سال پانزدهم، شماره‌های ۵۵ و ۵۶، بهار و تابستان
۱۴۰۲
صفحات:۱ تا۱۴

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

Improving Opinion Aspect Extraction Using Domain Knowledge and Term Graph

MohammadReza Shams*, Ahmad Baraani**, Mehdi Hashemi ***

* Assistant Professor, Department of Computer Engineering, Shahreza Higher Education Center, University of Isfahan, Iran
** Professor, Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Iran
*** Master's degree, Department of Software Engineering, Faculty of Computer Engineering, University of Isfahan, Iran

Abstract

With the advancement of technology, analyzing and assessing user opinions, as well as determining the user's attitude toward various aspects, have become a challenging and crucial issue. Opinion mining is the process of recognizing people's attitudes from textual comments at three different levels: document-level, sentence-level, and aspect-level. Aspect-based Opinion mining analyzes people's viewpoints on various aspects of a subject. The most important subtask of aspect-based opinion mining is aspect extraction, which is addressed in this paper. Most previous methods suggest a solution that requires labeled data or extensive language resources to extract aspects from the corpus, which can be time consuming and costly to prepare.

In this paper, we propose an unsupervised approach for aspect extraction that uses topic modeling and the Word2vec technique to integrate semantic information and domain knowledge based on term graph. The evaluation results show that the proposed method not only outperforms previous methods in terms of aspect extraction accuracy, but also automates all steps and thus eliminates the need for user intervention. Furthermore, because it is not reliant on language resources, it can be used in a wide range of languages.

Keywords: Text mining, Opinion mining, Word2Vec, Aspect Extraction, Domain Knowledge, Term Graph

بهبود استخراج جنبه‌های متن با استفاده از دانش دامنه و گراف کلمات

محمدرضا شمس*×، احمد براآنی**، مهدی هاشمی ***

* استادیار گروه مهندسی کامپیوتر، مرکز آموزش عالی شهرضا، دانشگاه اصفهان، ایران

** استاد گروه مهندسی نرم افزار، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، ایران

*** کارشناسی ارشد، گروه مهندسی نرم افزار، دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، ایران

تاریخ دریافت: ۱۴۰۰/۱۰/۰۳ تاریخ پذیرش: ۱۴۰۱/۰۵/۳۱

نوع مقاله: پژوهشی

چکیده

با گسترش روزافزون علم و فناوری، تحلیل نظرات کاربران و تعیین نحوه نگرش کاربر به موضوع‌های مختلف به یک امر مهم تبدیل شده است. نظرکاوی فرایند استخراج نگرش افراد از روی نظرات نوشته شده است که در سه سطح سند، جمله و جنبه قابل انجام است. در سطح جنبه، نظر افراد در خصوص جنبه‌های مختلف یک موضوع بررسی می‌شود. مهم‌ترین زیر بخش نظرکاوی جنبه‌گرا، استخراج جنبه است که موضوع اصلی این پژوهش می‌باشد. در بسیاری از روش‌های ارائه شده برای استخراج جنبه، راه حل مورد نظر نیاز به مجموعه یادگیری اولیه و یا منابع زبانی وسیع دارند که تهیه چنین داده‌هایی بسیار زمان‌بر و پرهزینه است.

در این مقاله، رویکردی بدون نظارت برای استخراج جنبه مبتنی بر مدل موضوعی و بردار کلمات پیشنهاد می‌شود که از ایجاد گراف کلمات برای ادغام اطلاعات معنایی و دانش دامنه استفاده می‌کند. نتایج ارزیابی‌ها نشان از این دارد که روش پیشنهادی نه تنها باعث بهبود دقت استخراج جنبه در مقایسه با سایر روش‌های پیشین شده است، بلکه تمامی مراحل به صورت خودکار و بدون دخالت کاربر انجام می‌شود و بدلیل عدم وابستگی به منابع زبانی، در زبان‌های مختلف قابل اجرا می‌باشد.

واژگان کلیدی: متن‌کاوی، نظرکاوی، بردار کلمات، استخراج جنبه، دانش دامنه، گراف کلمات

× نویسنده مسئول: محمدرضا شمس m.r.shams@shr.ui.ac.ir

۱. مقدمه

امروزه با توجه به فراگیرشدن شبکه‌های اجتماعی، ابزارهای گفتگو و چت، وبلاگ‌های شخصی، مقاله‌های علمی، سایت‌ها و فروشگاه‌های اینترنتی، متن‌های تولید شده توسط کاربران که شامل بیان نظرات و احساسات کاربر در مورد برخی از موضوعات روز است به طور قابل توجهی افزایش یافته است. همین امر سبب شده است تا تحلیل و بررسی انسانی این حجم زیاد از اطلاعات غیرممکن باشد. بنابراین استخراج خودکار نظر، از متن‌های برخط با هدف کشف موضوعات مربوط به نظر کاربر، به یک امر ضروری تبدیل شده است [۱]. درواقع با استفاده از نتایج این کاوش می‌توان به این نتیجه رسید که کاربران مختلف در مورد یک موضوع چگونه فکر می‌کنند، چه پیشنهادی در مورد آن موضوع دارند و نقاط قوت و ضعف آن را در چه چیزی می‌بینند.

نظرکاوی تعیین نگرش نویسنده نسبت به یک موضوع است که با استفاده از تکنیک‌های پردازش زبان طبیعی[۱] و یادگیری ماشین[۲] انجام می‌شود [۲].

نظرکاوی به سه سطح اصلی به‌نام‌های سطح سند، سطح جمله و سطح جنبه تقسیم می‌شود. هدف اصلی تحلیل احساسات در سطح جمله و سند، بررسی کلی و پیدا کردن احساس عمومی جمله یا سند است، درحالی‌که هدف آن در سطح جنبه، پیدا کردن احساس کاربر در مورد هر جنبه است [۳].

نظرکاوی جنبه‌گرا از سه بخش اصلی به نام‌های استخراج جنبه[۳]، شناسایی کلمه یا نظر قطبی برای هر جنبه استخراج شده و در نهایت دسته‌بندی و خلاصه‌سازی جنبه‌ها تشکیل شده است [۴]. در بخش استخراج جنبه، تمام جنبه‌های هر نظر استخراج می‌شوند که هدف اصلی این پژوهش می‌باشد. به‌طور مثال در جمله «دیروز این لپتاپ رو خریدم. صفحه‌نمایش خوبی دارد ولی کیبوردش خوب نیست.» کلمه‌ی «لپتاپ» موضوع نظر و دو کلمه‌ی «صفحه‌نمایش» و «کیبورد» به عنوان جنبه‌های این موضوع شناخته می‌شوند.

روش‌های استخراج جنبه به‌طور کلی به سه نوع دسته‌بندی به نام‌های دسته‌بندی نظارتی، بدون‌نظارت و نیمه‌نظارتی تقسیم می‌شوند. دسته‌بندی بدون‌نظارت از هیچ نوع داده برچسب گذاری‌شده‌ای برای استخراج جنبه استفاده نمی‌کند. دسته‌بندی بانظارت از داده‌های برچسب گذاری‌شده برای استخراج جنبه‌ها استفاده می‌کند و در نهایت دسته‌بندی نیمه‌نظارتی از هر دو نوع داده برچسب گذاری‌شده و بدون برچسب گذاری برای استخراج جنبه استفاده می‌کند ولی معمولاً تعداد داده‌های برچسب خورده مورد استفاده کمتر است.

در روش‌های بانظارت، تهیه داده‌های برچسب خورده اولیه مناسب برای هر دامنه و زبان بسیار سخت، زمان‌بر و پرهزینه است که همین امر سبب می‌شود تا استفاده از این دسته از روش‌ها محدود شود. بر همین اساس بیشتر پژوهش‌ها از روش بدون‌نظارت و یا نیمه‌نظارتی استفاده کرده‌اند [۵].

مدل موضوعی[۴] یک روش بدون نظارت برای طبقه‌بندی اسناد است. در روش‌های مبتنی بر مدل موضوعی، اسناد ترکیبی از موضوع‌ها می‌باشند، درحالی‌که یک موضوع توزیعی احتمالی بر روی تمام لغات است. مدل موضوعی، مدلی تولیدی بر روی اسناد بوده که فرایند احتمالاتی ساده‌ای را جهت تولید اسناد مشخص می‌کند. سپس برای هر لغت درون سند، موضوعی را به تصادف براساس این توزیع انتخاب کرده و لغتی را از آن موضوع انتخاب می‌کند. به عبارتی مدل موضوعی خوشه‌هایی از کلمات ایجاد می‌کند که هر خوشه بیانگر یک موضوع می‌باشد. مدل موضوعی هیچ‌گونه فرضی درباره‌ی ترتیب ظاهرشده لغات در سند نداشته و ترتیب کلمات را کاملا نادیده می‌گیرد. تنها اطلاعات موثر در مدل، تعداد دفعات تولید لغات در مدل می‌باشد. این فرض در مدل موضوعی به عنوان کیسه لغات[۵] شناخته می‌شود و به مدل‌هایی که بر پایه این فرض ساخته می‌شوند، مدل‌های کیسه‌ای می‌گویند [۶] و [۷].

تخصیص پنهان دیریکله[۶] یا LDA یک شبکه بیز و مدل تولیدی احتمالاتی و یکی از محبوب‌ترین روش‌های مدل موضوعی است. هر سند از موضوع‌های مختلف تشکیل شده است و هر موضوع نیز دارای کلمات مختلفی است که به آن تعلق دارد. هدف LDA یافتن موضوعاتی است که یک سند براساس کلمات موجود در آن، به آن‌ها تعلق دارد. ضعف اصلی روش‌های مبتنی بر LDA این است که به مجموعه داده بزرگی نیاز دارد تا نتایج قابل قبولی ارائه دهند و در صورتی که مجموعه داده کوچک باشد دقت این الگوریتم‌ها به طور محسوسی کاهش پیدا می‌کنند. همچنین این دسته از الگوریتم‌ها، عموماً موضوع‌ها و جنبه‌های متناقضی را شناسایی می‌کنند که غیرقابل قبول است.

دسته دیگر، روش‌های مبتنی بر الگوریتم بردار کلمات[۷] هستند. در این دسته به کمک شبکه عصبی، برای نمایش تمام کلمات، یک بردار با اندازه ثابت و کوچک در نظر گرفته می‌شود و در فاز آموزش مدل، با اعداد مناسب برای هر کلمه این بردار محاسبه و ایجاد می‌شود. به

---

[1] Natural Language Processing

[2] Machine Learning

[3] Aspect Extraction

[4] Topic model

[5] Bag of Words

[6] Latent Dirichlet allocation

[7] Word2vec

طور خلاصه، این الگوریتم از یکی از دو روش کیف لغات پیوسته[1] و یا Skip-gram برای ساخت بردارهای کلمات استفاده می‌کند. هر دو روش یک شبکه عصبی ساده هستند که بدون وجود لایه پنهانی زیاد که در اغلب روش‌های شبکه عصبی وجود دارد، بردارهای مورد نیاز را به کمک چند قانون ساده ایجاد می‌کنند. در روش کیف لغات پیوسته(CBOW)، ابتدا به ازای هر کلمه یک بردار با طول مشخص و با اعداد تصادفی ایجاد می‌شود. سپس به ازای هر کلمه از یک سند یا متن، تعدادی مشخص از کلمات بعد و قبل آن به شبکه عصبی داده می‌شود (به غیر از خود کلمه فعلی) و سپس با استفاده از شبکه عصبی، بردار کلمه فعلی تولید می‌شود (یا به عبارتی از روی کلمات قبل و بعد یک لغت، آن لغت حدس زده می‌شود) و بردار جدید با مقادیر قبلی بردار لغت جایگزین می‌شوند. زمانی که این کار بر روی تمام لغات در تمام متون انجام گیرد، بردارهای نهایی لغات، همان بردارهای مطلوب ما هستند. روش Skip-gram برعکس این روش کار می‌کند. به این صورت که بر اساس یک کلمه داده شده، تلاش می‌کند تا چند کلمه قبل و بعد آن را تشخیص دهد و با تغییر مداوم اعداد بردارهای کلمات، نهایتاً به یک وضعیت باثبات می‌رسد که همان بردارهای نهایی می‌باشند.

مشکل استفاده از بردار کلمات به تنهایی این است که هر کلمه‌ای که نیاز باشد برای آن بردار ایجاد شود باید به صورت دستی انتخاب شود که اینکار باعث می‌شود هزینه اجرا برای تعداد زیاد کلمات به شدت بالا برود. همچنین بردار کلمات صرفا کلمات مشابه را دسته‌بندی می‌کند که نتیجه اینکار از هدف این پژوهش که استخراج جنبه است بسیار متفاوت است.

در نتیجه در این مقاله روشی مبتنی بر ترکیب LDA و بردار کلمات ارائه می‌شود که به آن دانش دامنه گفته می شود و به منظور استخراج جنبه‌ها، گرافی از کلمات به صورت $G = (V, E)$ تشکیل می‌شود که یک گراف بدون جهت است و رأس‌های آن (V) نمایانگر یک کلمه است و هر یال آن (E) نمایانگر رابطه بین دو گره است. هر یال و رابطه بین دو گره با استفاده از معیار شباهت بین کلمات ایجاد شده و درنهایت خوشه‌بندی می‌شود تا جنبه‌ها استخراج شوند.

از مزایای این روش پیشنهادی در مقایسه با مطالعات پیشین، می‌توان به این اشاره کرد که دانش مورد نیاز برای خوشه‌بندی از ترکیب مدل‌های موضوعی و بردار کلمات ایجاد می‌گردد تا از نقاط قوت هر دو روش در کنار هم استفاده شود. در ضمن تمامی مراحل به صورت کاملاً خودکار و بدون دخالت کاربر انجام می‌شود و دانش مورد استفاده نیز به صورت خودکار به مدل اضافه می‌شود. همچنین این روش نیاز به داده برچسب خورده‌ی اولیه ندارد و کاملا به صورت بدون نظارت پیاده‌سازی شده است، به همین جهت هزینه اولیه آن بسیار کم است. علاوه بر این، در این روش از هیچ منبع زبانی استفاده نمی‌شود و به همین دلیل به راحتی در زبان‌های مختلف قابل اجرا می‌باشد.

مهم‌ترین نوآوری‌های روش پیشنهادی به صورت زیر قابل دسته‌بندی است:

- استفاده از بردار کلمات به منظور غنی‌سازی نتایج الگوریتم LDA
- استفاده از شباهت بین کلمات به عنوان دانش دامنه
- روش پیشنهادی کاملا مستقل از زبان بوده و بر روی هر زبانی قابل اجراست.
- در روش پیشنهادی تمامی مراحل کاملا خودکار بوده و بدون کمک کاربر انجام می‌شوند.

در ادامه این مقاله، در بخش ۲ مطالعات پیشین انجام شده در حوزه استخراج جنبه در نظرکاوی جنبه‌گرا در سه زیر بخش روش‌های مبتنی بر لغت‌نامه و هستان‌شناسی، مدل‌های موضوعی و مبتنی بر گراف مورد بررسی قرار گرفته است. در بخش ۳ گام‌های روش پیشنهادی به صورت مفصل شرح داده شده است. در بخش ۴ نتایج ارزیابی روش پیشنهادی شرح داده شده است و در نهایت در بخش پایانی جمع‌بندی و نتیجه‌گیری مقاله آمده است.

۲. مطالعات پیشین

همانطور که در فصل قبل اشاره شد، روش‌های استخراج جنبه به سه نوع دسته‌بندی به نام‌های دسته‌بندی بانظارت، دسته‌بندی بدون نظارت و دسته‌بندی نیمه‌نظارتی تقسیم می‌شوند.

در روش‌های بانظارت، الگوریتم‌هایی که مورد استفاده قرار می‌گیرند از مجموعه داده‌ی برچسب گذاری‌شده‌ای برای آموزش مدل استفاده می‌کنند. اکثر روش‌های بانظارت از الگوریتم‌های یادگیری ماشین استفاده می‌کنند. به‌طور کلی روش‌های بانظارت با توجه به اینکه به داده‌های برچسب‌گذاری شده احتیاج دارند بسیار پرهزینه هستند. مجموعه داده‌های برچسب خورده برای بسیاری از موضوعات در زبان‌های مختلف دردسترس نیست که همین امر سبب می‌شود تا دقت این روش‌ها برای مجموعه داده‌های مختلف مناسب نباشد. از طرف دیگر این روش‌ها بر روی هر دامنه اطلاعاتی نمی‌توانند اجرا شوند چرا که برچسب‌گذاری همه دامنه‌های موجود به صورت دستی و با کمک انسان، امری بسیار پرهزینه و غیرمعقول است. باید توجه شود که بعضی از کلمات ممکن است در دو دامنه مختلف معنای بسیار متفاوتی ارائه دهند. به طور مثال کلمه «غیرقابل پیش‌بینی»

[1] Continuous bag-of-words (CBOW)

در دامنه فیلم، مثبت تلقی می‌شود ولی در دامنه کارکرد اتومبیل، معنای منفی دارد.

روش‌های بدون نظارت، از داده‌های بدون برچسب برای استخراج جنبه استفاده می‌کنند و الگوریتم‌هایی که در این روش‌ها استفاده می‌شوند نیازی به آموزش اولیه ندارند. روش‌های نیمه‌نظارتی از هر دو داده برچسب‌خورده و بدون برچسب برای استخراج جنبه از متن استفاده می‌کنند. باتوجه به مشکلات و هزینه‌های زیادی که در روش‌های بانظارت وجود داشت، روش‌های بدون نظارت و نیمه‌نظارتی محبوبیت قابل توجهی بدست آوردند و بیشتر پژوهش‌های موجود در بخش استخراج جنبه به این دو روش اختصاص یافته‌اند. مطالعات حوزه بدون نظارت و نیمه نظارتی به سه روش کلی به نام‌های روش‌های مبتنی بر لغت‌نامه و هستان شناسی، روش‌های مبتنی بر مدل موضوعی و روش‌های مبتنی بر گراف تقسیم می‌شود.

۱،۲. روش‌های مبتنی بر لغت‌نامه و هستان‌شناسی

در این روش‌ها، از اطلاعات موجود در لغت‌نامه و یا هستان‌شناسی برای استخراج جنبه استفاده می‌شود. در پژوهش انجام شده در مقاله [۸] روشی با ترکیب الگوریتم احتمالاتی LDA با یک لغت‌نامه از کلمات مترادف، برای استخراج جنبه از مجموعه نظرات به زبان چینی معرفی شد. آن‌ها همچنین عبارت‌های اسمی را به عنوان جنبه فرض کرده و با ترکیب آن‌ها با خروجی الگوریتم LDA یک لیست تحت عنوان جنبه‌های کاندید ایجاد کردند و سپس آن لیست را با لغت‌نامه کلمات مترادف گسترش دادند.

در مقاله [۹] یک روش نیمه‌نظارتی برای استخراج جنبه معرفی شد. در این روش آن‌ها به صورت دستی یک لیست از جنبه‌ها تهیه کردند و با استفاده از این لیست و لغت‌نامه وردنت[1] جنبه‌های مشابه از نظرات مختلف استخراج شد.

هستان‌شناسی‌جنبه، مفاهیم موجود در یک دامنه و همچنین روابط بین آنها را به شکل یک ساختار درختی سلسله مراتبی نشان می‌دهد [۱۰]. در مقاله [۱۰] با استفاده از دانش معنایی و شباهت معنایی، درخت هستان‌شناسی جنبه ساخته شد. در این روش ابتدا عبارت‌هایی که نقش اسم در جمله دارند به عنوان کاندیدی برای جنبه استخراج می‌شود و سپس کلماتی که در نظرهای مختلف از یک حد آستانه مشخص بیشتر ظاهر شدند به عنوان جنبه شناسایی می‌شوند.

در مقاله [۱۱] یک روش با استفاده از روابط معنایی بین مفاهیم، ویژگی‌ها و افراد در هستان‌شناسی برای استخراج جنبه‌ها معرفی شد. در ابتدا نظراتی که هیچ جنبه صریحی در آن‌ها وجود نداشت را استخراج کردند. آن‌ها شش نوع رابطه وابستگی بین کلمات نظر و جنبه‌های مرتبط با آن درنظر گرفتند و سپس با استفاده از این روابط معنایی و نظرات استخراج شده در هستان‌شناسی، جنبه‌های نظرات استخراج شد. همچنین در [۱۲] از یک هستان‌شناسی در زبان عربی برای بهبود استخراج جنبه کمک گرفته شده است.

روش‌های مبتنی بر لغت‌نامه و هستان‌شناسی به راحتی قابل اجرا هستند و استفاده از آن‌ها باعث بهبود دقت در استخراج جنبه شده است. اما دقت این دسته از روش‌ها بسیار وابسته به لغت‌نامه و روابطی است که از قبل تعریف شده است. باتوجه به افزایش سریع متون آنلاین و گسترش داده‌ها، تهیه و به‌روزرسانی لغت‌نامه برای هر دامنه و زبان بسیار پرهزینه و زمان‌بر خواهد بود.

۲،۲. روش‌های مبتنی بر مدل موضوعی

روش‌های مبتنی بر مدل موضوعی به صورت گسترده برای استخراج و دسته‌بندی جنبه‌ها در نظرکاوی جنبه‌گرا مورد استفاده قرار گرفته است. روش‌های مدل موضوعی فرض می‌کنند که هر سند از ترکیبی از موضوعات یا جنبه‌های مختلف و هر موضوع از توزیعی از کلمات تشکیل شده است. به‌طور کلی در این دسته از روش‌ها، الگوریتم‌های احتمالاتی بر روی اسناد مختلف اعمال می‌شود و سپس خوشه‌هایی از کلمات ایجاد می‌شود که هر خوشه نمایانگر یک موضوع در سند است، به‌طوری‌که هر موضوع با توزیع احتمالی بر روی کلمات مشخص می‌شود.

تاکنون الگوریتم متفاوتی برای مدل‌سازی موضوعی ارائه شده است که از معروف‌ترین آن‌ها می‌توان به مدل آنالیز احتمالی معنایی مخفی[2] (pLSA) [۱۳]، تخصیص دیریکله پنهان (LDA) [۱۴] و روش ماشین بولتزمن محدود[3] (RBM) [۱۵] اشاره نمود. پژوهش‌های بسیاری با استفاده از این روش‌ها انجام شده است. همه این مدل‌ها یک متغیر پنهان (موضوع) را بین سند و کلمات برای تحلیل توزیع معنایی شناسایی می‌کنند.

در سال‌های اخیر، پژوهش‌های مختلفی با استفاده از مدل موضوعی مبتنی بر دانش دامنه انجام شده است. در این پژوهش‌ها از دانش‌های مختلف برای راهنمایی مدل موضوعی استفاده شده است تا موجب بهبود دقت استخراج جنبه شوند. در پژوهش‌های مختلف این دانش‌ها به دو صورت خودکار و نیمه‌خودکار استخراج و استفاده شده‌اند. بدیهی است که روش‌هایی که به صورت غیرخودکار از دانش دامنه استفاده می‌کنند نیازمند هزینه اولیه می‌باشند.

[1] WordNet

[2] Probabilistic latent semantic analysis

[3] Restricted Boltzmann machine

در مقاله [۱۶]، یک روش مبتنی بر مدل موضوعی با استفاده از الگوریتم تخصیص دیریکله پنهان و دو محدودیت «کلمات پیوند لازم» و «کلمات پیوند ناپذیر» ارائه شده است. در این روش از ترکیب محدودیت‌های ذکر شده با ارتباطات معنایی، دانش دامنه برای جنبه‌های صریح ایجاد شده، و از این دانش برای استخراج جنبه بهره گرفته شده است. محدودیت پیوند-لازم به این معنی است که دو کلمه باید در یک موضوع باشند ولی محدودیت پیوند ناپذیر به این معنی است که دو کلمه نمی‌توانند در یک موضوع مشابه باشند. در مقاله [۱۷] نیز یک روش مشابه این رویکرد برای داده‌های حجیم، با استفاده از مدل موضوعی و استخراج دانش به صورت خودکار معرفی شده است.در [۱۸] دانش مورد نیاز برای بهبود الگوریتم تخصیص دیریکله پنهان از طریق شکستن جملات به بخش‌های مختلف به ترتیبی که هر بخش در خصوص یک جنبه باشد استخراج شده است و در [۱۹] این دانش با استفاده از یادگیری عمیق و استفاده از مدل-های زبانی از پیش آموزش دیده شده ایجاد گردیده است.

در مقاله [۲۰] از یک مدل احتمالی برای استخراج جنبه استفاده شده است. در این روش، نظرات با استفاده از مدل زبانی مولد[1] احتمالی مدل‌سازی شدند. این نظرات نشان‌دهنده ارتباط بین جملات و جنبه‌ها با استفاده از متغیرهای پنهان است. آن‌ها ابتدا به صورت دستی جنبه‌های صریح موجود در نظرات را مشخص و برچسب‌گذاری کردند و این دانش به صورت نیمه‌خودکار برای آموزش مدل استفاده شده است و درنهایت با استفاده از مدل نهایی، کل جنبه‌ها استخراج می‌شوند.

در مقاله [۲۱] مدلی برای استخراج جنبه مبتنی بر مدل موضوعی ارائه شد. در این مقاله با ترکیب روابط هم‌رخدادی به عنوان دانش دامنه و الگوریتم تخصیص دیریکله پنهان، استخراج جنبه برای هر سند انجام شد که باعث بهبود دقت استخراج جنبه‌ها شد. در این روش، در ابتدا جنبه‌های اولیه بر اساس الگوریتم LDA شناسایی می‌شوند و سپس با یک روند تکرارشونده ، دانش دامنه به صورت خودکار و با استفاده از روابط هم‌رخدادی و جنبه‌های مشابه هر موضوع مرتبط استخراج می‌شود. با هر تکرار این مرحله کیفیت دانش بهتر شده و باعث بهبود کیفیت جنبه‌های استخراج شده خواهد شد. درنهایت دانش دامنه استخراج شده به مدل LDA تزریق می‌شود و جنبه‌های موجود در اسناد با استفاده از مدل جدید استخراج می‌شود. این روش وابستگی به زبان ندارد و بر روی هر مجموعه زبانی قابل اجرا است.

در [۲۲] نویسندگان ساختاری کامل بر مبنای مدل آنالیز احتمالی معنایی مخفی پیشنهاد کرده‌اند که همه زیربخش‌های نظرکاوی جنبه‌گرا از جمله استخراج جنبه را پوشش می‌دهد.

در سال‌های اخیر روش‌های مبتنی بر مدل موضوعی به صورت گسترده برای استخراج جنبه و دسته‌بندی آن‌ها استفاده شده است. اما یک ضعف این روش‌ها این است که به مجموعه داده بزرگی نیاز دارند تا نتایج قابل قبولی ارائه دهند و در صورتی که مجموعه داده کوچک باشد دقت این الگوریتم‌ها به طور محسوسی کاهش پیدا می‌کنند. همچنین این دسته از الگوریتم‌ها، عموماً موضوع‌ها و جنبه‌های متناقضی را شناسایی می‌کنند که با قضاوت انسانی همخوانی ندارد.

۳،۲. روش‌های مبتنی بر گراف

در این دسته از روش‌ها، نظرات به گراف کلمات تبدیل می‌شوند. گره‌های این گراف را می‌تواند کلمه، جمله و یا نظر تشکیل دهد. درنهایت از الگوریتم‌های خوشه‌بندی گراف استفاده می‌شود تا جنبه‌های موجود در سند، استخراج شوند.

در مقاله [۲۳] یک روش مبتنی بر گراف برای استخراج جنبه معرفی شد. در این روش، کلمه‌های نظر به عنوان گره‌های گراف در نظر گرفته شد و سپس این گراف با یک گراف از مجموعه جنبه‌های صریح ترکیب شد و در آخر برای وزن‌دهی به یال‌های این گراف از هم‌رخدادی بین کلمات استفاده شد. در نهایت با استفاده از یک الگوریتم خوشه‌بندی، جنبه‌ها استخراج شدند.

در مقاله [۲۴] یک رویکرد مبتنی بر گراف و با ادغام اطلاعات معنایی بدست آمده از مدل موضوعی و روابط هم‌رخدادی بین کلمات معرفی شد. در این روش، با استفاده از انواع روابط موجود در گراف متنی و مدل موضوعی، چالش نحوه استفاده از هم‌رخدادی در مدل موضوعی را به خوبی برطرف شد. در این روش، کلمات، گره‌های گراف را تشکیل می‌دهند و وزن هر یال براساس ادغام الگوریتم LDA به عنوان یک مدل موضوعی و روابط هم‌رخدادی بین کلمات محاسبه می‌شود. در این روش برای ساخت گراف اولیه، هر دو گره که رابطه هم‌رخدادی نزدیکی و بیشتر از یک حد آستانه مشخصی داشته باشند، به یکدگیر متصل می‌شوند که این گره‌ها کلمات پرتکرار هم‌رخداد نامیده می‌شوند.

در مقاله [۲۵] یک روش مبتنی بر گراف برای شناسایی موضوع و استخراج جنبه معرفی شده است. در این روش، با استفاده از یک الگوریتم به نام گراف کلیدی و براساس رابطه هم‌رخدادی بین کلمات، سند به گراف تبدیل می‌شود. سپس با استفاده از روش‌های تشخیص

[1] Generative language model

جامعه، گراف به چندین جامعه تقسیم می‌شود و درنهایت هر جامعه به‌عنوان یک موضوع درنظر گرفته می‌شود و گره‌های عضو در جامعه به عنوان جنبه‌های موضوع شناخته می‌شوند.

به‌طور کلی در این پژوهش‌ها از دانش‌های محدودی مانند هم‌رخدادی کلمات و توزیع کلمات استفاده شده است و سایر دانش‌های موجود در متن نادیده گرفته شده است، درحالی‌که استفاده از دانش‌های دیگر، درصورتی که دانش صحیح باشد می‌تواند باعث بهبود دقت تشخیص و استخراج جنبه‌ها شود.

۳. راه‌حل پیشنهادی

در این پژوهش روشی ارائه شده است که از مدل موضوعی و شباهت بین کلمات به عنوان دانش دامنه برای ساخت گراف کلمات و سپس استخراج صحیح جنبه‌ها استفاده می‌کند. در شکل ۱، شبه کد روش پیشنهادی آمده است. ورودی روش پیشنهادی مجموعه‌داده‌ای از متون مختلف در موضوعات متفاوت است و خروجی آن جنبه‌های متعلق به هر موضوع است. در روش پیشنهادی، پس از انجام پیش-پردازش‌های مرسوم (خط ۱)، با استفاده از روش LDA به عنوان یکی از روش‌های اصلی مدل موضوعی، جنبه‌های اولیه مجموعه داده استخراج می‌شوند (خطوط ۲ تا ۴). سپس مجموعه‌ای از ترکیب همه جنبه‌های یک موضوع در خط ۵ شکل می‌گیرد.

در مرحله بعد و در خطوط ۶ تا ۱۰ الگوریتم، با استفاده از یک روش مبتنی بر شبکه عصبی به‌نام بردار کلمات، نتایج مرحله اول غنی‌سازی می‌شود و کلمات مشابه برای هر کلمه از نتایج مرحله اول، بدست می‌آید. از آن‌جا که در روش بردار کلمات شباهت بین کلمات محاسبه می‌شود ولی نتایج مرحله اول به صورت احتمال است، لازم است تا نتیجه بردار کلمات پس از ترکیب در خط ۱۱ به احتمال تبدیل شود؛ در مرحله سوم با استفاده از یک نسبت ریاضی احتمال هر کلمه محاسبه می‌شود (خط ۱۲). سپس نتایج مرحله اول و سوم تجمیع می‌شود (خط ۱۳) و در مرحله بعد در خطوط ۱۴ تا ۱۸، شباهت بین کلمات در هر جنبه به عنوان دانش مورد استفاده، محاسبه می‌شود و سپس گراف کلمات تشکیل می‌شود (خط ۱۹).

درنهایت در خط آخر شبه‌کد پیشنهادی، یک الگوریتم خوشه‌بندی روی گراف اعمال می‌شود تا جنبه‌های موجود در هر دامنه به‌طور صحیح استخراج شود. هر خوشه شناسایی شده، یک جنبه از موضوع در نظر گرفته می‌شود و هر خوشه به عنوان جنبه، شامل کلماتی مرتبط و با معنی است.

Algorithm: Proposed aspect extraction method

Input:
Documents, $D = \{D_1, D_2, ..., D_n\}$

Output:
Aspects for each domain, $T = \{T_1, T_2, \ldots, T_n\}$

```
1-  Preprocessing(D)
2-  For each domain corpus Di ∈ D
3-      Ai ←LDA(Di)
4-  End for
5-  A ← Ui Ai
6-  For each domain corpus Di ∈ D
7-      For each word Wi ∈ A
8-          Bi ←Word2Vec (Wi, Di)
9-      End for
10- End for
11- B ← Ui Bi
12- B ← SimToProbability (B)
13- C ← Merge (A, B)
14- For each word Wi ∈ C
15-     For each word Xi ∈ C
16-         S ← Similarity (Wi, Xi)
17-     End for
18- End for
19- G ← CreateGraph (C)
20- T ← SpectralClustering (G)
```

شکل ۱. شبه کد روش پیشنهادی

در ادامه این بخش، هر یک از گام‌های روش پیشنهادی به طور کامل شرح داده می‌شود. اما پیش از آن و برای درک بهتر از روش پیشنهادی مثالی از اجرای الگوریتم ذکر می‌شود. این مثال، مربوط به اجرای روش پیشنهادی بر روی اسنادی در زمینه«لپتاپ» است. در مرحله اول الگوریتم LDA روی اسناد اعمال می‌شود و جنبه‌های اولیه آن ساخته می‌شوند.

به عنوان مثال، یک جنبه نمونه از خروجی این مرحله شامل کلماتی مانند «صفحه‌نمایش، کیفیت، حافظه، تصویر، اروپا و ...» است که واضح است بعضی از کلمات آن مانند «حافظه» و «اروپا» به اشتباه در این گروه قرار گرفته‌اند و ارتباط معنایی زیادی با سایر کلمات (که مربوط به صفحه‌نمایش است) ندارد. در مرحله بعد الگوریتم بردار کلمات بر روی اسناد اولیه و نتایج خروجی مرحله قبل اعمال می‌شود. نتیجه خروجی این مرحله، گروهی از کلمات برای هر لغت در مرحله اول است که این کلمات بیشترین شباهت را به آن لغت دارند. به‌طور مثال در این مرحله برای کلمه «صفحه‌نمایش» کلماتی نظیر «رزولوشن، روشنایی، کیبورد، مانیتور، ناراحت و ...» شناسایی می‌شوند. باز هم واضح است که بعضی از کلمات آن مانند «کیبورد» و «ناراحت» به اشتباه در این دسته قرار گرفته‌اند.

در مرحله بعد نتیجه این دو مرحله با یکدیگر تجمیع می‌شود و سپس در مرحله پنجم شباهت بین هر دو کلمه محاسبه می‌شود و یک ماتریس ایجاد می‌گردد. به عنوان مثال، وزن بین دو کلمه «صفحه‌نمایش، رزولوشن» بیشتر از «صفحه‌نمایش، اروپا» تعیین می‌شود. در قدم بعد، با استفاده اطلاعات موجود یک گراف کلمات

تشکیل می‌شود که هر گره‌ی آن بیانگر یک کلمه مانند «صفحه‌نمایش» است و هر یال بین دو گره برابر با وزنی است که در مرحله قبل به عنوان شباهت آن دو کلمه محاسبه شده است. همزمان، یال بین کلماتی که مشابهت کمتری با یکدیگر دارند حذف می‌شوند. در نهایت یک الگوریتم خوشه‌بندی روی گراف اعمال می‌شود که نتیجه آن تشکیل تعدادی خوشه از گراف کلمات می‌باشد که هر خوشه بیانگر یک جنبه است به‌طوری‌که هر جنبه را گروهی از کلماتی تشکیل داده‌اند که بیشترین ارتباط معنایی را با یکدیگر دارند.

۱،۳. اعمال الگوریتم LDA پایه

در مرحله اول، الگوریتم LDA بر روی هر دامنه از مجموعه داده اجرا می‌شود. LDA فرض می‌کند که هر سند جنبه‌های مختلفی را نمایش می‌دهد، به عبارت دیگر یعنی هر سند از کلماتی تشکیل شده است که هر یک متعلق به یک جنبه است و نسبت جنبه‌های داخل یک متن با همدیگر متفاوت است. همچنین در این روش فرض می‌شود که هر جنبه توزیعی روی مجموعه کلمات است. به عبارت دیگر، کلماتی که در یک جنبه دارای احتمال بالایی هستند، کلمات مربوط به آن جنبه می‌باشند. نتیجه خروجی این الگوریتم شامل جنبه‌های شناسایی شده است که در هر جنبه، کلمات مربوط به آن وجود دارد.

۲،۳. غنی سازی نتایج مرحله اول با ا ستفاده از بردار کلمات

به منظور غنی‌سازی جنبه‌های استخراجی از LDA، الگوریتم بردار کلمات بر روی هر دامنه از مجموعه داده اجرا می‌شود. در این روش به کمک شبکه عصبی یک بردار با اندازه کوچک و ثابت برای نمایش تمام لغات و متون در نظرگرفته شده و با اعداد مناسب در فاز آموزش مدل، این بردار برای هر لغت محاسبه می‌شود. در این بردار هر ستون، فقط یک عدد را نمایش می دهد که نماینده میزان شباهت با کلمه یا ویژگی مورد نظر است. نتیجه خروجی این الگوریتم برای هر کلمه از خروجی مرحله اول، یک گروه کلمات با شباهت زیاد به آن کلمه می‌باشد.

۳،۳. تبدیل شباهت به احتمال در نتایج مرحله قبل

نتایج حاصل از الگوریتم LDA اعمال شده در مرحله اول شامل جنبه‌های شناسایی شده است که در هر جنبه، کلمات مربوط به آن با احتمالی مشخص به آن موضوع (جنبه) وجود دارد. درحالی که نتایج حاصل از غنی‌سازی صورت گرفته با استفاده از الگوریتم بردار کلمات در مرحله دوم دسته‌های از کلمات است که هر دسته متشکل از میزان شباهت کلمات به یک لغت از خروجی مرحله اول است.

برای اینکه بتوان از نتیجه خروجی دو مرحله قبل در ادامه الگوریتم استفاده شود لازم است تا نسبت تشابه خروجی مرحله دوم به احتمال تغییر پیدا کند تا بتوان نتایج مرحله اول و دوم رو تجمیع کرد. برای تبدیل نمودن شباهت به احتمال، با استفاده از رابطه (۱) میزان شباهت هر کلمه W به مجموع همه کلمات دسته مورد نظر (M) تقسیم می‌شود.

$$p(W_i.M) = \frac{Sim(W_i)}{Sum\ of\ all\ Sim(W)\ in\ M} \qquad (۱)$$

۴،۳. تجمیع نتایج حاصل از مرحله اول و دوم

در این مرحله، نتایج حاصل از دو مرحله قبل تجمیع می‌شوند. همانطور که در مرحله اول و دوم توضیح داده شد، در این روش پیشنهادی از دو الگوریتم پایه برای استخراج موضوعات و جنبه‌های مختلف استفاده شده است. برای استفاده از نتایج مراحل قبل لازم است تا نتایج آن‌ها را با یکدیگر تجمیع شود. به‌طور مثال، نتایج حاصل از الگوریتم تخصیص پنهان دریکله بر روی مجموعه داده لپتاپ، شامل ۳۰ جنبه می‌باشد که هر جنبه از ۱۵ کلمه که با احتمال بیشتری متعلق به آن موضوع است تشکیل شده است.

در مرحله دوم با اعمال الگوریتم بردار کلمات روی مجموعه داده لپتاپ و تمام لغات مرحله قبل اعمال می‌شود و ۴۵۰ دسته لغت بدست می‌آید. با انجام اینکار کلمات مشابه به هر کلمه از کلمات بدست آمده از جنبه‌های ابتدایی بدست می‌آید و غنی‌سازی جنبه‌ها انجام می‌شود. سپس برای ادامه پیش‌برد الگوریتم لازم است تا تمام دسته کلمات بدست آمده از مراحل قبل با یکدیگر تجمیع شوند که با اینکار ۴۸۰ دسته شامل کلمات بدست می‌آید که هر دسته شامل ۱۵ لغت می‌باشد.

به‌طورکلی خروجی این مرحله شامل تمام جنبه‌های مختلفی است که توسط هر دو الگوریتم در هر دامنه شناسایی شده است و هر جنبه نیز شامل کلماتی مرتبط و مشخص است که متعلق به آن جنبه هستند.

۵،۳. محاسبه شباهت بین کلمات

در این مرحله برای هر کلمه، دانشی از روی شباهت بین آن کلمه و سایر کلمات ایجاد می‌شود. این دانش به این صورت است که یک کلمه با کلمه‌ی دیگر چند بار در یک جنبه کنار هم شناسایی شده‌اند و نسبت آن به تعداد کل آن کلمه در تمام جنبه‌های شناسایی شده اولیه، چقدر است. معیار محاسبه شده برای هر جنبه، عددی در بازه ی [۰-۱] می‌باشد. معیار مشابهت هر دو جنبه به یکدیگر از طریق رابطه (۲) محاسبه می‌شود.

$$\mathrm{Sim}(w_1.w_2) = \frac{\text{Number of co\_occurrences of } w_1 \text{ and } w_2 \text{ in the aspects}}{\text{Number of aspects with } w_1 + \text{ Number of aspects with } w_2} \quad (۲)$$

در این رابطه، شباهت بین دو کلمه به صورت نسبت تعداد دفعاتی که کلمه w_1 و کلمه w_2 در یک جنبه T_i کنار هم آمده، به جمع تعداد دفعاتی که هر کدام از آن کلمه‌ها در یک جنبه T حضور داشته‌اند تعریف می‌گردد. به بیان ساده، با استفاده از این معیار شباهت، کلماتی که در جنبه‌ها به صورت مشترک قرار دارند، وزن بیشتری می‌گیرند و به احتمال زیاد ارتباط معنایی بیشتری بین آن‌ها برقرار است.

۶،۳. ساخت گراف کلمات

در این مرحله، یک گراف کلمات از روی تمامی کلمات موجود در مجموعه دامنه و دانش افزوده شده در مرحله قبل ساخته می‌شود. این گراف، یک گراف بدون جهت و نیمه کامل است که رأس‌های آن از تمامی کلمات موجود در مجموعه داده تشکیل شده است و یال‌های آن بر حسب معیار شباهت بین هر دو کلمه، وزن‌دهی شده است.

در ماتریس ساخته شده در مرحله‌ی قبل، بعضی از کلمات با یگدیگر شباهت کمتری دارند و احتمال اینکه در یک جنبه به یکدیگر مرتبط باشند بسیار کم است. درنتیجه برای حذف ارتباط بین این کلمات نامرتبط و ساخت یک گراف بهینه، این گراف با یک حد آستانه T هرس می‌شود. این هرس مطابق رابطه (۳) انجام می‌شود.

$$weight_{w1.w2} = \begin{cases} 0 & Sim_{w1.w2} < \delta \\ Sim_{w1.w2} & Sim_{w1.w2} \geq \delta \end{cases} \quad (۳)$$

به منظور پیدا کردن مقدار بهینه حد آستانه، این الگوریتم با مقادیر متفاوت اجرا می‌شود و سپس با ارزیابی نتیجه تشخیص جوامع بر روی گراف، بهترین مقدار برای حد آستانه جهت هرس گراف انتخاب می‌شود.

۷،۳. اعمال الگوریتم خوشه‌بندی

در این گام، گراف وزن‌دار هرس شده با استفاده از الگوریتم‌های خوشه‌بندی پردازش می‌شود. به‌طور کلی خوشه‌بندی به هدف استخراج بخش‌هایی از داده‌ها انجام می‌شود که با یکدیگر شباهت زیادی دارند. به زبان ساده‌تر، در خوشه‌بندی، داده‌ها در گروه‌هایی تقسیم می‌شوند که داده‌های مربوط به هر گروه دارای ویژگی‌های نزدیک به هم باشند و داده‌های با ویژگی‌هایی متفاوت در دو گروه مختلف قرار داشته باشند.

برای تشخیص این خوشه‌ها از انواع الگوریتم‌های خوشه‌بندی می‌توان استفاده کرد. در این پژوهش از الگوریتم خوشه‌بندی طیفی استفاده شده تابهترین خوشه‌های ممکن شناسایی شود. البته همان‌گونه که اشاره شد از سایر روش‌های خوشه‌بندی گراف نیز می‌توان در این گام استفاده کرد.

الگوریتم خوشه‌بندی طیفی در ابتدا یک ماتریس وابستگی ایجاد می‌کند و سپس ماتریس لاپلاسین[1] گراف را ایجاد می‌کند تا در ادامه بردارهای ویژه مربوط به آن شناسایی شود. در نهایت یک الگوریتم پایه خوشه‌بندی بر روی بردارهای ویژه شناسایی شده، اعمال می‌شود تا خوشه‌بندی نهایی ایجاد شود.

از مزیت‌های این روش می‌توان گفت که در واقع الگوریتم خوشه‌بندی طیفی باعث می‌شود که خوشه‌ها به نحوی شناسایی شوند که نقاط نزدیک به هم در یک خوشه قرار گیرند. لازم به ذکر است که درصورتی که الگوریتمی با دقت بیشتری ساخته شود، به راحتی می‌توان از آن الگوریتم در این مرحله استفاده کرد.

۴. ارزیابی

برای ارزیابی روش پیشنهادی، ابتدا از یک مجموعه داده به زبان انگلیسی استفاده شده است و سپس کارایی روش در زبان فارسی نیز سنجیده شده است. مجموعه داده انگلیسی، توسط چن و لیو در مقاله [۲۶] معرفی و جمع‌آوری شده است و سپس به صورت جامع در بسیاری از پژوهش‌ها مورد استفاده قرار گرفته است. در این مجموعه داده، اطلاعات ۵۰ موضوع مختلف از سایت آمازون استخراج شده است، به‌طوری‌که هر دامنه شامل حداقل ۱۰۰۰ نظر می‌باشد. باتوجه به نیاز الگوریتم ارائه شده، پیش‌پردازش لازم بر روی این داده‌ها انجام شده و هر نظر تبدیل به یک سبد کلمات شده است. منظور از این سبد کلمات این است که کلمات اضافی، فعل‌ها و غیره در هر نظر حذف شده است و تنها کلمات اصلی موجود در متن، باقی مانده است. جهت استفاده از این مجموعه داده در روش پیشنهادی ابتدا لازم است تا لیست کلمات همه دامنه‌ها با یکدیگر تجمیع شود و در نهایت یک لیست از تمام کلمات موجود در تمام دامنه‌ها ایجاد شود.

برای ارزیابی این پژوهش از چهار معیار به نام‌های انسجام موضوع[2]، معیار پیمانگی[3]، دقت[4] و معیار اندازه‌گیری توافق[5] استفاده شده است.

۱،۴. انسجام موضوع

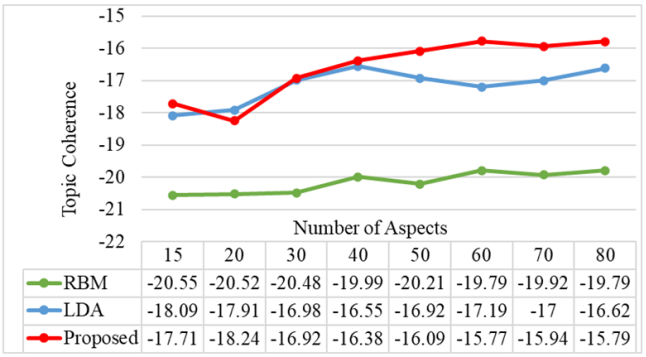
کلمات موجود در یک جمله یا سند باید دارای انسجام معنایی باشند. در واقع هر کلمه باید دارای یک ارتباط منطقی با کلمات بعدی خود باشد. معیار انسجام موضوع با اندازه‌گیری میزان هم‌رخدادی بین

[1] Laplacian
[2] Topic Coherence
[3] Modularity
[4] Accuracy
[5] Kappa

کلمات، شباهت معنایی را محاسبه می‌کند. به‌طور کلی هرچه کلمات یک جنبه بیشتر با هم تکرار شده باشند و از طرف دیگر کلمات خاص‌تر باشند، انسجام جنبه‌های استخراج شده برای یک موضوع بیشتر است و درنتیجه الگوریتم دقت بالاتری دارد. این معیار با استفاده از رابطه (۴) محاسبه می‌شود [۲۱] و [۲۷].

$$Coherence = \sum_{i<j} \log \frac{p(w_i.w_j)}{p(w_i)p(w_j)} \qquad (۴)$$

جهت ارزیابی، روش پیشنهادی با الگوریتم‌های پایه در این زمینه یعنی $\mathcal{LDA}$ و $\mathcal{RBM}$ مقایسه شده است. هر سه الگوریتم بر روی موضوع لپتاپ از مجموعه داده‌ای که در بخش قبل به‌طور مفصل به آن پرداخته شد، اجرا شده است. سپس میانگین معیار انسجام موضوع، برای هر سه الگوریتم برای تعداد متفاوت جنبه‌های استخراجی گزارش شده است. نتیجه این ارزیابی در شکل ۲ نمایش داده شده است.



	15	20	30	40	50	60	70	80
RBM	-20.55	-20.52	-20.48	-19.99	-20.21	-19.79	-19.92	-19.79
LDA	-18.09	-17.91	-16.98	-16.55	-16.92	-17.19	-17	-16.62
Proposed	-17.71	-18.24	-16.92	-16.38	-16.09	-15.77	-15.94	-15.79

شکل ۲. ارزیابی روش پیشنهادی بر اساس معیار انسجام موضوع

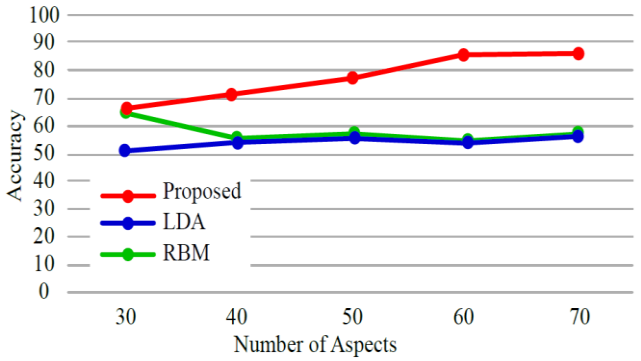
همانطور که گفته شد، هر چه میزان معیار انسجام موضوع بیشتر باشد، نتیجه بهتر ارزیابی می‌شود. همانطور که در شکل ۲ مشخص است، معیار انسجام موضوع در روش پیشنهادی به طور میانگین ۴ واحد از الگوریتم RBM و یک واحد از الگوریتم LDA بیشتر است. این نتیجه نشان‌دهنده این است که روش پیشنهادی باعث بهبود صحت استخراج جنبه شده است.

۲,۴. دقت و معیار اندازه‌گیری توافق

برای این ارزیابی، جنبه‌های استخراج شده نهایی در سه الگوریتم پیشنهادی، LDA و RBM در اختیار سه فرد مختلف قرار داده شد تا به صورت دستی عملیات برچسب‌گذاری بر روی پنج جنبه منتخب یکسان در هر سه الگوریتم انجام شود. خروجی‌های برچسب‌گذاری شده نشان‌دهنده این هستند که آیا کلمات مشخص شده برای هر یک از این پنج جنبه، مرتبط با آن جنبه هستند یا خیر. در مرحله بعد پاسخ اکثریت از بین سه فرد به عنوان گزینه نهایی انتخاب و دقت جنبه‌های استخراج شده در هر سه روش محاسبه شده است. محاسبه دقت طبق رابطه (۵) انجام شده است.

$$\text{Accuracy} = \frac{Number\ of\ correctly\ labeled\ data}{\text{Number of all data}} \qquad (۵)$$

طبق رابطه (۵) نسبت داده‌هایی که به درستی در هر جنبه قرار گرفته‌اند به تعداد کل نمونه‌ها، دقت هر روش را تعیین می‌کند. در شکل ۳ نتایج حاصل از روش پیشنهادی و دو روش پایه در تعداد جنبه‌های مختلف آورده شده است.



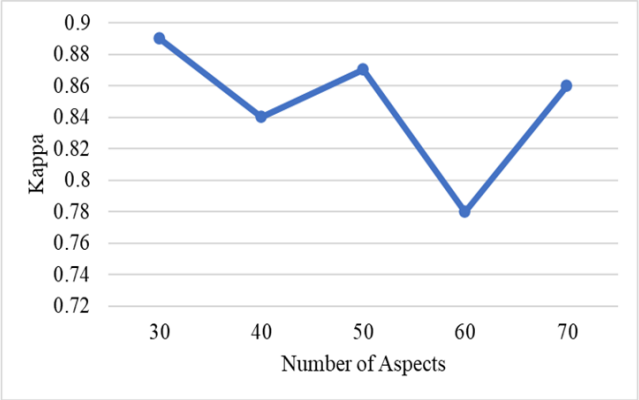
شکل ۳. ارزیابی دقت روش پیشنهادی در کلمات هر جنبه

همانطور که در شکل ۳ مشخص است، دقت استخراج جنبه در روش پیشنهادی به اندازه قابل توجهی از الگوریتم‌های پایه بیشتر است. این نتیجه نشان‌دهنده این است که روش پیشنهادی باعث بهبود دقت استخراج جنبه شده است.

همچنین برای ارزیابی توافق برچسب‌گذاری بین سه فرد بررسی کننده از معیار کاپا [۲۸] استفاده شده است. معیار کاپا با استفاده از رابطه (۶) محاسبه می‌شود.

$$Kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \qquad (۶)$$

در رابطه (۶)، Pr(a) نسبت تعداد دفعاتی است که داوران موافق هستند و Pr(e) نسبت تعداد دفعاتی است که انتظار می‌رود بر اساس شانس توافق کنند. اگر بیش از دو داور وجود داشته باشد، مقدار متوسط زوج‌های کاپا محاسبه می‌شود. به کمک این معیار نتایج نهایی برچسب‌گذاری سه فرد مورد بررسی قرار گرفت.



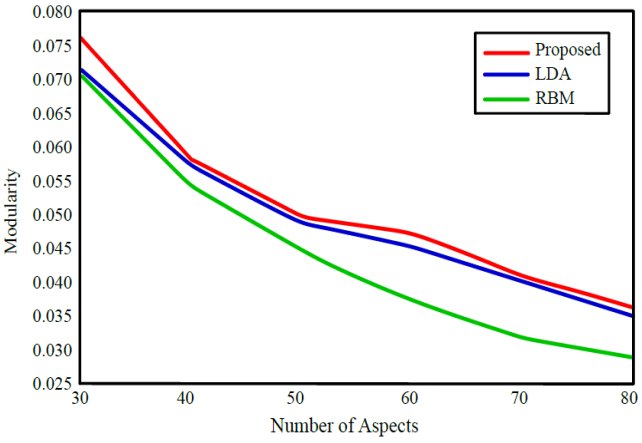
شکل ۴. میزان توافق میان برچسب‌های انتخابی با استفاده از معیار کاپا

همانطور که در شکل ۴ مشخص است، معیار کاپا بر روی نمونه‌های خروجی برچسب‌گذاری هر سه فرد در هر پنج حالت مختلف بالاتر از ۷۸ درصد می‌باشد که نشان‌دهنده این است که برچسب‌گذاری انجام شده توسط هر سه فرد به هم نزدیک بوده و توافق بین این سه فرد خوب ارزیابی می‌شود، پس برچسب‌گذاری قابل اعتماد است.

۳،۴. معیار پیمانگی

معیار پیمانگی به عنوان یک معیار برای اندازه‌گیری کیفیت خوشه‌بندی گراف است. اگر تعداد یال‌های درون خوشه‌ای بهتر از حالت گراف تصادفی نباشد، پیمانگی برابر صفر است. حداکثر مقدار پیمانگی زمانی به دست می‌آید که تمام رئوس هر خوشه به هم متصل باشند و خوشه‌ها توسط هیچ یالی به یکدیگر متصل نشوند. معرفی کامل این معیار در [۲۹] آمده است.

معیار پیمانگی بر روی خوشه‌های بدست آمده از روش پیشنهادی و دو الگوریتم پایه LDA و RBM بر روی مجموعه داده با موضوع لپتاپ اعمال شده است. نتیجه این ارزیابی در شکل ۵ و برای تعداد متفاوت خوشه نشان داده شده است.

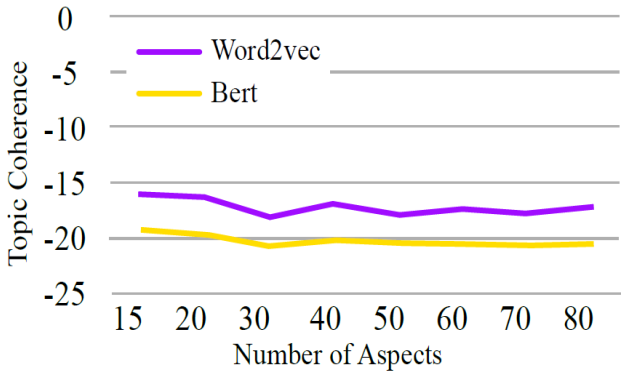


شکل ۵. ارزیابی خوشه‌های بدست آمده با استفاده از معیار پیمانگی

همانطور که در شکل ۵ مشخص است، معیار پیمانگی با استفاده از اعمال روش پیشنهادی در موضوع لپتاپ، بیشتر از الگوریتم‌های پایه است و درنتیجه خوشه‌های بهتری در گراف روش پیشنهادی ایجاد شده است. البته باید توجه کرد که از آن‌جایی که کلمات معرف در یک حوزه خاص (لپتاپ)، گره‌های گراف هستند، اکثر این کلمات احتمال تکرار زیادی با یکدیگر دارند، پس وجود یال بین خوشه‌های مختلف نیز بسیار محتمل است. همین دلیل باعث شده معیار پیمانگی در حالت کلی برای روش پیشنهادی و سایر روش‌های مورد مقایسه کم باشد. نکته بعد در مورد شکل ۵، این است که با افزایش تعداد جنبه‌ها احتمال تکرار کلمات در جنبه‌های مختلف زیاد شده پس عملاً وزن یال‌های بین جنبه‌ها زیاد شده و معیار پیمانگی کاهش می‌یابد.

۴،۴. بررسی اثربخشی بردار کلمات در روش پیشنهادی

روش پیشنهادی در این مقاله، به صورت چارچوب کلی است که می‌توان در هر یک از گام‌های آن از روش‌ها و الگوریتم‌های مختلفی استفاده کرد. به منظور بررسی اثربخشی بردار کلمات در روش پیشنهادی، به بررسی استفاده از الگوریتم Bert[1] در ساخت بردار کلمات پرداخته می‌شود. Bertیک مدل زبانی از پیش‌ساخته توسط گوگل است که از روی مجموعه وسیعی از متون زبان انگلیسی آموزش دیده است [۳۰]. در شکل ۶، روش پیشنهادی در این مقاله با جایگزینی بردار کلمات با مدل زبانی Bert ارزیابی شده است.



شکل ۶. مقایسه نتایج با استفاده از الگوریتم بردار کلمات و Bert با معیار انسجام موضوع

همانطور که در شکل ۶ مشخص است، معیار انسجام موضوع با استفاده از الگوریتم بردار کلمات نتایج بهتری دارد. از دلایل این امر می‌توان به این اشاره کرد که الگوریتم بردار کلمات بر روی مجموعه داده هدف، آموزش داده شده است اما الگوریتم Bert به دلیل این‌که نیاز به پردازش بسیار زیادی دارد از مدل پیش آموزش دیده استفاده

[1] Bidirectional Encoder Representations from Transformers

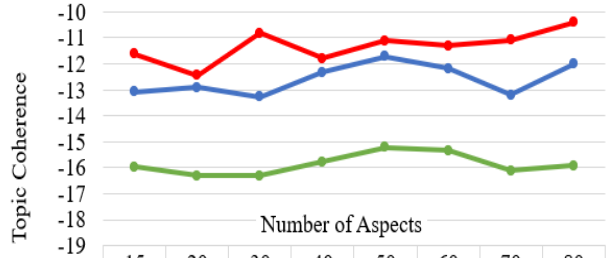
شده که همین امر سبب می‌شود تا نتایج الگوریتم بردار کلمات بهتر گردد.

۵،۴. ارز یابی روش پیشـــنهادی در مجمو عه داده فارسی

همان‌گونه که بیان شد روش‌پیشنهادی برای استخراج جنبه از هیچ منبع زبانی خاص استفاده نمی‌کند و صرفاً بر پایه الگوریتم‌های احتمالاتی شکل گرفته است. به همین دلیل به سادگی می‌توان از آن در زبان‌های مختلف بهره گرفت. به منظور اثبات صحت این ادعا در این بخش، روش پیشنهادی بر روی زبان فارسی بررسی و ارزیابی شده است.

مجموعه داده مورد استفاده مربوط به بیش از ۳۰۰۰ هزار سند در زمینه لپتاپ به زبان فارسی است که در مقاله [۲۱] معرفی و مورد استفاده قرار گرفته است. نتایج مربوط به اجرای روش پیشنهادی و الگوریتم‌های پایه در شکل ۷ آمده است.

منطبق بر شکل ۷، روش پیشنهادی در زبان فارسی بهتر از روش‌های پایه‌ای عمل کرده است و جنبه‌های منسجم‌تری در تعداد مختلف جنبه‌ها ایجاد نموده است. این ارزیابی نشان می‌دهد که استفاده از روش پیشنهادی در زبان‌هایی نظیر فارسی که از نظر منابع زبانی محدودیت دارند، می‌تواند بسیار راهگشا و کاربردی باشد. در بخش بعد به جمع‌بندی این مقاله پرداخته خواهد شد.



	15	20	30	40	50	60	70	80
RBM	-15.98	-16.31	-16.32	-15.78	-15.21	-15.33	-16.12	-15.92
LDA	-13.09	-12.91	-13.28	-12.32	-11.72	-12.19	-13.2	-12.02
Proposed	-11.63	-12.44	-10.81	-11.8	-11.12	-11.3	-11.09	-10.41

شکل ۷. ارزیابی طبق معیار انسجام موضوع در مجموعه داده فارسی

۵. جمع‌بندی

در اغلب مطالعات انجام شده در حوزه استخراج جنبه تنها از رابطه هم‌رخدادی به عنوان دانش دامنه استفاده شده است و سایر اطلاعات و دانش‌های موجود در متن نادیده گرفته شده است. به همین منظور در این پژوهش از شباهت بین کلمات به عنوان دانش دامنه در کنار مدل موضوعی و گراف کلمات استفاده شده است که همین امر موجب بهبود دقت استخراج جنبه شده است. درکنار این موضوع، استفاده از تجمیع نتایج الگوریتم LDA و بردار کلمات، باعث انسجام بیشتر جنبه‌های استخراجی شده است. همچنین برخلاف بسیاری از مطالعات پیشین، روش پیشنهادی نیازی به منابع زبانی ندارد از این‌رو وابسته به زبان خاصی نیست و بر روی هر مجموعه زبانی قابل اجرا می‌باشد. برای اثبات صحت این ادعا، روش پیشنهادی علاوه بر انگلیسی بر روی مجموعه داده‌ای به زبان فارسی نیز ارزیابی شده است.

روش پیشنهادی با دقت مناسبی که در استخراج جنبه‌ها دارد می‌تواند در فرایندهای مختلف پردازش زبان طبیعی از جمله نظرکاوی سطح جنبه، خلاصه‌سازی متن و تولید خودکار متن، کاربرد فراوانی داشته باشد. همچنین عدم وابستگی به زبان نوشته، روش پیشنهادی را در زبان‌های از جمله فارسی که منابع زبانی محدود دارند کارآمد می‌سازد. البته لازم به ذکر است که مطالعات انجام شده در حوزه استخراج جنبه هنوز در ابتدای راه قرار دارد. در آینده و برای بهبود ساختار پیشنهادی می‌توان از سایر اطلاعات متنی نیز به عنوان دانش دامنه بهره گرفت. به عنوان مثال از اطلاعات متنی مانند n-گرام و یا افزودن کلمات مرتبط (که می‌تواند از مدل‌های زبانی از پیش آموزش دیده شده بدست آید)، اطلاعات معنایی هر زبان و غیره به عنوان دانش دامنه می‌توان استفاده کرد و اثر هر یک را در نتایج مورد بررسی قرار داد. همچنین روش پیشنهادی در قالب یک چارچوب کلی پیشنهاد شده که در هر یک از گام‌های آن می‌توان از روش مختلف جهت افزایش دقت بهره گرفت. برای مثال در استخراج جنبه‌های اولیه می‌توان از هر مدل موضوعی جدیدی به جای LDA استفاده کرد. همچنین در ایجاد بردار کلمات و خوشه‌بندی گراف نیز، روش‌های دیگری می‌تواند جایگزین روش مورد استفاده در چارچوب ارائه شده در این مقاله گردد.

مراجع

[1] R. Wang, D. Zhou, M. Jiang, J. Si, and Y. Yang, “A Survey on Opinion Mining: From Stance to Product Aspect,” *IEEE Access*, vol. 7, pp. 41101–41124, 2019, doi: 10.1109/ACCESS.2019.2906754.

[2] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, “Current State of Text Sentiment Analysis from Opinion to Emotion Mining,” *ACM Comput. Surv.*, vol. 50, no. 2, pp. 1–33, Mar. 2018, doi: 10.1145/3057270.

[3] M. Tubishat, N. Idris, and M. A. M. Abushariah, “Implicit aspect extraction in sentiment analysis: Review, taxonomy, oppportunities, and open challenges,” *Inf. Process. Manag.*, vol. 54, no. 4, pp. 545–563, Jul. 2018, doi: 10.1016/j.ipm.2018.03.008.

[4] A. García-Pablos, M. Cuadros, and G. Rigau, “W2VLDA: Almost unsupervised system for Aspect Based Sentiment Analysis,” *Expert Syst. Appl.*, vol. 91, pp. 127–137, 2018, doi: 10.1016/j.eswa.2017.08.049.

[5] W. Zhang, X. Li, Y. Deng, L. Bing, and W.

feature identification in Chinese reviews using explicit topic mining model," *Knowledge-Based Syst.*, vol. 76, pp. 166–175, Mar. 2015, doi: 10.1016/j.knosys.2014.12.012.

[17] Z. Chen and B. Liu, "Mining topics in documents: Standing on the Shoulders of Big Data," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, 2014, pp. 1116–1125, doi: 10.1145/2623330.2623622.

[18] B. Ozyurt and M. A. Akcayol, "A new topic modeling based approach for aspect extraction in aspect based sentiment analysis: SS-LDA," *Expert Syst. Appl.*, vol. 168, p. 114231, Apr. 2021, doi: 10.1016/j.eswa.2020.114231.

[19] M. Venugopalan and D. Gupta, "An enhanced guided LDA model augmented with BERT based semantic strength for aspect term extraction in sentiment analysis," *Knowledge-Based Syst.*, vol. 246, p. 108668, Jun. 2022, doi: 10.1016/j.knosys.2022.108668.

[20] S. K. Karmaker Santu, P. Sondhi, and C. Zhai, "Generative Feature Language Models for Mining Implicit Features from Customer Reviews," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, Oct. 2016, pp. 929–938, doi: 10.1145/2983323.2983729.

[21] M. Shams and A. Baraani-dastjerdi, "Enriched LDA (ELDA): Combination of latent Dirichlet allocation with word co-occurrence analysis for aspect extraction," *Expert Syst. Appl.*, vol. 80, pp. 136–146, 2017, doi: 10.1016/j.eswa.2017.02.038.

[22] M. Shams, N. Khoshavi, and A. Baraani-Dastjerdi, "LISA: Language-Independent Method for Aspect-Based Sentiment Analysis," *IEEE Access*, vol. 8, pp. 31034–31044, 2020, doi: 10.1109/ACCESS.2020.2973587.

[23] A. Bagheri, M. Saraee, and F. de Jong, "Care more about customers: Unsupervised domain-independent aspect detection for sentiment analysis of customer reviews," *Knowledge-Based Syst.*, vol. 52, pp. 201–213, Nov. 2013, doi: 10.1016/j.knosys.2013.08.011.

[24] C. Zhang, H. Wang, L. Cao, W. Wang, and F. Xu, "A hybrid term–term relations Lam, "A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.01054.

[6] T. A. Rana and Y.-N. Cheah, "Aspect extraction in sentiment analysis: comparative analysis and survey," *Artif. Intell. Rev.*, vol. 46, no. 4, pp. 459–483, Dec. 2016, doi: 10.1007/s10462-016-9472-z.

[7] P. P. Tribhuvan, S. G. Bhirud, and R. R.Deshmukh, "Product Features Extraction for Feature Based Opinion Mining using Latent Dirichlet Allocation," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 10, pp. 128–131, Oct. 2017, doi: 10.26438/ijcse/v5i10.128131.

[8] B. Ma, D. Zhang, Z. Yan, and T. Kim, "An LDA and Synonym Lexicon Based Approach to Product Feature Extraction from Online Consumer Product Reviews," *J. Electron. Commer. Res.*, vol. 14, no. 4, pp. 304–314, 2013, doi: 10.1016/j.im.2015.02.002.

[9] A. K. Samha, Y. Li, and J. Zhang, "Aspect-Based Opinion Extraction from Customer reviews," Apr. 2014, [Online]. Available: http://arxiv.org/abs/1404.1982.

[10] A. Konjengbam, N. Dewangan, N. Kumar, and M. Singh, "Aspect ontology based review exploration," *Electron. Commer. Res. Appl.*, vol. 30, pp. 62–71, Jul. 2018, doi: 10.1016/j.elerap.2018.05.006.

[11] F. Lazhar and T. G. Yamina, "Mining explicit and implicit opinions from reviews," *Int. J. Data Mining, Model. Manag.*, vol. 8, no. 1, p. 75, 2016, doi: 10.1504/IJDMMM.2016.075966.

[12] S. Behdenna, F. Barigou, and G. Belalem, "An Ontology-Based Approach to Enhance Explicit Aspect Extraction in Standard Arabic Reviews," *Int. J. Comput. Digit. Syst.*, vol. 11, no. 1, pp. 277–287, Jan. 2022, doi: 10.12785/ijcds/110123.

[13] T. Hofmann, "Probabilistic Latent Semantic Analysis," Jan. 2013, [Online]. Available: http://arxiv.org/abs/1301.6705.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[15] N. Zhang, S. Ding, J. Zhang, and Y. Xue, "An overview on Restricted Boltzmann Machines," *Neurocomputing*, vol. 275, pp. 1186–1199, Jan. 2018, doi: 10.1016/j.neucom.2017.09.065.

[16] H. Xu, F. Zhang, and W. Wang, "Implicit

analysis approach for topic detection," *Knowledge-Based Syst.*, vol. 93, pp. 109–120, Feb. 2016, doi: 10.1016/j.knosys.2015.11.006.

[25] H. Sayyadi and L. Raschid, "A Graph Analytical Approach for Topic Detection," *ACM Trans. Internet Technol.*, vol. 13, no. 2, pp. 1–23, Dec. 2013, doi: 10.1145/2542214.2542215.

[26] Z. Chen and B. Liu, "Topic Modeling using Topics from Many Domains, Lifelong Learning and Big Data," in *Proceedings of the 31st International Conference on Machine Learning - ICML '14*, 2014, vol. 32, pp. 703–711.

[27] D. Newman, Y. Noh, E. Talley, S. Karimi, and T. Baldwin, "Evaluating topic models for digital libraries," in *Proceedings of the 10th annual joint conference on Digital libraries - JCDL '10*, 2010, p. 215, doi: 10.1145/1816123.1816156.

[28] M. L. McHugh, "Interrater reliability: the kappa statistic.," *Biochem. medica*, vol. 22, no. 3, pp. 276–82, 2012, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23092060

[29] M. E. J. Newman, "Modularity and community structure in networks," *Proc. Natl. Acad. Sci.*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006, doi: 10.1073/pnas.0601602103.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Oct. 2018, [Online]. Available: http://arxiv.org/abs/1810.04805.