شناسایی ویژگی‌های صریح محصولات در نظرات زبان فارسی

عاطفه محمدی[۱]، محمدرضا پژوهان[۲]، محمدعلی نعمت‌بخش[۳]

[۱] دانشجوی دکتری گروه مهندسی کامپیوتر، دانشگاه یزد، یزد
atefehmohammadi@stu.yazd.ac.ir

[۲] استادیار، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد
pajoohan@yazd.ac.ir

[۳] دانشیار، گروه مهندسی کامپیوتر، دانشگاه یزد، اصفهان
nematbakhsh@eng.ui.ac.ir

چکیده

نظراتی که در وبلاگ‌ها، فروم‌ها و وب سایت‌های نظری در مورد محصولات ارائه می‌شود، محققان را علاقه‌مند به زمینه‌ی نظرکاوی کرده است. اخیرا تحقیقات زیادی در زمینه نظرکاوی مبتنی بر ویژگی در زبان انگلیسی شده است تا بتوانند ویژگی‌های صریح محصول و نظرات مرتبط با آن ویژگی‌ها را شناسایی کنند. شناسایی ویژگی‌های محصول در زبان فارسی با چالش‌هایی روبه‌رو است. در پژوهش‌های پیشین برای استخراج ویژگی‌های صریح، تاکید بر نقش لغوی کلمات داشتند که قادر به متمایز کردن صفت‌ها به عنوان بخشی از اسم یا کلمات احساس نبوده است. روش ارائه شده شامل مرحله‌ی استخراج ویژگی‌های صریح محصول است. هدف این تحقیق، ارائه‌ی راه‌کاری است که بتواند بدون نظارت و بدون نیاز به مجموعه داده‌ی آموزشی برچسب‌خورده، نظرکاوی مبتنی بر ویژگی را برای محصولات در زبان فارسی انجام دهد. ارزیابی‌های انجام شده، نشان می‌دهند روش ارائه شده در مرحله‌ی استخراج ویژگی‌های صریح محصولات از بازخوانی و دقت بیشتری نسبت به کارهای قبلی برخوردار است.

کلمات کلیدی

ویژگی صریح، ویژگی ضمنی، قوانین انجمنی،نظرات، هم‌رخداد.

۱- مقدمه

اغلب افراد در فرایند تصمیم‌گیری در مورد محصولات و سرویس‌های جدید، نیاز به دانستن نظرات افرادی دارند که در مورد آن محصول یا سرویس تجربه ی قبلی دارند [۴]. مصرف‌کنندگان و مشتریان با مقایسه‌ی نقدها و نقطه نظرات می‌توانند بهترین محصولی را انتخاب کنند که مطابق با سلیقه‌ی آن‌ها است. از طرف دیگر فروشندگان محصولات علاقه دارند که نظر مشتریان را در مورد کالا یا محصولات خود بدانند تا بتوانند کیفیت محصولات خود را ارزیابی کرده و با رفع نقاط ضعف خود، کیفیت محصولات خود را ارتقا بخشیده و جایگاه خود را در بازار رقابت حفظ کنند. با رشد چشم‌گیر تجارت الکترونیک، بیشتر محصولات از طریق وب به فروش می‌رسند؛ از این‌رو تعداد افرادی که تمایل به خرید اینترنتی دارند، روزبه‌وز بیشتر می‌شود [۵]. با افزایش حجم کالاهای فروخته‌شده از طریق اینترنت، حجم نظراتی که افراد در مورد هر محصول می‌دهند به‌طور مرتب در حال افزایش است. حجم وسیع نظرات منجر به پیدایش زمینه‌ی جدیدی به نام نظرکاوی یا تحلیل احساسات شده است.

اطلاعات متنی به‌طور عمده به دو دسته‌ی اصلی حقایق و نظرات تقسیم‌بندی می‌شوند. حقایق که شامل عبارات واقعی درباره‌ی موجودیت‌ها، رویدادها و ویژگی‌هایشان است. در مقابل نظرات شامل عبارت‌های ذهنی هستند که احساسات افراد و ارزیابی‌هایشان را در مورد موجودیت‌ها، رویدادها و ویژگی‌هایشان توصیف می‌کنند [۶]. تجزیه و تحلیل احساس یا نظرکاوی،

بررسی محاسباتی نظرات و نگرش‌های افراد درباره‌ی موجودیت‌ها و ویژگی هایشان است [۷]. هدف اصلی نظرکاوی، استخراج، دسته‌بندی و خلاصه‌سازی نظرات و نگرش‌های افراد درباره‌ی ویژگی‌های مختلف یک موجودیت یا سرویس خاص است [۸].

تجزیه و تحلیل احساس در سه سطح سند، جمله و ویژگی مورد بررسی قرار می‌گیرد. تحلیل احساسات در سطح سند و جمله در بسیاری از مواقع مفید است اما جزییات لازم را در اکثر موارد ایجاد نمی‌کنند؛ زیرا دقیقا مشخص نمی شود که افراد به چه مشخصه‌ای از محصول تمایل و یا عدم تمایل بیشتری دارند. پس لازم است تا برای بدست آوردن جزییات بیشتر از تجزیه و تحلیل احساس در سطح ویژگی استفاده شود. نظرکاوی مبتنی بر ویژگی، نسبت به نظرکاوی در سطح جملات و سند، دارای پیچیدگی‌های بیشتر و همچنین نتایج غنی‌تر بوده و جزییات، کامل‌تر و دقیق‌تر بررسی می‌شوند [۵]. هدف اصلی تجزیه و تحلیل در سطح ویژگی، کشف احساسات روی ویژگی‌های مختلف یک محصول است. جمله‌ی «کیفیت تماس سامسونگ خوب است، اما عمر باتریش کوتاه است.»، دو ویژگی «سامسونگ» به نام «کیفیت تماس» و «عمر باتری» را ارزیابی می‌کند. احساس مربوط به «کیفیت تماس»، مثبت و احساس مربوط به «عمر باتری»، منفی است.

روش‌های تشخیص ویژگی به طور کلی به دو دسته‌ی بانظارت و بدون‌نظارت دسته‌بندی می‌شوند [۵،۹]. روش‌های تشخیص ویژگی بانظارت نیاز به یک مجموعه داده‌ی آموزشی برچسب خورده دارند. ایجاد یک مجموعه داده‌ی برچسب خورده اغلب پرهزینه و نیازمند کار انسانی زیادی است. از آن-جایی‌که بسیاری از داده‌های در دسترس عموم بدون برچسب هستند، لازم است که مدل‌هایی برای کار با داده‌های بدون برچسب توسعه داده شوند[۵].

در تحقیقات گذشته روش‌های متعددی جهت استخراج ویژگی‌ها مطرح شده است [۶] به عنوان مثال یکی از ایده‌هایی که در این زمینه وجود دارد این است که عبارت‌های اسمی پر تکرار در جملات، نشان دهنده‌ی ویژگی‌ها هستند. چون زمانی که افراد درباره‌ی ویژگی نظر می‌دهند مکررا آن ویژگی را به کار می‌برند [۶] ولی مشکلی که وجود دارد این است که ویژگی‌ها می‌توانند به دو صورت صریح و ضمنی در جملات ظاهر شوند. در صورتی‌که جمله‌ای به‌طور آشکار حاوی ویژگی باشد، آن ویژگی، ویژگی صریح ولی در صورتی‌که جمله‌ای حاوی ویژگی نباشد ولی توسط شاخص‌هایی قابل استنباط باشد آن ویژگی، ویژگی ضمنی نامیده می‌شود [۱۰]. به عنوان مثال در جمله‌ی «خوبی این گوشی ارزان بودن آن است.»، صفت «ارزان» به‌طور ضمنی به ویژگی «قیمت» اشاره دارد. پس ویژگی «قیمت» یک ویژگی ضمنی است.

یک روش کاوش قوانین انجمن هم رخداد* جهت تشخیص ویژگی‌های ضمنی توسط های و همکاران [۱۱] ارائه شده است. در این روش، اسم یا عبارات اسمی به‌عنوان ویژگی صریح و صفت‌ها و فعل‌ها به‌عنوان کلمه‌ی احساس در نظر گرفته می‌شوند. در مرحله‌ی اول از این روش، دو مجموعه از واژه‌ها (کلمه‌ی احساس و ویژگی) از جملات صریح موجود در پیکره استخراج شده و یک ماتریس هم‌رخداد† که هر عنصر آن، بیان‌گر تعداد هم‌رخدادی کلمه و ویژگی صریح در جمله است، ایجاد می‌شود. در مرحله‌ی دوم، پس از گروه‌بندی ویژگی‌های صریح، برای کلمه‌ی احساس داده شده که دارای ویژگی صریح نیست یک لیست تطبیق‌یافته از قوانین مورد جست‌وجو قرار می‌گیرد تا قوانینی انتخاب شوند که دارای خوشه‌ی ویژگی با بیشترین وزن تکرار هستند و بر این اساس، کلمه‌ی نماینده‌ی خوشه به عنوان ویژگی ضمنی در نظر گرفته می‌شود.

های و همکاران [۹] یک روش کلی برای استخراج کلمه‌ی نظر و ویژگی توسط تجزیه و تحلیل وابستگی آماری ارائه داده‌اند. ویژگی‌ها معمولا به‌صورت اسم یا عبارت اسمی اتفاق می‌افتند و تمایل دارند که فاعل یا مفعول یک جمله باشند. به سادگی، انتخاب اسم یا عبارت اسمی به عنوان ویژگی کاندید، بازخوانی خوبی می‌دهد اما در عوض منجر به استخراج تعداد زیادی کاندید نادرست می‌شود که ممکن است روی فرایند استخراج ویژگی تأثیر منفی بگذارد. به همین دلیل، برای شناسایی ویژگی‌ها از پارسر وابستگی استفاده کردند. روش جدید با مجموعه‌ی کوچکی از ویژگی‌های بذری شروع می‌کند و به‌طور مرتب با کاوش ارتباطات وابستگی «ویژگی-نظر»، «ویژگی-ویژگی» و «نظر-نظر» بسط داده می‌شود. دو مدل وابستگی به نام آزمون نسبت احتمال‡ و تجزیه نهان معنایی§ پیشنهاد شدند تا وابستگی دوبه‌دو، بین دو واژه را محاسبه کنند. مطابق با آن، دو روش راه‌انداز به نام راه‌انداز مبتنی بر آزمون نسبت احتمال و راه‌انداز مبتنی بر تجزیه نهان معنایی ارائه دادند که هر دوی آن‌ها به یک مجموعه بذری اولیه نیاز دارند تا فرایند استخراج ویژگی و نظر را راه‌اندازی کنند.

باقری و همکاران [۵] مدل بدون ناظر تشخیص ویژگی و احساس را ارائه دادند که قادر به استخراج ویژگی‌های صریح و ضمنی بر روی زبان انگلیسی است. در مدل آن‌ها، اسم‌ها بیان‌گر ویژگی و صفت‌ها، قیدها و فعل‌ها بیان‌گر کلمات احساس هستند. ویژگی‌های چندکلمه‌ای توسط روشی به نام FMLR** شناسایی می‌شوند. سپس، دو قانون حذفی در جهت تکمیل فرایند پیش‌پردازش استفاده شده است. ویژگی‌ها، بر اساس معیار A-score رتبه-بندی می‌شوند. پس از نهایی شدن لیست ویژگی‌ها از دو روش هرس پشتیبان بالا مجموعه و زیر مجموعه برای حذف ویژگی‌های غیر مفید و افزونه استفاده می‌شود. حال برای تشخیص ویژگی ضمنی، با داشتن مجموعه کلمات احساس و ویژگی‌های صریح، گرافی برای این کلمات و ویژگی‌ها ترسیم می‌کند. برای ترسیم گراف، هر کلمه‌ی احساس در قالب یک گره در نظر گرفته می‌شود. سپس، لیست ویژگی‌های صریح وابسته به آن از طریق جملات در مجموعه متون استخراج شده و این ویژگی‌ها نیز در قالب گره-هایی از گراف رسم شده و به گره‌های کلمات احساس متصل می‌شوند. وزن اولیه‌ی هر یال بر اساس تعداد هم‌رخدادی ویژگی و کلمه‌ی احساس در جملات مشخص می‌شود و در نهایت از گراف به‌دست آمده برای تشخیص ویژگی ضمنی استفاده می‌شود.

کوان و همکاران [۱۲] یک روش بدون نظارت برای استخراج ویژگی محصول جهت تعیین نظر مبتنی بر ویژگی پیشنهاد دادند که کلمات مختص دامنه را به عنوان دانش دامنه‌ای برای استخراج ویژگی مختص دامنه به‌کار می‌گیرد. ایده‌ی اصلی آن‌ها این است که ویژگی‌های دامنه‌ای محصول را از طریق تخمین وزن‌هایشان در دامنه‌های مختلف استخراج کنند که اندازه‌گیری تشابه را برای تخمین وابستگی ویژگی‌های کاندید و کلمات دامنه به کار گرفتند. برای هر ویژگی محصول، یک بردار دامنه‌ای بر اساس این تشابه به-

* Rule Mining Co-occurrence Association Approach
† Co-occurrence Matrix
‡ Likelihood Ratio Tests
§ Latent Semantic Analysis
** Frequency Modified Left Right

دست می‌آید. سپس ویژگی‌های مختص دامنه، توسط اندازه‌گیری بردارهای ویژگی‌های دامنه و بردار دامنه‌ای یک موجودیت استخراج می‌شوند. به عنوان مثال، ویژگی «کیفیت تصویر» وابستگی بیشتری به موجودیت دامنه‌ای «دوربین» نسبت به "mp3" در پیکره‌ی نظرات دوربین دارد. تنها منابع خارجی به کار برده شده، پیکره‌ی دامنه‌ای قیاسی است به همین دلیل روش پیشنهادی‌شان کلی و بدون نظارت است.

برای استخراج ویژگی از نقدهای محصول، یک روش مبتنی بر قانون توسط پوریا و همکارانش [۱۳] ارائه شد. آن‌ها برای تشخیص ویژگی‌های صریح و ضمنی از دانش عمومی و درخت وابستگی جمله استفاده کردند که منجر به افزایش دقت در دو مجموعه‌ی داده‌ی مورد استفاده شد. روش آن‌ها به‌طور کامل بدون نظارت است و دقت آن به دقت تجزیه کننده‌ی وابستگی و پیکره‌ی نظرات بستگی دارد.

باباعلی و همکاران [۱] یک روش بدون نظارت نظرکاوی مبتنی بر ویژگی برای محصولات در زبان فارسی ارائه دادند که شامل مراحل شناسایی ویژگی‌های صریح، ویژگی‌های ضمنی و تعیین جهت معنایی نظرات است. مرحله‌ی استخراج ویژگی صریح، ابتدا ویژگی‌های جذاب و پرتکراری که بسیاری از مردم در مورد آن‌ها نظر داده‌اند را می‌یابد. سپس، ویژگی‌هایی که به ندرت راجع به آن‌ها نظر داده شده و می‌تواند برای برخی از مشتریان بالقوه جالب باشد را پیدا می‌کند. در مرحله‌ی استخراج ویژگی‌های پرتکرار، ویژگی-های تک و چندکلمه‌ای توسط الگوریتم اپریوری تغییر یافته‌ی پیشنهادی با حداقل پشتیبان یک درصد به‌دست می‌آیند. برخی از این ویژگی‌های پرتکرار ایجاد شده توسط الگوریتم پیشنهادی، ویژگی‌های مفید یا واقعی نیستند. بنابراین برای حذف این ویژگی‌های نادرست، دو شیوه‌ی هرس تراکم و هرس افزونگی پیشنهادی معرفی کرده‌اند که لیست ویژگی‌های خروجی را بهبود می‌دهند. در پایان این مرحله، از اثر کلمات احساس برای شناسایی ویژگی-های کم‌تکرار استفاده کرده‌اند. روش پیشنهادی، ویژگی‌های ضمنی را با استفاده از کاوش قوانین انجمنی هم‌رخداد [۱۱] با ایجاد یک‌سری تغییرات استخراج می‌کند. این تغییرات شامل استفاده از لغت‌نامه‌ی کلمات مترادف نظر به جای خوشه‌بندی و نیز استفاده از بالاترین اطمینان به‌جای انتخاب خوشه‌ی ویژگی با بالاترین وزن تکرار است.

۲- مطالب اصلی مقاله

۲-۱- طرح مسئله

تشخیص ویژگی صریح: استخراج ویژگی‌های صریح در زبان فارسی با چالش‌هایی روبه‌رو است که این چالش‌ها موجب کاهش دقت و بازخوانی شده است؛ از جمله این چالش‌ها به موارد زیر می‌توان اشاره کرد:

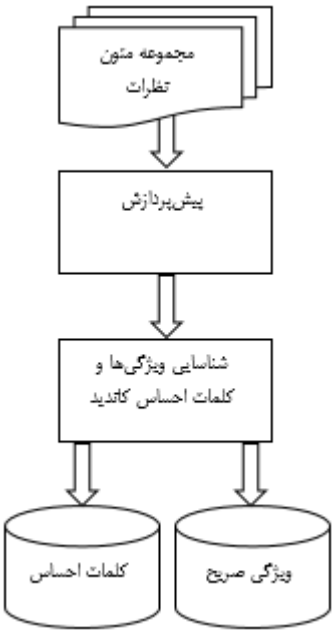
- گاهی اوقات در جملات مشاهده می‌شود که صفت و اسم در کنار یک‌دیگر قرار گرفته و از ترکیب آن‌ها یک اسم ایجاد می‌شود، در این‌صورت صفتی که با اسم ترکیب شده است نباید به عنوان احساس آن اسم تلقی شود. به عنوان مثال در جمله‌ی «سخت افزار آن حرف ندارد»، از ترکیب صفت «سخت» و اسم «افزار»، اسم جدیدی به نام «سخت افزار» ایجاد می‌شود که در این صورت، صفت «سخت» نباید به عنوان احساس جمله و یا ویژگی در نظرگرفته شود. همین‌طور اسم و صفت‌هایی هستند که اگر کنار هم قرار بگیرند، تشکیل یک اسم می‌دهند مثلا در جمله‌ی نظری «سیستم خنک کننده عالی است»، صفت «خنک کننده» نباید به عنوان احساس جمله و یا ویژگی تلقی شود زیرا از ترکیب اسم «سیستم» و صفت «خنک کننده»، اسم جدیدی به نام «سیستم خنک کننده» ایجاد می‌شود.
- در متون نظرات، بعضی از ویژگی‌ها ممکن است شامل چندین کلمه باشند. به عنوان مثال به ویژگی‌های «کیفیت صفحه نمایش» و «عمر باتری» می‌توان اشاره کرد. مدل پیشنهادی باید بتواند این ویژگی‌ها را با دقت خوبی شناسایی و استخراج کند.

در این پژوهش سعی شده که در طی فرایند استخراج ویژگی‌های صریح و کلمات احساس محصول، چالش‌های ذکر شده در بالا حل شوند. در نظر نگرفتن ارتباط ساختاری جملات، یکی دیگر از مشکلات روش‌های موجود است. در کارهای گذشته در زبان فارسی، توجهی به ارتباط ساختاری جملات نکرده و تنها از اطلاعات آماری برای هم‌رخدادی دو کلمه (ویژگی-ویژگی و ویژگی-کلمه‌ی احساس) استفاده می‌کردند که در مدل پیشنهادی سعی در برطرف نمودن این چالش‌ها شده است.

۲-۲- روش پیشنهادی

چارچوب روش پیشنهادی شامل مرحله‌ی اصلی تشخیص ویژگی‌های صریح و احساس هر محصول است.

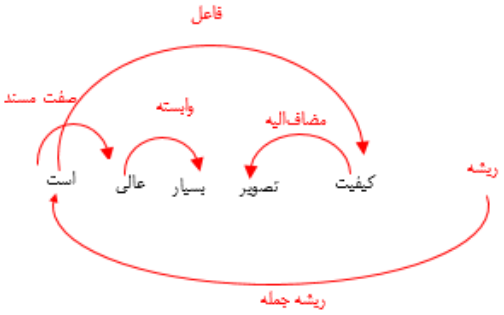
شکل (۱)، روش استخراج ویژگی‌ها و کلمات احساس را در حالت کلی نشان می‌دهد. همان‌طور که در شکل مشخص است، تمامی متون نظرات، قبل از استخراج کلمات احساس و ویژگی‌های محصول بایستی پیش‌پردازش شوند.



شکل (۱) روش کلی جهت استخراج ویژگی‌های صریح و کلمات احساس

۲-۲-۱- پیش‌پردازش

مجموعه نظرات مرتبط با هر محصول به عنوان ورودی این مرحله توسط ابزار ریشه‌یاب دانشگاه مشهد [۲] نرمال‌سازی می‌شوند. سپس، جملات بر اساس علائم نگارشی مانند نقطه «.»، علامت تعجب «!»، علامت سوال «؟» و علامت دو نقطه «:» به یک یا چند جمله شکسته می‌شوند. در مرحله‌ی بعدی از ابزار هضم [۳] جهت ریشه‌یابی، تعیین برچسب لغوی کلمات و ایجاد گراف وابستگی بین کلمات استفاده شده است. برچسب‌زنی هم برای کلمات ریشه‌یابی شده و هم برای کلمات اصلی (کلمات ریشه‌یابی نشده) موجود در جملات انجام می‌پذیرد و در پایگاه داده ذخیره می‌شوند. از آنجایی‌که کارایی ابزار هضم در ایجاد گراف وابستگی جملات طولانی مناسب نیست، جملات طولانی به جملات کوتاه‌تر به طوری شکسته شده‌اند که تنها حاوی یک فعل باشند. به عنوان مثال، جمله‌ی نظری «کیفیت تصویر بسیار عالی است ولی دوربین مناسبی ندارد» را در نظر بگیرید. این نظر به صورت « کیفیت/N تصویر/N بسیار/ADV عالی/AJ است/V ولی/CONJ دوربین/N مناسب/AJ نداشت/V» برچسب زده می‌شود. جهت تعیین نقش نحوی و گراف وابستگی جملات، نظر بالا به دو جمله‌ی «کیفیت/N تصویر/N بسیار/ADV عالی/AJ است/V» و «ولی/CONJ دوربین/N مناسب/AJ نداشت/V» شکسته می‌شوند. نقش نحوی جمله‌ی اول به صورت «کیفیت/SBJ تصویر/MOZ بسیار/APREMOD عالی/MOS است/ROOT» است و گراف وابستگی همان جمله، در شکل (۲) آمده است.



شکل (۲) گراف وابستگی جمله

همان‌طور که در شکل (۲) مشاهده می‌شود، کلمه‌ی «کیفیت» به کلمه‌ی «است» و کلمه‌ی «تصویر» به کلمه‌ی «کیفیت» وابسته است.

۲-۲-۲- شناسایی ویژگی‌ها و کلمات احساس کاندید

پس از پیش‌پردازش جملات، تمامی صفات و اسم‌ها به منظور استخراج کلمات احساسات و ویژگی‌های صریح کاندید هر محصول مورد بررسی قرار می‌گیرند. اسم‌ها و صفات به ترتیب بیان‌گر ویژگی و احساسات هستند که شرایط استخراج هر کدام به‌طور جداگانه بررسی می‌شوند.

برای استخراج ویژگی‌های صریح هر جمله، برچسب کلمه و نقش نحوی آن کلمه مدنظر قرار می‌گیرد. به عنوان مثال برای جمله‌ی «قیمت آن عادلانه است»، نقش نحوی و لغوی آن به ترتیب به صورت «قیمت/N آن/PRO عادلانه/AJ است/V» است (نقش‌های لغوی و نحوی کلمات با علامت «/» از هم جدا شده‌اند). پس برای هر جمله، تمامی کلماتی که دارای شرایط زیر باشند، به لیست ویژگی‌های کاندید اضافه می‌شوند:

۱- کلماتی که دارای نقش نحوی «sbj» و «moz» و برچسب «N» و یا «res» هستند به لیست ویژگی‌های کاندید اضافه می‌شوند (لازم به ذکر است که ابزار هضم، اسم‌های فارسی مثل «صفحه»، «باتری» و غیره را با برچسب «N» و اسم‌های انگلیسی مثل «nfc»، «ram » و غیره را با برچسب «res » مشخص می‌کند.).

۲- کلماتی هم که دارای نقش نحوی «sbj» و برچسب «aj» باشند به لیست ویژگی‌های کاندید اضافه می‌شوند. ریشه‌ی کلماتی مثل «کارایی» و «زیبایی» به ترتیب به صورت «کارا» و «زیبا» است که به دلیل این‌که دارای برچسب «aj» هستند، نمی‌توانند صفت در نظر گرفته شوند. به عنوان مثال، ریشه‌ی جمله‌ی «کارایی این گوشی قابل قبول است.» به صورت «کارا این گوشی قابل قبول است.» است که در این‌جا کلمه‌ی «کارا» به ترتیب دارای نقش نحوی و لغوی، «sbj» و «aj» است.

۳- کلماتی هم که دارای نقش نحوی «posdep» و برچسب «N» و یا «res» هستند به شرطی به لیست ویژگی‌های کاندید اضافه می‌شوند که بعد از حرف ربط «و» آمده باشند و حرف ربط هم دارای نقش نحوی «NCONJ» باشد. به عنوان مثال، در جمله‌ی «فاقد/P/MOS فوکوس/N/NEZ و/CONJ/NCONJ فلاش/N/POSDEP هست/V/ROOT» کلمه‌ی «فلاش» به لیست ویژگی‌های کاندید اضافه می‌شود زیرا دارای نقش نحوی «POSDEP» و برچسب «N» است و بعد از حرف ربط «و» آمده است که این حرف دارای نقش نحوی «NCONJ» است.

پس از استخراج ویژگی‌های کاندید محصول، صفات کاندید استخراج می‌شوند. در صورتی‌که کلمه‌ای در جمله‌ی ریشه‌یابی شده صفت و در جمله‌ی اصلی (جمله‌ای که عملیات ریشه‌یابی انجام نشده است) هم صفت باشد، مشابه با استخراج ویژگی‌های محصول نقش نحوی آن کلمه نیز مدنظر قرار می‌گیرد. اگر آن کلمه دارای شرایط زیر باشد، به لیست صفات کاندید اضافه می‌شود:

۱- نقش نحوی کلمه، «mos»، «npostmod»، «obj» و یا «adv» باشد.

۲- اگر نقش نحوی آن «root» بود باید بررسی شود که کلمه‌ی مورد بررسی در آخر جمله باشد، زیرا بعضی از جملات دارای فعل نیستند. به عنوان مثال، در جمله‌ی «خیلی از صفحه نمایش این گوشی راضیم.»، چون حاوی هیچ فعلی نیست، کلمه‌ی «راضیم» توسط ابزار به عنوان ریشه‌ی جمله تشخیص داده می‌شود.

۳- کلماتی هم که دارای نقش نحوی «posdep» هستند به شرطی به لیست صفت‌های کاندید اضافه می‌شوند که بعد از حرف ربط «و» آمده باشند و این حرف دارای نقش نحوی «AJCONJ» باشد. به عنوان مثال، در جمله‌ی «تاچ/Ne/MOS عالی CONJ/AJCONJ/وAJ/NPOSTMOD/ روان/AJ/POSDEP است/V/VCl» کلمه‌ی «روان» بنابر دلیل ذکر شده به لیست صفات کاندید اضافه می‌شود.

در صورتی که کلمه‌ای در جمله‌ی ریشه‌یابی شده، صفت و در جمله‌ی بدون ریشه‌یابی شده، اسم تشخیص داده شده باشد، اسم و یا صفت بودن آن

کلمه نیز بررسی می‌شود. اگر آن کلمه دارای یکی از شرایط زیر باشد، صفت در غیر این‌صورت، ویژگی در نظر گرفته می‌شود.

- اگر کلمه‌ی کاندید بعد از اسم و قبل از فعل بیاید
- اگر کلمه‌ی کاندید بعد از قید آمده باشد و بعد از کلمه‌ی کاندید صفتی نباشد.
- اگر کلمه‌ی کاندید با پسوند «یی» در جمله‌ی اصلی نیامده باشد (مثل زیبایی).

ممکن است نقش نحوی برخی از کلمات، به علت خطای ابزار صحیح نباشد. به همین دلیل از لیست ویژگی‌ها و صفات کاندید بدست آمده برای بررسی استخراج ویژگی‌ها یا صفات کاندید بیشتر استفاده می‌شود. به عنوان مثال، فرض کنید که سه جمله‌ی زیر در پایگاه داده‌ی نظرات موجود است:

- «کیفیت خیلی خوبی دارد.»
- «کیفیت صفحه عالی است.»
- «صفحه نمایشش چندان هم خوب نبود.»

اگر از جمله اول، ویژگی کاندید «کیفیت» و از جمله‌ی دوم، ویژگی‌های کاندید «کیفیت» و «صفحه» و از جمله‌ی سوم هم ویژگی کاندید «نمایش» استخراج و به لیست ویژگی‌های کاندید مرتبط با هر جمله اضافه شده باشند ولی ویژگی «صفحه» از جمله‌ی سوم استخراج نشده باشد، با استفاده از لیست ویژگی‌های کاندید، این کلمه استخراج شده و به لیست ویژگی‌های کاندید این جمله اضافه می‌شود. زیرا ویژگی «صفحه نمایش» یک ویژگی معنادار است که بعدا در مرحله‌ی استخراج ویژگی‌های چندکلمه‌ای ایجاد می‌شود. در گام بعدی، تمامی کلمات مانع[††] از لیست صفات و ویژگی‌های کاندید حذف می‌شوند.

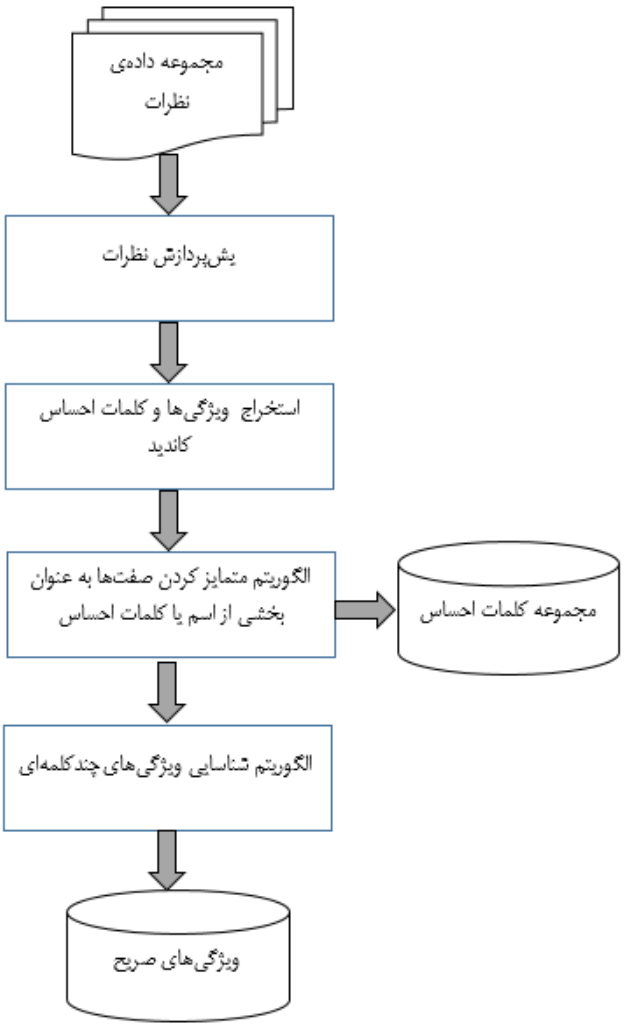
۳-۲-۲- استخراج ویژگی‌های صریح و کلمات احساس

پس از استخراج ویژگی‌ها و صفات کاندید، بایستی چالش‌هایی که در طی فرایند استخراج ویژگی صریح و کلمات احساس اتفاق می‌افتند، بررسی و حل شوند. شکل (۳)، گام‌های اصلی به منظور حل این چالش‌ها و تکمیل فرایند استخراج ویژگی صریح و کلمات احساس را نشان می‌دهد. مرحله‌ی استخراج ویژگی‌های صریح و کلمات احساس در بخش‌های زیر نسبت به کارهای قبلی متمایز است:

۱- الگوریتم متمایز کردن صفت‌ها به عنوان بخشی از اسم یا کلمات احساس
۲- الگوریتم شناسایی ویژگی‌های چندکلمه‌ای

الگوریتم متمایز کردن صفت‌ها به عنوان بخشی از اسم یا کلمات احساس

در زبان فارسی ممکن است که صفتی که قبل از اسم ظاهر می‌شود، بخشی از اسم بوده و بیان‌گر احساس جمله و یا ویژگی نباشد. به عنوان مثال، صفت «سخت» در جمله‌ی «سخت افزار این گوشی حرف ندارد.»، بیان‌گر احساس جمله یا ویژگی نیست و بخشی از اسم است. ولی در جمله‌ی «در کل اگر گوشی ارزان قیمت و کارآمد می‌خواهید بخرید، توی خرید این گوشی شک نکنید.»، صفت «ارزان» به عنوان احساس ویژگی «قیمت» است و نمی‌تواند با ویژگی «قیمت» ترکیب شود. گاهی مواقع هم ممکن است که صفت بخشی از اسم بوده ولی بیان‌گر احساس ویژگی و یا جمله باشد و با هم تشکیل کلمه‌ی احساس را بدهند.



شکل (۳) روش پیشنهادی بدون ناظر تشخیص ویژگی‌های صریح و کلمات احساس

به عنوان مثال، در جمله‌ی «باتری این گوشی کم مصرف است»، از ترکیب صفت «کم» با اسم «مصرف»، صفت «کم مصرف» ایجاد می‌شود که بیان‌گر احساس ویژگی «باتری» است. هدف اصلی در این قسمت، تفکیک صفت‌هایی است که با اسم ترکیب شده و تشکیل ویژگی کاندید می‌دهند و یا به عنوان احساس جمله و یا ویژگی به کار می‌روند. بدین منظور، برای هر کدام از جملات با توجه به لیست ویژگی‌ها و صفات کاندید، ترکیب «صفت+اسم» را در نظر می‌گیرد و اگر ترکیب به‌دست آمده در دو شرط زیر صدق کند، با هم ترکیب می‌شوند:

- ترکیب صفت+ اسم در پیکره‌ی نظرات زیاد رخ داده باشد.
- مقدار معیار LRT [۹] این دو از حدآستانه‌ی ∂ بیشتر باشد. (LRT وابستگی بین دو واژه را در پیکره‌ی نظرات محاسبه می‌کندهر چقدر که مقدار LRT بزرگ‌تر باشد، نشان می‌دهد که دو واژه بیشتر به‌هم وابسته هستند.)

†† Stop word

اگر ترکیب به‌دست آمده شرط زیر را برآورده کند، تشکیل ویژگی می‌دهد و صفت مربوطه از لیست صفات کاندید آن جمله حذف شده و مقدار ویژگی قبلی با مقدار جدید جایگزین می‌شود:

- اگر صفت یا قید مقداری(خیلی، بسیار و غیره) بعد از ترکیب این دو و قبل از فعل در پیکره‌ی نظرات آمده باشد.

در غیر این‌صورت، اگر این ترکیب شرایط زیر را برآورده کند تشکیل صفت می‌دهد و ویژگی مربوطه از لیست ویژگی‌های کاندید حذف شده و مقدار صفت با مقدار جدید جایگزین می‌شود:

- اگر بعد از این ترکیب این دو در پیکره‌ی نظرات، صفت یا قیدی نباشد.
- اگر قبل از ترکیب این دو، قید مقدار وجود داشته باشد.

به عنوان مثال، برای جمله‌ی «سیستم‌عامل کم مصرف است»، اگر صفت «کم» در لیست صفات کاندید و ویژگی «مصرف» در لیست ویژگی‌های کاندید باشد و دو شرط ترکیب صفت و اسم را برآورده کند، با هم ترکیب می‌شوند. اگر مقدار ترکیبی در دو شرط فوق صدق کند در این‌صورت، صفت «کم» با «کم مصرف» جایگزین و ویژگی «مصرف» از لیست ویژگی‌های کاندید آن جمله حذف می‌شود.

گاهی اوقات ممکن است که ترکیب «اسم+ صفت»، تشکیل اسم دهد که برای تشخیص چنین حالت‌هایی ، با توجه به لیست ویژگی‌ها و صفات کاندید هر جمله، ترکیب «اسم+صفت» در نظرگرفته می‌شود و اگر این ترکیب سه شرط زیر را برآورده کند، صفت مربوطه از لیست صفات کاندید حذف شده و ویژگی جدید با اسم قبلی جایگزین می‌شود:

- ترکیب صفت+ اسم زیاد در پیکره‌ی نظرات رخ داده باشد.
- مقدار معیار LRT این دو از حدآستانه‌ی λ بیشتر باشد.
- اگر صفت یا قید مقدار بعد از ترکیب این دو و قبل از فعل در پیکره‌ی نظرات آمده باشد.

در این مرحله، امکان استخراج ویژگی‌های چندکلمه‌ای مانند «عمر باتری» و «صفحه نمایش» و غیره وجود ندارد که در گام بعدی، سعی در حل این چالش شده است.

الگوریتم شناسایی ویژگی‌های چندکلمه‌ای

در متون نظرات، بعضی از ویژگی‌ها ممکن است شامل چندکلمه باشند. به عنوان مثال «کیفیت صفحه نمایش» و «عمر باتری» نمونه‌هایی از این ویژگی‌ها هستند. ویژگی‌های کاندید از هر جمله در صورتی یک ویژگی چندکلمه‌ای تشکیل می‌دهند که یکی از شرایط زیر را برآورده کنند:

۱- اگر چندکلمه در گراف وابستگی به هم وابسته باشند. به عنوان مثال در جمله‌ی «عمر باتریش بسیار عالی است» ویژگی‌های «عمر» و «باتری» در لیست ویژگی‌های کاندید وجود دارند و در گراف وابستگی این جمله، این دو کلمه به هم وابسته هستند. در نتیجه «عمر باتری» به لیست ویژگی‌های کاندید اضافه و دو ویژگی تک‌کلمه‌ای «عمر» و «باتری» از لیست ویژگی‌های کاندید این جمله حذف می‌شوند.

۲- ممکن است تشخیص بعضی از ویژگی‌های چندکلمه‌ای توسط گراف وابستگی امکان‌پذیر نباشد؛ در این‌صورت از معیار LRT برای تشخیص ویژگی‌های چندکلمه‌ای استفاده می‌شود. اگر مقدار معیار LRT برای دو ویژگی کاندید از حد آستانه‌ی λ بیشتر باشد، در این‌صورت با هم ترکیب می‌شوند.

ممکن است الگوریتم استخراج ویژگی‌های چندکلمه‌ای، برخی از ویژگی‌های چندکلمه‌ای موجود در جملات را درست تشخیص نداده باشد، برای این منظور تمامی جملات مجددا بررسی می‌شوند تا در صورتی که ویژگی چندکلمه‌ای موجود در آن‌ها تشخیص داده نشده است، با استفاده از این لیست تشخیص داده شود و به لیست ویژگی‌های کاندید مرتبط با آن جمله اضافه شود.

همچنین، ویژگی‌هایی که تعداد تکرار کمتر یا مساوی دو دارند در صورتی‌که جزء ویژگی‌های چندکلمه‌ای نباشند از لیست ویژگی‌های کاندید حذف می‌شوند.

۳-۲- ارزیابی

در این مقاله، از مجموعه نظرات جمع‌آوری شده در مقاله‌ی باباعلی [۱] (که این مجموعه نظرات از سایت digikala جمع‌آوری شده است) جهت ارزیابی سیستم استفاده شده است. این مجموعه نظرات، مربوط به سه نوع گوشی و لپ‌تاپ می‌باشد. در جدول (۱) اطلاعات مربوط به هر کالا به تفکیک گروه ارائه شده است.

جدول (۱) اطلاعات آماری مجموعه نظرات

گروه کالا	تعداد نظرات	تعداد جملات
تلفن همراه	۳۹۴	۱۸۴۷
لپ‌تاپ	۱۶۸	۶۶۵
تعداد کل	۵۶۲	۲۵۱۲

جهت ارزیابی کارایی روش پیشنهادی در استخراج ویژگی‌های صریح همانند کارهای قبلی از معیارهای دقت، بازخوانی و معیار F استفاده شده است. برای این‌کار، ویژگی‌های صریح استخراج شده توسط روش پیشنهادی با ویژگی‌هایی که برای هر جمله به صورت دستی مشخص شده، مقایسه و معیارهای مورد نظر محاسبه شده است. در آزمایش‌های انجام شده مقدار حد آستانه‌ی ∂ گفته شده در بخش قبل برابر ۲,۵ در نظر گرفته شده است.

مقادیر معیارهای محاسبه شده برای روش پیشنهادی در تشخیص ویژگی صریح در مقایسه با نتایج بدست آمده در روش ارائه شده توسط هو [۱۴] و باقری [۵] به‌ترتیب در جدول (۲) و (۳) نشان داده شده است. روش باقری و همکارش [۵] که برای زبان انگلیسی ارائه شده است و قادر به استخراج ویژگی‌هایی مانند «سیستم خنک کننده» نیست.

جدول (۲) مقادیر معیارهای ارزیابی برای تشخیص ویژگی صریح برای روش پیشنهادی (روش ۱) در مقایسه با روش هو [۱٤] (روش ۲)

گروه کالا	دقت		بازخوانی		معیار F	
	روش ۱	روش ۲	روش ۱	روش ۲	روش ۱	روش ۲
تلفن همراه	۰,۸۵	۰,۷۰	۰,۸۸	۰,۷۲	۰,۸۶	۰,۷۳
لپ‌تاپ	۰,۸۰	۰,۶۹	۰,۸۷	۰,۷۰	۰,۸۳	۰,۷۰
تعداد کل	۰,۸۳	۰,۷۰	۰,۸۸	۰,۷۱	۰,۸۵	۰,۷۲

جدول (۳) مقادیر معیارهای ارزیابی برای تشخیص ویژگی صریح برای روش پیشنهادی (روش ۱) در مقایسه با روش باقری [۵] (روش ۲)

گروه کالا	دقت		بازخوانی		معیار F	
	روش ۱	روش ۲	روش ۱	روش ۲	روش ۱	روش ۲
تلفن همراه	۰٫۸۵	۰٫۷۵	۰٫۸۸	۰٫۸۳	۰٫۸۶	۰٫۸۵
لپ‌تاپ	۰٫۸۰	۰٫۷۲	۰٫۸۷	۰٫۸۱	۰٫۸۳	۰٫۸۰
تعداد کل	۰٫۸۳	۰٫۷۴	۰٫۸۸	۰٫۸۲	۰٫۸۵	۰٫۸۳

۳- نتیجه گیری

در این مقاله، روشی جهت استخراج ویژگی‌های صریح مجموعه نظرات کاربران در زبان فارسی ارائه شده است. در روش پیشنهادی از گراف وابستگی و قواعد نحوی زبان فارسی جهت تشخیص ویژگی‌های صریح و کلمات احساس و حل چالش‌های مربوط به ویژگی صریح استفاده شده است که باعث شده ویژگی‌های صریح با دقت بیشتری استخراج شوند. از جمله مسائلی که در ادامه‌ی مسیر این پژوهش پیشنهاد می‌شود، بهبود مباحث برچسب‌گذاری، ریشه‌یابی است که در نتایج نظرکاوی، تأثیر مستقیمی دارد. از موارد دیگر به کاوش زمانی در متون نظری می‌توان اشاره کرد. با توجه به تغییر نظرات کاربران در طول زمان، پیگیری آن‌ها از طریق نظرکاوی، زمینه‌ی پژوهشی جذابی در بین بسیاری از سازمان‌ها یا شرکت‌ها است.

مراجع

[۱] م. باباعلی و م. نمعت‌بخش، **" استخراج ویژگی‌های محصول در زبان فارسی،"** سومین همایش زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، ۱۳۹۳.

[۲] ابزارهای پردازش زبان‌های طبیعی، آزمایشگاه فناوری وب دانشگاه فردوسی مشهد (wtlab.um.ac.ir)، ۱۳۹۱.

[۳] م. ایمانی و م. خلاش، ابزار پردازش زبان فارسی، (http://www.sobhe.ir/hazm)، 1392.

[4] E. Lloret, A. Balahur, J. M. Gómez, A. Montoyo, and M. Palomar, "*Towards a unified framework for opinion retrieval, mining and summarization,*" Journal of Intelligent Information Systems, vol. 39, pp. 711-747, 2012.

[5] 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 Systems, vol. 52, pp. 201-213, 2013.

[6] B. Liu, "*Sentiment analysis and subjectivity*," Handbook of natural language processing, vol. 2, pp. 627-666, 2010.

[7] B. Liu, "*Sentiment analysis and opinion mining*," Synthesis Lectures on Human Language Technologies, vol. 5, pp. 1-167, 2012.

[8] G. Vinodhini and R. Chandrasekaran, "*Sentiment analysis and opinion mining: a survey*," International Journal, vol. 2, 2012.

[9] Z. Hai, K. Chang, and G. Cong, "*One seed to find them all: mining opinion features via association*," in Proceedings of the 21st ACM international conference on Information and knowledge management, 2012, pp. 255-264.

[10] H. Xu, F. Zhang, and W. Wang, "*Implicit feature identification in Chinese reviews using explicit topic mining model*," Knowledge-Based Systems, vol. 76, pp. 166-175, 2015.

[11] Z. Hai, K. Chang, and J.-j. Kim, "*Implicit feature identification via co-occurrence association rule mining*," in Computational Linguistics and Intelligent Text Processing, ed: Springer, 2011, pp. 393-404.

[12] C. Quan and F. Ren, "*Unsupervised product feature extraction for feature-oriented opinion determination*," Information Sciences, vol. 272, pp. 16-28, 2014.

[13] S. Poria, E. Cambria, L.-W. Ku, C. Gui, and A. Gelbukh, "*A rule-based approach to aspect extraction from product reviews*," in Proceedings of the Second Workshop on Natural Language Processing for Social Media (SocialNLP), 2014, pp. 28-37.

[14] M. Hu and B. Liu, "*Mining opinion features in customer reviews*," in AAAI, 2004, pp. 755-760.