تشخیص ویژگی‌های ضمنی با استفاده از قواعد نحوی زبان فارسی و خوشه‌بندی صفات

عاطفه محمدی[۱]، دانشجوی دکتری؛ مهدی یزدیان دهکردی[۲]، استادیار؛ محمدعلی نعمت‌بخش[۳]، دانشیار

۱- گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران، atefehmohammadi@stu.yazd.ac.ir
۲- گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران، yazdian@yazd.ac.ir
۳- دانشکده مهندسی کامپیوتر، دانشگاه اصفهان، اصفهان، ایران، nematbakhsh@eng.ui.ac.ir

چکیده: به‌طور معمول وقتی فردی قصد خرید یک محصول برخط را دارد، نظرات و یادداشت‌های نوشته‌شده توسط سایر افراد در مورد محصول را بررسی می‌کند و این مسئله تاثیر بسزایی در تصمیم‌گیری فرد برای خرید دارد. تجزیه و تحلیل نظرات کاربران که با عنوان تجزیه و تحلیل احساس یا نظرکاوی شناخته می‌شود، یکی از داغ‌ترین موضوعات تحقیقاتی در علم کامپیوتر است. هدف اصلی نظرکاوی، استخراج نظرات افراد درباره‌ی ویژگی‌های یک موجودیت یا کالا است. در این تحقیق، یک راه‌کار نظرکاوی بدون نظارت و مبتنی‌بر استخراج ویژگی‌های ضمنی برای محصولات در زبان فارسی ارائه شده است. استخراج ویژگی‌های ضمنی یکی از مراحل دشوار در تحلیل احساسات مبتنی بر ویژگی می‌باشد. در بیشتر پژوهش‌های پیشین از اطلاعات آماری برای ایجاد ماتریس هم‌رخداد و سپس تشخیص ویژگی‌های ضمنی استفاده شده است. در این پژوهش در کنار اطلاعات آماری، از قواعد نحوی زبان و خوشه‌بندی صفات جهت بهبود ماتریس هم‌رخداد بین کلمات (ویژگی‌ها و احساسات) بهره گرفته شده است. ارزیابی های انجام‌شده بر روی داده‌های واقعی و استخراج‌شده از نظرات کاربران در سایت دیجی‌کالا، نشان می‌دهند که روش ارائه‌شده در این مقاله به نرخ بازخوانی و دقت بهتری نسبت به کارهای قبلی دست یافته است.

واژه‌های کلیدی: نظرکاوی، ویژگی ضمنی، ماتریس هم‌رخداد، خوشه‌بندی.

Identification of Implicit Features using Persian Language Rules and Sentiments Clustering

A. Mohammadi[1], PhD Student; M. Yazdian-Dehkordi[2], Assistant Professor;
M. Nematbakhsh[3], Associated Professor

1- Faculty of Engineering, Yazd University, Yazd, Iran, Email: atefehmohammadi@stu.yazd.ac.ir
2- Faculty of Engineering, Yazd University, Yazd, Iran, Email: yazdian@yazd.ac.ir
3- Faculty of Computer Engineering, University of Isfahan, Isfahan, Iran, Email: nematbakhsh@eng.ui.ac.ir

Abstract: Typically, when someone wants to buy an online product, he/she reviews the comments and notes written by others about the product. Clearly, it has a profound impact on the person's decision to buy or not to buy the product. Sentiment analysis or opinion mining is one of the hot research topics in computer science. The main purpose of sentiment analysis is to extract opinions of individuals about the characteristics of an entity such as a product. In this research, an unsupervised opinion mining is proposed for Persian products based on implicit feature extraction which is a critical step in sentiment analysis. In most previous studies, statistical information is utilized to create a co-occurrence matrix and determine the implicit features. In this paper, we benefit from syntactic rules and sentiment clustering in conjunction with statistical information to construct an efficient co-occurrence matrix between features and sentiment words. The evaluation results provided on a real-world dataset, extracted from Digikala website, indicates that the proposed method achieves higher recall and precision compared to the previous studies.

Keywords: Opinion mining, implicit feature, co-occurrence matrix, clustering.

تاریخ ارسال مقاله: ۱۳۹۷/۰۵/۰۶
تاریخ اصلاح مقاله: ۱۳۹۷/۰۹/۲۵و ۱۳۹۷/۱۱/۱۳
تاریخ پذیرش مقاله: ۱۳۹۷/۱۱/۲۰
نام نویسنده مسئول: مهدی یزدیان دهکردی
نشانی نویسنده مسئول: ایران، یزد، صفائیه، بلوار دانشگاه، دانشگاه یزد، پردیس فنی مهندسی، گروه مهندسی کامپیوتر.

۱- مقدمه

با رشد چشم‌گیر محبوبیت وب، تعداد نظرات برخط در مورد محصولات مختلف با سرعت بالایی در حال افزایش است. تعداد زیادی از وبلاگ‌ها، تالارهای پرسمان و وب سایت‌هایی مانند دیجی‌کالا به کاربران اجازه می‌دهند تا نظراتشان را در مورد محصولات و سرویس‌های مختلف بیان کنند. استخراج دانش از نظرات ارائه شده در مورد محصول توسط یک سیستم نظرکاوی [۱، ۲] انجام می‌شود و پیشنهاد یک محصول مناسب به کاربر توسط یک سیستم توصیه‌گر [۳، ۴] انجام می‌شود.

نظرکاوی، تولیدکنندگان را قادر می‌سازد تا بتوانند ارزیابی مناسبی از رضایت کاربران داشته و با توجه به آن محصولات خود را ارتقا دهند [۱]. از آن‌جایی که بررسی تمامی نظرات برای کاربران امکان‌پذیر نیست، ارائه روش‌هایی جهت استخراج دانش از نظرات و ارائه آن به صورت خلاصه برای کاربران اهمیت زیادی پیدا کرده است.

از این‌رو، بسیاری از محققان، الگوریتم‌های موثری را برای کشف دانش از نظرات، تحت عنوان نظرکاوی[۱]، ارائه کرده‌اند. امروزه تجزیه و تحلیل احساس[۲] (که نظرکاوی هم نامیده می‌شود)، مطالعه‌ی نظرات، احساسات و نگرش‌های افراد درباره‌ی موجودیت‌هایی مانند محصولات و سرویس‌ها است.

تجزیه و تحلیل احساس در سه سطح سند، جمله و ویژگی مورد بررسی قرار می‌گیرد [۲، ۵، ۶]. در تجزیه و تحلیل احساس در سطح سند، گرایش کلی سند با فرض این‌که هر سند حاوی نظراتی روی یک موجودیت منفرد است، به صورت مثبت یا منفی تعیین می‌شود. تجزیه و تحلیل احساس در سطح جمله[۳]، با فرض این‌که هر جمله، دارای یک ویژگی یا موجودیت است گرایش هر جمله، اعم از مثبت، منفی یا خنثی را مشخص می‌کند. در صورتی‌که، یک جمله ممکن است حاوی نظراتی درباره‌ی چندین ویژگی از یک موجودیت باشد. تجزیه و تحلیل احساس در سطح سند و جمله به‌طور دقیق مشخص نمی‌کنند که افراد به چه مشخصه‌ای از محصول تمایل و یا عدم تمایل بیشتری دارند. در تجزیه و تحلیل احساس در سطح ویژگی یا موجودیت، هر نظر شامل احساس مثبت یا منفی و مجموعه‌ای از ویژگی‌ها است. درک اهمیت هدف‌های حاوی نظر به درک بهتر مسئله‌ی تجزیه و تحلیل احساس کمک می‌کند. به‌عنوان مثال، جمله‌ی نظری «گرچه ظاهرش جالب نبود، اما من هنوز عاشق این گوشی هستم.»، دارای لحنی مثبت است؛ اما نمی‌توانیم بگوییم که این جمله به‌طور کامل دارای گرایش مثبت است. در واقع جمله دارای گرایش مثبت برای «گوشی» و دارای گرایش منفی برای «ظاهر گوشی» است. هدف اصلی تجزیه و تحلیل احساس در سطح ویژگی، کشف احساسات مربوط به موجودیت‌ها و یا ویژگی‌های آن‌ها است [۷].

روش‌های تشخیص ویژگی به‌طور کلی به دو دسته‌ی بانظارت[۴] و بدون نظارت[۵] تقسیم می‌شوند [۷، ۸]. روش‌های بانظارت، روش‌هایی هستند که در آن‌ها، مجموعه داده‌ی آموزشی دارای برچسب صحیح و از قبل مشخص هستند. از جمله این تکنیک‌ها به بیز ساده، ماشین برداری پشتیبان[۶] و حداکثر انتروپی[۷] می‌توان اشاره کرد. از آن‌جایی که مسئله‌ی استخراج بانظارت ویژگی از متون، یک مسئله‌ی طبقه‌بندی است؛ این مدل‌ها برای استخراج ویژگی‌های صریح[۸] به‌کار می‌رود. در ابتدا، مجموعه‌ای از برچسب‌ها با داشتن مجموعه‌ای از کلمات در یک جمله تهیه می‌شوند که ویژگی یا ویژگی نبودن یک کلمه را مشخص می‌کند. در گام بعدی، فرایند برچسب‌گذاری واژگان بر روی جملات انجام می‌شود. پس از ایجاد این مجموعه داده، مدل حداکثر آنتروپی به‌عنوان مثال آموزش داده می‌شود و بر روی داده‌های بدون برچسب آزمایش می شود. هدف از به‌کارگیری مجموعه‌ی آموزشی، تشخیص و استخراج ویژگی‌های مناسب توسط سیستم است [۹، ۱۰]. هرچند که تهیه‌ی این مجموعه داده وظیفه‌ای سخت و طاقت‌فرسا و نیازمند صرف زمان زیادی است اما در عوض نتایج تولیدشده توسط این روش از دقت بیشتری برخوردار هستند. مدل‌های بانظارت وابسته به دامنه بوده و در صورت انتقال به دامنه‌ی دیگر بایستی مجددا آموزش ببینند. مدل‌های بدون نظارت، نیاز به مجموعه داده‌ی آموزشی برچسب‌خورده ندارند و در مقایسه با روش‌های بانظارت از دقت کمتری برخوردار هستند [۷، ۸].

ویژگی‌ها می‌توانند به دو روش صریح و ضمنی ظاهر شوند. اسم یا عبارات اسمی که به‌طور صریح در جملات رخ می‌دهند، ویژگی صریح نامیده می‌شوند. به‌عنوان مثال در جمله‌ی «کیفیت ساخت خوبی دارد.»، «کیفیت ساخت» به عنوان ویژگی صریح استخراج می‌شود. ویژگی که اسم یا عبارت اسمی نباشد را ویژگی ضمنی[۹] می‌نامند که به‌طور معمول به شکل صفت یا قید ظاهر می‌شود. به‌عنوان مثال در جمله‌ی «این گوشی خیلی گران است.»، به‌طور ضمنی به ویژگی «قیمت» اشاره دارد.

در زمینه تشخیص ویژگی‌های صریح کارهای زیادی نسبت به ویژگی‌های ضمنی انجام شده است. متاسفانه طبق بررسی‌های ما، در زبان فارسی تعداد کارهای انجام شده بر روی ویژگی‌های ضمنی انگشت‌شمار بوده است. در داده‌های واقعی وب، وجود ویژگی‌های ضمنی در نظرات کاربران غیرقابل اجتناب است. بنابراین انجام پژوهش در حوزه ویژگی‌های ضمنی نیز از اهمیت زیادی برخوردار است. در روش‌های استخراج ویژگی ضمنی با توجه به عدم وجود ویژگی در جملات نسبت به روش استخراج ویژگی صریح دارای چالش بیشتری هستند. در این پژوهش، برای تشخیص ویژگی‌های ضمنی از ماتریس هم‌رخداد[۱۰] استفاده می‌شود. در پژوهش‌های پیشین از اطلاعات آماری برای ایجاد ماتریس هم‌رخداد استفاده شده است. در این پژوهش برای استفاده از اطلاعات آماری، علاوه‌بر معیار LRT، از معیار فاصله‌ی ویژگی صریح از کلمه‌ی احساس، قواعد نحوی زبان، گراف وابستگی جملات و

خوشه‌بندی صفات، برای ایجاد کاراتر ماتریس هم رخداد استفاده شده است؛ تا با به‌کارگیری این ماتریس بتوان ویژگی‌های ضمنی مناسب‌تری را استخراج نمود. برای ارزیابی روش ارائه شده، از داده‌های واقعی که از سایت معروف و پربازدید فروشگاه اینترنتی دیجی‌کالا استخراج شده است، استفاده گردید. نتایج به‌دست آمده بر روی این داده‌ها نشان می دهند که روش ارائه شده توانسته است با استخراج ویژگی‌های ضمنی مناسب‌تر، کارایی بهتری نسبت به روش‌های پیشین به همراه داشته باشد.

در ادامه‌ی پژوهش، در بخش ۲ مروری بر تاریخچه موضوع و کارهای پیشین صورت‌گرفته است. در بخش ۳، به ارائه روش پیشنهادی و مراحل انجام کار پرداخته‌شده و در بخش ۴، نتایج ارزیابی روش پیشنهادی بیان خواهند شد و در نهایت در بخش ۵، جمع بندی و کارهای آینده شرح داده می‌شوند.

۲- مروری بر کارهای پیشین

همان‌گونه که در بخش مقدمه عنوان شد، روش‌های استخراج ویژگی در سیستم‌های نظرکاوی را می‌توان به دو دسته عمده شامل روش‌های بانظارت و بدون نظارت تقسیم‌بندی کرد.

در روش بانظارت داده‌های آموزش دارای برچسب هستند و به‌طور معمول از طبقه‌بندی‌کننده[11] برای ساخت سیستم استفاده می‌شود.

شیخ حسنی و منصوری‌زاده [۱۱] با استفاده از میدان تصادفی شرطی (CRF)[12]، جنبه‌های نهفته در متن نظرات را استخراج کرده‌اند. با توجه به انتخاب مدل میدان تصادف شرطی، آن‌ها هر جمله از نظر را به کلمات تشکیل‌دهنده‌اش بخش‌بندی کرده و برای کلماتی غیر از کلمات لیست ایست‌واژه[13] و لیست غیر الفبایی، یک بردار ویژگی ساختند. به این صورت که بر اساس مجموعه دادگان برچسب‌خورده، کلاس هر کلمه را مشخص کرده و به‌دنبال آن، ویژگی‌های استخراج‌شده را قرار داده‌اند. سپس، کلمه‌های هر جمله را با یک خط خالی از هم جدا نموده و مجموعه داده‌های آموزشی را به‌عنوان بردار ورودی به طبقه‌بندی‌کننده CRF داده‌اند.

در سال ۲۰۱۷ اصغر و همکاران [۱۲]چارچوبی پیشنهاد داده‌اند که شامل چهار ماژول اصلی استخراج ویژگی- احساس، گروه‌بندی ویژگی، دسته‌بندی ویژگی- احساس و خلاصه‌سازی ویژگی است. آن‌ها تعدادی الگوی جدید ارائه داده‌اند که در آن اسم، عبارات اسمی و فعل‌ها، بیان‌گر ویژگی‌های کاندید هستند. پس از استخراج ویژگی‌ها، گروه‌بندی ویژگی‌ها را بر اساس اندازه‌گیری شباهت معنایی پیشنهاد داده‌اند. به‌منظور انتساب امتیاز به نظرات داده‌شده در باره‌ی ویژگی های استخراج‌شده، یک تکنیک ترکیبی برای امتیازدهی احساسات پیشنهاد داده شد که مفاهیم مبتنی بر پیکره و مبتنی بر لغت را ترکیب و در نهایت نقاط مثبت و منفی هر ویژگی را بیان می‌کند.

لیانگ و همکاران [۱۳] در سال ۲۰۱۷ از یادگیری عمیق که یک روش جدید و موفق در حوزه متن‌کاوی بوده، برای استخراج ویژگی‌های متنی استفاده کرده‌اند. تفاوت یادگیری عمیق نسبت به روش‌های مرسوم در این حوزه این است که یادگیری عمیق برای یادگیری ویژگی‌ها نیازمند داده‌های زیاد است.

عسگریان و همکاران [۱۴]، یک wordnet فارسی کامل، با دقت و بازخوانی مناسب ارائه داده‌اند. آن‌ها با به‌کارگیری روش‌های انتخاب ویژگی شناخته شده مانند CHI[14]، IG[15]، MI[16]، ۲۰۰۰ ویژگی با بالاترین امتیاز را استخراج کرده‌اند و سپس با کمک روش‌های یادگیری ماشین، یک طبقه‌بندی احساس برای نظرات متنی فارسی ارائه داده‌اند. در برخی از روش‌ها از روش یادگیری نیمه‌نظارتی نیز استفاده شده است.

روش‌های بانظارت نیازمند داده‌های آموزشی برچسب‌دار هستند که تهیه آن‌ها در داده‌های حجیمی مانند وب بسیار پرهزینه است. به‌علاوه این روش‌ها محدود به دامنه نوع محصولات هستند. در مقابل روش‌های بانظارت، روش‌های بدون نظارت برای استخراج ویژگی نیز ارائه شده‌اند.

در سال ۲۰۱۸، دراگونی [۱۵] برای استخراج ویژگی از یک پیکره‌ی نظرات بزرگ استفاده کرده‌اند که شامل ۳۵ میلیون نظر است. در این روش برای استخراج ویژگی‌ها از یک تکنیک پردازش زبان طبیعی بی‌نظارت استفاده شده که شامل دو گام است. در گام اول، برچسب‌زن کلمات و وابستگی‌های بین کلمات از متن استخراج شده و در گام بعدی، واژه‌هایی که به‌عنوان اسم هستند به کلمات احساس، مرتبط می‌شوند. سپس، ویژگی‌های صریح خوشه‌بندی می‌گردند.

در روش ارائه شده توسط سو و همکاران [۱۶]، صفت‌ها به‌عنوان احساس و اسم، عبارات اسمی و عبارت‌های فعلی به‌عنوان ویژگی‌های کاندید محصول در نظر گرفته می‌شوند. در گام بعدی، ویژگی‌های کاندید، هرس و ویژگی‌های محصول و کلمات احساس[17] دسته‌بندی می‌شوند. در گام آخر، ارتباطات احساسی پنهان[18] بین گروه‌های ویژگی محصول و کلمات نظر کاوش شده و یک مجموعه‌ی انجمنی[19]، با شناسایی n تا از قوی‌ترین پیوند احساسی بین دو گروه از اشیاء داده‌ای، ایجاد و با استفاده از مجموعه‌ای از پیش ساخته‌شده، ویژگی‌های ضمنی شناسایی می‌گردند.

یک روش کاوش قوانین انجمن هم رخداد[20] جهت تشخیص ویژگی‌های ضمنی توسط های و همکاران [۱۷] ارائه شد. کلمات احساس و ویژگی‌ها از جملات صریح موجود در پیکره استخراج و یک ماتریس هم‌رخداد ایجاد می‌شود. سپس برای هر کلمه‌ی احساس، قواعد انجمنی از ماتریس هم‌رخداد ساخته می‌شوند و با استفاده از اندازه‌گیری‌های پشتیبان و اطمینان، قوانین ضعیف به‌طور مناسبی هرس می‌گردند. برای کلمه‌ی احساس داده‌شده‌ای که دارای ویژگی صریح نیست، یک لیست تطبیق‌یافته از قوانین انجمنی مورد جست‌وجو قرار گرفته تا ویژگی ضمنی مرتبط از آن استخراج شوند.

باقری و همکاران [۷] یک مدل بدون ناظر تشخیص ویژگی و احساس را ارائه داده‌اند که قادر به استخراج ویژگی‌های صریح و ضمنی بر روی زبان انگلیسی است. در این مدل، اسم‌ها بیان‌گر ویژگی و

صفت‌ها، قیدها و فعل‌ها بیان‌گر کلمات احساس هستند. ویژگی‌های چندکلمه‌ای توسط روشی به نام FMLR[۲۱] شناسایی 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. در نهایت، از یک گراف برای تشخیص ویژگی ضمنی استفاده می‌شود.

مدل ارائه شده توسط شوتن و همکاران [۱۸] از ماتریس هم‌رخداد برای تشخیص ویژگی‌های ضمنی استفاده می‌کند. ابتدا داده‌ی آموزشی، بررسی شده و سه لیست ایجاد می‌شوند. ۱) لیستی به نام F از تمامی ویژگی‌های ضمنی ۲) یک لیست به نام O از تمامی کلمات (به‌جای به‌کارگیری کلمات احساس، از همه‌ی کلمات برای تشخیص ویژگی ضمنی استفاده می‌کند.) و ۳) یک ماتریس به نام V از هم‌رخدادی بین ویژگی‌های ضمنی و کلماتی که با هم در یک جمله ظاهر می‌شوند. در داده‌ی آزمایشی، برای هر ویژگی ضمنی f_i بر اساس مقادیر لیست‌های F، O و ماتریس C یک نمره محاسبه‌شده که از مجموع هم‌رخداد هر کلمه تقسیم بر تعداد تکرار آن کلمه در جملات به‌دست می‌آید.

باباعلی و همکارش [۱۹] یک روش نظرکاوی بدون نظارت مبتنی بر ویژگی برای محصولات در زبان فارسی ارائه داده‌اند که شامل مراحل شناسایی ویژگی‌های صریح، ویژگی‌های ضمنی و تعیین جهت معنایی نظرات است. در مرحله‌ی استخراج ویژگی صریح، ابتدا ویژگی‌های جذاب و پرتکرار و سپس ویژگی‌های نادر شناسایی می‌شوند. در مرحله‌ی استخراج ویژگی‌های پرتکرار، ویژگی‌های تک و چندکلمه‌ای با حداقل پشتیبان یک درصد به‌دست می‌آیند. سپس، ویژگی‌های ضمنی با استفاده از کاوش قوانین انجمنی هم‌رخداد ارائه‌شده توسط های [۱۷] استخراج می‌شوند.

با وجود تحقیقات بسیاری که در زمینه نظر کاوی مبتنی بر ویژگی‌های صریح انجام شده، اما هنوز در حوزه نظرکاوی مبتنی بر ویژگی های ضمنی با چالش های بسیاری روبه‌رو بوده و نیازمند پژوهش‌های بیشتری در این زمینه است.

۳- روش پیشنهادی

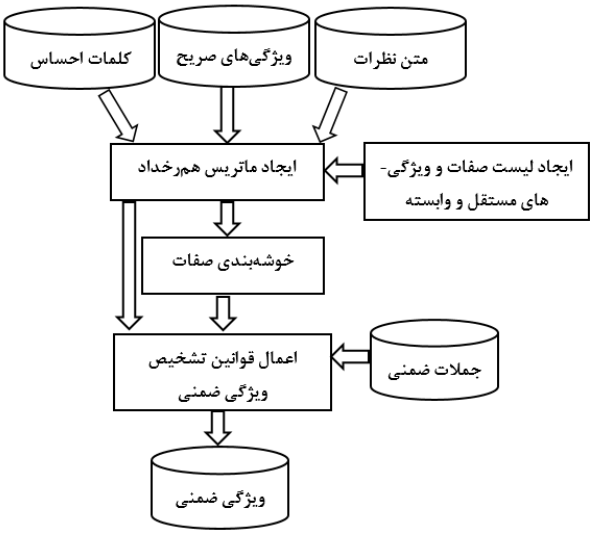
از آن‌جایی که تمرکز این پژوهش بر استخراج ویژگی‌های ضمنی است، در ابتدا به اختصار به توصیف ویژگی‌های ضمنی پرداخته و سپس فرایند پیشنهادی ارائه می‌شود.

ویژگی ضمنی، ویژگی است که به‌وضوح در جملات ظاهر نمی‌شود و با استفاده از کلمات احساس در جمله‌ی دارای احساس، قابل‌شناسایی است. ویژگی‌های ضمنی را می‌توان با استفاده از ویژگی‌های صریح شناخته‌شده از نظرات استخراج نمود. در جدول ۱، نمونه‌هایی از ویژگی‌های ضمنی آورده شده است.

جدول ۱: سه مثال برای ویژگی ضمنی

نظر	ویژگی ضمنی
خیلی گران است	قیمت
اصلا به صرفه نیست	قیمت
خیلی خوشم اومد، خیلی شیک است	ظاهر

فرایند کلی فلوچارت الگوریتم پیشنهادی که به استخراج ویژگی ضمنی می‌پردازد در شکل ۱، آورده شده است و در ادامه نیز به تفصیل هر یک از مراحل خواهیم پرداخت.



شکل ۱: فلوچارت روش پیشنهادی استخراج ویژگی ضمنی

۳-۱- پیش‌پردازش جملات

در ابتدا، لازم است که متون نظرات پیش‌پردازش شوند. نظرات مربوط به محصول توسط ابزار هضم[۲۲] [۲۰] نرمال سازی شده، ریشه‌یابی و سپس برچسب‌زده می‌شوند. به‌عنوان مثال خروجی این عملیات برای جمله‌ی "دوربین مناسبی دارد ولی کیفیت ساخت خوبی ندارد" به صورت «دوربین/Ne مناسب/AJ داشت/V ولی/CONJ کیفیت/Ne ساخت/Ne خوب/AJ نداشت/V» خواهد بود. سپس از ابزار هضم، برای تعیین گراف وابستگی جملات استفاده می‌شود که این گراف، وابستگی بین کلمات یک جمله را از نظر نحوی مشخص می‌کند. جهت بهبود دقت گراف وابستگی، نظرات طولانی با استفاده از کلماتی که برچسب «V» دارند به یک یا چند جمله شکسته می‌شوند که تنها حاوی یک فعل باشند. به‌عنوان مثال، نظر بالا به دو جمله‌ی «دوربین/Ne مناسب/AJ داشت/V» و «ولی/CONJ کیفیت/Ne ساخت/Ne خوب/AJ نداشت/V» شکسته می‌شود. این گراف وابستگی و نقش نحوی استخراج‌شده از ابزار هضم، در پایگاه داده ذخیره می‌شوند. به‌عنوان مثال نقش نحوی جمله‌ی اول به‌صورت « دوربین/SBJ مناسب/NPOSTMOD داشت/ROOT» مشخص می‌شود. گراف وابستگی همان جمله، به‌صورت شکل ۲ آورده شده است.

شکل ۲: نمونه گراف وابستگی جمله

پس از تعیین نقش نحوی و تشکیل گراف وابستگی، لیست صفات و ویژگی‌های صریح مستقل و وابسته ایجاد می‌شوند. در ابتدا، برای هر یک از جملات پیکره که حاوی ویژگی یا صفت هستند، دو لیست از صفات به نام لیست صفات مستقل و وابسته و دو لیست از ویژگی‌ها به نام لیست ویژگی‌های مستقل و وابسته ایجاد می‌شوند. از این لیست‌ها در فرایند تشکیل ماتریس هم‌رخداد استفاده شده است. در ادامه هر یک به اختصار بیان شده‌اند.

- صفات مستقل و وابسته: تمامی صفت‌هایی که در یک جمله رخ می‌دهند، صفت مستقل خواهند بود مگر اینکه صفتی (صفت‌هایی) بعد از حرف ربط «و» و یا علامت «ویرگول» آمده باشند و قبل از آن حروف نیز، یک صفت باشد. در این‌صورت، اولین صفت قبل از حرف ربط «و» و یا علامت « ویرگول» صفت مستقل و صفت‌های بعد از آن‌ها، صفت‌های وابسته خواهند بود. به‌عنوان مثال در جمله‌ی «دوربین و باتری خوب است»، «خوب» صفت مستقل و یا در جمله‌ی «ظاهر گوشی زیبا و جذاب است»، "جذاب" صفت مستقل و "زیبا" صفت وابسته است.
- ویژگی‌های صریح مستقل و وابسته: دسته‌بندی ویژگی‌ها دقیقا مشابه فرایند ارائه‌شده برای دسته‌بندی صفات است، یعنی همه‌ی ویژگی‌ها مستقل خواهند بود مگر این‌که ویژگی (ویژگی‌هایی) بعد از حرف ربط «و» و یا علامت «ویرگول» بیاید و قبل از آن حروف نیز یک ویژگی آمده باشد. به‌عنوان مثال در جمله‌ی «دوربین و باتری خوب است»، "دوربین" ویژگی مستقل و "باتری" ویژگی وابسته است.

۳-۲- ایجاد ماتریس هم‌رخداد

اولین گام از فرایند کلی استخراج ویژگی ضمنی، ایجاد ماتریس هم‌رخداد است. برای ایجاد ماتریس هم‌رخداد ابتدا ویژگی‌های صریح و احساس به‌عنوان ورودی به سیستم داده می‌شوند. این اطلاعات از نتیجه پژوهش‌های قبلی استخراج شده است.

در ماتریس هم‌رخداد برای هر ویژگی صریح یک سطر و برای هر کلمه‌ی احساس یک ستون در نظر گرفته می‌شود. برای هر زوج «ویژگی صریح-کلمه‌ی احساس» موجود در این ماتریس، معیار آزمون نسبت درستنمایی[۲۳] (LRT) [۲۱] و تعداد رخداد ویژگی صریح و کلمه‌ی احساس با هم محاسبه می‌شوند.

نسبت درستنمایی یا LRT از نسبت احتمال رخداد دو واژه به احتمال کل نظرات محاسبه می‌شود که میزان وابستگی بین دو واژه را نشان می‌دهد. هر چه مقدار این معیار بزرگتر باشد، وابستگی دو واژه به یکدیگر بیشتر خواهد بود [۲۱]. عنصر (i,j) در ماتریس هم‌رخداد میزان LRT و همچنین تعداد رخداد ویژگی صریح iام با احساس jام را نشان می‌دهد. استفاده از معیار LRT منجر به نتایج آماری بهتری در تجزیه و تحلیل متن شده است و دلایل استفاده از آن به اختصار در ادامه آورده شده است:

- برای متن‌های کوتاه نیز نتایج خوبی را تولید می‌کند.
- در محاسبه‌ی وابستگی بین دو واژه، به وابستگی بین کلمات نادر و همچنین کلمات پرتکرار اهمیت می‌دهد [۲۱].

برخلاف سایر روش‌ها در ایجاد ماتریس هم‌رخداد، که ویژگی‌های صریح و کلمات احساس تنها از طریق اطلاعات آماری به ماتریس اضافه می شوند، در این‌جا، همه‌ی ویژگی‌های صریح و احساسات موجود در جملات بر اساس روابط وابستگی بین کلمات، قواعد نحوی زبان فارسی و اطلاعات آماری به این ماتریس افزوده می شوند. در روش پیشنهادی برای استفاده از اطلاعات آماری، علاوه‌بر معیار LRT، از معیار فاصله‌ی ویژگی صریح از کلمه‌ی احساس، بر خلاف سایر روش‌ها استفاده شد که بهبود خوبی در نتایج ارزیابی داشت. به‌عبارتی، برای هر زوج «ویژگی صریح-کلمه‌ی احساس»، در صورتی که معیار LRT این دو زوج از حد آستانه بیشتر باشد و فاصله‌ی این دو زوج در محدوده [۴و۱-] باشد به ماتریس هم‌رخداد اضافه می‌شوند.

شبه کد ایجاد ماتریس هم‌رخداد در شکل ۳ آورده شده که شامل سه مرحله می‌باشد که در ادامه آورده شده است:

۱- ایجاد گراف وابستگی
۲- تعیین قواعد نحوی زبان
۳- استخراج اطلاعات آماری بین ویژگی‌های صریح و کلمات احساس

```
Input: opinion Set {O}, feature set {F}
Output: co-occurrence matrix M_OF
for each opinion word o_i in {O} do
  for each feature f_j in {F} do
    if (o_i and f_j are identified dependent by graph) then
      add o_i and f_j to M_of;
    end
    else if ( o_i and f_j are identified dependent by syntactic rules then
      add o_i and f_j to M_of;
    end
    sort LRT( o_i and f_j) to descending order ;
    else if (o_i and f_j are identified dependent by LRT criterion and are in range of (-1,4)) then
      add o_i and f_j to M_of ;
    end;
  end;
end;
return M_of co-occurrency matrix;
```

شکل ۳: شبه کد پیشنهادی تشکیل ماتریس هم‌رخداد

در مرحله‌ی اول، گراف وابستگی جملات در نظر گرفته می‌شود. در این گراف، در صورتی‌که یک کلمه‌ی احساس به یک ویژگی صریح وابسته باشد، در قالب زوج «ویژگی صریح- کلمه‌ی احساس» به ماتریس هم‌رخداد اضافه خواهد شد. برای تشکیل زوج‌های مذکور، از لیست‌های مستقل و وابسته که قبلا ایجاد شده‌اند، استفاده شده است. به‌عنوان مثال در جمله‌ی «ظاهر گوشی زیبا و جذاب است.» در

صورتی‌که گراف تشخیص دهد که کلمه‌ی احساس «زیبا» به ویژگی «ظاهر» وابسته است، با استفاده از لیست صفات وابسته، علاوه‌بر زوج «ظاهر-زیبا»، زوج «ظاهر-جذاب» هم به ماتریس اضافه خواهد شد؛ چرا که صفت «جذاب» به‌عنوان وابسته‌ی صفت «زیبا» دسته‌بندی شده است. برای ویژگی‌های صریح وابسته‌ی موجود در جمله نیز روال مشابهی انجام می‌شود. به‌عنوان مثال، از جمله‌ی «دوربین و باتری خوب است.»، دو زوج «دوربین-خوب» و «باتری-خوب» به ماتریس هم‌رخدادی اضافه خواهند شد؛ چرا که «باتری» ویژگی وابسته‌ی «دوربین» است.

ممکن است که ابزار هضم نتواند برخی روابط وابستگی بین کلمات احساس و ویژگی‌های صریح را نشان دهد. در مرحله‌ی دوم، قواعد نحوی زبان بررسی می‌شوند. به‌عنوان نمونه، در زبان فارسی، هر صفتی که بلافاصله پس از یک ویژگی آورده شود یا یک قید مقدار (مانند خیلی، بسیار، کمی و غیره) بین صفت و ویژگی بیاید، آن صفت مربوط به ویژگی است. صفت‌ها و ویژگی‌هایی که از قاعده‌ی فوق پیروی کنند، به‌صورت زوج «ویژگی صریح-کلمه‌ی احساس» به ماتریس هم‌رخدادی اضافه می‌شوند. به‌عنوان مثال، جمله‌ی «صفحه نمایش و کیفیت تصویر خیلی خوب است.» را در نظر بگیرید. بر اساس این قانون نحوی، صفت «خوب» به ویژگی «کیفیت تصویر» وابسته است، زیرا این صفت، بلافاصله بعد از ویژگی و قید مقدار آمده است؛ در نتیجه، زوج «کیفیت تصویر-خوب» به ماتریس هم‌رخدادی اضافه می‌شود، علاوه‌بر این، زوج «صفحه نمایش-خوب» به‌دلیل وابسته‌بودن ویژگی «کیفیت تصویر» به ویژگی «صفحه نمایش» نیز به ماتریس هم‌رخدادی اضافه خواهد شد.

مرحله‌ی سوم، تمامی زوج‌های ویژگی صریح و صفت‌هایی که در دو مرحله قبلی استخراج نشده‌اند، بررسی می‌شوند. در ابتدا معیار LRT برای هر زوج ویژگی صریح-کلمه‌ی احساس محاسبه می‌شود که نشان‌دهنده‌ی میزان هم‌رخداد واژه‌ها در جملات پیکره است. پس از محاسبه‌ی LRT هر زوج ویژگی صریح-کلمه‌ی احساس، لیست زوج ویژگی صریح-کلمه‌ی احساس به‌ترتیب نزولی با توجه به مقدار LRT مرتب شده و لیست مرتب به‌دست‌آمده جهت افزودن زوج ویژگی صریح-کلمه‌ی احساس به ماتریس هم‌رخدادی، از بالا پیمایش می‌شود. سپس آن عنصر در صورتی به ماتریس هم‌رخدادی اضافه می شود که در بازه‌ی [۴و۱-] باشد. در محاسبه‌ی بازه کلمات مانع[۲۴] نادیده گرفته می شوند. دلیل استفاده از بازه این است که صفتی که برای توصیف احساس یک ویژگی به‌کار می‌رود، در نزدیکی آن ویژگی ظاهر می‌شود. نتایج به‌دست‌آمده نشان می‌دهد که استفاده از بازه باعث افزایش دقت و بازخوانی در ارزیابی نتایج می‌شود. هر بار که عنصری از این لیست به ماتریس افزوده می‌شود سایر زوج ویژگی صریح-کلمات احساس موجود در لیست که شامل ویژگی یا کلمه‌ی احساس اضافه شده به ماتریس باشند، از لیست حذف می‌شوند. به‌عنوان مثال، در جمله‌ی «دوربین چندان هم خوب نبود ولی باتری آن نسبتا عالی است.»، مقدار LRT تمامی زوج های «دوربین-خوب»، «دوربین- عالی»، «باتری-خوب» و «باتری-عالی» را محاسبه و به‌ترتیب نزولی مرتب می‌کند و زوج با بالاترین مقدار LRT را در صورتی‌که در بازه‌ی [۴و۱-] باشد به ماتریس هم‌رخدادی اضافه می‌کند. بر اساس جمله فوق دو زوج ویژگی «دوربین-خوب» و «باتری-عالی» به ماتریس اضافه می‌شوند. دلیل انتخاب مقدار ۱- ، این است که گاهی اوقات، صفتی که برای بیان احساس یک ویژگی به‌کار می‌رود، قبل از آن ویژگی ظاهر می‌شود. به‌عنوان نمونه در جمله‌ی نظری «در کل اگر یک گوشی ارزان قیمت و با امکانات عالی می‌خواهید توی خرید این گوشی شک نکنید.»، صفت «ارزان» که برای بیان احساس ویژگی «قیمت» به‌کار می‌رود، قبل از آن ظاهر می‌شود. در شکل ۴، مثالی برای تشکیل ماتریس هم‌رخداد آورده شده است.

(۱) تاچ ، سرعت و قدرت گوشی واقعا رضایت بخش و عالی است! (۲) هزینه‌اش یکم گران است ولی ظاهر زیبا و قشنگی دارد.
(۳) دوربین چندان هم خوب نبود ولی باتری آن به نسبت عالی است. (۴) قیمتش خیلی ارزان است.
(۵) قیمتش کمی گران است ولی دوربینش عالی است. (۶) هزینه‌اش یکم گران است.

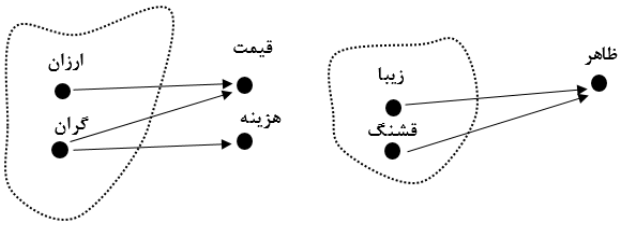
	گران	زیبا	ارزان	عالی	رضایت‌بخش	قشنگ	خوب
تاچ				۱	۱		
سرعت				۱	۱		
قدرت				۱	۱		
هزینه	۲						
ظاهر		۱				۱	
دوربین				۱			۱
قیمت	۱		۱				
باتری				۱			

شکل ۴: مثالی برای تشکیل ماتریس هم‌رخداد

۳-۳- خوشه‌بندی صفات

در گام بعدی، صفت‌های موجود در ماتریس هم‌رخدادی بر اساس شباهت کسینوسی خوشه‌بندی می‌شوند. درخوشه‌بندی صفت‌ها، صفت‌هایی که از نظر مفهومی و معنایی به هم نزدیک هستند، در یک خوشه قرار می‌گیرند. در ابتدا، برای هر صفتی که در ماتریس هم‌رخدادی موجود است، تعداد تکرار آن صفت با تمامی ویژگی‌های

صـــریح به‌دســـت می‌آید. یعنی برای هر صـــفت، یک بردار از ویژگی‌ها به‌دســـت می‌آید که هر عنصـــر این بردار بیان‌گر تعداد هم‌رخدادی آن صـــفت با ویژگی‌های صـــریح موجود در جملات نظرات اســـت. پس از نرمال‌کردن بردار هر صفت، شباهت کسینوسی را برای هر دو صفت کاندید به‌دست آورده و در صورتی که شباهت کسینوسی برای دو صفت کاندید بزرگ‌تر و یا مســـاوی مقدار ۰/۹۵ باشـــد، در یک خوشـــه قرار می‌گیرند. خوشـــه‌های به‌دســـت‌آمده در پایگاه داده ذخیره می‌شـــوند تا برای تشـخیص ویژگی ضـمنی مورد اسـتفاده قرار گیرند. در شـکل ۵، خوشه‌بندی معادل برای ماتریس هم‌رخدادی موجود در شکل ۴ نشان داده شـده اسـت. در این شـکل، صـفت‌های «ارزان» و «گران» در یک خوشه و صفت‌های «زیبا» و «قشنگ» در خوشه‌ی دیگر قرار می‌گیرند زیرا بردار ویژگی خوشه‌ی اول شامل «قیمت» و «هزینه» و بردار ویژگی خوشه‌ی دوم شامل «ظاهر» است.



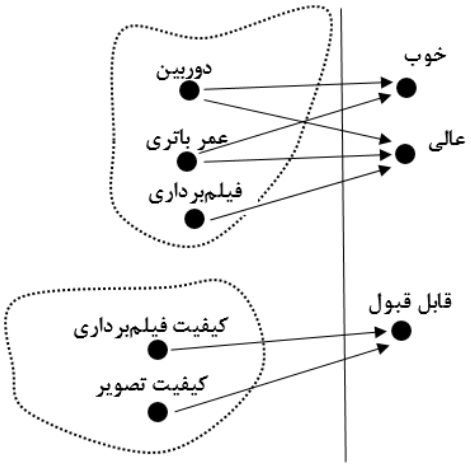
شکل ۵: نمونه خوشه‌بندی احساسات بر اساس بردار ویژگی‌ها

خوشه‌بندی ویژگی‌ها بر اساس بردار احساسات کاربردی نیست زیرا صـفت‌های زیادی هسـتند که برای بیان احسـاس ویژگی‌های مختلفی به‌کار می‌روند. به‌عنوان مثال همان‌طور که در جدول ۲، نشان داده شده است، احساس‌هایی مانند «قابل قبول»، «خوب»، «عالی» و غیره برای توصیف ویژگی‌های مختلفی به‌کار می‌روند.

جدول ۲: مثال‌هایی برای احساس مشترک مرتبط با ویژگی‌ها

جمله	احساس مرتبط با ویژگی
دوربین خیلی خوبی داره من که خیلی دوستش دارم	خوب
عمر باتریش خیلی خوبه، واقعا سامسونگ گل کاشته	خوب
عمر باتریش عالی است	عالی
کیفیت تصویرش قابل قبول است	قابل قبول
دوربین عالیه	عالی
فیلم‌برداریش عالیه	عالی
کیفیت فیلم‌برداریش قابل قبول است	قابل قبول

با توجه به جدول ۲، خوشه‌بندی ویژگی‌ها بر اساس بردار احساس به‌صورت شـکل ۶ اسـت. همان‌طور که مشـاهده می‌شـود، ویژگی‌های «دوربین»، «عمر باتری» و «فیلم‌برداری» در یک خوشـه و ویژگی‌های «کیفیت فیلم‌برداری» و «کیفیت تصـویر» در خوشـه‌ی دیگر قرار می‌گیرند. پس خوشـه‌بندی ویژگی‌ها بر اسـاس بردار احسـاس، نتایج درسـتی ایجاد نکرده و کمکی به حل مسـئله‌ی تشـخیص ویژگی‌های ضمنی نمی‌کند. از این‌رو از دسته‌بندی احساسات بر اساس بردار ویژگی پیشنهاد شده است.



شکل ۶: نمونه خوشه‌بندی ویژگی‌ها بر اساس بردار احساس

۳-۴- اعمال قوانین تشخیص ویژگی ضمنی

پس از انجام مراحل فوق، از صفات خوشه‌بندی‌شده و ماتریس هم‌رخداد برای تشخیص ویژگی‌های ضمنی استفاده می شود. تشخیص ویژگی ضـمنی برای جملاتی انجام می‌شـود که فاقد ویژگی صـریح و دارای کلمه‌ی احسـاس اسـت. برای هر صـفت مانند s در این‌گونه جملات، از چهار قانون برای استخراج ویژگی ضمنی مناسب استفاده می شود که در ادامه توضیح داده شده‌اند و شبه کد آن نیز در شکل ۷ نیز آورده شده است.

- در صورتی که صفت s فقط با یک ویژگی صریح مانند a در ماتریس هم‌رخدادی رخ داده باشد، ویژگی a به‌عنوان ویژگی ضمنی جمله‌ای که دارای احسـاس و فاقد ویژگی اسـت انتخاب می‌شـود. به‌عنوان مثال در جمله‌ی نظری «خیلی گران است» در صورتی‌که تنها زوج «قیمت-گران» در ماتریس هم‌رخدادی وجود داشـته باشـد، ویژگی ضمنی«قیمت» برای صفت «گران» انتخاب می‌شود.
- در صـورتی که صـفت s، در مجموعه صـفت‌های خوشـه‌بندی‌شـده موجود باشد، خوشه‌ی دربردارنده‌ی صفت s انتخاب می‌شود. سپس ویژگی صـریح مشـترک بین تمامی صـفت‌های داخل خوشـه از ماتریس هم‌رخدادی انتخاب و به‌عنوان ویژگی ضـمنی صـفت s انتسـاب داده می‌شـود. به‌عنوان مثال اگر صـفت‌های «ارزان»، «گران» و «عادلانه» در یک خوشـه باشـند و صـفت «ارزان» و «گران» با ویژگی‌های «قیمت» و «هزینه» و صـفت «عادلانه» با ویژگی «قیمت» در ماتریس هم‌رخدادی موجود باشـند. در این‌صورت اگر جمله‌ی ضمنی «خیلی گران است.» را داشته باشیم، ویژگی مشترک بین این سه صفت یعنی «قیمت» به عنوان ویژگی ضمنی صفت مربوطه تعیین می‌شود.
- اگر برای صـفت s موجود در جمله، دو ویژگی a_1 و a_2 در ماتریس هم‌رخدادی وجود داشته باشند. در این حالت، در صورتی که تعداد هم‌رخدادی صفت s با ویژگی a_1 ، بیش‌تر از دو بار باشد و علاوه‌بر

این مقدار LRT آن دو هم از حد آستانه‌ی ۷ بیشتر باشد و همچنین تعداد هم‌رخدادی صفت s با ویژگی a_2، کم‌تر از دوبار باشد، ویژگی a_1 به‌عنوان ویژگی ضمنی صفت s استخراج می‌شود.

- اگر صفت s با بیش از دو ویژگی در ماتریس هم‌رخدادی وجود داشته باشد، دیگر نمی‌توان ویژگی مشخصی را برای صفت s در نظر گرفت. به‌عنوان مثال در جمله‌ی «گوشی خیلی عالی است.»، احساس «عالی» با ویژگی‌های مختلفی در ماتریس هم‌رخدادی اتفاق می‌افتد. پس نمی‌توان ویژگی مشخصی را برای این جمله تعیین کرد. در این شرایط، این صفت برای کلیت محصول در نظر گرفته می شود. به‌عنوان مثال، اگر نظر فوق در مورد گو شی موبایل باشد، صفت «عالی» به «گوشی» نسبت داده می‌شود.

Input: M_{of}, G_O
Output: IM_{of}
$IS\leftarrow$ implicit sentence set with no feature ;
For each opinion words o_i in IS **do**
  if (o_i just occurred candidate feature (cf) in M_{of}) **then**
    Add o_i and cf to IM_{of};
  end;
  else if (o_i is in G_O cluster) **then**
    $cf\leftarrow$Find common features between all sentiments in G_o ;
    add o_i and cf to IM_{of};
  end;
  else if (o_i occurred with features f_1, f_2 in M_{of} matrix) **then**
    if (co-occurrence_number of o_i with f_1 >=2 and LRT(o_i, f_1)>=5)
      and (co-occurrence_number of o_i with f_2<2) **then**
        add o_i and f_1 to IM_{of};
    end;
    else if (co-occurrence_number of o_i with f_2>=2 and LRT(o_i, f_2)>=5)
      and (co-occurrence_number of o_i with f_1<2) **then**
        add o_i and f_2 to IM_{of};
    end;
  end;
  else if (o_i occurred with more than two features) **then**
    add o_i and «کلیت محصول» to IM_{of};
end;
return IM_{of};

شکل ۷: شبه کد تشخیص ویژگی مناسب برای جمله‌ی ضمنی

۴- نتایج و ارزیابی روش پیشنهادی

در این بخش، ابتدا به معرفی مجموعه داد گان و معیار های ارز یابی پرداخته و سپس عملکرد روش پیشنهادی، ارزیابی می‌شود.

۴-۱- نحوه آماده‌سازی مجموعه داده‌ها

در این مقاله، از مجموعه نظرات جمع‌آوری‌شده در مقاله‌ی باباعلی [۱۹] که از سایت دیجی‌کالا در سال ۱۳۹۴جمع‌آوری گردیده، جهت ارزیابی سیستم ا ستفاده شده ا ست. این مجموعه نظرات، مربوط به سه نوع تلفن همراه و سه نوع لپ‌تاپ [۲۲] ا ست. اطلاعات مربوط به این داده ها در جدول ۳ ارائه شده است.

جدول ۳: اطلاعات آماری مجموعه داده‌ها

گروه کالا	شرح کالا	تعداد نظرات	تعداد جملات
تلفن همراه	-Fujitsu LifeBook AH-532-B -Huawei Ascend Y51 -Samsung Galaxy Young S6310	۳۹۴	۱۸۴۷
لپ‌تاپ	- ASUS K46CB-C - ASUS N550JV - Nokia Lumia 520	۱۶۸	۶۶۵
تعداد کل	۶	۵۶۲	۲۵۱۲

تمامی جملات مرتبط با محصولات توسط ابزار هضم نرمال شده و فعل‌های محاوره (مانند میشه) به فعل رسمی (مانند می‌شود) تبدیل شده و همچنین تا حدودی غلط‌های املایی اصلاح می‌شوند. سپس نقش لغوی، نحوی و گراف وابستگی تمامی جملات توسط این ابزار استخراج می‌گردد. این مجمو عه داده ها به‌عنوان ورودی الگوریتم پیشنهادی استفاده می‌شود.

روش پی شنهادی ما بر روی ویژگی‌های ضمنی عمل می‌کند. برای ارزیابی صحت خروجی، از دو کاربر خبره خواسته شد تا با خواندن تمامی نظرات، ویژگی های ضمنی هر جمله را (در صورت وجود) استخراج کنند. این ویژگی‌ها با ویژگی‌های ضمنی استخراج‌شده توسط روش پی شنهادی ما مقای سه شده ا ست. به‌علاوه، برای ا ستخراج قواعد رایج نحوی زبان فارسی (مرحله ۲ از روش پیشنهادی) از چند مرجع معتبر در این حوزه استفاده شده است.

۴-۲- معیارهای ارزیابی

ارزیابی یک مدل نظرکاوی، بر اساس نمونه‌های آزمونی است که توسط فرد خبره برچسب‌گذاری شده باشد. برای ارزیابی مدل، باید برچسبی را که مدل برای متون نظرات (ویژگی‌های ضمنی) در نظر گرفته است، با برچسب فرد خبره مقایسه کرد. بنابراین چهار حالت وجود دارد که جدول ۴، شرایط وقوع این حالت‌ها را نشان می‌دهد.

جدول ۴: وقوع حالت‌های مختلف برای مدل تجزیه و تحلیل متون نظرات

	برچسب‌گذاری توسط فرد خبره		
برچسب‌گذاری شده توسط مدل		YES	NO
	YES	TP	FP
	NO	FN	TN

پارامترهای معرفی شده در جدول ۴ به‌صورت زیر برای استخراج ویژگی‌های ضمنی تعریف می‌شوند:

- پارامتر TP: تعداد ویژگی‌هایی است که به‌درستی توسط مدل مورد نظر به‌عنوان ویژگی استخراج شده‌اند.

- پارامتر FP: مربوط به تعداد ویژگی‌هایی است که به‌صورت نادرست توسط مدل مورد نظر به‌عنوان ویژگی استخراج شده‌اند.
- پارامتر FN: تعداد ویژگی‌هایی است که ویژگی ضمنی هستند اما توسط مدل مورد نظر به‌عنوان ویژگی ضمنی تشخیص داده نشده‌اند.
- پارامتر TN: تعداد ویژگی‌هایی است که ضمنی نیستند و به‌درستی توسط مدل مورد نظر به‌عنوان ویژگی استخراج نشده‌اند.

جهت ارزیابی کارایی روش پیشنهادی در استخراج ویژگی ضمنی همانند کارهای قبلی از معیارهای دقت، بازخوانی و معیار F استفاده شده است که معیارهای ارزیابی به‌صورت زیر قابل محاسبه هستند:

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (۱)$$

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (۲)$$

$$F - measure = \frac{2 * \mathrm{Precision} * \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \quad (۳)$$

در ادامه با معیارهای فوق نتایج روش پیشنهادی ارزیابی شده است.

۴-۳- ارزیابی نتایج

در آزمایش‌های انجام شده روش پیشنهادی، مقدار حد آستانه‌ی **۷** که در استخراج ویژگی‌های ضمنی در بخش ۴-۳ استفاده گردید، به‌صورت تجربی برابر ۵ در نظر گرفته شده است.

مقادیر معیارهای محاسبه شده برای روش پیشنهادی در تشخیص ویژگی ضمنی در مقایسه با نتایج به‌دست‌آمده در روش ارائه‌شده توسط باباعلی [۱۹]، باقری [۷] و دنینگ [۲۱] در جدول ۵ نشان داده شده است.

جدول ۵: مقایسه روش پیشنهادی با روش‌های پیشین

		تلفن همراه	لپ‌تاپ
دقت	**روش پیشنهادی**	**۰/۷۵**	**۰/۷۵**
	باقری [۷]	۰/۶۷	۰/۶۹
	باباعلی [۱۹]	۰/۶۰	۰/۶۴
	دنینگ [۲۱]	۰/۵۹	۰/۶۰
بازخوانی	**روش پیشنهادی**	**۰/۶۵**	**۰/۶۹**
	باقری [۷]	۰/۴۹	۰/۵۰
	باباعلی [۱۹]	۰/۴۲	۰/۴۵
	دنینگ [۲۱]	۰/۳۸	۰/۴۰
معیار F	**روش پیشنهادی**	**۰/۶۹**	**۰/۷۱**
	باقری [۷]	۰/۵۷	۰/۵۸
	باباعلی [۱۹]	۰/۴۹	۰/۵۲
	دنینگ [۲۱]	۰/۴۶	۰/۴۸

نتایج نشان می‌دهد که روش پیشنهادی ما کارایی تشخیص ویژگی ضمنی را نسبت به روش‌های پیشین [۷، ۱۹، ۲۱] بهبود داده است. زیرا روش ما برخلاف روش‌های قبلی علاوه‌بر در نظرگرفتن معیار فاصله در اطلاعات آماری از ارتباطات نحوی زبان و گراف وابستگی بین کلمات یک جمله برای ایجاد ماتریس هم‌رخداد کاراتر بهره برده است. به‌عنوان مثال جمله‌ی «با این‌که قیمتش خیلی ارزان است ولی در کل گوشی خوبی است»، را در نظر بگیرید. روش‌های قبلی در صورتی‌که احساس «خوب» بیش‌تر از احساس «ارزان» با ویژگی «قیمت» رخ داده باشد، احساس «خوب» را به ویژگی «قیمت» نسبت می‌دهند که نادرست است. اما در روش پیشنهادی ما به دلیل استفاده از ارتباطات نحوی زبان، صفت «ارزان» را به ویژگی «قیمت» نسبت خواهد داد.

علیرغم اینکه روش پیشنهادی ما نسبت به روش‌های قبلی بهبود مناسبی داشته است، در برخی موارد به‌درستی ویژگی ضمنی را مشخص نکرده است. در مواردی، این مسئله به‌علت زیادنبودن تعداد قواعد نحوی زبان فارسی رخ داده است. شایان ذکر است که در این پژوهش از قواعد نحوی رایج زبان فارسی استفاده شده‌است. دلیل دیگر برخی از خطاها نیز نرم‌افزار هضم بوده است که در برخی از موارد مانند تشخیص روابط وابستگی بین کلمات یک جمله، دقت بالایی نداشته است. علاوه‌بر این، برخلاف سایت‌های انگلیسی مانند آمازون، تعداد نظرات کاربران در سایت دیجی‌کالا کم است. این مسئله نیز می‌تواند باعث کاهش کارایی در تشکیل ماتریس هم‌رخداد و درنتیجه کاهش دقت در تشخیص ویژگی ضمنی شود.

۵- نتیجه‌گیری

در این مقاله، روشی جهت استخراج ویژگی ضمنی مجموعه نظرات کاربران در زبان فارسی ارائه گردید. در این روش یک معیار جهت محاسبه فاصله‌ی صفت مرتبط با ویژگی در اطلاعات آماری پیشنهاد شده است. علاوه‌بر این گراف وابستگی، قواعد نحوی زبان فارسی و خوشه‌بندی صفات برای تعیین بهتر هم‌رخدادی ویژگی‌ها و کلمات احساس پیشنهاد شده است. ارزیابی انجام‌شده بر روی نظرات زبان فارسی گرفته‌شده از سایت دیجی‌کالا نشان می‌دهد که این روش بهبود ایجاد ماتریس هم‌رخداد و در نتیجه کارایی تشخیص ویژگی‌های ضمنی را به‌همراه داشته است.

مراجع

[1] J. J. Li, H. Yang and H. Tang, "Feature Mining and Sentiment Orientation Analysis on Product Review," in International Conference on Management Information and Optoelectronic Engineering, pp. 79-84, 2016.

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

[۳] مصطفی. رجب‌زاده و رضا. رافع، «ارائه یک سیستم توصیه‌گر ترکیبی برای تجارت الکترونیک»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، صفحه ۱۷۷-۱۶۳، ۱۳۹۴.

[۴] سیامک. عبداله‌زاده، محمدعلی. بالافر و لیلی. محمدخانلی، «استفاده از خوشه‌بندی و مدل مارکوف جهت پیش‌بینی درخواست آتی کاربر در وب»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، صفحه ۱۷۷-۱۶۳، ۱۳۹۴.

[5] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," ACM Computing Surveys (CSUR), vol. 50, no. 25, pp. 1-33, 2017.

[6] E. Breck and C. Cardie, *Opinion Mining and Sentiment Analysis*, in The Oxford Handbook of Computational Linguistics 2nd edition, 2017.

[7] 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.

[8] 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, pp. 255-264, 2012.

[9] N. Jakob and I. Gurevych, "Extracting opinion targets in a single- and cross-domain setting with conditional random fields," in Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 1035-1045, 2010.

[10] G. Somprasertsri and P. Lalitrojwong, "Automatic product feature extraction from online product reviews using maximum entropy with lexical and syntactic features," in IEEE International Conference on Information Reuse and Integration, pp. 250-255, 2008.

[۱۱] مائده. شیخ حسنی، محرم. منصوری‌زاده و میرحسین. دزفولیان، «نظر کاوی جنبه‌گرا به کمک استخراج روابط معنایی»، بیست و دومین کنفرانس ملی سالانه انجمن کامپیوتر ایران، دانشگاه صنعتی شریف، ۱۳۹۵.

[12] M. Z. Asghar, A. Khan, S. R. Zahra, S. Ahmad, and F. M. Kundi, "Aspect-based opinion mining framework using heuristic patterns," Cluster Computing, https://doi.org/10.1007/s10586-017-1096-9, pp. 1-19, 2017.

[13] H. Liang, X. Sun, Y. Sun, and Y. Gao, "Text feature extraction based on deep learning: a review," EURASIP journal on wireless communications and networking, vol. 1, pp. 1-12, 2017.

[14] E. Asgarian, M. Kahani, and S. Sharifi, "The impact of sentiment features on the sentiment polarity classification in Persian reviews," Cognitive Computation, vol. 10, pp. 117-135, 2018.

[15] M. Dragoni, "Computational advertising in social networks: an opinion mining-based approach," in Proceedings of the 33rd Annual ACM Symposium on Applied Computing, pp. 1798-1804, 2018.

[16] Q. Su, X. Xu, H. Guo, Z. Guo, X. Wu, and X. Zhang, "Hidden sentiment association in chinese web opinion mining," in Proceedings of the 17th international conference on World Wide Web, pp. 959-968, 2008.

[17] Z. Hai, K. Chang, and J.-j. Kim, "Implicit feature identification via co-occurrence association rule mining," in International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, Springer, pp. 393-404, 2011.

[18] K. Schouten and F. Frasincar, "Implicit Feature Extraction for Sentiment Analysis in Consumer Reviews," in International Conference on Applications of Natural Language to Data Bases Information Systems, Lecture Notes in Computer Science, Springer, pp. 228-231, 2014.

[۱۹] مرضیه. باباعلی و محمدعلی. نمعت‌بخش، "استخراج ویژگی‌های محصول در زبان فارسی"، سومین همایش زبان‌شناسی رایانشی، دانشگاه صنعتی شریف، ۱۳۹۳.

[۲۰] محسن. ایمانی و مجتبی. خلاش، ابزار پردازش زبان فارسی، (http://www.sobhe.ir/hazm)، ۱۳۹۲.

[21] T. Dunning, "Accurate methods for the statistics of surprise and coincidence," Computational linguistics, vol. 19, pp. 61-74, 1993.

[22] Digikala Dataset (2018, December 13), Retrieved February 2, 2019.

[23] https://www.uplooder.net/files/701edb674a0ef75695b47de35db298d4/Data.rar.html.

زیرنویس‌ها

[1] Opinion mining

[2] Sentiment analysis

[3] Sentence Level

[4] Supervised

[5] Unsupervised

[6] Support Vector Machine (SVM)

[7] Maximum Entropy

[8] Explicit features

[9] Implicit feature

[10] Co-occurrence matrix

[11] Classifier

[12] Conditional Random Field (CRF)

[13] Stop list

[14] Chi-square

[15] Information Gain

[16] Mutual Infomation

[17] Opinion word

[18] Hidden sentiment links

[19] Association set

[20] Rule mining co-occurrence association approach

[21] Frequency Modified Left Right

[22] Hazm

[23] Likelihood Ratio Tests (LRT)

[24] Stop words