش می‌یابد.

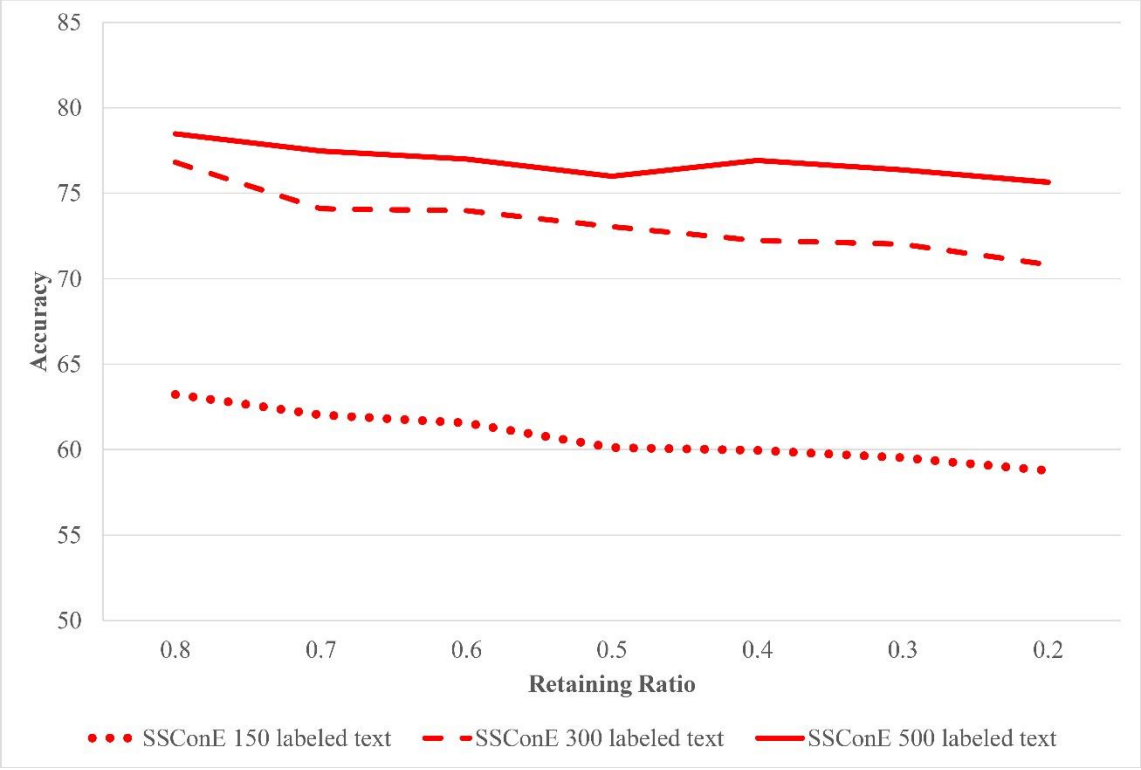


شکل ۴-۹. دقت طبقه بندی اسناد Reuters-21578 با روش پیشنهادی (SSConE) در نسبت‌های مختلف نگهداری و تعداد متغیر داده‌های برچسب‌دار.

در کاربرد طبقه‌بندی اسناد، اگر تعداد داده‌های استفاده شده در مرحله آموزش زیاد باشد، خوشه‌های بیشتری کشف می‌شود. با این حال اگر تعداد خوشه‌ها خیلی زیاد یا خیلی کم باشد، عملکرد طبقه‌بندی راضی‌کننده نیست. زیرا اگر تعداد خوشه‌ها بسیار کم باشد، هر خوشه شامل بیش از یک دسته اسناد

[۷۳] Bias

است. همچنین، اگر تعداد خوشه‌ها خیلی زیاد باشد، تعداد اسناد در هر خوشه کاهش می‌یابد و پیدا کردن دقیق ساختار و محدوده‌ی کلی خوشه‌ها چالش برانگیز است.

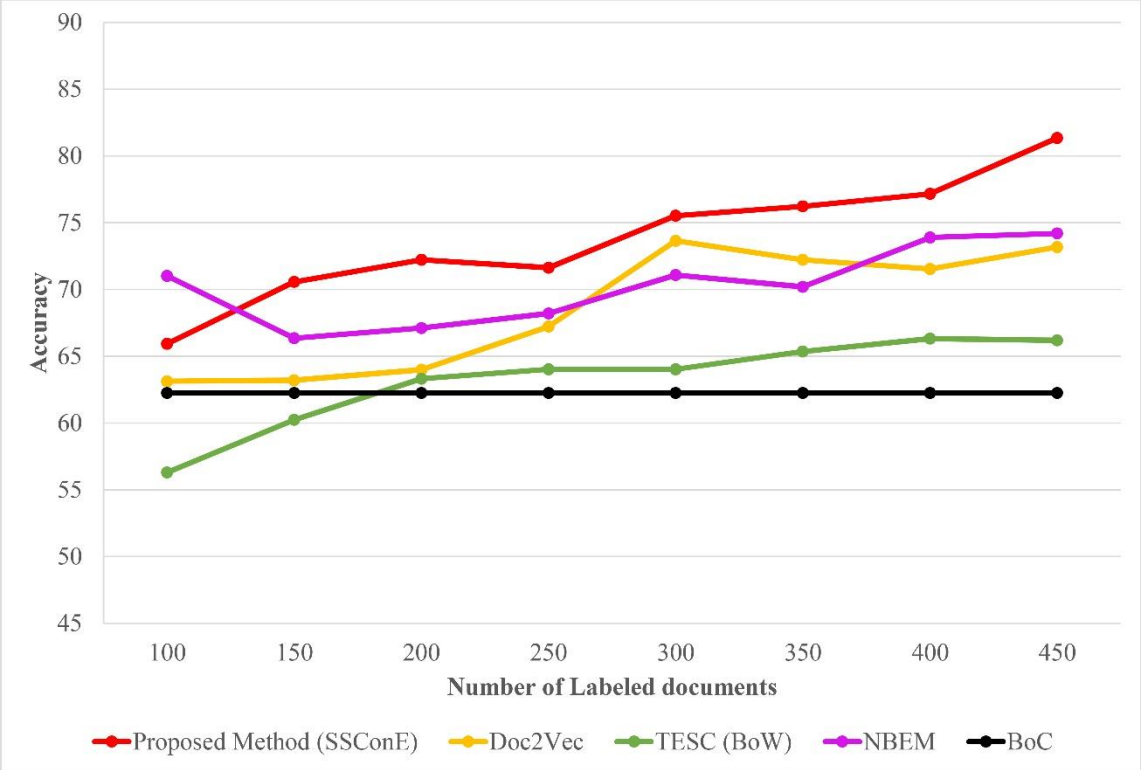


شکل ۴-۱۰. دقت طبقه بندی اسناد 20NewsGroup با روش پیشنهادی (SSConE) در نسبت های مختلف نگهداری و تعداد متغیر داده‌های برچسب دار.

۴-۵-۲-۲ مقایسه با روش‌های دیگر

برای مشخص شدن برتری روش‌های پیشنهادی نسبت به سایر روش‌ها، آزمایشی برای کاربرد طبقه‌بندی اسناد بر روی دو پایگاه داده مذکور انجام می‌دهیم. شکل (۴-۱۱) دقت طبقه‌بندی روش پیشنهادی را با NBEM، TESC (BoW)، Doc2Vec و BoC مقایسه می‌کند. محور عمودی دقت طبقه‌بندی را نشان می‌دهد و محور افقی تعداد اسناد برچسب‌دار است که با نسبت نگهداری ۰,۸ برای مجموعه‌ی آموزش استفاده می‌شوند. همانطور که از شکل (۴-۱۱) و (۴-۱۲) برداشت می‌شود، در هر دو مجموعه‌ی داده اسناد، روش پیشنهادی دقت بالاتری را به همراه دارد، که با افزایش تعداد داده‌های برچسب‌دار، دقت طبقه‌بندی نیز افزایش می‌یابد. این افزایش دقت با ارزیابی‌های قبلی سازگار است و تأیید می‌کند که

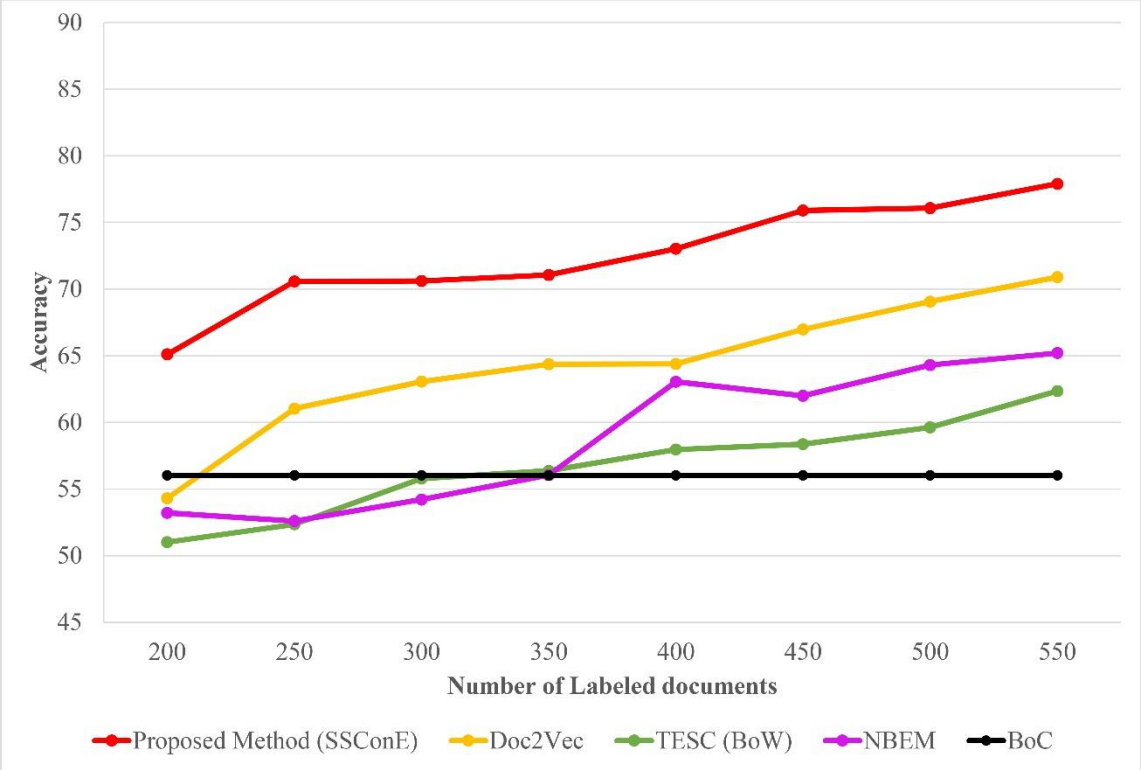
وقتی تعداد داده‌های دارای برچسب کم باشد، مدل نسبت به مابقی روش‌ها برتر است، در حالی که می‌تواند از مزیت مجموعه‌ی داده برچسب‌دار بزرگتر نیز بهره مند شود. سایر روش‌ها نیز با داشتن داده‌های برچسب‌دار بیشتر، افزایش دقت را تجربه می‌کنند.



شکل ۴-۱۱. دقت طبقه بندی اسناد Reuters-21578 با روش پیشنهادی در مقایسه با NBEM، BOW، Doc2Vec و BoC با نسبت نگهداری ۰,۸.

برای مجموعه داده‌ی Reuters-21578 ، پس از روش پیشنهادی، NBEM و Doc2Vec با یکدیگر رقابت نزدیکی دارند. بدترین عملکرد در طبقه‌بندی مجموعه داده‌های Reuters-21578 مربوط به TESC (BoW) و BoC است. البته باید به این نکته توجه شود که عملکرد BoC به عنوان یک رویکرد بدون نظارت، به تعداد داده‌های برچسب‌دار وابسته نیست. مطابق با شکل (۴-۱۲) برای مجموعه داده‌های گروه‌های 20NewsGroup، روش پیشنهادی دقتی را ارائه می‌دهد که در تمام شرایط حداقل ۵ درصد بهتر از تمام روش‌های دیگر است. Doc2Vec صاحب مقام دوم است و سایر روش‌ها در عین رقابت نسبت به یکدیگر، مقادیر دقت کمتری تولید می‌کنند. این نشان می‌دهد که استفاده از مدل‌سازی مفهومی برای نمایش اسناد برای طبقه‌بندی اسناد نیز موثر است. در پایان نکته‌ی حائز اهمیت این است که روش

پیشنهادی، به دلیل استفاده از رویکرد نیمه‌نظارت شده، حتی با مقدار کمی داده برچسب‌دار، می‌تواند به طور قابل توجهی دقت طبقه‌بندی را در مقایسه با BoC که از مفاهیم نیز استفاده می‌کند، افزایش دهد.



شکل ۴-۱۲. دقت طبقه بندی اسناد 20NewsGroup با روش پیشنهادی در مقایسه با NBEM، BOW، Doc2Vec و BoC با نسبت نگهداری ۰,۸.

۴-۶ جمع‌بندی

در این فصل پایگاه‌های داده و مدل ارزیابی به عنوان پیش نیازات مسئله‌ی ارزیابی و آزمایشات و نتایج مربوط به آن‌ها به طور مفصل مورد بررسی قرار گرفت. عملکرد دقیق هر دو روش پیشنهادی SSConE و SSClusE در کاربرهای خوشه‌بندی و طبقه‌بندی اسناد با سایر روش‌ها مقایسه و گزارش شد. همچنین تحلیل جامعی به منظور برجسته کردن برتری‌های روش‌های پیشنهادی ارائه شد. با مشاهده نمودارهای استخراج شده مشخص شد که مهمترین دلایل برتری روش‌های پیشنهادی، استفاده توام از رویکرد

نیمه‌نظارتی و نمایش مفهومی اسناد است که منجر به ایجاد خوشه‌بندی‌های با کیفیت‌تری شده است. در فصل بعد نتیجه‌گیری از نتایج و پیشنهادهایی برای کارهای آینده مورد بحث قرار خواهند گرفت.

فصل۵ : نتیجه‌گیری

۵-۱ جمع‌بندی پایان‌نامه

در این پژوهش، یک روش نیمه‌نظارتی مبتنی بر مفهوم برای خوشه‌بندی اسناد ارائه شده است. فرض اصلی این بود که نوع نمایش اسناد بر کیفیت خوشه‌بندی و کاربرد طبقه‌بندی اسناد تأثیر می‌گذارد. به منظور پوشش ضعف روش‌های سابق نمایش متن، برای اسناد نمایشی مبتنی بر مفهوم را اتخاذ کردیم. این مفاهیم بر اساس شباهت‌های معنایی بین کلمات پیکره متنی و با کمک خوشه‌بندی آن‌ها استخراج می‌شوند. این روش بر محدودیت‌ها و ضعف‌های سایر روش‌های نمایش متن از قبیل کیسه‌کلمات، فرکانس کلمه-معکوس فرکانس سند و تعبیه اسناد غلبه می‌کند. علاوه بر این، روش‌های پیشنهادی این مزیت را دارند که اسناد را مبتنی بر مفاهیم در یک فضای برداری با طول منطقی توصیف کنند، این در حالی است که برخلاف روش‌هایی مانند تعبیه اسناد و شبکه‌های عصبی که تفسیرپذیری کمی داشتند، نمایش مبتنی بر مفاهیم قابلیت تفسیرپذیری زیادی دارند و به راحتی اجزای تشکیل‌دهنده سند را نشان می‌دهند.

همچنین در ادامه پس از نمایش اسناد به صورت مفهومی، از رویکرد نیمه‌نظارتی برای خوشه‌بندی اسناد استفاده شده است. تعداد محدودی داده‌های دارای برچسب در کنار داده‌های بدون برچسب برای تنظیم مراکز خوشه استفاده می‌شود. در این تحقیق، دو روش برای استفاده از داده‌های دارای برچسب در فرآیند خوشه‌بندی معرفی شده است. در روش اول، اسناد برچسب‌دار در زمان نمایش اسناد و استخراج مفهوم به کار گرفته شده‌اند، در حالی که در روش دوم، داده‌های دارای برچسب در مرحله خوشه‌بندی واقعی اسناد استفاده می‌شوند. روش اول برای کشف ساختار پنهان در فضای تعبیه شده کلمات از داده‌های برچسب‌دار به عنوان ناظر بهره می‌برد و مفاهیم برچسب خورده را ایجاد می‌کند. داده‌های برچسب‌دار امکان کشف بهتر رابطه بین اجزای موجود در متن را فراهم می‌کنند. در روش دوم، مفاهیم موجود در اسناد از طریق یک فرآیند بدون نظارت کشف می‌شوند. به همین علت مفاهیم کشف شده برچسب ندارند

و برای هدایت خوشه‌بندی به نتیجه‌ی بهتر، از الگوریتمی نیمه‌نظارتی مشابه با روش اول، در خوشه‌بندی بردارهای اسناد استفاده شده است. تفاوت در ورودی الگوریتم نیمه‌نظارتی است که در روش اول بردارهای کلمات وارد می‌شوند تا مفاهیم را بسازند، اما در روش دوم بردارهای سند به الگوریتم وارد شده تا خوشه‌های سند را ایجاد کنند.

برای بررسی عملکرد سیستم، آزمایشاتی بر روی دو پایگاه داده رایج در زمینه آنالیز متن انجام گرفته است. از پایگاه داده‌ی Reuters-21578 تعداد ۲۱۲۳ سند از چهار کلاس مختلف و از پایگاه داده‌ی 20Newsgroup تعداد ۲۳۳۹ سند در چهار کلاس متفاوت مورد آزمایش قرار گرفتند. همینطور از سه معیار $NMSE$، NMI و دقت برای ارزیابی روش پیشنهادی بهره برده شده است. تنها به ارزیابی روش‌ها در کاربرد خوشه‌بندی بسنده نشده و برای نمایش قدرت روش‌های پیشنهادی در کاربرد طبقه‌بندی اسناد نیز آزمایشات صورت گرفته است. گروهی از آزمایشات برای ارزیابی عملکرد روش‌های پیشنهادی انجام شده است. همانطور که در آزمایشات نشان داده شده است، حتی با تعداد کمی داده‌ی برچسب‌دار، روش ارائه شده نسبت به سایر روش‌ها برتری خود را نشان می‌دهد. نتایج تجربی نشان داد که روش پیشنهادی بهتر از سایر روش‌های بدون نظارت و نیمه‌نظارتی است. در آزمایشات مربوط به کیفیت خوشه‌بندی، روش‌های ارائه شده مقدار $NMSE$ کمتر و NMI بیشتری نسبت به سایر روش‌ها از خود نشان دادند، که این امر یعنی کیفیت بالاتر خوشه‌های حاصله و فاصله‌ی درون خوشه‌ای کمتر. همینطور با توجه به نمودارهای بخش دقت طبقه‌بندی مشاهده شد که در اکثر شرایط روش‌های پیشنهادی ۵ درصد بهتر از دیگر مدل‌ها دسته‌بندی اسناد را انجام می‌دهند. با افزایش تعداد اسناد برچسب‌دار در آموزش مدل، دقت و کیفیت بهتری در عملکرد مدل پیشنهادی در شرایط آزمون مشاهده شده است.

۵-۲ پیشنهادات برای کارهای آینده

اگرچه نتایج آزمایشات در مقایسه با روش‌های پیشین رضایت‌بخش و امیدوارکننده است، اما می‌توان مطالعات بیشتری را برای این تحقیق انجام داد. جهت‌گیری‌های پژوهش در کارهای آینده به شرح ذیل می‌باشد:

- فازی‌سازی در مفاهیم تولید شده ممکن است مدل واقعی‌تری از مولفه‌های متن ارائه دهد. بهره بردن از مدل فازی می‌تواند مفاهیمی خالص‌تر و متمایزکننده‌تری را برای نمایش اسناد استخراج کند.
- اعمال پیش‌پردازش‌های بیشتر مانند استفاده از N-Grams بجای توکن در استخراج کلماتِ اسناد، ممکن است نتایج بهتری را ایجاد کند.
- بررسی پارامتریک مدل Word2Vec و بهینه‌سازی مقادیر سایز پنجره، تعداد تکرارها، طول بردارها، مرتب‌سازی کلمات، ورود دسته‌ای کلمات و غیره باعث بهبود موثری در بردار کلمات تولیدشده خواهد داشت.
- در بخش خلاصه‌سازی کلمات، مقدار ثابتی برای تعداد کلمه خروجی در نظر گرفته شده است، که نیاز به یافتن مقدار بهینه برای هر کاربرد و پایگاه داده مخصوص به خود دارد.
- با اینکه بردارهای سند طول منطقی دارند اما باز هم در برخی شرایط تنکی در آن‌ها دیده می‌شود که نیاز به بررسی بیشتر برای حذف مفاهیم غیرضروری و حذف صفرهای بردار سند دیده می‌شود.
- ایجاد ساختار درختی برای مفاهیم و انتخاب مهمترین مفاهیم که بتوان بهترین و متمایزکننده-ترین بردار سند را برای متون موجود در پایگاه داده ایجاد کرد.

- اعمال روش‌های پیشنهادی بر روی پایگاه داده‌های متون کوتاه و جریان داده‌های شبکه‌های اجتماعی به منظور خوشه‌بندی و طبقه‌بندی آن‌ها با حجم کمی داده‌ی برچسب‌دار که نیاز اساسی این حوزه‌ها می‌باشد.
- به دلیل آنکه خوشه‌بندی پیش‌نیاز اکثر کاربردهای داده‌کاوی و یادگیری ماشین است، از این رو استفاده از روش‌های پیشنهادی در تمامی کاربردهای دیگر از جمله سیستم‌های توصیه‌گر ممکن است سودمند واقع شود.

۵-۳ منابع

[1] R. Forsati, M. Mahdavi, M. Kangavari, B. Safarkhani, Web page clustering using harmony search optimization, in: Can. Conf. Electr. Comput. Eng., 2008: pp. 1601–1604. https://doi.org/10.1109/CCECE.2008.4564812.

[2] M. Steinbach, G. Karypis, V. Kumar, A Comparison of Document Clustering Techniques, Retrieved from the University of Minnesota Digital Conservancy, 2000.

[3] W. Zhang, T. Yoshida, X. Tang, Q. Wang, Text clustering using frequent itemsets, Knowledge-Based Syst. 23 (2010) 379–388. https://doi.org/10.1016/j.knosys.2010.01.011.

[4] Y. Xiao, B. Liu, J. Yin, Z. Hao, A multiple-instance stream learning framework for adaptive document categorization, Knowledge-Based Syst. 120 (2017) 198–210. https://doi.org/10.1016/j.knosys.2017.01.001.

[5] D.C. Edara, L.P. Vanukuri, V. Sistla, V.K.K. Kolli, Sentiment analysis and text categorization of cancer medical records with LSTM, Journal of Ambient Intelligence and Humanized Computing (2019) 1–17. https://doi.org/10.1007/s12652-019-01399-8.

[6] S.Z. Sajadi, M.A.Z. Chahooki, S. Gharaghani, K. Abbasi, AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders, BMC Bioinforma. 2021 221. 22 (2021) 1–19. https://doi.org/10.1186/S12859-021-04127-2.

[7] T.A. Almeida, T.P. Silva, I. Santos, J.M. Gómez Hidalgo, Text normalization and semantic indexing to enhance Instant Messaging and SMS spam filtering, Knowledge-Based Syst. 108 (2016) 25–32. https://doi.org/10.1016/j.knosys.2016.05.001.

[8] Y. Djenouri, A. Belhadi, P. Fournier-Viger, J.C.W. Lin, Fast and effective cluster-based information retrieval using frequent closed itemsets, Inf. Sci. (Ny). 453 (2018) 154–167. https://doi.org/10.1016/j.ins.2018.04.008.

[9] S. Joty, G. Carenini, R.T. Ng, Topic segmentation and labeling in asynchronous conversations, J. Artif. Intell. Res. 47 (2013) 521–573. https://doi.org/10.1613/jair.3940.

[10] J. Paparrizos, L. Gravano, Fast and accurate time-series clustering, ACM Trans. Database Syst. 42 (2017) 1–49. https://doi.org/10.1145/3044711.

[11] Y. Li, H. Guo, Q. Zhang, M. Gu, J. Yang, Imbalanced text sentiment

classification using universal and domain-specific knowledge, Knowledge-Based Syst. 160 (2018) 1–15. https://doi.org/10.1016/j.knosys.2018.06.019.

[12] M. Mohd, R. Jan, M. Shah, Text document summarization using word embedding, Expert Syst. Appl. 143 (2020) 112958. https://doi.org/10.1016/j.eswa.2019.112958.

[13] W. Zhang, T. Yoshida, X. Tang, A comparative study of TF*IDF, LSI and multi-words for text classification, Expert Syst. Appl. 38 (2011) 2758–2765. https://doi.org/10.1016/j.eswa.2010.08.066.

[14] S. Fouzia Sayeedunnissa, A.R. Hussain, M.A. Hameed, Supervised Opinion Mining of Social Network Data Using a Bag-of-Words Approach on the Cloud, Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), in: J.C. Bansal, P. Singh, K. Deep, M. Pant, A. Nagar (Eds.), Springer India, India, 2013: pp. 299–309.

[15] A. Jacovi, O.S. Shalom, Y. Goldberg, Understanding Convolutional Neural Networks for Text Classification, arXiv preprint arXiv:1809.08037; 2018.

[16] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: 52nd Annu. Meet. Assoc. Comput. Linguist. ACL 2014 - Proc. Conf., Association for Computational Linguistics (ACL), 2014: pp. 655–665. https://doi.org/10.3115/v1/p14-1062.

[17] Q. Le, T. Mikolov, Distributed Representations of Sentences and Documents, Proceedings of the 31st International Conference on Machine Learning, (2014) 1188–1196.

[18] H.K. Kim, H. Kim, S. Cho, Bag-of-concepts: Comprehending document representation through clustering words in distributed representation, Neurocomputing. 266 (2017) 336–352. https://doi.org/10.1016/j.neucom.2017.05.046.

[19] F. Gagliardi Cozman FGCOZMAN, M. Cesar Cirelo, Semi-Supervised Learning of Mixture Models, ICML, (2003).

[20] X. Luo, F. Liu, S. Yang, X. Wang, Z. Zhou, Joint sparse regularization based Sparse Semi-Supervised Extreme Learning Machine (S3ELM) for classification, Knowledge-Based Syst. 73 (2015) 149–160. https://doi.org/10.1016/j.knosys.2014.09.014.

[21] W. Zhang, Y. Yang, Q. Wang, Using Bayesian regression and EM algorithm with missing handling for software effort prediction, Inf. Softw. Technol. 58 (2015) 58–70. https://doi.org/10.1016/j.infsof.2014.10.005.

[22] R. Dara, S.C. Kremer, D.A. Stacey, Clustering unlabeled data with SOMs improves classification of labeled real-world data, in: Proc. Int. Jt. Conf. Neural Networks, 2002: pp. 2237–2242. https://doi.org/10.1109/ijcnn.2002.1007489.

[23] W. Zhang, X. Tang, T. Yoshida, TESC: An approach to TExt classification using Semi-supervised Clustering, Knowledge-Based Syst. 75 (2015) 152–160.

https://doi.org/10.1016/j.knosys.2014.11.028.

[24] W.H.I.T.E.P.A. Per, Big Data : Beyond the Hype Why Big Data Matters to You, By Datastax Corp. (2013) 1–16. http://www.datastax.com/wp-content/uploads/2011/10/WP-DataStax-BigData.pdf.

[25] M.A. Syakur, B.K. Khotimah, E.M.S. Rochman, B.D. Satoto, Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster, IOP Conf. Ser. Mater. Sci. Eng. 336 (2018) 012017. https://doi.org/10.1088/1757-899X/336/1/012017.

[26] A. Alrabea, A. V. Senthilkumar, H. Al-Shalabi, A. Bader, Enhancing K-Means Algorithm with Initial Cluster Centers Derived from Data Partitioning along the Data Axis with PCA, J. Adv. Comput. Networks. 2 (2013) 137–142. https://doi.org/10.7763/jacn.2013.v1.28.

[27] W.H. Gomaa, A.A. Fahmy, A Survey of Text Similarity Approaches, Int. J. Comput. Appl. 68 (2013) 975–8887.

[28] P. Rodríguez, M.A. Bautista, J. Gonzàlez, S. Escalera, Beyond one-hot encoding: Lower dimensional target embedding, Image Vis. Comput. 75 (2018) 21–31. https://doi.org/10.1016/J.IMAVIS.2018.04.004.

[29] S. Fouzia Sayeedunnissa, A.R. Hussain, M.A. Hameed, Supervised opinion mining of social network data using a bag-of-words approach on the cloud, in: Adv. Intell. Syst. Comput., Springer Verlag, 2013: pp. 299–309. https://doi.org/10.1007/978-81-322-1041-2_26.

[30] C. Manning, H. Schutze, Foundations of statistical natural language processing, MIT press; 1999 May 28.

[31] R. Baeza-Yates, B. Ribeiro-Neto, Modern information retrieval, New York: ACM press; 1999 May 15.

[32] L. Wu, S.C.H. Hoi, N. Yu, Semantics-preserving bag-of-words models and applications, IEEE Trans. Image Process. 19 (2010) 1908–1920. https://doi.org/10.1109/TIP.2010.2045169.

[33] P. Kolari, A. Java, T. Finin, T. Oates, A. Joshi, Detecting spam blogs: A machine learning approach, Proceedings of the national conference on artificial intelligence (Vol. 21, No. 2, p. 1351); MIT Press; 1999.

[34] A.Huang, Similarity measures for text document clustering, Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008) (Vol. 4, pp. 9-56); 2008.

[35] T. Hofmann, Probabilistic latent semantic indexing, Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retrieval, SIGIR 1999. (1999) 50–57. https://doi.org/10.1145/312624.312649.

[36] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-

ASI1>3.0.CO;2-9.

[37] DM. Blei, AY. Ng, MI. Jordan, Latent Dirichlet Allocation, The Journal of machine Learning research; 2003.

[38] S. Lai, K. Liu, S. He, J. Zhao, How to generate a good word embedding, IEEE Intell. Syst. 31 (2016) 5–14. https://doi.org/10.1109/MIS.2016.45.

[39] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: 1st Int. Conf. Learn. Represent. ICLR 2013 - Work. Track Proc., International Conference on Learning Representations, ICLR, 2013.

[40] D. Guthrie, B. Allison, W. Liu, L. Guthrie, Y. Wilks, A Closer Look at Skip-gram Modelling, InLREC (Vol. 6, pp. 1222-1225); 2006.

[41] C. Xing, D. Wang, X. Zhang, C. Liu, Document classification with distributions of word vectors, in: 2014 Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. APSIPA 2014, Institute of Electrical and Electronics Engineers Inc., 2014. https://doi.org/10.1109/APSIPA.2014.7041633.

[42] P.D. Turney, P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, J. Artif. Intell. Res. 37 (2010) 141–188. https://doi.org/10.1613/JAIR.2934.

[43] L. Cao, F. Wang, Robust latent semantic exploration for image retrieval in social media, Neurocomputing. 169 (2015) 180–184. https://doi.org/10.1016/J.NEUCOM.2015.02.082.

[44] Z. Cui, X. Shi, Y. Chen, Sentiment analysis via integrating distributed representations of variable-length word sequence, Neurocomputing. 187 (2016) 126–132. https://doi.org/10.1016/J.NEUCOM.2015.07.129.

[45] B. Xue, C. Fu, Z. Shaobin, A study on sentiment computing and classification of sina weibo with Word2vec, Institute of Electrical and Electronics Engineers Inc., 2014.

[46] L. Van der Maaten, G.Hinton, Visualizing data using t-SNE, Journal of machine learning research; 2008.

[47] A.M. Dai, C. Olah, Q. V. Le, Document Embedding with Paragraph Vectors, arXiv preprint arXiv:1507.07998; 2015.

[48] P. Li, K. Mao, Y. Xu, Q. Li, J. Zhang, Bag-of-Concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base, Knowledge-Based Syst. 193 (2020) 105436. https://doi.org/10.1016/j.knosys.2019.105436.

[49] X. Luo, S. Shah, Concept embedding-based weighting scheme for biomedical text clustering and visualization, Appl. Informatics. 5 (2018) 1–19. https://doi.org/10.1186/s40535-018-0055-8.

[50] C. Jia, M.B. Carson, X. Wang, J. Yu, Concept decompositions for short text clustering by identifying word communities, Pattern Recognit. 76 (2018) 691–

703. https://doi.org/10.1016/j.patcog.2017.09.045.

[51] S. Basu, I. Davidson, K. Wagstaff, Constrained clustering: Advances in algorithms, theory, and applications, CRC Press; 2008.

[52] S. Basu, M. Bilenko, R.J. Mooney, Comparing and Unifying Search-Based and Similarity-Based Approaches to Semi-Supervised Clustering, Proceedings of the ICML; 2003.

[53] M. Lu, X.J. Zhao, L. Zhang, F.Z. Li, Semi-supervised concept factorization for document clustering, Inf. Sci. (Ny). 331 (2016) 86–98. https://doi.org/10.1016/j.ins.2015.10.038.

[54] I. Diaz-Valenzuela, V. Loia, M.J. Martin-Bautista, S. Senatore, M.A. Vila, Automatic constraints generation for semisupervised clustering: experiences with documents classification, Soft Comput. 20 (2016) 2329–2339. https://doi.org/10.1007/s00500-015-1643-3.

[55] P. Li, Z. Deng, Use of distributed semi-supervised clustering for text classification, J. Circuits, Syst. Comput. 28 (2019) 1–13. https://doi.org/10.1142/S0218126619501275.

[56] M.A. Masud, J.Z. Huang, M. Zhong, X. Fu, Generate pairwise constraints from unlabeled data for semi-supervised clustering, Data Knowl. Eng. 123 (2019) 101715. https://doi.org/10.1016/j.datak.2019.101715.

[57] H. Gan, Y. Fan, Z. Luo, Q. Zhang, Local homogeneous consistent safe semi-supervised clustering, Expert Syst. Appl. 97 (2018) 384–393. https://doi.org/10.1016/j.eswa.2017.12.046.

[58] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed Representations of Words and Phrases and their Compositionality, Adv. Neural Inf. Process. Syst; 2013.

[59] C. Buchta, M. Kober, I. Feinerer, K. Hornik, Spherical k-Means Clustering, Journal of statistical software 50, no. 10; (2012).

[60] S. Robertson, Understanding inverse document frequency: On theoretical arguments for IDF, Journal of documentation. 60; (2004) 503–520. https://doi.org/10.1108/00220410410560582.

[61] A. Strehl, J. Ghosh, R. Mooney, Impact of Similarity Measures on Web-page Clustering, Workshop on artificial intelligence for web search (AAAI 2000); 2000.

واژه نامه

Data Mining	داده کاوی
Topic Modeling	مدل سازی موضوعی
Text Representation	بازنمایی متن، نمایش متن
Spell Correction	تصحیح املا
Stopwords	کلمات توقف، کلمات بی اثر
Text Canonicalization	متعارف کردن متن
Tokenization	نشانه گذاری
Stemming	ریشه یابی
Part Of Speech	نقش دستوری
Chunking	تکه تکه کردن
Shallow Parsing	تجزیه جزیی
Named Entity Recognition	شناسایی موجودیت های اسم دار
Token	نشانه
Inflectional	عطفی
Morphological	ریخت شناسی
Lexical Knowledge Base	پایگاه دانش لغوی
Numerical Feature Vectors	بردارهای ویژگی عددی
Semantic Quality	کیفیت معنایی

Statistical Quality	کیفیت آماری
Bag-of-Words	کیسه کلمات
Concept	مفهوم
Mixture Model	مدل مخلوط
Components	اجزا
Sparsity	تُنُکی
Term Frequency	رخداد کلمه
Normal Mutual Information	اطلاعات متقابل عادی
Mean Squared Error	خطای میانگین مربعات
Parse	تجزیه
Word Embedding	تعبیه کلمات
Distributed Hypothesis	فرضیه‌ی توزیع
Node	گره
Semantic Word Communities	انجمن های لغت معنایی
Pairwise Constraints	محدودیت های جفت
Retaining Ratio	نسبت نگهداری

جدول اختصارات

BoW	Bag of Words	کیسه کلمات
CNN	Convolutional Neural Networks	شبکه عصبی کانوولوشن
RNN	Recurrent Neural Networks	شبکه عصبی بازگشتی
TF-IDF	Term Frequency and Inverse Document Frequency	رخداد کلمه-معکوس رخداد سند
NMI	Normal Mutual Information	اطلاعات متقابل عادی
NMSE	Normal Mean squared error	میانگین مربعات خطای عادی شده
BoC	Bag of Concepts	کیسه مفاهیم
SSConE	Semi-Supervised Concept Extraction	استخراج نیمه نظارتی مفهوم
SSClusE	Semi-Supervised Cluster Extraction	استخراج نیمه نظارتی خوشه

Abstract: Text clustering is used in various applications of text analysis. In the clustering process, the method of document representation has a significant impact on the results. Some popular document representation methods based on the Bag-of-Words (BoW), depend on the word frequencies and can produce large and sparse document vectors. In addition, embedding-based methods such as Doc2Vec, which preserve the proximity information, suffer from low interpretability. These challenges have been largely addressed by the introduction of concept-based methods. The existing semi-supervised document clustering methods do not use the more recent concept-based representation of documents. Therefore, this paper proposes a concept-based semi-supervised document clustering approach that uses both labeled and unlabeled data to yield a higher clustering quality. The documents are represented based on the concepts extracted from the set of embedded words in the corpus. This representation preserves the proximity information of documents and improves interpretability. The semi-supervised clustering process uses unlabeled data to capture the overall structure of the clusters and a small number of labeled data to adjust the centroid of clusters. We also propose the notion of semi-supervised concepts and a new method of clustering documents based on the weights of the concepts. The clustering results of this method are compared with the actual semi-supervised clustering of documents. Experiments on two sets of text data, Reuters-21578 and 20-NewsGroup, demonstrate that the proposed method performs at least 10% better in terms of clustering quality and at least 5% in text classification accuracy than several existing semi-supervised and unsupervised approaches.

keywords: Machine learning, Semi-supervised, Concept extraction, Word embedding, Document clustering, Concept-based representation

Shahrood University of Technology

Faculty of Computer Engineering and Information Technology

M.Sc. Thesis in Artificial Intelligence Engineering

Semi-supervised text clustering using word embeddings

By:

Seyed Mojtaba Sadjadi

Supervisor:

Dr. Hoda Mashayekhi

Advisor:

Dr. Hamid Hassanpour

June 2021