سال سیزدهم، شماره‌های۴۹و۵۰، پاییز و زمستان ۱۴۰۰
صص: ۸۵_۱۰۲

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

ارائه‌ی مدلی برای عقیده‌کاوی در سطح ویژگی سند برای نظرات کاربران هتل‌ها

شهریار محمدی * **الهام خلج** **

*دانشیار، دانشکده مهندسی صنایع، گروه مهندسی فناوری اطلاعات، دانشگاه صنعتی خواجه نصیرالدین طوسی

** کارشناس ارشد دانشکده مهندسی صنایع، گروه مهندسی فناوری اطلاعات، دانشگاه صنعتی خواجه نصیرالدین طوسی

تاریخ دریافت: ۱۳۹۹/۱۱/۰۴ تاریخ پذیرش: ۱۴۰۰/۰۳/۲۹

نوع مقاله: پژوهشی

چکیده

امروزه بررسی نظرات و عقاید کاربران در بستر اینترنت بخش مهمی از فرآیند تصمیم‌گیری مردم در رابطه با انتخاب محصول یا استفاده از خدمات را شامل می‌شود. با وجود اینترنت و دسترسی ساده به وبلاگ‌های مربوط به نظرات در زمینه صنعت گردشگری و هتلداری، منابع غنی و عظیمی از عقاید بصورت متن موجود می‌باشد که می‌توان از روش‌های متن کاوی برای کشف دانش نهفته در این متون استفاده کرد. با توجه به اهمیت نظرات و عقاید کاربران در صنایع، به‌ویژه صنعت گردشگری و هتلداری، مباحث عقیده‌کاوی و تحلیل احساسات مورد توجه متصدیان امور قرار گرفته است. در این مقاله یک روش ترکیبی و جدید بر اساس یک رویکرد رایج در تحلیل احساسات، استفاده از واژگان و الگوریتم ژنتیک برای تولید ویژگی‌هایی برای طبقه‌بندی بار احساسی نظرات ارائه شده است. بدین‌صورت که دو روش ساخت فهرست واژگان یکی با استفاده از روش‌های آماری و دیگری با استفاده از الگوریتم ژنتیک ارائه شده‌است. واژگان فوق الذکر با فرهنگ واژگان احساس عمومی و استاندارد لیو بینگ آمیخته می‌شوند. نتایج نشان می‌دهد روش پیشنهادی از روش‌های پایه براساس واژه‌نامه‌های احساسی روی این مجموعه داده بهتر عمل کرده و معیارهای ارزیابی صحت، دقت، بازخوانی و معیار F با استفاده از روش پیشنهادی، به‌ترتیب ۹۴،۶۵، ۹۴،۵۳، ۹۳،۸۹ و ۹۵،۱۷ می‌باشند.

واژه‌های کلیدی: متن کاوی، عقیده کاوی، تحلیل احساسات در سطح ویژگی، داده کاوی، الگوریتم ژنتیک، طبقه بندی.

۱- مقدمه

تجزیه و تحلیل احساسات[۱] با هدف کشف خودکار نگرش اساسی انسان‌ها نسبت به یک موجودیت انجام می‌شود. در حال حاضر، تجزیه و تحلیل احساسات از داده‌های متنی به طور گسترده‌ای برای ارزیابی رضایت مشتری و تجزیه و تحلیل‌ها استفاده می‌شود [۱] و [۲].

تحلیل احساس و عقیده کاوی[۲] با استفاده از الگوریتم‌های داده‌کاوی و متن‌کاوی به صورت سیستماتیک و بدون نیاز به مطالعه تمامی متن‌های موجود، عقاید، احساسات، ارزیابی‌ها، رفتارها و گرایش‌های کاربران را که به‌صورت داده‌های متن بیان شده‌اند را آنالیز می‌کند. افزایش اهمیت تحلیل‌احساس با رشد رسانه‌های اجتماعی مانند توئیتر،

نویسنده مسئول: شهریار محمدی mohammadi@kntu.ac.ir

[۱] Sentiment Analysis

[۲] Opinion Mining

شبکه‌های اجتماعی، نظرسنجی‌های آنلاین، وبلاگ‌ها و همچنین سهولت بازیابی آنلاین نظرات کاربران همزمان شده‌است. در پژوهش پیرمحمدیانی و محمدی بیان شده-است که امروزه سیستم‌های تحلیل‌احساس تقریبا در همه-ی زمینه‌ها مورد استفاده قرار می‌گیرند، زیرا آرا و عقاید در تمام فعالیت‌های انسانی مهم بوده و تاثیر کلیدی بر فرآیندهای تصمیم‌گیری دارند[۳]. در مطالعه‌ی ژانگ و همکاران[۳] با اشاره به تاثیر توسعه‌ی سریع فناوری‌های وب ۲,۰ و محتوای تولید شده توسط کاربر[۴] ، به تحلیل و بررسی نظرات آنلاین درباره‌ی سفر در صنعت گردشگری پرداخته شده است. همچنین وب‌سایت‌هایی نظیر[۵]TripAdvisor، [۶]Expedia که گردشگران نظرات، عقاید و تجربیات خود در استفاده از خدمات ارایه شده را به اشتراک می‌گذارند، معرفی شده است، این نوع وبسایت‌ها محبوب‌ترین منابع اطلاعاتی برای کسب اطلاعات در راستای تهیه برنامه‌ی سفر و نحوه‌ی رزرو بلیط و هتل هستند [۴]. براساس پژوهش پابلوس و همکاران[۷] فن‌آوری تجزیه و تحلیل متن مبتنی بر پردازش زبان طبیعی[۸] می‌تواند بطور خودکار مقادیر زیادی از بررسی‌ها و نظرات مشتری را از منظر مفهوم کلمه، به‌کارگیرد. این روش به طور گسترده‌ای در شناسایی موضوع و استخراج مفهوم نظر و متن، مورد استفاده قرارمی‌گیرد [۵].

در اکثر مقالات مطالعه شده، پایه و اساس تحلیل احساس مبنی برشمارش کلمات احساسی و تعیین بار مثبت و یا منفی کل متن است[۲، ۶، ۷] و [۸، ۹] نقطه ضعف این روش آن است که ممکن است در یک متن چند ویژگی بیان شده باشد و محاسبه‌ی بار معنایی کلی نتایج دقیقی به همراه نداشته باشد. به عنوان مثال جمله‌ی " دسترسی به مراکز خرید از هتل خوب است اما غذاها کیفیت مناسبی ندارند." در این نظر دو ویژگی مکان هتل و غذا مورد نظر بوده پس طبقه‌بندی کلی نظر مناسب نیست و توجه به هر دو ویژگی مناسب‌تر است.

اخیرا این موضوع مورد توجه دانشمندان فناوری اطلاعات قرار گرفته و روش‌های مختلفی را مورد بررسی قرار داده‌اند [۸،۱۰، ۱۱، ۱۲] اما کمتر از الگوریتم های فراابتکاری استفاده شده و همچنین با توجه به میزان سودآوری صنعت گردشگری و تاثیر مستقیم تجربه و نظر سایر گردشگران برای انتخاب دوباره خدمات ارایه‌شده و یا پیشنهاد انتخاب به سایرین، لزوم پرداختن بیشتر به این موضوع احساس می‌شود. بنابراین در این مقاله صنعت گردشگری و هتلداری مورد بررسی قرار گرفته است که در مقاله‌های قبلی کمتر مورد توجه بوده همچنین در این مقاله علاوه بر تعیین مثبت و یا منفی بودن نظرات در داده های متنی ویژگی‌های خاص مربوط به هتل مورد توجه قرارگرفته، دراین مقاله ازروش پیشنهادی جدید ترکیب الگوریتم ژنتیک و فرهنگ واژگان احساسی استفاده شده است. تحلیل احساسات مشتریان در محیط‌هایی مثل خدمات گردشگری و هتلداری، بیمه، موسسات مالی و بانک‌ها، خرده‌فروشی‌ها، شرکت‌های تجارت الکترونیک و فروش آنلاین و... می‌تواند بسیار کاربردی باشد[۵].

۱-۱ هدف از تحقیق

این پژوهش به‌دنبال راه حلی برای بهبود تحلیل حجم عظیمی از نظرات متنی می‌باشد که معمولا ساختار نیافته یا نیمه‌ساختار هستند و برای انجام این کار سعی می‌شود طبقه‌بندی رتبه‌ی احساسی واژگان دقیق‌تر از روش‌های قبلی محاسبه گردد. سؤالاتی که در این پژوهش به آن‌ها پرداخته می‌شود به شرح زیر هستند:

- ✓ تأثیر بهره‌گیری از الگوریتم ژنتیک بر بهبود پارامترهای ارزیابی طبقه‌بندی واژگان احساسی به چه صورت است؟
- ✓ آیا طبقه بندی واژگان احساسی با روش ترکیبی بیان شده که ویژگی‌های خاصی را نیز در نظر می‌گیرد، نتایج بهتری از روش‌های پایه به همراه دارد؟

[۳] Zhang et al.

[۴] user generated content

[۵] www. TripAdvisor.com

[۶] expediagroup.com

[۷] Pablos et al.

[۸] Natural language processing (NLP)

۲- مروری برتحقیقات پیشین

عقیده‌کاوی یک فن‌آوری است که به طور خودکار با استفاده از ابزار و نرم افزارهای تجزیه و تحلیل متن، از جمله زبان‌رایانه و پردازش زبان طبیعی، دانسته‌های کامنت‌های آنلاین را استخراج می‌کند. این نرم‌افزارها نظرها، ارزیابی‌ها، نگرش‌ها و احساس مردم را نسبت به سازمان‌ها، اشخاص، افراد، موضوع‌ها، اقدام‌ها و ویژگی‌های آن‌ها را تجزیه و تحلیل می‌کند [۱، ۲، ۱۳].

با توجه به گزارش UNEP[۹] در سال۲۰۱۶، برای هر دو کشور پذیرنده گردشگر و فرستنده‌ی گردشگر، صنعت-گردشگری کمک چشمگیری به تولید منافع اقتصادی آنان می‌کند. توسعه‌ی صنعت گردشگری تا حد زیادی به رضایت مشتریان در خدمات هتل وابسته بوده و در نتیجه تخصص دراین زمینه یکی از عوامل مهم برای توسعه‌ی این صنعت است. مطالعه‌ی ژانگ و همکاران[۱۰] نشان می‌دهد اثربخشی کیفیت و خدمات ارایه شده در هتل‌ها در تقویت صنعت‌گردشگری سهم بسزایی دارد[۶]. طبق پژوهش بالاز و همکاران[۱۱]، نظرات نقش اساسی را در فرآیند تصمیم-گیری افراد و سازمان‌ها دارند زیرا تأثیر عمیقی بر روی نگرش و اعتقادات افراد می‌گذارد. عقیده‌کاوی و تحلیل-احساسات باعث می‌شود تا مشاغل تجارت الکترونیکی بتوانند دانش بیشتری از مشتریان و محصولات خود کسب کنند بدون اینکه هزینه نظرسنجی‌ها را بپردازند [۷].

کلیه تکنیک‌های مورد استفاده برای استخراج نظر و عقیده می‌توانند به دو طبقه اصلی تقسیم شوند:

- ✓ رویکردهای مبتنی بر واژه‌ها: این روش واژه‌های دارای باراحساسی متن را با تکیه بر یک فرهنگ واژگان احساسی و رویکرد دانش‌زبانی، طبقه‌بندی می‌کند که شامل یک رویکرد مبتنی بر بدنه و یک رویکرد مبتنی بر فرهنگ لغت است.
- ✓ رویکرد یادگیری ماشین: از الگوریتم‌های یادگیری ماشین بهره‌می‌برد و می‌تواند به سه گروه تقسیم شود: یادگیری نظارت شده، یادگیری نیمه نظارت شده و یادگیری بدون نظارت [۱، ۲، ۱۳].

در مقاله‌ی سینگ و همکاران[۱۲] از این رویکرد برای کشف دانسته‌های نظرات متنی مربوط به محصولات مختلف با استفاده از یک روش طبقه‌بندی برای تجزیه و تحلیل، استفاده می‌شود[۱۵].

۲-۱ سطوح عقیده کاوی

- استخراج افکار در سطح سند، احساسات غالب و کلی را به‌جای موارد و جزئیات مطرح در مباحث در نظر می‌گیرد [۱۳]. وظیفه‌ی عقیده کاوی درسطح سند تعیین قطبیت کلی یک سند است که شامل چند جمله است.
- عقیده کاوی در سطح جمله، بطور ویژه متمرکز بر هر جمله است. خواه جمله بیان‌شده گرایش مثبت، منفی یا خنثی داشته باشد. طبقه‌بندی ذهنی یکی دیگر از وظایف در سطح جمله است که بخش‌های ذهنی و عینی اسناد را استخراج می‌کند. مسئله تجزیه و تحلیل مبتنی بر سطح جمله نیز به‌همین صورت تعریف میشود با این تفاوت که نتیجه تشخیص احساس برای هر جمله به طورجداگانه بررسی میشود [۱۶].
- استخراج ویژگی‌های ارائه‌شده در یک متن یا نظر و بیان گرایش احساس مثبت و یا منفی بر روی آن ها، تجزیه‌وتحلیل احساسات و یا عقیده کاوی در سطح ویژگی سند نامیده می‌شود[۹].

۲-۲ رویکردهای متفاوت برای عقیده‌کاوی و تحلیل احساسات

رویکردهای مبتنی بر واژگان و نظارت نشده[۱۳]:

این رویکرد، در تعیین گرایش مثبت یا منفی متن بااستفاده از مجموعه قوانین و اکتشاف‌های حاصل از دانش و قواعد زبان شناسی مورد استفاده است. اقدامات معمول برای اجرای مرحله اول، علامت‌گذاری هرکلمه وتعیین گرایش احساسی مربوط به آن با کمک یک فرهنگ واژگان احساسی و در مرحله دوم،

[۹] United Nations Environment Program

[۱۰] Zhang et al.

[۱۱] Balazs et al.

[۱۲] Singh et al.

[۱۳] Unsupervised Lexicon-based Approaches

ترکیب کلمات و تحلیل احساسی و تاثیر ترکیب کلمات و درآخر، بررسی این‌که ترکیب‌ها ۱۴چگونه بر قطبیت و گرایش تأثیر می‌گذارند و این را در نمره احساسات نهایی منعکس می‌کنند. در نهایت مراحل بعدی شامل جمع بندی و مصورسازی نظر به کمک نرم افزار می‌باشد.

رویکردهای مبتنی بر یادگیری با نظارت۱۵: با نام روش‌های مبتنی بر یادگیری ماشین یا روش‌های آماری برای طبقه‌بندی احساسات شناخته می‌شوند و از الگوریتم‌های داده‌کاوی تشکیل‌شده که الگوهای زیربنایی را از داده‌های آموزش داده شده یا برچسب گذاری‌شده یاد می‌گیرند، سپس در مرحله بعدی الگوریتم برای طبقه‌بندی داده‌های جدید بدون برچسب کلاس پیش بینی می‌شود، و سپس با استفاده از بازنمایی کلاس‌های پیدا شده توسط الگوریتم به عنوان ورودی برای عقیده کاوی استفاده می‌شوند.

رویکردهای مبتنی بر مفهوم: این رویکرد شامل استفاده از علم هستی شناسی برای پشتیبانی از عقیده کاوی و تحلیل احساسات است. هستی شناسی به عنوان مدلی تعریف می شود که دانش یک حوزه¬ی معین را برای کامپیوتر با دستورهای اگر و آنگاه مفهوم سازی می کند. معمولاً به صورت نمودارهایی ارائه می شود که در آن مفاهیم مدنظر، به گره¬های مرتبط با هم و بصورت متصل کشیده می شوند.

جدول ۱. مروری کلی برمهمترین مقالات مطالعه شده مختص صنعت هتل‌داری

هدف و نتیجه کلی مقاله	الگوریتم و روش مقاله	حوزه‌ی‌مورد مطالعه	مقاله	شماره
ارایه یک تکنیک جهت خلاصه کردن چند متن برای شناسایی جملات کلیدی ومهم از نظرات هتل را ارائه می‌دهد	از شباهت‌های محتوا واحساسات برای تعیین تشابه دو جمله استفاده‌شد. برای شناسایی تعدادی جملات برگزیده[۱] ، از الگوریتم خوشه‌بندی مدیود[۱] استفاده‌شده است.	داده‌های متنی نظرات از ۲ هتل از سایت Tripadvisor.com موردمطالعه قرارگرفته‌است.	[۹]	۱
بهبود نتای حاصل از طبقه‌بندی نظرات متنی	استفاده از الگوریتم‌های چند جمله ای بیز ساده[۱] و بیز ساده‌ی برنولی[۱] جهت طبقه‌بندی نظرات متنی	داده‌های متنی نظرات هتل‌های اروپایی جمع آوری‌شده از وبسایت هتل ها	[۱۲]	۲
کشف مفاهیم و واژه‌های کلیدی در تحلیل احساسات	انجام دقیق و مرحله به مرحله تکنیک‌های تحلیل احساس را برای پیش پردازش متن وتجزیه و تحلیل سپس تولید مصورسازی نتایج	داده‌های متنی نظرات هتل‌های اروپایی جمع آوری شده از وبسایت هتل‌ها	[۱۳]	۳
طبقه بندی مجموعه داده‌های مورد بررسی به نظرات + یا -	استفاده از روش یادگیری‌ماشین بدون نظارت	استفاده از دو مجموعه‌داده از دو وبسایت از Epinions و Slashdot	[۲]	۴
تشخیص نظرات واقعی کاربران از نظرات اسپم یا دروغین توسط ربات یا افراد مغرض	استفاده از الگوریتم‌های یادگیری‌ماشین با نظارت	این مجموعه‌داده شامل ۸۰۰ بررسی صادقانه و ۸۰۰ بررسی فریبنده در بین ۲۰ هتل محبوب شیکاگو توزیع شده‌است.	[۶]	۵
مقایسه داده‌های متنی سه آژانس گردشگری چین و یافتن نقاط قوت و ارتباط آن‌ها با دیگر عوامل	روش‌های تجزیه و تحلیل معنایی و مصورسازی با نرم افزار گفی[۱]	داده‌های سه آژانس گردشگری آنلاین چین	[۲۸]	۶
یافتن کلمه‌های موضوعی مانند اتاق مهمان (مثلاً نمای هتل، راحتی)، مکان (به عنوان مثال، نزدیک فرودگاه) و حمل و نقل (به عنوان مثال، شاتل، پارکینگ)	تجزیه و تحلیل آماری کلمات موضوعی	داده‌های جمع آوری شده از وبسایت هتل	[۱۵]	۷
باید زمینه را در طبقه بندی کلمه‌های‌موضوعی در نظر گرفت. جذابیت بدنی، حس شوخ طبعی و باتجربه‌بودن راهنماهای تور از عوامل مهم مؤثر بر تعامل آن‌ها با گردشگران است .	تجزیه و تحلیل آماری کلمات موضوعی	داده‌های جمع آوری شده از وبسایت هتل	[۷]	۸

[۱۴] N-grams

[۱۵] Supervised Learning-based Approaches

۳-۲ مراحل اصلی فرآیند

شامل جمع آوری داده ها، پیش پردازش متن، فرآیند اصلی، جمع بندی نتایج و تجسم بوسیله ی نمودارها و شکل ها است[۵]. مرحله جمع آوری داده ها: در حال حاضر برای دستیابی به این کار دو رویکرد وجود دارد.

اول از طریق رابط برنامه‌نویسی وب سایت[۱۶] و دوم استفاده از خزنده‌های وب[۱۷] به منظور دستیابی به داده‌ها از وب سایت‌های مورد نظر است.

مرحله پیش پردازش متن: متداول ترین تکنیک ها عبارتند از:

۱- نشانه گذاری[۱۸] : که عملکرد آن باعث جدا کردن رشته متن کامل به لیستی از کلمات جداگانه می‌شود.

۲- یافتن ریشه و بن‌کلمه [۱۹]: برای مثال واژه‌های شخص، اشخاص، شخصیت به بن آن‌ها یعنی شخص تبدیل می‌شوند .

۳- حذف کلمات بی‌اثر[۲۰]: عملکرد آن باعث حذف واژگانی که برای ساخت زبان کاربرد داشته اما در محتوای معنایی آن تاثیر ندارند می‌شود. برخی از این کلمات درزبان انگلیسی ، a ، the و هستند.

۴- بخش‌بندی جمله[۲۱]: عملکرد آن باعث تبدیل پاراگراف‌ها به جملات می‌شود.

۵- برچسب گذاری بخشی از گفتار[۲۲]: عملکرد آن باعث برچسب خوردن هرکلمه، یک جمله یا بخشی از گفتار است. مانند صفت، اسم، فعل، ضرب‌المثل یا پیشگفتار. کاربرد این روش به عنوان ویژگی فرآیند یادگیری‌ماشین استفاده می‌شود.

۳- روش پیشنهادی

در این مقاله، دو فرهنگ واژگان نظر آگاه به‌زمینه با روش-های FBSA[۲۳] و ALGA[۲۴] بر روی مجموعه داده آموزشی نظرات ساخته می‌شود [۱۷]. سپس، بر روی هر مجموعه‌داده، با استفاده از هرکدام از این دو فرهنگ واژگان نظر، ویژگی‌هایی محاسبه

می‌شوند. این ویژگی‌ها در کنار ویژگی‌های محاسبه شده با فرهنگ واژگان نظر عام منظوره لیوبینگ[۲۵] قرار می-گیرند و به

این ترتیب، برای ویژگی‌های حاصل، انتخاب ویژگی توسط آزمون t [۲۶]صورت می‌گیرد.

آزمون t برای تعیین اختلاف میانگین یک گروه با یک مقدار پیش فرض و یا میانگین‌های دو گروه به کار می روند. در واقع، آزمون t یک نوع آمار استنباطی است برای تعیین اینکه آیا بین میانگین دو گروه اختلاف معنی داری وجود دارد یاخیر، چرا که

ممکن است در ویژگی‌های خاصی مرتبط باشند. در انتها، مدلی برای دسته‌بندی از روی مجموعه داده‌های آموزشی ساخته می‌شود و بر روی مجموعه داده آزمایشی اعمال می‌شود.

روش FBSA مبتنی بر ایجاد فرهنگ واژگان احساسی با استفاده از کامنت‌ها و نظرات است که با تناوب و میزان تکرار کلمات، بار احساسی آن‌ها مشخص میشود و روش ALGA

فرهنگ واژگان احساسی نظرات و عقاید بیان شده‌ی متنی را با استفاده از الگوریتم ژنتیک می‌سازد [۱۲] .

۳-۱ تولید واژگان توسط FBSA

در روش پیشنهادی، در این بخش که از تولید واژگان توسط FBSA استفاده می‌شود، از روش تحلیل عبارات مبتنی بر فرکانس به دلیل تخمین دقیق رتبه‌های کلمات مثبت ومنفی استفاده شده است. در روش FBSA برای

---

۱۶ API

۱۷ Web crawlers

۱۸ Tokenization

۱۹ Stemming or lemmizitation

۲۰ Stopword Removal

۲۱ Sentence Segmentation

۲۲ Part-of-Speech (POS)

۲۳ Frequency Based Sentiment Analysis (FBSA)

۲۴ Adaptive Lexicon learning using a Genetic Algorithm (ALGA)

۲۵ Bing Liu’s Opinion Lexicon English.

۲۶ t-test

یافتن بار احساسی واژگان تمام کلمات در نظر گرفته می‌شود و یکی از مزیت‌های این روش عدم حذف کلمات توقف[۲۷] است زیرا می‌تواند در نظرات هتل‌ها تعیین کننده باشد. بدین صورت که رتبه‌ی عبارت برای هر کلمه براساس فرکانسی از کلمات در مجموعه داده‌های آموزشی محاسبه می‌شود. اگر فرض شود که p مجموعه داده داشته باشیم، D۱ تا Dp شامل نظرات و برچسب‌ها هستند که نیمی از آنها شامل نظرات مثبت و نیمی از آن‌ها دارای نظرات منفی هستند. بنابراین، مجموعه داده‌های آموزش و تست با مدل اعتبارسنجی k-دسته[۲۸] تقسیم می‌شوند. در این روش نمونه اصلی به‌طور تصادفی به زیرنمونه‌هایی با اندازه k تقسیم شده و در هر مرحله یک زیر نمونه مورد تحلیل قرار می‌گیرد.

از زیرنمونه‌های k، که در هرمرحله بصورت تصادفی بدست آمده‌اند، یک زیرنمونه منفرد به‌عنوان داده‌های آزمایشی برای اعتبارسنجی الگوریتم و مدل ذخیره شده و زیرنمونه‌های دیگر که درواقع تعداد آن‌ها حالا k-۱ شده‌است نقش داده‌های آموزشی را خواهند داشت. فرایند اعتبارسنجی، که k بار تکرار می‌شود، هر بار بصورت تصادفی مجموعه داده‌ای را انتخاب می‌کند، با هر یک ازاین نمونه‌های k دقیقا یک بار داده‌ها اعتبارسنجی می‌شوند. نتایج k می‌تواند برای برآورد میانگین مورد استفاده قرار بگیرد. در این روش همه مشاهدات برای آموزش و اعتبار مورد استفاده قرار می‌گیرند، و هر مشاهده برای اعتبارسنجی به‌طور دقیق استفاده می‌شود و این مورد برتری این روش نسبت به نمونه‌گیری تصادفی تکراری است. لازم به ذکر است در این مقاله K برابر با ۱۰ در نظر گرفته شده است. بنابراین، برای هر کلمه w_j در مجموعه D_iداده‌های آموزشی، دو مقدار تجمعی تعریف می‌شود: فرکانس‌های مثبت و منفی[۲۹].

$$ferq_+(w_I, D_i)=\sum_{k=1}^{n_i} TF(w_I,R_{I,k}) \qquad (1)$$

در رابطه‌ی (۱) اگر کلاس $R_{I,k}$ مثبت باشد، از این رابطه استفاده می‌شود.

$$ferq_-(w_I, D_i)=\sum_{k=1}^{n_i} TF(w_I,R_{I,k}) \qquad (2)$$

در رابطه‌ی (۲) اگر کلاس $R_{I,k}$ منفی باشد، از این رابطه استفاده می‌شود.

در این روابط، ni تعداد رکوردها در Di ، $R_{I,k}$ شماره رکورد k در مجموعه داده Di و $TF(w_I,R_{I,k})$ تعداد رخدادها در wi در $R_{I,k}$ است. در دیگر کلمات، $ferq_+(w_I, D_I)$ و $ferq\_(W_I, D_I)$ تعداد رخدادی از wi در رکوردهای مثبت و منفی در Di مجموعه داده هستند.

در این روش، Di رکوردهای مجموعه داده‌های آموزشی را شامل می‌شود و رکوردهای تست را در برنمی‌گیرد. مقایسه freq + وfreq − تنها زمانی معنی‌دار است که تعداد سوابق مثبت و منفی برابر است زیرا رکوردهای تست در نظر گرفته نشده و بنابراین، داده‌های آموزشی و تست برابر نیست.

به همین دلیل، از ضریب مبتنی بر سوابق در کلاس‌های مثبت و منفی استفاده شده است و سپس فرکانس نرمال با استفاده از رابطه‌ی (۳) محاسبه می‌شود:

$$ferq_-(w_I, D_i)=\frac{n_p(i)}{n_N(i)}.ferq_-(w_I, D_i) \qquad (3)$$

بنابراین، $n_p(i)$ و $n_N(i)$ تعداد رکوردهای مثبت ومنفی را در Di مجموعه داده نشان می‌دهند. در رابطه‌ی (۴) از رتبه بندی عبارت برای هر کلمه در واژگان استفاده می‌شود که این مقدار برای هر کلمه محاسبه می‌شود:

$$Score(w_I, D_i) = \frac{ferq_+(w_I, D_i)-ferq_-(w_I, D_i)}{ferq_+(w_I, D_i)+ferq_-(w_I, D_i)} \qquad (4)$$

در رابطه‌ی فوق مقدار رتبه بین ۱- تا ۱+ است. هر چقدر این عدد به ۰ نزدیکتر باشد، این کلمه به فاعل یا همان کلمه‌ی ویژگی، نزدیکتر است. رتبه‌ی نزدیک به ۱+ مثبت بودن کلمه و رتبه‌ی نزدیک به ۱- منفی بودن کلمه را نشان می‌دهد. بطور مثال اگر کلمه‌ی Love ۲۸ بار در عبارت بصورت جداگانه ظاهر شود، و ۳ بار بصورت منفی که ۲ بار در یک نظر و ۱ بار در نظر دیگر ، درمجموع بصورت $ferq_+(love, D_i) = ۲۸$ و $ferq\_(love, D_i) = ۳$ نمایش داده می‌شود. پس از این مرحله، همان طور که در رابطه‌ی (۴) نشان داده شد،

[۲۷] Stop-Word

[۲۸] K-Fold

[۲۹] positive and negative frequencies

رتبه‌ی کلمه پس از نرمال شدن محاسبه می‌شود. در ادامه پنج ویژگی برای طبقه بندی مجموعه داده‌ها بصورت زیر معرفی شده اند:

- Fpos: تعداد رتبه‌های کلمات مثبت در نظرات
- Fneg: تعداد رتبه‌های کلمات منفی در نظرات
- Pwords: تعداد کلمات مثبت دررکورد براساس رتبه
- Nword: تعداد کلمات منفی در رکورد براساس رتبه
- Score: مجموع همه‌ی رتبه‌ها در نظرات

در این روش، تولید واژگان مبتنی بر تحلیل عبارات فرکانسی به معنی این است که واژگان تولید شده در عبارت برای کل متن استفاده می‌شود. از آنجا که استثنائی برای مواردی است که نیاز است طبقه بندی شود، واژگان براساس مجموعه داده‌های آموزشی تولید می‌شوند.

هر رکورد در مجموعه داده (هر نظر) باید به یک بردار ویژگی تبدیل شود. فرض شود که یک رکورد شامل "It is good" است، بردار ویژگی توسط جدول (۲) محاسبه می‌شود.

جدول ۲. بردار ویژگی It is good [۱۷]

FPos	FNeg	PWords	NWoreds	Score
۰,۸۴۹	-۰,۱۸۶	۲	۱	۰,۶۶۳

در روش پیشنهادی هر رکورد به چندین ویژگی براساس طول جمله تبدیل می‌شود که در جدول(۲) به پنج ویژگی تقسیم شده است. سپس این مدل برای مجموعه آموزشی ایجاد و برای مجموعه تست بکارمی‌رود.

هسته برداررویژگی از ویژگی‌های فوق الذکر ساخته شده که به‌صورت زیر است و در جدول (۳) نشان داده شده است. این جدول درواقع، نمونه‌ای از محاسبه ویژگی‌ها برای یک رکورد خاص است که که با روش FBSA ایجاد شده است.

جدول ۳. واژگان ساده ایجاد شده توسط روشFBSA [۱۷]

It	For	Is	Good	Of
+۰,۲۲۸	+۰,۳۰۷	-۰,۱۸۶	+۰,۰۶۲۱	-۰,۰۷۴

واژگان نیز براساس داده‌های آموزشی ایجاد می‌شود و برای محاسبه‌ی ویژگی‌ها براساس داده آموزش و تست بکار می‌رود. در این روش تفاوت بین کلمات بدون هشتگ(#) و با هشتگ نیز در نظر گرفته می‌شود، کلماتی که باهشتگ در نظر گرفته می‌شوند، نشان دهنده تاکید و مهم بودن است. لذا، در برخورد با چنین کلماتی به دو صورت رفتار می‌شود و این کلمات در مجموعه آموزشی بسته به مثبت یا منفی بودن نظر ۲ بار شمارش می شوند .

در این مرحله فهرست واژگان احساسی مبتنی بر تحلیل فرکانس ساخته شد، در ادامه به تولید واژگان براساس ABALGA پرداخته می‌شود تا در نهایت این واژگان باهم ادغام شوند.

۳-۲ نحوه‌ی استخراج ویژگی‌ها و جنبه‌های مختلف مطرح شده در متن نظرات

در این مقاله از روش IOB- encoding که در کتابخانه-ی nltk.corpus.reader با دستور import ConllChunkCorpusReader فراخوانی می شود که، برای استخراج جنبه‌های مختلف و صریح بیان شده درهر توئیت استفاده شده است. در این روش کلمات به کار رفته در توئیت‌ها برچسب گذاری می‌شوند که در آن B-POS نشانه‌ی جنبه‌ای است که در ابتدای توئیت شناسایی شده B، نشانه‌ی شروع جنبه‌ی جدید و O نشانه‌ی عدم شناسایی به عنوان جنبه و ویژگی می‌باشد. نمونه‌ای از خروجی در جدول (۴) نمایش داده شده است.

جدول ۴. توئیت های برچسب گذاری شده با IOB-encoding

Words:	**Bathroom**	was	clean	,	but	**bed**	is	not comfort
Labels:	**B-POS**	O	O	O	O	**B**	O	O	O

در این مقاله از مجموعه داده‌ی مورد نظر ۱۰۰۰ توئیت بررسی شد و پنج ویژگی با جنبه‌ی اصلی مختص هتل شناسایی شد و به صورتی که در جدول (۵) نمایش داده شده، دسته‌بندی گردید تا بتوان جنبه‌های مهم را شناسایی و بار احساسی کلمات مطرح شده را در نظرات حول این ویژگی‌ها شناسایی کرد .

جدول ۵. دسته بندی جنبه‌های استخراج شده

Value	Location	Service	Meal	Room
Price Amount Rate Cheap Worth Low Money Economic Fee expensive	Railway View Airport Mall Far Close Near Metro distance market	Check-in Check-out Staff Ticket Transport	Breakfast Lunch Dinner Coffee Tea Drink Restaurant Bar	Bed Bathroom View Shower Air condition Bedsheets Tv Furniture

۳-۳ تولید واژگان توسط ABALGA

در این مقاله یادگیری فهرست واژگان تطبیق شده با ویژگی‌ها، با استفاده از الگوریتم ژنتیک و در سطح توییت طراحی شده‌است. هر توییت ممکن است یک جمله کوتاه یا یک جمله طولانی از ۱۴۰ کاراکتر تا ۲۸۰ کاراکتر باشد. روش پیشنهادی برای نظرات کاربران صنعت هتلداری استفاده می‌شود و در این مقاله متن نظرهای کوتاه‌تر و یا بلندتر مدنظر قرار نگرفته است.

در الگوریتم پیشنهادی، یک توالی ژنی تعریف می‌شود تا زمانی که کلمات موجود در مجموعه آموزش، مشغول یادگیری واژگان هستند، ادامه می‌یابد. هر ژن در توالی ژن ذکر شده دارای نمره‌ای برای کلمه مربوطه است. به عبارت دیگر، درصدد بهینه سازی رتبه‌ی احساسی واژگان هستیم که این بهینه سازی با عملیات کراس آور و جهش که در الگوریتم ژنتیک در جهت ارائه بهترین راه حل صورت می‌گیرد، انجام می‌شود که درادامه تشریح شده است. در ALGA ، رتبه‌های کلمات، ژن‌ها هستند. کمبود ALGA در مشکلات مبتنی بر جنبه از ساختار نظرات ناشی می‌شود.

بطور مثال در تحلیل نظرات، گاهی به هر توئیت یک برچسب نسبت داده می‌شود که این برچسب برای کل کلمات در نظر گرفته می‌شود. از طرفی دیگر هر جمله ممکن است جنبه‌های مختلفی داشته باشد که ناشی از بخش‌های مختلف نظرمطرح شده باشد که نیاز به تجدید نظر دارد لذا همانطور که در بخش قبلی توضیح داده شد، جنبه‌های مختلف بیان شده درهر توئیت استخراج می‌شود.

۳-۳-۱ ادبیات تحقیق (الگوریتم ژنتیک)

امروزه الگوریتم ژنتیک جایگاه ویژه‌ای در میان الگوریتم‌های بهینه سازی برای حل مسائل پیچیده دارد زیرا از لحاظ محاسباتی ساده، در عین حال قدرتمند است، همچنین در هر مرحله فضای جستجو در مجموعه‌ی داده محدود نمی‌شود[۱۹].

الگوریتم‌های فراابتکاری همچون الگوریتم ژنتیک، یکی از الگوریتم های جستجو به حساب می‌آید و از طریق تعامل با اعضا، در پی یافتن جواب بهینه‌ی سراسری هستند. در همه این الگوریتم ها، جواب‌های بهتر، شانس بیشتری برای حضور در تکرارهای بعدی الگوریتم و تولید نسل بعد دارند که این ویژگی خاص " حیات مناسب‌ترین[۳۰]" موجب یافتن نتایج بهتر است[۱۸].

شرط پایان الگوریتم، رسیدن به حداکثر تعداد تکرار از پیش تعیین شده، عدم بهبود جواب در چند تکرار پیاپی می‌باشد[۲۰].

۳-۳-۲ پیش پردازش

مراحل پیش پردازش در ABALGA شامل موراد زیر است:

- ✓ جداسازی کلمات در هر بررسی[۳۱]
- ✓ حذف کلمات توقف
- ✓ واژهای فیلتر شده براساس برچسب‌های مثبت یا منفی
- ✓ لمس کردن کلمات باقی مانده (به جز اصطلاح جنبه)
- ✓ ساخت مجموعه ای از کلمات ریشه[۳۲] از مرحله قبل
- ✓ حذف کلمات با فرکانس کمتر از ۳
- ✓ گرفتن پنجره با اندازه ثابت در حدود اصطلاحات

بعنوان نمونه‌ای از پیش پردازش که در موارد بالا ذکر شد، منظور از جداسازی، تشخیص مرز کلمات در متون است،

[۳۰] Survival of the fittest

[۳۱] tokenization of words in each review

[۳۲] lemmatized

بدین صورت که متن را به دنباله‌ای از کلمات تبدیل می-کند.

در مورد دوم، حذف کلمات توقف، برخی از کلمات همانند is ، the و... که ارزش احساساتی ندارند، از جمله حذف می‌شوند.

در مرحله‌ی فیلتر کردن حذف یا نگهداری برخی از کلمات می‌تواند در نظر گرفته شود. همچنین در مراحل بعد بازگردان شکل کلمه به حالت ریشه و بن، حذف کلمات کمتر از سه حرفی که ارزش محاسباتی ندارند.

در مرحله‌ی آخر، گرفتن پنجره با اندازه پنج در حدود جنبه یا اصطلاح room می‌تواند در توئیت “The room was clean and I satisfied the view is good.” بصورت زیر در نظر گرفته می‌شود.در این مرحله تعداد کلمات توقف هم شمارش می‌شود و جدول (۶) نشان داده شده است.

جدول۶. پنجره‌ی کلمات در حدود جنبه

[5words before]Aspect [5words after]						
The	**room**	was	clean	and	i	satisfied

در ادامه الگوریتم ABALGA شامل ساختار کروموزوم‌ها[۳۳]، تابع ایجاد[۳۴]، برازندگی[۳۵] و تابع جهش[۳۶] و تابع کراس آور[۳۷] و انتخاب والد ها تشریح خواهند شد.

۳-۳-۳ ساختار کروموزوم ها

کروموزم‌ها در در الگوریتم ABALGA اطلاعات زیر را نگهداری می‌کنند:

- ژن‌ها
- برازندگی
- سن
- استراتژی

موازی سازی با مجموعه ای از کلمات ریشه ذکر شده در بخش پیش پردازش، به عنوان یکی از پارامترهای ورودی در ABALGA ، یک توالی ژنی به طول یکسانی در این مجموعه قرار دارد. هر ژن برای یک کلمه ریشه‌ی مربوطه در ورودی دارای یک رتبه‌ی شناور در محدوده ۱- تا ۱ است. در مرحله‌ی اول رتبه‌ها به‌صورت تصادفی به هر ژن بصورت عددی در بازه‌ی ۱ - تا ۱ داده می‌شود.

تابع برازندگی برای از بین بردن محاسبات زائد و سرعت بخشیدن به‌کار می‌رود و در کروموزوم‌ها ذخیره می‌شود.

در روش پیشنهادی برای هر توالی ژن، تابع برازندگی فقط یک بار در جهت سرعت بخشیدن به روند، محاسبه می‌شود مگر اینکه با جهش یا کراس آور تغییر کند[۱۹].

پارامتر سن تعداد دفعاتی را که توالی فعلی پس از انجام جهش یا کراس آور شکست خورده است، را شمارش می-کند.

پس از رسیدن به حداکثر سنی، الگوریتم، کروموزوم‌ها را از مخزن پدر حذف می‌شوند و کروموزوم فرزند ایجاد می-شود. این روند بعد از عمل جهش و کراس آور بوجود می-آید[۱۹].

کروموزوم همچنین اطلاعات استراتژی را که یکی از توابع ایجاد، جهش یا کراس آور است، ذخیره می‌کند و نشان می‌دهد که کدام عملکرد منجر به کروموزوم فعلی شده است.

۴-۳-۳ تابع ایجاد

تابع ایجاد بعنوان ورودی طولی از کلمات را می‌گیرد و لیستی با همان طول را تولید می‌کند که حاوی مقادیر شناور تصادفی در دامنه ۱- تا۱+ است که در واقع رتبه کلمات در مجموعه کلمات است. بطور مثال مقدار ژن مربوطه واژه ”polite“ در کروموزم ۱ می‌تواند ۰,۳۱+ باشد که نشانگر درجه احساسات نسبتاً مثبت برای واژه‌ی مذکور است ، در حالی که در کروموزوم ۲ ، مقدار ژن مربوطه می‌تواند ۰,۴۶- باشد ، که نشان دهنده رتبه احساسات منفی است. نمونه‌ای از این توالی در جدول (۷) نشان داده شده است.

جدول۷. نمونه ای از توالی های ژن ها در الگوریتم ژنتیک

کلمات	near	Far	clean	warm	danger	fair	Polite
توالی ۱	+۰,۴۵	-۰,۷۸	+۰,۵۳	+۰,۲۵	-۰,۴۲	+۰,۲۳	+۰,۸۰

[۳۳] Chromosome structure

[۳۴] Create

[۳۵] Fitness

[۳۶] Mutate function

[۳۷] Crossover function

توالی ۲	+۰.۳۲	-۰.۵۱	+۰.۳۲	+۰.۴۱	-۰.۱۷	-۰.۳۶	-۰.۱۴
توالی ۳	-۰.۰۲۱	-۰.۶۵	+۰.۴۱	-۰.۳۶	-۰.۴۵	+۰.۰۴۸	+۰.۰۱۲
توالی ۴	+۰.۶۳	+۰.۱۲	+۰.۲۳	-۰.۵۲	+۰.۳۱	-۰.۲۲	+۰.۷۹

۳-۳-۵ برازندگی

در روش پیشنهادی تابع پاداش و مجازات[۳۸] الگوریتم ABALGA متفاوت از ALGA است. عملکرد پاداش و مجازات دوباره در مقایسه با تابع پاداش و مجازات منفرد درALGA طراحی شده است. با دادن Dm مجموعه داده و با استفاده از Lk واژگان برای هر کلمه‌ی ریشه یابی شده، lem(wi) در wiپنجره اطراف واژه‌ی جنبه‌ی مربوطه در الگوریتم ABALGA از رابطه‌ی (۵) محاسبه می‌شود:

(۵)

$$ABALGA(D_m,W_I,L_k)=\sum_{W_j} S_k(lem(w_i)W_j$$

بطوریکه $S_k(lem(w_i)W_j$ بصورت رابطه‌ی (۶) محاسبه می‌شود:

(۶)

$$S_k(lem(w_i)W_j = \begin{cases} -S_k(lem(w_i)W_j & if\ w_{j-1} \in in\ Negations \\ S_k(lem(w_i)W_j & otherwise \end{cases}$$

در رابطه‌ی (۵)، $S_k(lem(w_i)W_j$ رتبه‌ای از هر کلمه‌ی lemmatized ، wj در Wiپنجره اطراف واژه‌ی جنبه‌ی مربوطه است. در رابطه‌ی (۶) رتبه‌ی احساسات یک کلمه در یک پنجره را بیان می‌کند، اگر کلمه‌ای با بار منفی قبل از کلمه‌ی جاری باشد، در محاسبه رتبه‌ی پنجره معکوس می‌شود. لیستی از کلمات با بارمنفی در جدول(۸) نشان داده شده است.

جدول ۸. لیست کلمات منفی [۱۷]

Not	n't	No	Barely	rarely	Nerever	Hardly

برای تمام پنجره‌هایی که حاوی اصطلاح جنبه هستند، محاسبه می‌شود. اگر رتبه برای یک پنجره مثبت باشد، برچسب جنبه مربوط به عنوان مثبت و در غیر این صورت منفی شمارش می‌شود. بنابراین ، پیش بینی برچسب مبتنی بر ABALGA در پنجره به شرح زیر محاسبه می‌شود:

(۷)

$$Label(D_m,W_I,L_k) = \begin{cases} Positive\ if\ ABALGA(D_m,W_I,L_k) > 0 \\ negative\ if\ ABALGA(D_m,W_I,L_k) \le 0 \end{cases}$$

در روش پیشنهادی از تابع برازندگی برای قیاس کردن کروموزم‌ها استفاده می‌شود. الگوریتم ABALGA دارای دوهدف است: هدف اصلی این است که تعداد تطبیق‌ها را حداکثر کند. درواقع کروموزومی که تعداد تطبیق بالاتری داشته باشد نسبت به دیگر کروموزم‌ها ارجح خواهد بود. اگر حالت تطبیق‌ها محاسبه شده یکسان باشد، در تابع برازندگی، مقدار قدرمطلق تفاضل برای بدست آوردن رتبه استفاده می‌شود. دیگر هدف این است که به تعداد تطبیق na به صفر برسد.

این عملکرد پاداش و مجازات تجدید نظر شده ABALGA، در روش ما ، شبیه به ALGA [۱۸] است.

۳-۳-۶ تابع جهش

عملکرد جهش یک توالی ژنی و تابع برازندگی و همچنین پنجره‌های اطراف اصطلاحات را به عنوان پارامترهای ورودی در نظر می‌گیرد و برای بهبود تابع برازندگی خود سعی در تغییر تابع برازندگی قبلی دارد. واژگان کاندیدا در پنجره عبارت‌هایی هستند که برچسب محاسبه شده از رابطه‌ی (۷) با برچسب ابعاد واقعی آن مطابقت نداشته است.

۳-۳-۷ تابع کراس آور

توالی ژن به نام‌های والدین و اهداکننده‌ها و مقادیر تابع برازندگی آن‌ها، پارامترهای ورودی تابع کراس‌آور الگوریتم ABALGA است. هنگامی‌که ژن‌های والدین و اهداکننده یکسان هستند، از تابع ایجاد برای جایگزینی ژن‌های اهداکننده با توالی ژن جدید استفاده می‌شود زیرا توالی ژن‌های یکسان در مخزن والدین مطلوب نیستند. اگر این محدودیت ارضا نشود، سپس، به تعداد تصادفی از دفعات، در دامنه ۱ تا ۱۰ ، تابع در ژن‌های اهدا کننده که ژن‌های مشابه آن‌ها در ژن‌های والدینی که یکسان نیستند،

[۳۸] Reward and Penalty

قرار می‌دهد. سپس ژن‌های مستقر در نسخه‌ای از ژن‌های والدین استخراج و جایگزین می‌شوند. در صورت پیشرفت پس از این جایگزینی، تابع ژن‌های فعلی را برمی‌گردانند و در صورت عدم پیشرفت، تابع دوباره سعی می‌کند تا به یک آستانه حداکثر برسد، سپس در آن مرحله آخرین نسخه کپی شده ژن‌های والدین را برمی‌گرداند.

۳-۳-۸ انتخاب والدین

هدف اصلی سوق دادن جستجو به بخشهایی از فضا که امکان یافتن جوابهای با کیفیت بالاتر وجود دارد.در هر نسل تعدادی از عناصر جمعیت این فرصت را پیدا می کنند که تولید مثل کنند. به این عناصر که از میان جمعیت انتخاب می شوند، والدین می گویند.

به این منظور در ابتدا همه‌ی واژه‌ها به عنوان والد شناخته می‌شوند، سپس در هر مرحله با استفاده از تابع برازندگی که در قسمت قبلی توضیح داده شد هر کروموزومی که تعداد تطبیق بالاتری داشته باشد به عنوان والد انتخاب می شود[۱۹].

۳-۴ ادغام واژگان و طبقه بندی

به منظور طبقه بندی نظرات ، واژگان FBSA ، ABALGA، فرهنگ واژگان SentiWordNet، و فرهنگ واژگان احساسی لیو بینگ با یکدیگر ترکیب شده و برای رتبه دهی کلمات استفاده می‌شوند. بدین صورت که اندازه پنجره ۵ در نظر گرفته می‌شود و کلمات بیان شده در توئیت به عنوان دامنه‌ی ورودی به این پنجره‌ها و امتیازات شمارش شده‌ی زیر بعنوان ویژگی‌ها برای هر جنبه استخراج می‌شوند. بنابراین، ویژگی‌های استخراج شده برای هر پنجره بصورت زیر هستند:

ScoreSum: مجموع رتبه‌های کلمه در پنجره بر طبق واژگان

NormalizedScoreSum: مجموع رتبه کلمات در یک پنجره با توجه به واژگان استفاده شده طول، نشانه‌ها، از پنجره تقسیم می‌شود.

DistSum: مجموع، هر کلمه ، از نمرات تقسیم شده با فاصله آن از اصطلاحات.

PosCount: تعداد کلمات، در پنجره فعلی، با توجه به واژگان در حال استفاده، دارای رتبه‌ی مثبت است.

NegCount: تعداد کلمات، در پنجره فعلی، با توجه به واژگان در حال استفاده، دارای رتبه‌ی منفی است.

CountSum: تعداد کلمات با نمره مثبت منهای تعداد کلمات با نمره منفی ، در پنجره فعلی ، مطابق واژگان در حال استفاده.

PosSum: مجموع رتبه‌های مثبت کلمات در پنجره فعلی برای واژگان جاری

NegSum: مجموع رتبه‌های منفی کلمات در پنجره فعلی برای واژگان جاری

طبقه بندهای متعددی شامل درخت تصمیم ، درخت تصادفی ، جنگل تصادفی ، ماشین بردار پشتیبان، بیزین و طبقه بندی نزدیکترین همسایگی برای روش پیشنهادی در نظر گرفته می‌شود. در روش پیشنهادی، واژگان لیو بینگ بعنوان پشتیبان اصلی در الگوریتم FBSA یا ABALGAاستفاده می‌شود.

بنابراین؛ واژگان SentiWordNet بعنوان اخرین لایه برای رتبه دهی کلمه بصورت رابطه‌ی (۸)، (۹) و (۱۰) محاسبه می‌شود:

$$Pos_{score}(w_i) = \frac{\sum_i^n p(i)}{totalSyn(w_i)} \quad (۸)$$

$$Neg_{score}(w_i) = \frac{\sum_i^n N(i)}{totalSyn(w_i)} \quad (۹)$$

$$Obj_{score}(w_i) = \frac{\sum_i^n O(i)}{totalSyn(w_i)} \quad (۱۰)$$

در روابط فوق، p(i) ، N(i) و O(i) رتبه‌های مفعولی هستند. $totalSyn(w_i)$ در واقع تعداد Synset برای wi است.

درنهایت، رتبه نهایی برای هر کلمه با استفاده از واژگان SentiWordNet با استفاده از رابطه‌ی (۱۱) محاسبه می‌شود:

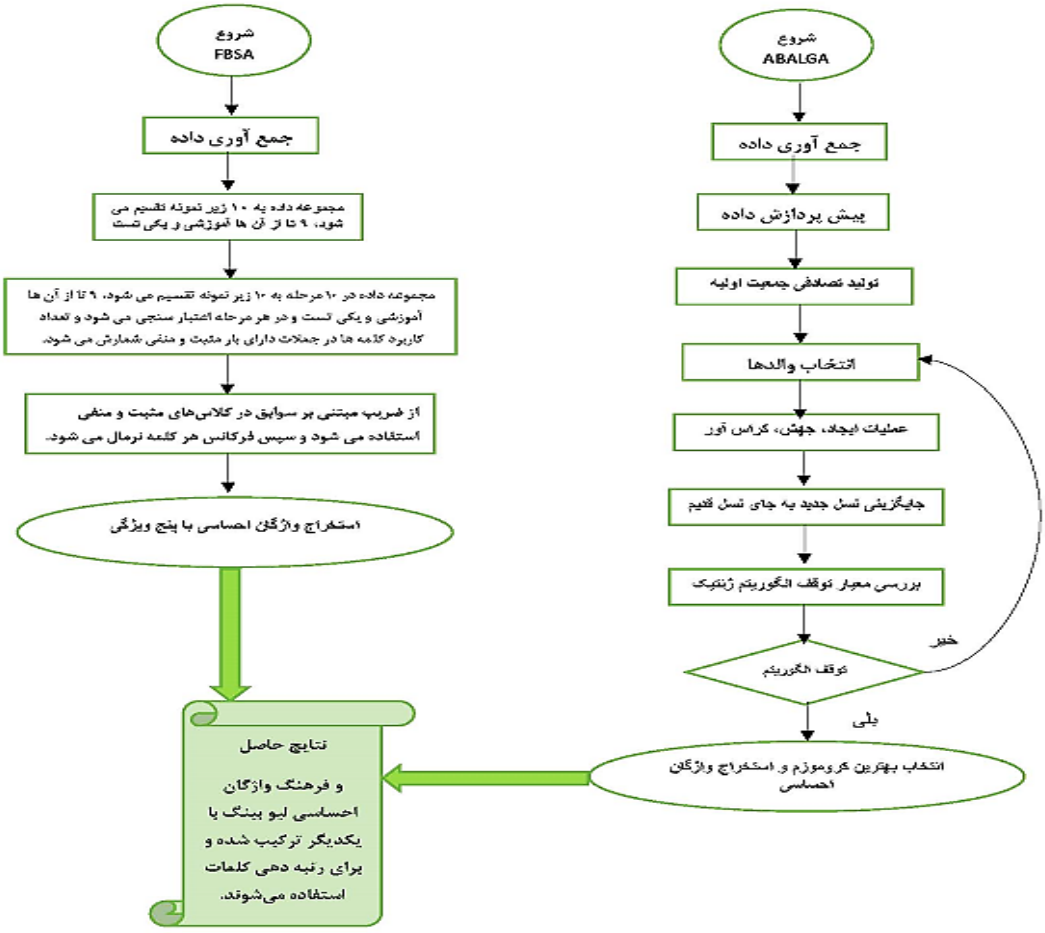
$$\text{word score}(w_i) = \begin{cases} Pos_{score} & \text{if } Pos_{score} > Neg_{score} \text{ and } Pos_{score} > Obj_{score} \\ Neg_{score} & \text{if } Neg_{score} > Pos_{score} \text{ and } Neg_{score} > Obj_{score} \\ Obj_{score} & \text{otherwise} \end{cases} \quad (۱۱)$$

ABALGA واژگان احساسی را به‌صورت پویا و FBSAاحساسات درون کلمات را بصورت استاتیک (ثابت) ضبط می‌کنند، رتبه کلمات احساسی بصورت پویا، ممکن است نادرست نشان داده شود و از سوی دیگر، واژگان استاتیک رتبه دقیق‌تری از احساسات ارائه می‌دهند

اما قادر به در نظر گرفتن رتبه‌ی احساسی در یک متن نیستند. از این‌رو، ادغام واژگان پویا و استاتیک معیارهای عملکرد را در مقایسه با استفاده از تنها یک نوع واژگان افزایش می‌دهد لذا این الگوریتم رتبه‌دهی احساسی واژگان در این مقاله برای استخراج ویژگی‌های کلمات در نظر گفته شده است.

در روش پیشنهادی، پس از پیش پردازش مجموعه داده‌ی نظرات کاربران، به استخراج ویژگی‌های متا پرداخته شد.در این روش، از یک روش ترکیبی و جدید با استفاده ازروش FBSAکه مبتنی بر ایجاد فرهنگ واژگان نظر با استفاده از میزان تکرار و تناوب کلمات است و ABALGA که فرهنگ واژگان احساس نظر را با استفاده ازالگوریتم ژنتیک می‌سازد استفاده و در مرحله‌ی نهایی به طبقه بندی مجموعه داده‌ی نظرات صنعت هتلداری پرداخته شد.

درشکل(۱) مراحل کلی بیان شده در روش پیشنهادی را نشان می‌دهد.



شکل ۱. مراحل کلی مدل پیشنهادی

۴- نتایج و تفسیر داده ها

در این بخش مجموعه داده‌ی مورد نظر مورد تحلیل قرار گرفت ونتایج حاصل از تحلیل هایی که توسط الگوریتم‌های مربوطه در مدل پیشنهادی بکار رفته بود با نتایج دیگر مقاله ها مقایسه شد.نتایج حاصل نشان داد که مدل پیشنهادی نسبت به مقاله-های قبلی عملکرد مناسبتری دارد.

۴-۱ مجموعه داده‌ها

در این مقاله مجموعه داده‌ای متنی مرتبط با نظرات کاربرانی که از خدمات هتل استفاده کرده‌اند، فراهم شده که شامل لیستی از نام هتل‌های اروپایی و نظرات وعقاید متنی به زبان انگلیسی توسط کاربرانی[۳۹] که قبلا از خدمات هتل مربوطه استفاده کرده‌اند می‌باشد که از سایت کگل[۴۰] فراهم آورده شده است.

سایت کگل به پژوهشگران این امکان را می‌دهد تا مجموعه داده های مناسب مقاله در زمینه علم داده را پیدا کنند همچنین اگر مجموعه داده ای دارند که نیاز به تحلیل دارد را منتشر کنند و با کمک افراد حرفه ای و متخصص در یک محیط دانش مبتنی بر وب ، دانش پنهان در مجموعه داده را کشف و مدل‌های مناسب و کاربردی بسازند.

مجموعه داده منتخب این مقاله شامل داده های متنی ۱۴۹۳ هتل با مقادیر رتبه‌بندی شده منحصربفرد در اروپا می‌باشد. این مجموعه داده‌ها شامل ۵۱۵۰۰۰ نظر مشتری است که همگی به زبان انگلیسی بیان شده‌اند. برای تحلیل بیشتر، مکان‌های جغرافیایی هتل‌های مختلف در این مجموعه داده در نظر گرفته شده که در قالب فایلی باپسوند CSV تهیه و تنظیم شده است.

مجموعه داده‌ی فوق الذکر دارای ۱۷ ستون است، که عناوین آن عبارت است از:

۱)Hotel_Address,۲)Additional_Number_of_Scoring,۳)Review_Date,۴)Average_Score,۵)Hotel_Name,۶)Reviewer_Nationality,۷)Negative_Review,۸)Review_Total_Negative_Word_Counts,۹)Total_Number_of_Reviews,۱۰)Positive_Review,۱۱)Review_Total_Positive_Word_Counts,۱۲)Total_Number_of_Reviews_Reviewer_Has_Given,۱۳)Reviewer_Score,Tags,۱۴)days_since_review,۱۵)lat,۱۶)lng, ۱۷)Sty

ما در این پژوهش تنها ستون نظرات مثبت و نظرات منفی کلاس‌بندی شده را در نظر می‌گیریم. که نمونه‌ی آن در شکل زیر نمایش داده شده‌است.

Positive_Review	Negative_Review
Location of hotel was good	I am so angry I made this post….
Very nice and amazing hotel	No negative
Very clean and staff are polite	Rooms are nice but elderly a bit difficult
Great location	My room was dirty
Foods are delicious	Im sad because of food..

شکل ۲. نمونه‌ای از مجموعه داده

۴-۲ معیارهای ارزیابی

در مطالعات قبلی صورت‌گرفته روی داده‌های متنی ازجمله روی داده‌های مربوط به صنعت گردشگری و هتل‌داری، از معیارهای ارزیابی صحت[۴۱]، بازخوانی[۴۲]، دقت[۴۳]و معیار F [۴۴] [۴۴] استفاده شده است، لذا برای مقایسه مدل پیشنهادی این مقاله با سایر روش‌ها از معیار های فوق که از معیارهای اصلی سنجش میزان دقت مدل‌ها می‌باشد، استفاده گردیده است.

۴-۲-۱ دقت، بازخوانی و ماتریس درهم ریختگی[۴۵]

ماتریس درهم ریختگی یک ماتریس مربعی Nدر N می-باشد که N همان تعداد برچسب ها وکلاس‌های مشخص شده در دسته بند تعریف شده است. پارامترهای مرتبط با دقت و بازخوانی توسط الگوریتم‌های داده‌کاوی بطور کلی و بخصوص در موضوع مقاله تحلیل احساسات و عقیده‌کاوی بصورت زیر تعریف می‌شوند:

[۳۹] ۵۱۵K hotel reviews in Europe

[۴۰] WWW.Kaggle.com

[۴۱] Accuracy

[۴۲] Recall

[۴۳] Precision

[۴۴] F-Measure

[۴۵] Confusion matrix

پارامتر[۴۶] TP بیان کننده‌ی تعداد ویژگی‌هایی است که به درستی توسط مدل انتخاب شده به عنوان ویژگی بازیابی شده‌اند .

پارامتر[۴۷] FP مبین تعداد ویژگی‌هایی است که به صورت نادرست توسط مدل انتخاب شده به عنوان ویژگی بازیابی شده‌اند.

پارمتر[۴۸] TN مربوط به تعداد ویژگی‌هایی است که به درستی توسط مدل انتخاب شده به عنوان ویژگی بازیابی نشده‌اند.

پارامتر[۴۹] FN مربوط به تعداد ویژگی‌هایی است که به نادرستی توسط مدل انتخاب شده به عنوان ویژگی بازیابی شده اند.

با استفاده از پارامترهای تعریف شده در بالا و اطلاعات بازیابی شده توسط الگوریتم های متن کاوی می‌توانیم کارایی مدل های مختلف را با استفاده از رابطه‌های زیر ارزیابی کنیم.

$$Precision = TP/ (TP +FP) \quad (۱۲)$$

$$Recall = TP/ (TP+FN) \quad (۱۳)$$

معیارPrecision دررابطه (۱۲) میزان دقت مدل انتخاب شده را مشخص می‌کند ومیزان برچسب‌های درست را نشان میدهد

معیار Recall دررابطه (۱۳) بیان کننده نسبت تعداد داده‌های متنی درست دسته‌بندی شده که توسط ما برچسب زده شده در یک کلاس خاص، به تعداد کل داده‌های متنی است که باید در همان کلاس خاص دسته‌بندی شوند .

۴-۲-۲ معیار ترکیبی F[۵۰]

برای ارزیابی عملکرد دسته‌بندها بسیار مورد استفاده قرار می‌گیرد و از ترکیب دو پارامتر دقت و بازخوانی حاصل می-شود.این معیار ترکیبی به صورت زیر بدست می‌آید:

$$F\text{-measure}= ۲*((precision*recall)/(precision+recall)) \quad (۱۴)$$

۴-۲-۳ صحت

صحت معیاری است که در روش‌های ارایه شده در داده کاوی برای طبقه بندی‌ها کاربرد داشته و به میزان نزدیکی پیش بینی‌های مدل پیشنهادی با میزان اندازه گیری شده با مقدار واقعی اشاره دارد و بوسیله‌ی رابطه‌ی (۱۵) به صورت تقسیم تعداد کلماتی که به درستی طبقه‌بندی‌شده اند به تعداد کل کلمات در مجموعه داده مورد نظرمحاسبه می‌شود.

$$Accurency = \frac{T_p+T_N}{N} \quad (۱۵)$$

۴-۳ نتایج آزمایش‌ها

۴-۳-۱ محیط آزمایش و شبیه سازی مدل پیشنهادی

الگوریتم مدل پیشنهادی این مقاله در محیط برنامه نویسی آنا کوندا[۵۱] که یک توزیع متن باز برای زبان‌های برنامه نویسی پایتون و R می‌باشدو در غالب نرم افزار اسپایدر[۵۲] و نرم افزار R پیاده‌سازی شده است.

۴-۳-۲ پارامترهای استفاده شده برای ارزیابی مدل پیشنهادی

برای ارزیابی نتایج حاصل شده از روش و مدل پیشنهاد شده با سایر روش‌ها از سه معیار بیان شده در قسمت قبلی یعنی دقت، صحت و معیار بازخوانی، مورد استفاده قرارگرفته اند.

۴-۳-۳ نتایج تحلیل روش پیشنهادی

در این مقاله، نظرات کاربران به خوب(مشتری راضی) و بد (مشتری ناراضی)تقسیم می‌شود. بدین شکل که رتبه کلی نظرات بد بصورت ratings <۵ و نظرات خوب بصورت ratings =<۵ می‌باشند که در شکل(۳) نشان داده شده است.

---

[۴۶] True Positive

[۴۷] False Positive

[۴۸] True Negative

[۴۹] False Negative

[۵۰] Hybrid F-measure

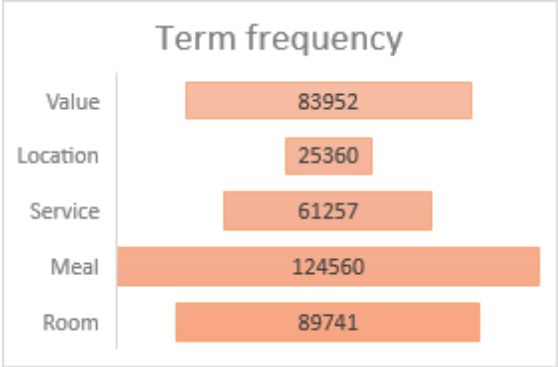
[۵۱] AnaConda

[۵۲] Spyder

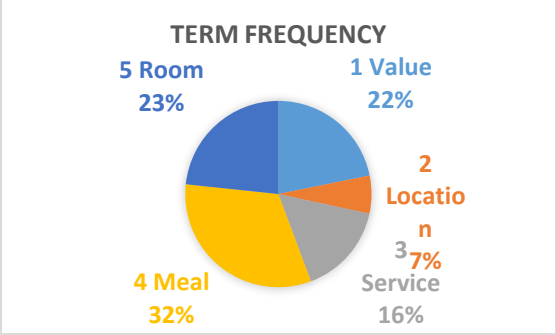
	review	is_bad_review
0	I am so angry that i made this post available...	1
1	No Negative No real complaints the hotel was g...	0
2	Rooms are nice but for elderly a bit difficul...	0
3	My room was dirty and I was afraid to walk ba...	1
4	You When I booked with your company on line y...	0

شکل (۳) دسته بندی نظرات خوب و بد

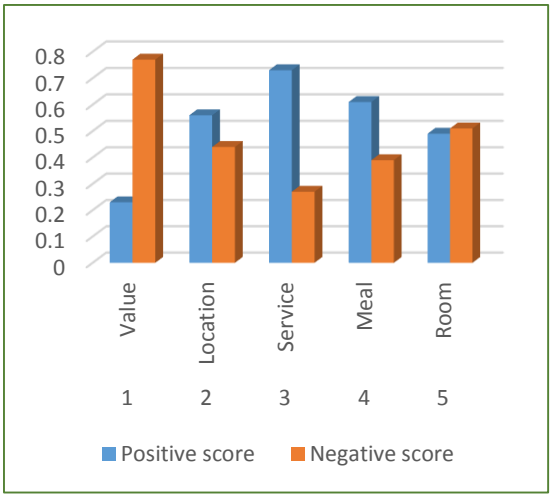
حال در ادامه با توجه به جنبه‌های استخراج شده باتوجه به اینکه کلمه‌های دسته‌بندی شده در هر گروه دقیقا چندبار در سند ظاهر می‌شود و اهمیت نسبی آن‌ها برای متصدیان امور جهت اتخاذ تصمیمات مقتضی در نمودار(۱) و نمودار(۲) به ترتیب تعداد و درصد تکرار آن‌ها و در نمودار(۳) امتیاز بار احساسی استنباط شده از آن‌ها را نشان داده شده است.



نمودار(۱) تعداد تکرار جنبه‌های استخراج شده

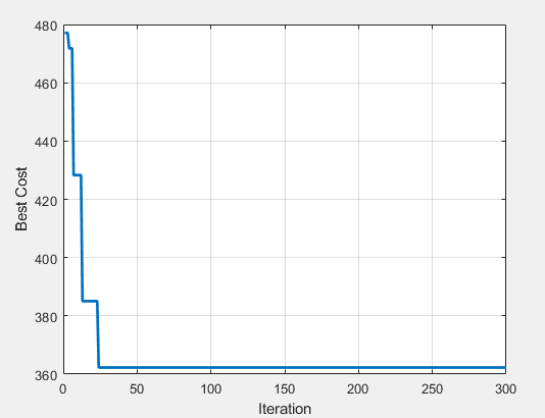


نمودار(۲) درصد تکرار جنبه‌های استخراج شده



نمودار (۳) امتیاز بار احساسی ویژگی‌های استخراج شده

پس از استخراج ویژگی‌های ترکیبی، الگوریتم ژنتیک جهت بهینه سازی واژگان بکاررفته که در شکل(۳) نشان داده شده است.



شکل (۳) همگرایی بهینه سازی ویژگی‌ها

شکل(۳) نشان می‌دهد الگوریتم ژنتیک پس از ۳۰۰ بار تکرار به بهترین تابع برازندگی خود می‌رسد و همگرا می-شود.

در جدول (۹) معیارهای دقت، صحت و F در مرحله‌ی آخر که ادغام وطبقه‌بندی می‌باشد مرحله به مرحله نشان داده شده است.

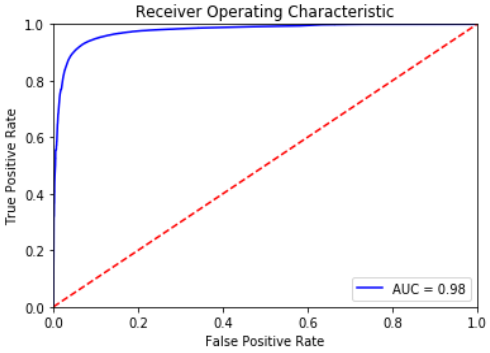
جدول(۹) معیارهای محاسبه شده در مرحله ی ادغام و طبقه بندی

نتایج مراحل ادغام و طبقه بندی	Precision	Recall	F-measure
ScoreSum	۹۰,٤	۹۰,٥	۹۰,٤
NormalizedScoreSum	۹۱,۸	۹۱,۸	۹۱,۸
DistSum	۹۰,٦	۹۰,۷	۹۰,٦
PosCount	۹۱,۹	۹۲,۰	۹۱,۹
NegCount	۹٤,۸	۹۳,۲	۹۳,۱
CountSum	۹۲,۰	۹۲,۰	۹۲,۰
PosSum	۹٤,۱	۹٤,۰	۹٤,۰

CountSum	۹۲,۰	۹۲,۰
PosSum	۹۴,۱	۹۴,۰
نتیجه نهائی	۹۵,۱۷	۹۳,۸۹

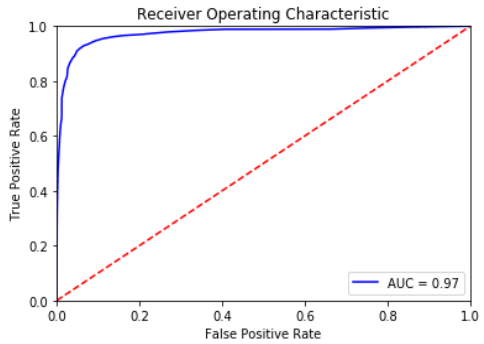
۴-۳-۴ مقایسه طبقه بندی‌ها در روش پیشنهادی

در این بخش به مقایسه طبقه‌بندهای مختلف مانند نایو بیزین ، ماشین بردار پشتیبان و رگرسیون لجستیک[۵۳] در روش پیشنهادی پرداخته می‌شود. شکل‌های (۴)، (۵) و(۶) نشان دهنده‌ی نرخ صحت تشخیص نظرات مثبت و منفی می‌باشد.

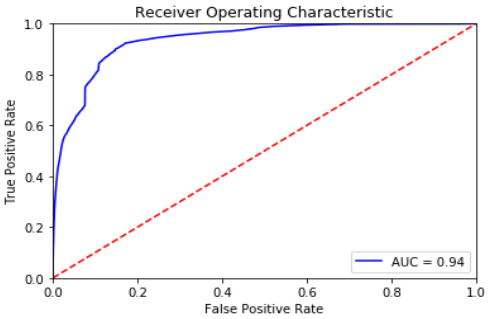


شکل (۴) منحنی ROC و نمرات AUC ازطبقه بندی NB

همانطور که از شکل‌های (۶)، (۷) و(۸) مشخص است، سطح زیر نمودار طبقه‌بند بیزین بیش‌تر بوده و نسبت به دو طبقه‌بند ماشین بردار پشتیبان و رگرسیون عملکرد بهتری دارد. درحالیکه روش ماشین بردار پشتیبان در رده دوم و روش رگرسیون در رده سوم قرار دارد.



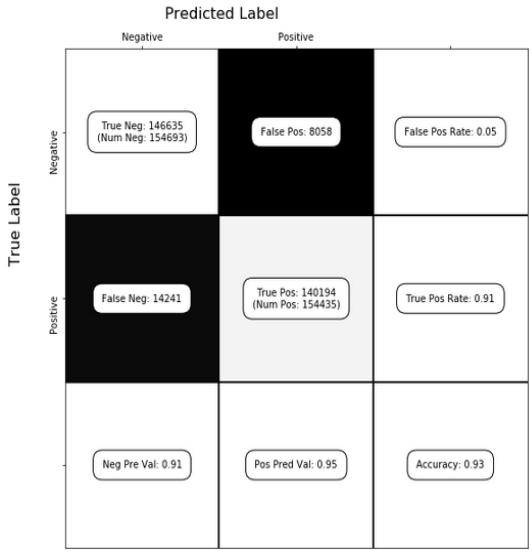
شکل (۵) منحنی ROC ونمرات AUC از ماشین بردار پشتیبان



شکل (۶) منحنی ROC و نمرات AUC از رگرسیون لجستیک

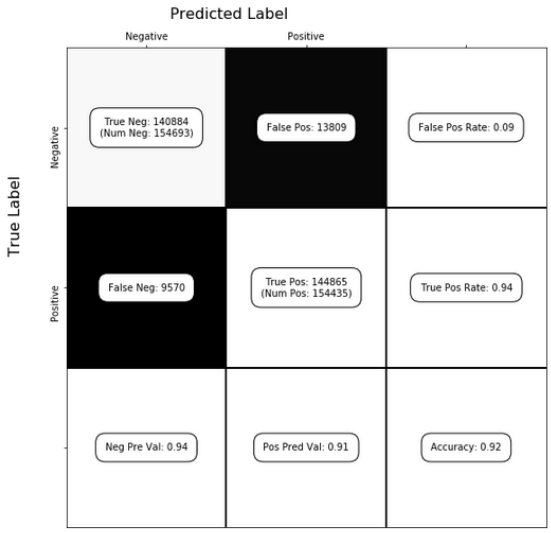
همانطور که در اشکال فوق نشان داده‌شده، سطح زیر نمودار الگوریتم ماشین بردار پشتیبان با وجود نزدیکی به نایوبیزین بیشتر است،بنابراین از عملکرد خوبی نسبت آن و نسبت به رگرسیون لیجستیک برخوردار است و در رده دوم قرار دارد.

شکل‌های (۷)، (۸)و (۹) صحت در کانفیوژن ماتریس را براساس برچسب‌های واقعی و پیش‌گویی شده نشان می-دهد.

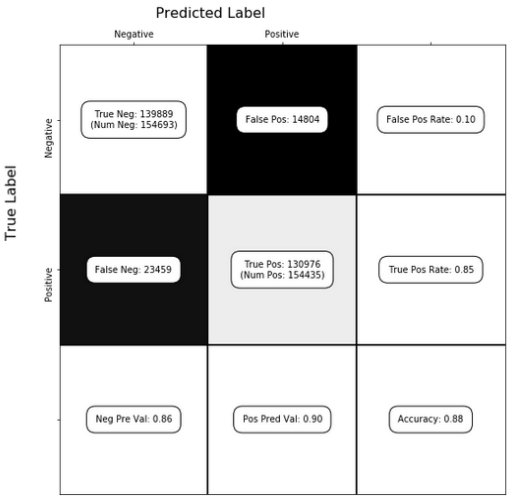


شکل (۷) ماتریس کانفیوژن با طبقه بند بیزین

[۵۳] Logistic Regression



شکل (۸) ماتریس کانفیوژن با طبقه بند ماشین بردار پشتیبان



شکل (۹) ماتریس کانفیوژن با طبقه بند رگرسیون

سپس همانطور که در جدول(۱۰) و (۱۱) نشان داده شده، به منظور آزمایش اینکه آیا الگوریتم ژنتیک انتخابی دارای پاسخ بهینه بوده یا خیر، نتایج استفاده از طبقه بندها را قبل و بعد از استفاده از این الگوریتم محاسبه می‌کنیم.

جدول (۱۰) معیارهای ارزیابی طبقه بندها قبل از استفاده از الگوریتم ژنیک پیشنهادی

طبقه بند	Accuracy	Precision	Recall	F-measure
بیزین	۹۰،۶۵	۹۱،۵۴	۹۲،۷۳	۹۲،۱۳
SVM	۸۹،۳۶	۹۰،۶۹	۸۸،۱۷	۸۹،۴۱
لجستیک رگریون	۸۱،۵۳	۸۲،۱۵	۸۳،۴۶	۸۲،۸۰

جدول (۱۱) معیارهای ارزیابی طبقه بندها بعد از استفاده از الگوریتم ژنتیک پیشنهادی

طبقه بند	Accuracy	Precision	Recall	F-measure
بیزین	۹۴،۶۵	۹۵،۱۷	۹۳،۸۹	۹۴،۵۳
SVM	۹۲،۷۹	۹۴،۵۹	۹۰،۷۷	۹۲،۶۴
لجستیک رگریون	۸۸،۲۹	۸۴،۴۴	۸۶،۴۶	۸۵،۴۴

۴-۳-۵ مقایسه واژگان مدل پیشنهادی با روش‌های دیگر

در این قسمت به تحلیل و بررسی و مقایسه معیارهای صحت، دقت و معیار F در مقالات مبتنی بر روش واژگان و روش پیشنهادی با سایر روش‌ها نظیر SentiWordNet [۱۵]، سامها [۱۶]، روش ترکیبی واژه نامه و corpus-based [۱۷] و یادگیری عمیق [۱۸] پرداخته‌ایم.

روش پیشنهادی خود را با نتایج چند مقاله که تنها با استفاده از روش محاسبه‌ی واژگان احساسی کارکرده‌اند مقایسه کردیم و نتایج نشان داد روش ترکیبی ما با با فرضیات مطرح شده و تعداد ویژگی های بیان شده، نسبت به روش این مقالات که بر اساس روش واژگان احساسی هستند، برتری قابل وضوحی دارد که در جدول (۱۲) به مقایسه کمی این روش‌ها پرداخته شده است.

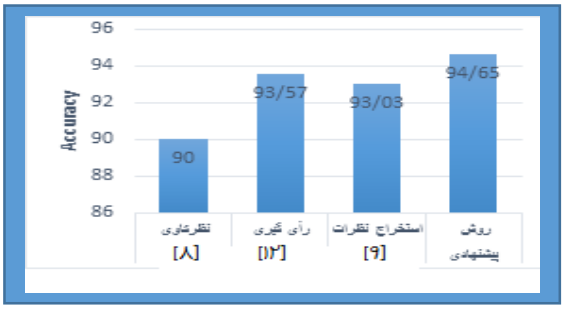
جدول (۱۲) مقایسه معیارهای ارزیابی واژگان با روش پیشنهادی

واژگان	Accuracy	Precision	F-measure
SentiWordNet [۲۱]	۷۹،۷۵	۹۰،۱	۷۸،۳
سامها [۲۲]	۵۶	۶۱	۶۰
ترکیب روش واژه‌نامه و corpus-based [۲۳]	۸۵	۷۳	۷۸

یادگیری عمیق [۲۴]	۸۳	۸۴,۷	۸۳,۶
روش پیشنهادی	۹۴,۶۵	۹۵,۱۷	۹۴,۵۳

۶-۳-۴ مقایسه تحلیل نظرات هتل هادرمدل پیشنهادی با روش‌های دیگر

در این بخش به مقایسه معیار صحت بدست آمده ازطبقه بندی مدل پیشنهادی با سایر روش‌ها روی مجموعه داده-های نظرات مربوط به هتل مانند، نظرکاوی[۸] روش رای گیریی [۱۸] و روش استخراج نظرات [۹] پرداخته شده است. در شکل (۱۰) این مقایسه نشان داده شده است.



شکل (۱۰)مقایسه تحلیل نظرات هتل روش پیشنهادی با سایر روش‌ها

۵- نتیجه گیری و پیشنهاد برای تحقیقات آتی

عقیده کاوی، تجزیه‌وتحلیل احساسات نیز نامیده می‌شود که فرآیندی برای کشف عقیده‌ی کاربران درباره موضوع یا محصول یا مسئله‌ی خاصی است. موضوع می‌تواند یک خبر، رویداد، محصول، فیلم، موقعیت هتل،خدمات ارایه شده در هتل و ... باشد. عقیده‌کاوی و تحلیل‌احساسات موضوعی تحقیقاتی در حوزه متن کاوی، پردازش زبان طبیعی و وب کاوی می‌باشد.

عقیده کاوی و تحلیل احساسات سیستماتیک زیرشاخه ای ازعلم داده‌کاوی است که برای کسب دانش نهفته ومخفی در مجموعه داده‌های متنی ساختار نیافته با حجم بالا به‌کار می‌رود. مجموعه داده‌های متنی می‌تواند نظرو عقیده مطرح‌شده یا کامنت‌های کاربران در شبکه‌های اجتماعی یا وبسایت های خاص مرتبط با موضوع مورد نظر باشد، بازخورد و تجربه‌ی مشتری در استفاده از هر محصول یا موضوع خاص یا هر تاپیک دیگری باشد.

این مقاله بر روی استخراج آرا و عقاید در قالب سند، جمله و تحلیل احساسات در سطح ویژگی بر روی مجموعه داده‌های نظرات کاربران هتل ارایه شد. در روش پیشنهادی، هدف، تعیین گرایش احساسی متن و طبقه‌بندی آن به دسته‌های مثبت و منفی است. در مدل ارایه شده در این مقاله علاوه بر تعیین گرایش کلی احساسی داده های متنی مورد آزمایش و تعیین قطبیت کلی مثبت یا منفی، ویژگی-های مهم بیان شده در آرا مهم کاربران هتل‌ها که بیشتر مدنظرشان بوده نیز استخراج شده است. آگاهی از این ویژگی‌ها، به هتلداران و یا سایر ذینفعان این امکان را می‌دهد که از مهم‌ترین شاخصه‌های مدنظر مشتریان آگاه شوند و از این دانش در راستای سیاست‌گذاری بهتر و در صورت لزوم تغییر رویکرد و در نهایت کسب سود بیشتر استفاده کنند.

در روش پیشنهادی علاوه بر طبقه‌بندی نظرات، مجموع ویژگی‌های بیان شده در نظرات کاربران هتل‌ها برای هر یک از هتل‌ها براساس رتبه‌های مثبت و منفی استخراج شد.

نتایج بدست آمده نشان دهنده‌ی دقت بالای مدل پیشنهادی این مقاله در انتخاب ویژگی‌ها می‌باشد. با تحلیل و بررسی نتایج مشخص شد نظرات متنی کاربرانی که ویژگی‌های شاخص‌تری برای هتل ابراز می‌کنند، تعداد صفات بیشتری برای وصف واژه‌ها و ویژگی‌های مد نظرشان استفاده می‌کنند، نتایج نشان می‌دهد که این روش بسیار بهتر از روش‌های مبتنی بر فرکانس واژه‌ها عمل می‌کند. علاوه بر این با مقایسه این روش با سایر روش‌ها دریافت شد که نتایج نشان دهنده بهبود نتیجه‌ی بدست آمده از طبقه بندی در این پژوهش نسبت به روش‌های پیشین است.

۵-۱ پیشنهادهای آینده

در کارهای آینده می‌توان از روش بیان شده در مجموعه داده‌های بزرگ‌تر و متنوع‌تر و نیز در مورد سایر خدمات و کالاهای مختلف و نیز روش مشابهی را برای داده‌های به زبان فارسی به کاربرد، مدل پیشنهاد شده تنها جنبه‌های صریح مطرح شده در متن نظر را در بررسی می‌کند می‌توان با یافتن جنبه‌های ضمنی مطرح شده مدل را بهبود داد تا تصویر جامعی از کلیه‌ی عقاید مطرح شده را دراختیار متصدیان امورقرار دهیم.

در این مقاله جنبه‌های مطرح شده در قالب پنج ویژگی بیان شد، می‌توان با افراد خبره مشورت کرد و جنبه‌های بیشتری را درنظر گرفت، همچنین میتوان مدل پیشنهادی این مقاله را با سایر مدل‌ها و الگوریتم ها ترکیب و نتایج بدست آمده را بررسی کرد.

در نهایت با طراحی شبکه واژگان احساسی جداگانه برای تحلیل هر دسته نظرات در موضوعات مختلف همانند رستوران، نقاط دیدنی، خدمات فرودگاهی و دیگر موارد، تاثیر آن را بر بهبـود نتایج بدست آمده بررسی نمود. با توجه به اینکه بار احساسی ومعنای ادراک شده از یک واژه در کالاها و خدمات مختلف با یکدیگر متفـاوت است، طراحی وساخت فرهنگ واژگان احساسی خاص منظوره برای زمینه های مختلف می‌تواند موجب بهبود نتایج حاصل شده از پژوهش‌های قبلی گردد.

مراجع

[۱] K. Ravi, V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications,"Knowledge-BasedSystems, vol.۸۹, pp.۱٤-٤٦, November ۲۰۱٥.

[۲] J. A. Balazs, J. D. Vel´asquez, "Opinion Mining and Information Fusion: A Survey," Information Fusion, vol.۲۷, pp۹٥-۱۱۰, ۲۰۱٦.

[۳] ر. پیرمحمدیانی و ش. محمدی " معیارهای ارزیابی ارزش اثرگذاری کاربران رسانه‌های اجتماعی چارچوبی براساس کاوش رسانه‌های اجتماعی،" در دوفصلنامه علمی فناوری اطلاعات و ارتباطات ایران، صفحات۱۰۹-۱۲۵، بهار و تابستان ۱۳۹۸.

[٤] Z. Zhang, Z. Zhang & Y. Yang, "The power of expert identity: How website recognized expert reviews influence travelers' online rating behavior," Tourism Management, vol.٥٥, pp۱٥–۲٤, August ۲۰۱٦.

[٥] G. Pablos, M. Cuadros, & M. T. Linaza, "Automatic analysis of textual hotel reviews," Information Technology & Tourism, vol.۱٦(۱), pp٤٥–٦۹, ۲۰۱٦.

[٦].M.R. Martinez-Torresa , S.L. Tora. (۲۰۱۹) A machine learning approach for the identification of the deceptive reviews in the hospitality sector using unique attributes and sentiment orientation. *Tourism Management*. .

[۷] C. Y. Tsai, M. T. Wang, & H. T. Tseng, " The impact of tour guides' physical attractiveness, sense of humor, and seniority on guide attention and efficiency." *Journal of Travel & Tourism Marketing*, vol.۳۳, pp۱–۱۳, ۲۰۱٥.

[۸]A. S. Mohammad & M. Al Kadri, "Using Lexicon-Based Opinion Mining to Gauge Customer Satisfaction," International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol.۹, February ۲۰۲۰

[۹]A. Ahania, M. Nilashib,O. Ibrahimc, L. Sanzognia & S. Weaven, " Market segmentation and travel choice prediction in Spa hotels through TripAdvisor's online reviews," International Journal of Hospitality Management, vol.۸۰, pp٥۲-۷۷, July ۲۰۱۹.

[۱۰] T. W. Lui, M. Bartosia, G. Piccoli, & V. Sadhya, "Online review response strategy and its effects on competitive performance," Tourism Management, vol.٦۷, pp۱۸۰-۱۹۰, August ۲۰۱۸.

[۱۲] M. E. Mowlaei, M. Saniee Abadeh, H. Keshavarz, " Aspect-Based Sentiment Analysis using Adaptive Aspect-Based Lexicons," Expert System with Applications, vol.۱٤۸, June ۲۰۲۰.

[۱۳] L. M. Domingo, J. C. Martín, G. Mandsberg. " Social media as a resource for sentiment analysis of Airport Service Quality (ASQ)," Journal of Air Transport Management, vol.۷۸, pp۱۰٦-۱۱٥, July ۲۰۱۹

[۱۴] D. Zhang, J. Tu, L. Zhou & Z. Yu, "Higher tourism specialization, better hotel industry efficiency?," *International Journal of Hospitality Management*, ۲۰۲۰.

[۱٥]T. Chinsha & J. Shibily, " A syntactic approach for aspect based opinion mining," Proceedings of the ۲۰۱٥ IEEE ۹th International Conference on Semantic Computing, pp۸۱-۸۸, March ۲۰۱٥ .

[۱۶]A. K. Samha, Y. Li & J. Zhang. "Aspect-Based Opinion Extraction from Customerreviews," Computation and Language, April ۲۰۱٤.

[۱۷] H. Keshavarz & M. Saniee Abadeh, "Accurate frequency-based lexicon generation for opinion mining," Journal of Intelligent and Fuzzy System, September ۲۰۱۷.

[۱۸] H. Keshavarz & M. Saniee Abadeh, "ALGA: Adaptive lexicon learning using genetic algorithm for sentiment analysys of microblogs," *Knowledge Based Systems*, vol.۱۲۲, pp۱-۱۶, April ۲۰۱۷

[۱۹] S. Katoch, S. S. Chauhan & V. Kumar, " A review on genetic algorithm: past, present, and future," *Multimedia Tools and Applications*, vol.۸۰, pp۸۰۹۱-۸۱۲۶, ۲۰۲۱.

[۱۲۰] م. امین طوسی و ه . عزتی "الگوریتم ژنتیک آگاه از بهترین عضو با کاربرد در رنگ‌آمیزی و بعدمتریک گراف،" در دوفصلنامه علمی فناوری اطلاعات و ارتباطات ایران، صفحات۱۴۳-۱۵۴، بهار و تابستان ۱۳۹۹.

[۲۱] T. Chinsha & J. Shibily, " A syntactic approach for aspect based opinion mining," *Proceedings of the ۲۰۱۵ IEEE ۹th International Conference on Semantic Computing*, pp۸۱-۸۸, March ۲۰۱۵.

[۲۲] A. K. Samha, Y. Li & J. Zhang. "Aspect-Based Opinion Extraction from Customerreviews," *Computation and Language*, April ۲۰۱۴.

[۲۳] M. Z. Asghar, A Khan, S. R. Zahra, S. Ahmad & F. M. Kundi, " Aspect-based opinion mining framework using heuristic patterns," *Cluster Computing, Springe*, vol.۲۲, pp۷۱۸۱-۷۱۹۹, August ۲۰۱۹

[۲۴] D. Tang, F. Wei, N. Yang, M. Zhou, T. Liu, & B. Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," *Learning Sentiment-SpecificWord Embedding for Twitter Sentiment Classification*, vol.۱, June ۲۰۱۴.

[۲۵] K. Sailunaz, R. Alhaj, " Emotion and Sentiment Analysis from Twitter Text," *Computational Science*, vol۳۶, September ۲۰۱۹.

[۲۶] Y. H. Hu, Y. L. Chen, & H. L. Chou, "Opinion mining from online hotel reviews – a text summarization approach," *Information Processing & Management*, vol.۵۳, pp۴۳۶–۴۴۹, March ۲۰۱۷.

[۲۷] J. Li, L. Xu, L. Tang, S. Wang, S., & L. Li, "Big data in tourism research: A literature review," *Tourism Management*, vol.۶۸, pp۳۰۱-۳۲۳, October ۲۰۱۸.

[۲۸] D. Zhang, J. Tu, L. Zhou & Z. Yu, "Higher tourism specialization, better hotel industry efficiency?," *International Journal of Hospitality Management*, ۲۰۲۰.

[۲۹] K. Cheng, J. Li, J. Tang, H. Liu, "Unsupervised Sentiment Analysis with Signed Social Networks," *Proceeding of the Thirty-First AAAI Conference on Artificial Intelligence*, vol.۳۱, pp۳۴۲۹-۳۴۳, ۲۰۱۷.