مدل‌سازی موضوعی و تحلیل احساسات مشتریان با استفاده از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشین

راضیه رحیمی[۱]، رضا قاسمی‌یقین[۲]
[۱]دانشجوی کارشناسی ارشد، دانشکده مهندسی نساجی، دانشگاه صنعتی امیرکبیر؛ razieh.rahimi@aut.ac.ir
[۲]استادیار، دانشکده مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر؛ yaghin@aut.ac.ir
نویسنده مسئول: رضا قاسمی‌یقین

چکیده

در این مقاله به بررسی و تجزیه‌وتحلیل نظرات مشتریان پوشاک زنان با استفاده از تکنیک‌های یادگیری ماشین و پردازش زبان طبیعی پرداخته می‌شود. مدل‌های یادگیری ماشین به کار گرفته‌شده شامل ماشین بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده چندجمله‌ای، بیز ساده مکمل، XGBoost و LightGBM هستند. به‌منظور برداری کردن و استخراج ویژگی متن نظرات از الگوریتم TF-IDF و Word2vec استفاده می‌شود. مدل‌سازی موضوعی با استفاده از روش تخصیص دیریکله پنهان و خوشه‌بندی K-Means انجام می‌شود. مجموعه داده مورداستفاده مربوط به نظرات پوشاک زنان و متغیر هدف امتیازهای مشتریان در نظرات است. مطالعه پیش‌رو در حالت دوکلاسه، سه‌کلاسه و پنج‌کلاسه انجام و در هر سه حالت بهترین عملکرد مربوط به مدل جنگل تصادفی با دقت ۰٫۹۸ در حالت دوکلاسه، ۰٫۹۵ در حالت سه‌کلاسه و ۰٫۹۱ در حالت پنج‌کلاسه برآورد می‌شود.

کلمات کلیدی: تجزیه‌وتحلیل احساسات، یادگیری ماشین، پردازش زبان طبیعی، متن‌کاوی، رفتار مشتری، مدل‌سازی موضوعی

Topic Modeling and Sentiment Analysis of Customers Using Natural Language Processing and Machine Learning Techniques

Razieh Rahimi[1], Reza Ghasemy Yaghin[2]
[1]Master's student, Department of Textile Engineering, Amirkabir University of Technology; razieh.rahimi@aut.ac.ir
[2]Assistant Professor, Department of Industrial Engineering & Management Systems, Amirkabir University of Technology; yaghin@aut.ac.ir
Corresponding author: Reza Ghasemy Yaghin

Abstract

Given that modeling and predicting customer behavior using data science helps companies gain a better understanding of customer behavior, this research focuses on analyzing customer reviews in the women's clothing domain within e-commerce. We employ machine learning techniques and natural language processing (NLP) to achieve this goal. The machine learning models used include Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, Multinomial Naive Bayes, Complement Naive Bayes, XGBoost, and LightGBM. To extract and vectorize text features from the reviews, we utilize the TF-IDF and Word2vec algorithms. We employ Topic Modeling using Latent Dirichlet Allocation (LDA) method and k-means clustering. The dataset consists of women's clothing reviews, with the target variable being customer ratings in those reviews. The study is conducted in binary, three-class, and five-class scenarios. The target variable, which originally has five classes (scores 1 to 5), is categorized into two-class and three-class modes. In the two-class mode, scores below 3 are class zero, while scores of 3 and above are class one. In the three-class mode, scores below 3 are class zero, scores equal to 3 are class one, and scores above 3 are class two. In all three cases, the Random Forest model performs best, achieving an accuracy of 0.98 in the binary case, 0.95 in the three-class case, and 0.91 in the five-class case. After performing the required preprocessing and feature engineering, principal component analysis (PCA) and T SNE are applied. After that, the scatter diagram of the data is drawn and the optimal number of clusters 3 is estimated using the elbow and silhouette diagrams. In the next step, by removing

provide a comprehensive understanding of the key themes and concerns customers have when considering womenswear items in each of the four topics.

Keywords: Sentiment analysis, Machine learning, Natural language processing, Text mining, Customer behavior

۱- مقدمه

کاربران اینترنت با نوشتن نظرات خود و رتبه‌بندی محصولات، از گیرنده‌های اطلاعات سنتی به ناشران اطلاعات تغییر پیدا کرده‌اند. این عمل منجر به ذخیره حجم قابل‌توجهی از اطلاعات ارزشمند توسط وب‌سایت‌های تجارت الکترونیک می‌شود که تجزیه‌وتحلیل آن‌ها به‌منظور به‌کارگیری تکنیک‌های بازاریابی مناسب، از توانایی‌های پردازش شناختی انسان‌ها پیشی می‌گیرد [۱, ۲]. رتبه‌بندی و نظرات نقش مهمی در درک احساسات مشتری دارند. تحلیل احساسات در تجارت الکترونیک بسیار مهم است و به کسب‌وکارها کمک می‌کند تا بازخورد مشتری را تجزیه‌وتحلیل و روندها را شناسایی کنند. تحلیل احساسات، استراتژی مورداستفاده برای استخراج و ارزیابی احساسات بیان‌شده در داده‌های متنی است [۳] و هدف آن استخراج مزایا و معایب محصول از نظرات مشتریان، پرداختن به تجزیه‌وتحلیل نظرات مثبت و منفی و استخراج اطلاعات برجسته به‌منظور بهبود محصول، انتخاب برای تولید و افزایش نرخ خرید است [۴]. برای تجزیه‌وتحلیل این اطلاعات و درک احساسات مشتری، می‌توان از روش‌های مبتی بر یادگیری ماشین و پردازش زبان طبیعی استفاده کرد.

به همین منظور در این پژوهش رویکردی مبتنی بر روش‌های پردازش زبان طبیعی و مدل‌های یادگیری ماشین بر اساس نظرات و رتبه‌بندی‌های مشتریان ارائه خواهیم کرد که به کمک آن می‌توان با آگاهی از احساسات مشتریان خود خدمات و استراتژی‌های بازاریابی مناسب را به کار گرفت. در این مقاله با استفاده از مجموعه داده‌ی نظرات پوشاک زنان در خریدهای الکترونیکی به بررسی رضایت یا عدم رضایت از خرید مشتریان، شناسایی عوامل موثر بر رفتار آنها از طریق شناسایی احساسات متن نظرات، مدل‌سازی موضوعی بر اساس دسته‌بندی محصولات به منظور شناسایی موضوعات کلیدی به همراه خوشه بندی مشتریان و در نهایت ارائه مدل‌های کلاس‌بندی احساسات با استفاده از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشین پرداخته می‌شود. درک احساسات نهفته در نظرات مشتریان

punctuation marks, stop words and words with less than three letters, converting the first letter of the words to lowercase and lemmatization, data cleaning was done. After that, topic modeling is done and each of the topics and words related to them are examined. In the next step, the topics are examined in different clusters. These analyzes

اهمیت فراوانی دارد و اطلاعات ارزشمندی در مورد رضایت مشتریان نسبت به محصولات یا خدمات شرکت ارائه می‌دهد. احساسات مثبت نشان‌دهنده‌ی خوشحالی و رضایت مشتریان و احساسات منفی نشان‌دهنده‌ی زمینه‌های نیازمند بهبود هستند. با شناسایی این احساسات، کسب و کارها می‌توانند استراتژی‌های خود را برای رسیدگی به نیازها و نگرانی‌های خاص مشتریان خود تنظیم کنند. به همین منظور هدف کلی مقاله پیش‌رو ارائه تکنیک‌ها و مدل‌هایی است که با ترکیب پردازش زبان طبیعی و یادگیری ماشین، اطلاعات و داده‌های شرکت‌های پوشاک را به بینش‌های عمیق و عملی برای رشد و موفقیت آن‌ها تبدیل و از قدرت بازخوردهای مشتریان برای بهبود محصولات، خدمات و تجربه کلی مشتری از خرید خود استفاده می‌کنند.

ساختار ادامه مقاله به این صورت است که در بخش ۲ به مرور ادبیات موضوع و بیان کارهای مرتبط پرداخته خواهد شد. در بخش ۳ روش‌شناسی و مبانی نظری بیان و در بخش ۴ به ارزیابی مدل‌های پیاده‌سازی‌شده و ارائه نتایج به‌دست‌آمده پرداخته خواهد شد. در بخش ۵ نتیجه‌گیری و پیشنهادهای آینده مطرح خواهد شد.

۲- مرور ادبیات موضوع

خلاصه کارهای مرتبط در جدول ۱ نشان داده‌شده است. همان‌طور که مشاهده می‌شود، مطالعات گذشته به منظور بررسی نظرات مشتریان پوشاک در خریدهای الکترونیک با مجموعه داده‌های متفاوت، متغیر هدف و مدل‌های یادگیری ماشین متفاوت انجام شده‌اند. از آنجایی که مجموعه داده مورد استفاده در این مقاله با عنوان نظرات تجارت الکترونیک پوشاک زنانه با مقالات [۵]، [۶]، [۷]، [۸]، [۳]، [۹]، [۱۰]، [۱۱]، [۱۲]، [۱۳] و [۱۴] یکسان است، در ادامه به بررسی بیشتر آن‌ها پرداخته می‌شود. این مجموعه داده در تحقیقات متعددی مورد استفاده قرار گرفته است که بیشتر بر تحلیل احساسات مشتریان و پیش‌بینی رفتارهای آن‌ها متمرکز هستند. برخی از این تحقیقات شامل تحلیل اکتشافی داده‌ها و پیش‌بینی احساسات با استفاده از مدل‌های یادگیری ماشین مانند

ماشین بردار پشتیبان و درخت تصمیم، همچنین مقایسه مدل‌های مختلف یادگیری ماشین برای تحلیل نظرات مشتریان است. در مقایسه با این تحقیقات، تحقیق حاضر با استفاده از مدل‌های مختلف یادگیری ماشین تلاش کرده است تا دقت مدل‌ها را در تحلیل و طبقه‌بندی نظرات مشتریان بهبود بخشد. علاوه بر این، استفاده از تکنیک‌های پردازش زبان طبیعی مانند TF-IDF و Word2Vec در استخراج ویژگی‌ها و مدل‌سازی موضوعی با روش تخصیص دیریکله پنهان، از ویژگی‌های خاص این تحقیق است که به دقت بیشتر در پیش‌بینی احساسات و دسته‌بندی نظرات کمک کرده است. بنابراین، این تحقیق با استفاده از ترکیب روش‌های مختلف و ارزیابی دقیق‌تر مدل‌ها، به نتایج متفاوتی از سایر تحقیقات مشابه دست یافته است. در مطالعه‌های لین [۵]، کوبروسلی و همکاران [۶]، مونیاسامی و همکاران [۷]، سونیل و شیرازی [۸]، کاسیمو و همکاران [۳]، لوکیلی [۹] و شتی و همکاران [۱۱] متغیر هدف در نظر گرفته به منظور پیاده‌سازی مدل‌های یادگیری ماشین ستون ویژگی شاخص پیشنهاد محصول در مجموعه داده و در حالت دو کلاسه (صفر برای پیشنهاد نکردن محصول توسط مشتری و یک برای پیشنهاد محصول توسط مشتری) انجام شده‌اند. در مطالعه شتی و همکاران [۱۲] و مطالعه یابس و همکاران [۱۳] متغیر هدف در نظر گرفته شده ستون ویژگی امتیاز محصولات و در حالت دو کلاسه با مرز امتیاز سه (امتیازهای برابر سه و بالاتر به عنوان کلاس ۱ و امتیازهای کمتر از سه به عنوان کلاس صفر) صورت گرفته‌اند. موضوع قابل‌توجه این است که مجموعه داده مورداستفاده در همه‌ی مطالعات یکسان است و همان‌طور که در ادامه با تجزیه‌وتحلیل داده‌های اکتشافی نشان خواهیم داد این مجموعه داده برای متغیرهای هدف شاخص پیشنهاد محصول و امتیاز مشتریان متعادل نیست. ازآنجایی‌که بسیاری از مدل‌های یادگیری ماشین، مانند رگرسیون لجستیک، هر زمان که اختلاف نسبت بین تعداد نقاط داده نامتعادل باشد، با مشکلات عملکردی مواجه می‌شوند [۸] و به‌منظور جلوگیری از غیرقابل‌اعتماد بودن نتایج و معیارهای ارزیابی مدل‌های پیاده‌سازی‌شده موضوع متعادل کردن مجموعه داده مهم است و باید موردتوجه قرار گیرد.

در میان کارهای مرتبط انجام‌شده چهار مطالعه بحث متعادل کردن مجموعه داده را مطرح کرده و موردبررسی قرار داده‌اند. کوبروسلی و همکاران [۶] به‌منظور انجام این کار به صورت تصادفی مجموعه‌های متعادل آموزش و آزمایش را انتخاب کرده‌اند. در این مطالعه توضیح دقیق‌تری از روش انجام کار ارائه نشده است. سونیل و شیرازی [۸] به‌منظور متعادل کردن مجموعه داده از روش افزایش تعداد داده‌ها در کلاس اقلیت استفاده کرده‌اند اما به تکنیک انجام این کار به صورت صریح اشاره‌ای نشده است. لوکیلی [۹] در بخش تجزیه‌وتحلیل اکتشافی به موضوع متعادل نبودن مجموعه داده با متغیر هدف شاخص پیشنهاد محصول پی برده است. با این حال در ادامه مقاله توضیحی مبنی بر استفاده از روش‌هایی برای رفع این نامتعادل بودن در نظر گرفته نشده است. شتی و همکاران [۱۱] به نامتعادل بودن مجموعه داده اشاره کرده‌اند اما راهکاری به‌منظور برطرف کردن آن ارائه نشده است. بااین‌حال آن‌ها بیان کرده‌اند که به دلیل نامتعادل بودن مجموعه داده معیار صحت قابل‌اعتماد نیست و بهتر است از معیارهای دقت و پوشش برای ارزیابی عملکرد مجموعه داده‌های نامتعادل استفاده شود.

در مطالعه پیش‌رو با در نظر گرفتن امتیاز محصولات به‌عنوان متغیر هدف و به‌کارگیری روش SMOTE برای رفع مشکل نامتعادل بودن مجموعه داده و به کمک تکنیک‌های یادگیری ماشین و پردازش زبان طبیعی به تجزیه‌وتحلیل احساسات در حالت دوکلاسه، سه‌کلاسه و پنج‌کلاسه پرداخته خواهد شد. روش SMOTE با استفاده از الگوریتم k-نزدیکترین همسایه به تولید داده‌های مصنوعی جدید می‌پردازد [۱۵] و در مقایسه با تکنیک‌های معمول افزایش تعداد داده‌ها در کلاس اقلیت مانند OverSampling، که ممکن است داده‌های تکراری زیادی به مجموعه اضافه و نتایج را غیرقابل اعتماد کنند انتخاب بهتری است. همچنین انجام مقاله در حالت‌های دو کلاسه، سه‌کلاسه و پنج‌کلاسه به بررسی دقیق‌تر و با جزئیات بیشتر رفتار و احساسات مشتریان کمک می‌کند. علاوه بر موضوعاتی که اشاره شد، در این مقاله مدل‌سازی موضوعی و برآورد موضوعات نهفته در متن نظرات با استفاده از روش تخصیص دیریکله پنهان نیز انجام می‌شود. نتایج حاصل از همه‌ی این موارد اطلاعات مهمی در رابطه با شناخت عوامل مؤثر بر رفتار مشتریان و تحلیل احساسات آن‌ها در اختیار ما قرار می‌دهد.

جدول ۱ خلاصه کارهای مرتبط

عنوان مقاله و سال انتشار	مجموعه داده	متغیر هدف	متعادل کردن مجموعه داده	مدل‌های یادگیری ماشین
تحلیل احساسات نظرات مشتریان تجارت الکترونیک بر اساس پردازش زبان طبیعی (۲۰۲۰) [۵]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	-	رگرسیون لجستیک، ماشین بردار پشتیبان، جنگل تصادفی، Xgboost وLightGBM
تحلیل آماری نظرات متنی تجارت الکترونیک با استفاده از روش‌های مبتنی بر درخت (۲۰۲۲) [۶]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	انتخاب مجموعه‌های متعادل آموزش (۵۸۴۰) و آزمایش (۲۵۰۴) به صورت تصادفی	جنگل تصادفی، XGBoost، درخت تصمیم و تقویت گرادیان
تجزیه‌وتحلیل نظرات آنلاین مشتریان با استفاده از تکنیک‌های یادگیری ماشین (۲۰۲۲) [۷]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	-	بیز ساده، رگرسیون لجستیک، ماشین بردار پشتیبان و شبکه عصبی
کلاس‌بندی نظرات مشتری با استفاده از تکنیک‌های یادگیری ماشین و یادگیری عمیق (۲۰۲۳) [۸]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	افزایش تعداد داده‌ها در کلاس اقلیت	رگرسیون لجستیک، ماشین بردار پشتیبان خطی، درخت تصمیم، جنگل تصادفی و AdaBoost
تحلیل احساسات قابل توضیح برای بازاریابی شخصی نساجی (۲۰۲۳) [۳]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	-	رگرسیون لجستیک
تجزیه‌وتحلیل احساسات نظرات محصول برای توصیه تجارت الکترونیک بر اساس یادگیری ماشین (۲۰۲۳) [۹]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	-	رگرسیون لجستیک، جنگل تصادفی، KNN و CatBoost
کاوش اصطلاحات رایج نظرات آنلاین در زمینه مد برای پیش بینی مفید بودن نظر (۲۰۲۳) [۱۶]	نظرات مشتریان پلتفرم مد آمازون	شاخص مفید بودن نظر	-	جنگل تصادفی، ماشین بردار پشتیبان، رگرسیون لجستیک، تقویت گرادیان تصادفی، مدل‌سازی موضوعی و تخصیص دیریکله پنهان
واکاوی نظرات آنلاین برای درک نیازهای مشتریان لباس تطبیقی (۲۰۲۳) [۱۷]	نظرات مشتریان آمازون، Silverts IZ-Adaptive, ResellerRating	-	-	تخصیص دیریکله پنهان (LDA) و مدل‌سازی موضوعی
تحلیل احساسات نظرات لباس زنانه در بسترهای تجارت الکترونیک: رویکرد یادگیری ماشینی (۲۰۲۳) [۱۸]	نظرات تجارت الکترونیک بنگلادش	احساسات	-	جنگل تصادفی، ماشین بردار پشتیبان، رگرسیون لجستیک، بیز ساده، تقویت گرادیان، حافظه کوتاه‌مدت طولانی، مدل‌سازی موضوعی و تخصیص دیریکله پنهان (LDA)
استفاده از مدل‌سازی موضوع برای استخراج انتظارات مشتریان: موردی از پوشاک زنان (۲۰۲۳) [۱۰]	نظرات تجارت الکترونیک پوشاک زنانه	-	-	مدل‌سازی موضوعی، تخصیص دیریکله پنهان (LDA)
بهینه‌سازی هایپرپارامتر مدل‌های یادگیری ماشین با استفاده از جستجوی شبکه‌ای برای تحلیل احساسات نظرات آمازون (۲۰۲۴) [۱۱]	نظرات تجارت الکترونیک پوشاک زنانه	شاخص پیشنهاد محصول	-	درخت تصمیم، رگرسیون لجستیک، ماشین بردار پشتیبان، بیز ساده، XGBoost و جنگل تصادفی
کاوش احساسات بازخورد داده‌های تجارت الکترونیک با استفاده از الگوریتم‌های یادگیری ماشین (۲۰۲۴) [۱۲]	نظرات تجارت الکترونیک پوشاک زنانه	امتیازهای مشتریان (در حالت دوکلاسه با مرز امتیاز ۳)	-	رگرسیون لجستیک، بیز ساده، بیز ساده چندجمله‌ای، بیز ساده برنولی، ماشین بردار پشتیبان، جنگل تصادفی و AdaBoosting

عوامل مؤثر بر توصیه‌ها برای رضایت از لباس زنان: رویکرد تخصیص دیریکله پنهان با استفاده از نظرات آنلاین (۲۰۲۴) [۱۴]	نظرات تجارت الکترونیک پوشاک زنانه	-	-	مدل‌سازی موضوعی، تخصیص دیریکله پنهان (LDA) و پردازش زبان طبیعی
این مطالعه: تحلیل احساسات نظرات تجارت الکترونیک با استفاده از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشین	نظرات تجارت الکترونیک پوشاک زنانه	امتیازهای مشتریان در حالت دوکلاسه، سه‌کلاسه و پنج-کلاسه	SMOTE	رگرسیون لجستیک، ماشین‌بردار-پشتیبان، درخت تصمیم، جنگل تصادفی، بیزساده مکمل، بیزساده چندجمله‌ای، XGBoost و LightGBM، مدل‌سازی موضوعی، تخصیص دیریکله پنهان (LDA) و پردازش زبان طبیعی

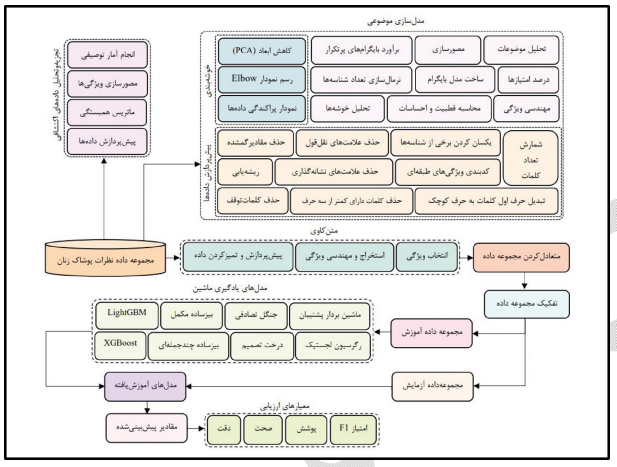
۳- روش‌شناسی

در شکل ۱ مراحل مطالعه انجام‌شده مشاهده می‌شود. در این مطالعه پس از انتخاب مجموعه داده مربوط به نظرات پوشاک زنان، به انجام تمیز کردن داده‌ها و پیش‌پردازش‌هایی چون مدیریت مقادیر گمشده، حذف ستون مربوط به شماره ردیف داده‌ها پرداخته می‌شود.

در مرحله‌ی بعد ابتدا به‌منظور شناخت و درک مجموعه داده به تجزیه‌وتحلیل اکتشافی انجام می‌شود که در ادامه موردبحث قرار خواهد گرفت. در بخش بعد، پس از انجام پیش‌پردازش به انتخاب ویژگی[۱]، استخراج[۲] و مهندسی ویژگی[۳] توسط الگوریتم TF-IDF، متعادل کردن مجموعه داده با استفاده از روش SMOTE و تفکیک آن به دو دسته‌ی مجموعه داده آموزش و آزمایش صورت می‌گیرد. درنهایت مدل‌های یادگیری ماشین موردنظر پیاده‌سازی و عملکرد آن‌ها مورد ارزیابی قرار می‌گیرد.

۳-۱- جمع‌آوری داده

مجموعه داده مورداستفاده در این مطالعه شامل داده‌های تجاری واقعی و با عنوان نظرات تجارت الکترونیک پوشاک زنانه[۴]، از وب-سایت Kaggle دانلود شده است [۱۹]. مجموعه داده شامل ۲۳۴۸۶ ردیف و ۱۰ ستون ویژگی است و هر ردیف آن بیان‌کننده شناسه قطعه در حال بررسی، سن مشتری صاحب نظر، عنوان نظر، متن نظر، امتیاز مشتری به محصول موردنظر از ۱ بدترین تا ۵ بهترین، شاخص پیشنهاد محصول توسط مشتری، تعداد مشتریانی که این نظر را مثبت و کمک‌کننده دانسته‌اند، نام بخش محصول سطح بالا[۵]، نام بخش محصول و نام کلاس محصول است. ستون-های ویژگی این مجموعه داده شامل متغیرهای عددی[۶]، رشته‌ای[۷]، عددی ترتیبی[۸]، عددی دودویی[۹] و طبقه‌ای[۱۰] می‌شود.



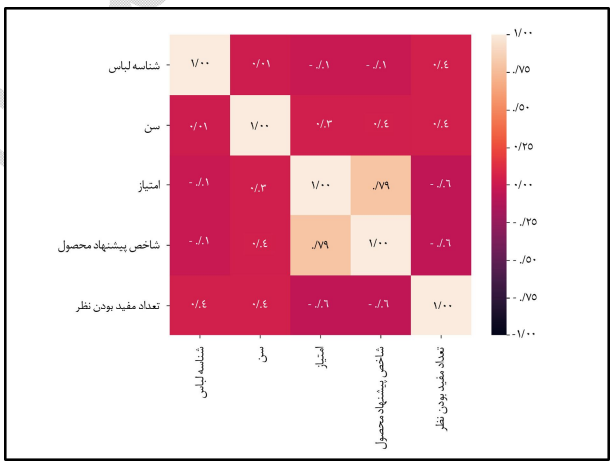
شکل ۱ روش‌شناسی مطالعه انجام‌شده

۳-۲ تجزیه‌وتحلیل داده‌های اکتشافی

به‌منظور انجام تجزیه‌وتحلیل داده‌های اکتشافی[۱۱] (EDA) پس از بررسی و حذف مقادیر داده گمشده[۱۲]، ابتدا محاسبه پارامترهای آمار توصیفی برای ویژگی‌های عددی انجام و سپس ماتریس همبستگی این ویژگی‌ها ترسیم می‌شود. در ادامه به تحلیل ویژگی‌ها پرداخته می‌شود. پس از حذف مقادیر داده گمشده تعداد ردیف‌های موردبررسی مجموعه داده برابر با ۱۹۶۶۲ در نظر گرفته

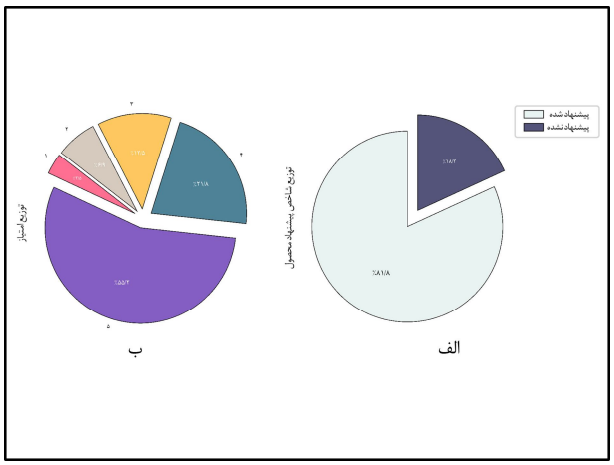
[1] Feature Selection
[2] Feature Extraction
[3] Feature Engineering
[4] Women's clothing e-commerce reviews
[5] Product high level division
[6] Numerical
[7] String
[8] Ordinal variable
[9] Binary variable
[10] Categorical
[11] Exploratory Data Analysis
[12] Missing values

می‌شود. میانگین گروه سنی افرادی که نظرات خود را در رابطه با پوشاک زنان خریداری‌شده به اشتراک گذاشته‌اند حدود ۴۳ سال و در بازه‌ی ۱۸ تا ۹۹ سال برآورد می‌شود. میانگین امتیازها برابر با ۴/۱ و میانگین شاخص پیشنهاد محصول ۰/۸۱ محاسبه شده و نشان‌دهنده‌ی آن است که در حالت کلی مشتریان از خرید خود رضایت دارند. تعداد مشتریانی که نظرات ثبت‌شده را مفید و کمک‌کننده دانسته‌اند نیز از ۰ تا ۱۲۲ عدد متغیر است. در شکل ۲ ماتریس همبستگی ویژگی‌های عددی نشان داده‌شده است. همان‌طور که مشاهده می‌شود، بیشترین مقدار ضریب همبستگی مربوط به ویژگی‌های امتیاز و شاخص پیشنهاد محصول و برابر با ۰/۷۹ برآورد می‌شود. شکل ۳-الف نحوه توزیع شاخص پیشنهاد محصول توسط مشتریان، شکل ۳-ب نحوه توزیع امتیازهای موجود در مجموعه داده و شکل ۴ امتیازهای مشتریان با توجه به گروه سنی و نام بخش را نشان می‌دهند.

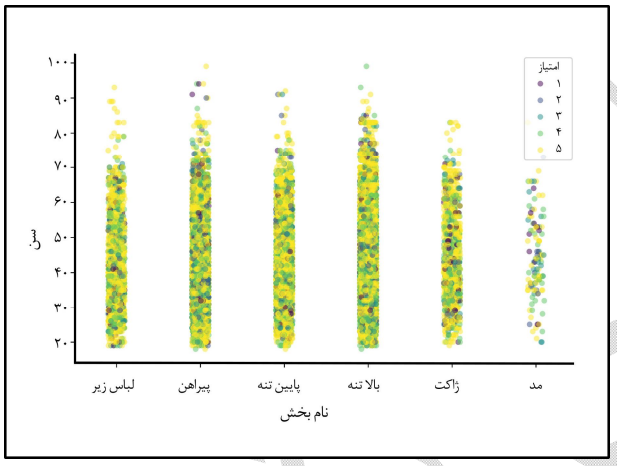


شکل ۲ ماتریس همبستگی ویژگی‌های عددی

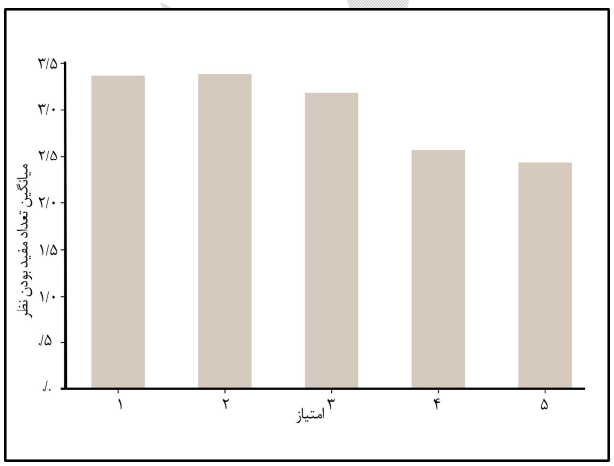
همان‌طور که مشاهده می‌شود، ۸۲/۲ درصد از مشتریان محصول خریداری‌شده را به دیگران پیشنهاد می‌کنند. امتیاز ۵ با مقدار ۵۵/۲ درصد دارای بیشترین سهم در تعداد کل مجموعه داده است و پس از آن به ترتیب امتیازهای ۴، ۳، ۲ و ۱ قرار می‌گیرند. بیشتر امتیازها توسط افراد در محدوده سنی ۲۰ تا ۷۵ سال ثبت شده‌اند و بیشترین نظرات دارای امتیاز ۳ و بالاتر هستند. شکل ۵ میانگین تعداد مفید بودن نظرات برای هر امتیاز را نشان می‌دهد. همان‌طور که مشاهده می‌شود بیشترین مقدار میانگین تعداد مفید بودن نظرات مربوط به امتیازهای ۱ تا ۳ برآورد می‌شود.



شکل ۳ الف) توزیع شاخص پیشنهاد محصول ب) نحوه توزیع امتیازها



شکل ۴ امتیازهای مشتریان با توجه به گروه سنی و نام بخش



شکل ۵ میانگین تعداد مفید بودن نظرات برای هر امتیاز

۳-۳ مدل‌سازی موضوعی

در این بخش ابتدا به بیان توضیح مختصری از روش تخصیص پنهان دیریکله[13] در مدل‌سازی موضوعی[14]، خوشه‌بندی[15] و چرایی استفاده از این دو روش در این مقاله پرداخته خواهد شد. در ادامه روش‌شناسی این بخش از مطالعه بیان و نتایج برآوردشده گزارش خواهد شد.

- تخصیص پنهان دیریکله: شکلی از یادگیری ماشین بدون‌نظارت است که برای کشف موضوعات پنهان از متن نظرات استفاده می‌شود. هدف این مدل یافتن گروهی از موضوعات است که کلمات مشاهده شده را در تمام اسناد به بهترین شکل توصیف کند. اطلاعات تولیدشده از این مدل شامل کلمات کلیدی مرتبط با هر یک از موضوعات و احتمال مرتبط شدن هر یک از نظرات متنی با هر موضوع است [۲۰].
- خوشه‌بندی K-means: الگوریتم خوشه‌بندی K-means از الگوریتم‌های یادگیری بدون‌نظارت است که برای طبقه‌بندی مجموعه‌ای از داده‌ها به کلاس‌هایی از داده های مشابه استفاده می‌شود. طبقه‌بندی مجموعه داده‌ی N مورد بر اساس ویژگی‌ها به k زیرمجموعه‌های مجزا با به حداقل رساندن فاصله بین آیتم داده و مرکز خوشه مربوطه انجام می‌شود [۲۱].

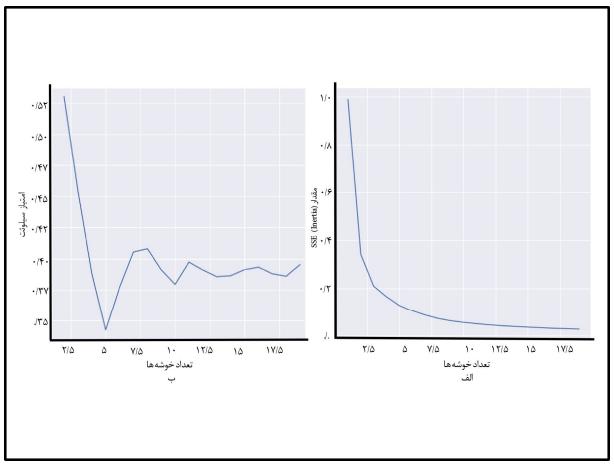
در این مقاله، برای انجام مدل‌سازی موضوعی از روش تخصیص دیریکله پنهان و خوشه‌بندی K-Means به‌طور ترکیبی استفاده شده است. تخصیص دیریکله پنهان به‌عنوان یک روش مبتنی بر احتمال، توانایی شبیه‌سازی توزیع موضوعات پنهان در اسناد مختلف را داراست و می‌تواند به‌طور مؤثری موضوعات غالب در متن‌های مشتریان را شناسایی کند. این روش به تحلیل ساختار معنایی داده‌ها کمک کرده و امکان استخراج ویژگی‌های دقیق موضوعی از متون نظرات مشتریان را فراهم می‌آورد. به‌علاوه، به دلیل سادگی در پیاده‌سازی، قابلیت تفسیر بالا و کارایی آن در تحلیل داده‌های متنی با حجم بالا، به‌عنوان یکی از روش‌های محبوب در پردازش زبان طبیعی شناخته می‌شود. به‌ویژه برای مسائلی که نیاز به شناسایی و استخراج موضوعات پنهان از مجموعه‌های بزرگ داده دارند، انتخابی مناسب و کارآمد است [۲۲]. در کنار آن، الگوریتم خوشه‌بندی K-Means برای تقسیم‌بندی اسناد به خوشه‌های مختلف، بر اساس توزیع‌های موضوعی به‌دست آمده از تخصیص پنهان دیریکله، به‌کار گرفته شده است. این ترکیب، داده‌ها را بر اساس شباهت‌های موضوعی به‌طور مؤثری گروه‌بندی کرده و تحلیل‌های دقیق‌تری را از نظرات مشتریان امکان‌پذیر می‌سازد. علاوه بر این، در این مطالعه، ارتباط موضوعات غالب در هر خوشه با سن مشتریان و تعداد کلمات در متن نظرات مشتریان بررسی شده است. این تحلیل‌ها نشان می‌دهند که چگونه سن و طول متن نظرات می‌تواند بر تمایل به صحبت در مورد موضوعات خاص تأثیر بگذارد و به درک بهتر الگوهای رفتاری و اولویت‌های مشتریان کمک می‌کند.

بررسی هر خوشه به‌طور ویژه مزایای زیادی از منظر استراتژی‌های بازاریابی و توسعه کسب و کار در بر دارد. با شناسایی ویژگی‌های خاص هر خوشه، کسب و کارها می‌توانند استراتژی‌های بازاریابی متناسب با نیازها و خواسته‌های خاص مشتریان در هر گروه را طراحی کنند. برای مثال، یک خوشه ممکن است تمایل زیادی به ویژگی‌های خاص محصول نشان دهد که نیاز به تبلیغات هدفمند و تشویق به خرید دارد، در حالی که خوشه‌ای دیگر ممکن است بیشتر به خدمات پس از فروش یا تجربه مشتری علاقه‌مند باشد. به‌علاوه، تحلیل موضوعات موجود در هر خوشه می‌تواند به کسب و کارها کمک کند تا پیشنهادات محصولات و خدمات را بر اساس ترجیحات خاص مشتریان هر خوشه سفارشی کرده و تجربه مشتری را بهبود بخشند. در نتیجه، این روش نه تنها به تحلیل دقیق‌تر رفتار مشتریان کمک می‌کند، بلکه می‌تواند راهکارهای مؤثری برای رشد کسب و کار و جذب مشتریان بیشتر ارائه دهد.

پس از انجام پیش‌پردازش‌هایی مانند مدیریت مقادیر گمشده، کدبندی ویژگی‌های طبقه‌ای، شمارش تعداد کلمات و تعداد کلمات یکتا، شناسه‌های لباس دارای فراوانی کمتر از یک درصد تعداد کل شناسایی و شناسه یکسانی به آن‌ها نسبت داده شد. سپس تعداد تکرار هر یک از شناسه‌های لباس نرمال شده و در ستونی با نام شناسه لباس جدید به مجموعه داده اضافه شد. در مرحله بعد میزان قطبیت و احساسات متن نظرات به همراه عنوان آن‌ها برآورد و در ستونی با عنوان امتیاز احساسات کل به مجموعه داده افزوده شد. به‌منظور بهبود عملکرد الگوریتم خوشه‌بندی K-Means، کاهش ابعاد و رسم مجموعه داده، داده‌ها با استفاده از تجزیه‌وتحلیل مؤلفه-

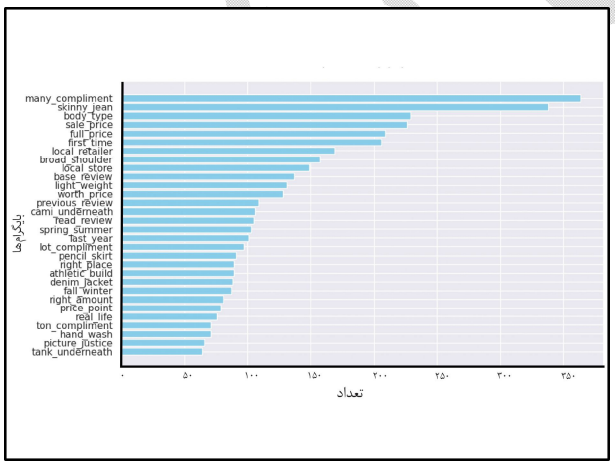
[13] Latent Dirichlet Allocation (LDA)
[14] Topic Modeling
[15] Clustering

های اصلی[16] و سپس تحلیل T-SNE[17] به ابعاد پایین‌تر کاهش یافتند و به عنوان ورودی الگوریتم K-Means در نظر گرفته شدند. T-SNE یک روش کاهش ابعاد غیرخطی است که برای نمایش ساختارهای پیچیده داده‌ها در فضای دو یا سه‌بعدی بسیار مؤثر است و امکان خوشه‌بندی مؤثرتر داده‌های پیچیده با حفظ روابط غیرخطی میان نقاط داده را فراهم می‌کند [۲۳]. پس از آن با استفاده از نمودار Elbow و معیار سیلوئت تعداد بهینه خوشه‌ها برآورد شد. روش elbow برای تعیین تعداد بهینه خوشه‌ها در الگوریتم K-Means بر اساس معیار مجموع مربعات خطا[18](SSE) عمل می‌کند. این معیار مجموع فاصله‌های مربعات داده‌ها تا نزدیک‌ترین مرکز خوشه را اندازه‌گیری و به‌عنوان شاخصی برای پراکندگی داده‌ها درون خوشه‌ها عمل می‌کند. برای این منظور، SSE برای تعداد خوشه‌های مختلف محاسبه و مقادیر به‌دست‌آمده در قالب یک نمودار رسم و نقطه‌ای که در آن کاهش SSE روند کندتری پیدا می‌کند به‌عنوان مقدار بهینه تعداد خوشه‌ها انتخاب شد. کاهش SSE به این معناست که داده‌ها به خوشه‌های مجزا و منسجم‌تری تقسیم می‌شوند [۲۴]. میانگین سیلوئت نیز به‌منظور ارزیابی کیفیت خوشه‌بندی داده‌ها استفاده می‌شود و نشان‌دهنده انسجام درون‌خوشه‌ای و جداشدگی بین خوشه‌ها است. برای محاسبه این مقدار، فاصله هر نقطه از نقاط دیگر در خوشه خود و نزدیک‌ترین خوشه محاسبه می‌شود [۲۵]. در این مطالعه، مقدار سیلوئت برای تعداد دو خوشه برابر با ۰٫۵۳ و برای تعداد سه خوشه برابر با ۰٫۴۵ به‌دست آمد. این تفاوت نشان‌دهنده این است که هر دو تعداد خوشه کیفیت مشابهی در خوشه‌بندی داده‌ها دارند. از طرفی، با توجه به نمودار elbow تعداد سه خوشه به‌عنوان تعداد بهینه خوشه‌ها با کمترین مجموع مربعات خطا و تقسیم‌بندی بهتر داده‌ها شناسایی شد. با توجه به این که نتایج روش elbow و امتیاز سیلوئت به‌طور کلی هم‌راستا هستند و تعداد سه خوشه هم‌زمان کیفیت بالاتری در خوشه‌بندی و انسجام بهتر داده‌ها ارائه می‌دهد، تعداد سه خوشه به‌عنوان تعداد بهینه خوشه‌ها در نظر گرفته شد. در نهایت عملیات خوشه‌بندی با استفاده از الگوریتم K-means انجام شد. شکل ۶- الف نمودار Elbow و شکل ۶- ب نمودار میانگین سیلوئت برای محاسبه تعداد خوشه‌های بهینه را نشان می‌دهند.



شکل ۶ الف) نمودار Elbow ب) نمودار میانگین سیلوئت

در مرحله بعد برای اطمینان حاصل کردن از اینکه کلمات به درستی بر اساس بخش گفتارشان[19] به صورت کلمه بیان می‌شوند، با استفاده از تابعی اسم، صفت، قید و یا فعل بودنشان مشخص شده است. پس از آن با حذف علامت‌های نگارشی، کلمات توقف و کلمات دارای کمتر از سه حرف، تبدیل حرف اول کلمات به حرف کوچک و ریشه‌یابی[20] تمیزسازی داده‌ها انجام شد. سپس با استفاده از کلمات اسم و صفت به ساخت مدل بایگرام[21] به کمک تکنیک Word2vec (برای استخراج ویژگی) و برآورد ۳۰ بایگرام پرتکرار پرداخته شد. این بایگرام‌ها در شکل ۷ نشان داده شده است.

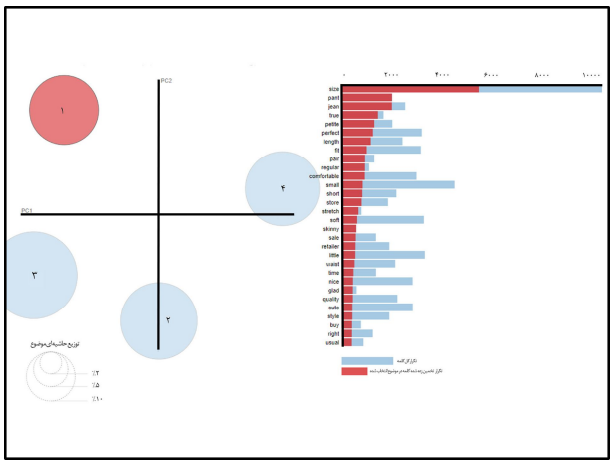


شکل ۷ ۳۰ بایگرام پرتکرار

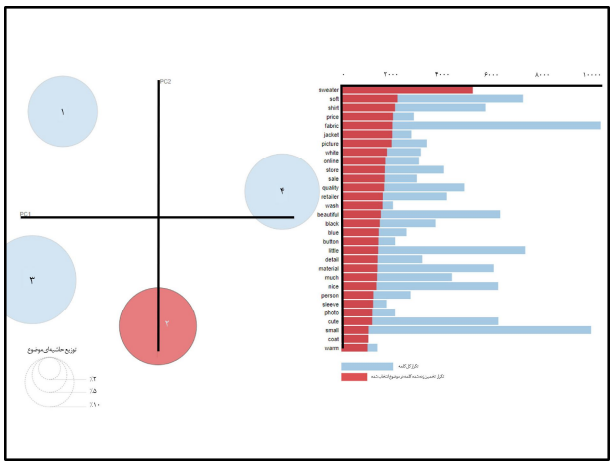
[16] Principal component analysis (PCA)
[17] T-distributed Stochastic Neighbor Embedding (t-SNE)
[18] Sum of Squared Errors (SSE)

[19] Part-of-speech
[20] Lemmatization
[21] Bigram

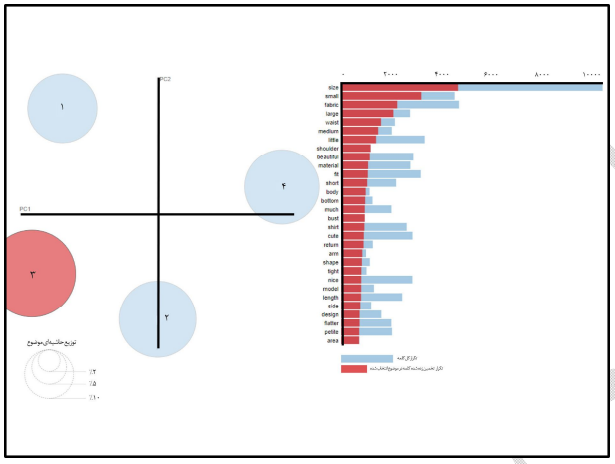
همانطور که مشاهده می‌شود، عبارت‌هایی مانند تمجید فراوان نشان-دهنده آن است که در حالت کلی مشتریان از خرید خود رضایت دارند. از طرفی نوع اندام، قیمت فروش، ارزش کالا با توجه به قیمت، مقایسه با فروشنده‌های محلی، وزن سبک و نحوه شست‌وشو از عبارت‌های پرتکرار و در نتیجه از موضوعاتی هستند که باید موردتوجه قرار گیرند. دسته‌بندی لباس‌ها مانند شلوار جین، دامن، ژاکت و غیره، تناسب و اندازه لباس‌ها و همچنین فصل‌های سال نیز از موضوعات مهم در نظر گرفته می‌شوند. شکل ۸ نمودار فاصله بین موضوعی و کلمات پرتکرار مربوط به موضوع یک، شکل ۹ نمودار بین موضوعی و کلمات پرتکرار مربوط به موضوع دو، شکل ۱۰ نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع سه و شکل ۱۱ نمودار فاصله بین موضوعی و کلمات پرتکرار مربوط به موضوع چهار را نشان می‌دهند. جدول ۲ به بررسی هر یک از موضوعات و کلمات مرتبط با آن‌ها به صورت مختصر می‌پردازد. این تحلیل‌ها درک جامعی از موضوعات کلیدی و نگرانی‌هایی که مشتریان هنگام بررسی اقلام لباس زنانه در هر یک از چهار موضوع دارند، ارائه می‌دهند. موضوع قابل‌توجه این است که در این جدول ترتیب دسته‌بندی هر یک از موضوعات بر اساس تعداد کلمات به‌کاررفته در آن دسته است. بنابراین در موضوع یک پراهمیت‌ترین مورد تناسب و اندازه، در موضوع دو پارچه و راحتی، در موضوع سه تناسب و اندازه و در موضوع چهار تناسب و راحتی است. پس از آن تخصیص دیریکله پنهان برای حالت‌های ۲ تا ۲۳ موضوعی آزمایش و با توجه به معیار coherence بهترین تعداد موضوع‌ها چهار در نظر گرفته شد.



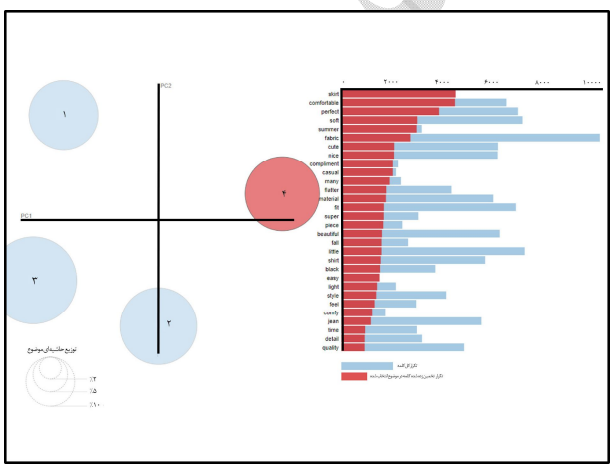
شکل ۸ نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع یک



شکل ۹ نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع دو



شکل ۱۰ نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع سه



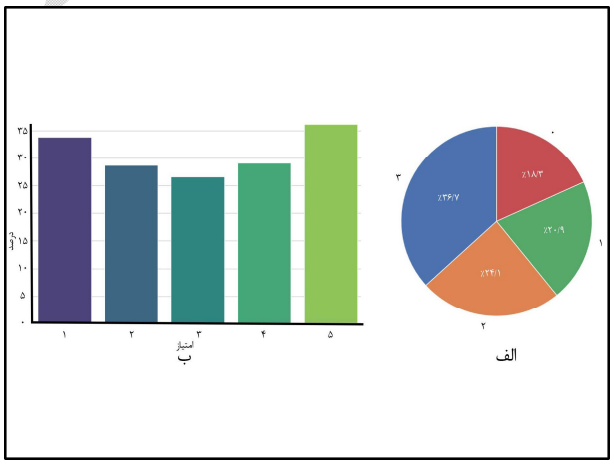
شکل ۱۱ نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع چهار

جدول ۲ تجزیه‌وتحلیل نتایج مدل‌سازی موضوعی

موارد بیان‌شده		کلمات استفاده‌شده	تحلیل
موضوعات	موضوع یک (متمرکز بر شلوار و جین)	تناسب و اندازه	"size," "petite," "regular," "small," "waist," "short," "usual"	عباراتی مانند اندازه، ریزه اندام، عادی، کوچک، کمر، کوتاه و معمول تاکید بر تناسب را برجسته می‌کند. خریداران اغلب به تناسب شلوار یا شلوار جین، نیاز به سایزهای مختلف (مثلاً کوچک) و رسیدن به تناسب کامل اشاره می‌کنند.
		راحتی	"comfortable," "stretch","soft", "skinny"	کلماتی مانند راحت، کشش و نرم نشان می‌دهد خریداران احساس لباس را در اولویت قرار می‌دهند.
		ظاهر و استایل	"pair," "perfect," "nice," "cute," "style," "quality"	کلماتی مانند جفت، عالی، خوب، ناز، استایل و کیفیت اهمیت ظاهر و استایل کلی شلوار و شلوار جین را نشان می‌دهد.
		تجربه خرید	"store," "sale," and "retailer"	عباراتی مانند فروشگاه، فروش و خرده‌فروش نشان می‌دهد که تجربه خرید، ازجمله مکان و نحوه خرید اقلام، نقش مهمی در رضایت مشتری دارد.
		مناسبت‌ها و کاربردها	"time," "buy," "right,", "glad"	کلماتی مانند زمان، خرید، درست و خوشحال نشان می‌دهد که مشتریان این موارد را بر اساس مناسب بودن آن‌ها برای مناسبت‌های مختلف و رضایت کلی از خرید خود بررسی می‌کنند.
	موضوع دو (متمرکز بر ژاکت و پیراهن)	پارچه و راحتی	"soft," "fabric," "material," "warm," "nice"	کلماتی مانند نرم، پارچه، مواد، گرم و خوب نشان می‌دهد که مشتریان برای مواد راحت و باکیفیت در ژاکت و پیراهن خود ارزش قائل هستند.
		ظاهر	"beautiful," "cute," "black," "blue," "button," and "detail"	عباراتی مانند زیبا، ناز، مشکی، آبی، دکمه و جزئیات نشان می‌دهند که جذابیت بصری و جزئیات طراحی برای مشتریان بسیار مهم است.
		قیمت و ارزش	"price," "sale," and "retailer"	کلماتی مانند قیمت، فروش و خرده‌فروش نشان می‌دهند که مشتریان هنگام بررسی این موارد، هزینه و ارزش پول را در نظر می‌گیرند. یافتن معاملات خوب و احساس اینکه کالا ارزش قیمت را دارد، عوامل مهمی هستند.
		تناسب و اندازه	"small" and "little"	بیان کلمات کوچک و کم نشان می‌دهد که تناسب و اندازه نیز از موارد مهم است. مشتریان اغلب در مورد اینکه آیا اقلام مطابق انتظار است و اینکه آیا اندازه دقیق است یا خیر بحث می‌کنند.
		تجربه خرید	"store," "online," "photo," and "picture"	عباراتی مانند فروشگاه، آنلاین، عکس و تصویر اهمیت تجربه خرید را چه در فروشگاه و چه آنلاین نشان می‌دهد. مشتریان درباره نحوه ظاهر شدن اقلام در عکس‌ها در مقابل حضوری و تجربه کلی خریدشان بحث می‌کنند.
	موضوع سه (متمرکز بر پیراهن و تاپ)	تناسب و اندازه	"size," "small," "large," "medium," "shoulder," "waist," "body," "bust," and "arm"	کلماتی مانند اندازه، کوچک، بزرگ، متوسط، شانه، کمر، بدن، سینه و بازو نشان‌دهنده تأکید زیادی بر نحوه تناسب پیراهن‌ها با اعضای مختلف بدن و نگرانی در مورد اندازه کلی است.
		پارچه و جنس	"fabric," "material," and "soft"	عباراتی مانند پارچه، مواد و نرم اهمیت کیفیت و راحتی پارچه را برجسته می‌کند. مشتریان اغلب در مورد احساس مواد و کیفیت کلی آن اظهارنظر می‌کنند.
		ظاهر و استایل	"beautiful," "cute," "nice," "design," "flatter," "shape," and "model"	کلماتی مانند زیبا، ناز، زیبا، طراحی، مطلوب‌تر، شکل و مدل اهمیت ظاهر و استایل را نشان می‌دهند. مشتریان به این فکر می‌کنند که پیراهن‌ها چگونه به نظر می‌رسند و آیا آن‌ها نسبت به فرم بدنشان جذاب هستند یا خیر.
		کاربردها و مناسبت‌ها	"return", "design"	کلمه بازگشت نشان می‌دهد که برخی از نظرات ممکن است در مورد روند بازگشت به دلیل تناسب یا مشکلات کیفیت بحث کنند. کلماتی مانند طراحی نشان می‌دهد که مشتریان پیراهن‌ها را برای مناسبت‌های مختلف در نظر می‌گیرند.

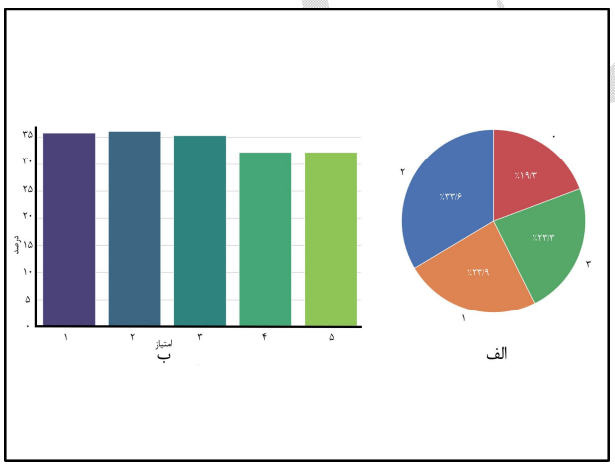
موضوع چهار (دامن و لباس‌های غیررسمی)	جزئیات	"short," "bottom," "tight," "length," "side," "area"	عباراتی مانند کوتاه، پایین، تنگ، طول، کنار و منطقه نشان می‌دهد که جزئیات خاصی در مورد تناسب و طرح پیراهن برای مشتریان مهم است. آن‌ها اغلب جنبه‌های خاصی از پیراهن را بیان می‌کنند که بر رضایت آن‌ها تأثیر می‌گذارد.
	تناسب و راحتی	"comfortable," "soft," "fit," "comfy," "perfect"	کلماتی مانند راحت، نرم، مناسب، راحت و کامل نشان می‌دهد که مشتریان به احساس و تناسب دامن‌ها اهمیت می‌دهند. تاکید بر یافتن مواردی است که هم راحت و هم مناسب باشند.
	ظاهر و استایل	"cute," "nice," "beautiful," "flatter," "style," "detail," and "quality"	کلماتی مانند ناز، زیبا، مسحورکننده، استایل، جزئیات و کیفیت اهمیت ظاهر و استایل را نشان می‌دهند. مشتریان از دامن‌هایی که زیبا به نظر می‌رسند و جزئیات طراحی جذابی دارند قدردانی می‌کنند.
	مناسبت‌ها و تطبیق-پذیری	"casual," "summer," "fall," "piece," "easy,"	عباراتی مانند کژوال، تابستان، پاییز، تکه و آسان نشان می‌دهد که مشتریان به دنبال دامن‌های همه‌کاره‌ای هستند که بتوانند در فصول مختلف و برای مناسبت‌های مختلف بپوشند. قابلیت ست کردن دامن‌ها با تاپ‌های مختلف مانند پیراهن و بلوز نیز مهم است.
	پارچه و جنس	"fabric," "material," "light," and "feel"	کلماتی مانند پارچه، مواد، سبک و حس نشان می‌دهد که کیفیت پارچه و احساس دامن‌ها دارای اهمیت هستند. مشتریان اغلب در مورد مناسب بودن مواد برای شرایط آب و هوایی مختلف و کیفیت کلی آن نظر می‌دهند.
	تجربه خرید و رضایت	"compliment," "time," and "glad"	کلماتی مانند تمجید، وقت و خوشحال نشان می‌دهد که مشتریان تحت تأثیر بازخوردهایی هستند که هنگام پوشیدن دامن و رضایت کلی آن‌ها از خرید دریافت می‌کنند. تجارب مثبت و تعریف و تمجید دیگران به رضایت آن‌ها کمک می‌کند.

درنهایت تجزیه‌وتحلیل مربوط به هر خوشه با توجه به درصد امتیازها، مصورسازی ویژگی‌های عددی و موضوعات اصلی در آن خوشه انجام شد. شکل‌های ۱۲، ۱۳ و ۱۴ به ترتیب درصد موضوعات و امتیازها در خوشه یک، دو و سه را نشان می‌دهند.



شکل ۱۲ الف) درصد موضوعات در خوشه یک ب) درصد امتیازها در خوشه یک

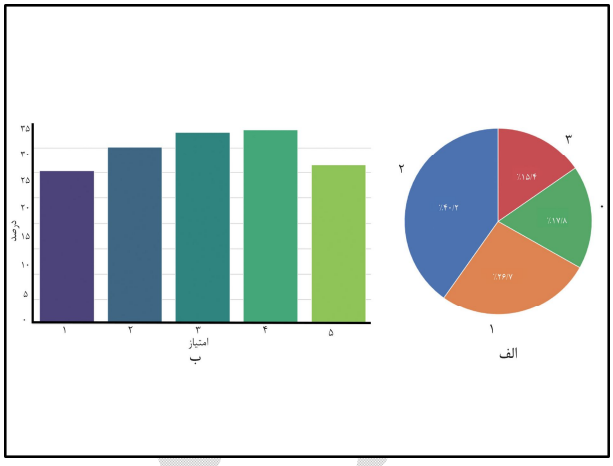
شکل ۱۵ نحوه توزیع برخی از ویژگی‌های عددی در هر سه خوشه را نشان می‌دهد. با توجه به نمودارهای گزارش‌شده، خوشه یک شامل زنان جوان و میان‌سال با گروه سنی ۲۵-۴۰ است که نظرات نسبتاً دقیقی (۱۰-۵۰ کلمه) نوشته‌اند و حدود ۴۷ درصد از نظرات آن‌ها مربوط به موضوع ۳ و متمرکز بر تناسب و راحتی، پارچه و جنس، ظاهر و استایل و جزئیات پیراهن و تاپ است. خوشه دو دربرگیرنده نظرات زنان میان‌سال و سالمند با گروه سنی ۶۰-۳۵ سال است.



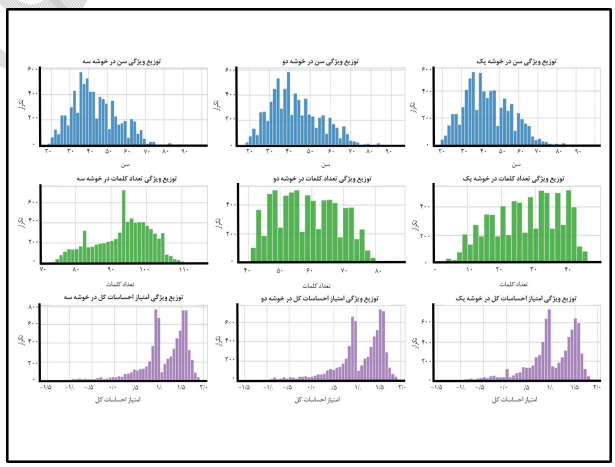
شکل ۱۳ الف) درصد موضوعات در خوشه دو ب) درصد امتیازها در خوشه دو

نظرات آن‌ها به صورت توصیفی با نزدیک به ۴۰-۹۰ کلمه و تقریبا ۳۳ درصد آن‌ها مربوط به موضوع ۲ است. همان‌طور که گفته شد این موضوع بیشتر روی ژاکت و پیراهن، راحتی، پارچه و ارزش و قیمت متمرکز شده است. خوشه سه شامل نظرات همه گروه‌های

سنی از ۲۵ تا ۶۰ سال است که نظرات بسیار دقیق با تعداد ۸۰-۱۱۰ کلمه نوشته‌اند. نزدیک به ۴۲ درصد از این نظرات مربوط به موضوع ۲ است.



شکل ۱۴ الف) درصد موضوعات در خوشه سه ب) درصد امتیازها در خوشه سه



شکل ۱۵ نحوه توزیع ویژگی‌های عددی در خوشه‌ها

۳-۴- انتخاب ویژگی

پس از بررسی میزان همبستگی ویژگی‌ها با ویژگی امتیاز به‌عنوان متغیر هدف و با توجه به صحت به‌دست‌آمده از اعمال مدل‌های یادگیری ماشین بر مجموعه داده تصمیم گرفته شد به‌منظور انجام این مطالعه از ویژگی متن نظرات و امتیاز استفاده شود.

۳-۵- پیش‌پردازش داده‌ها و استخراج ویژگی

در این مطالعه به‌منظور انجام پیش‌پردازش و استخراج ویژگی از الگوریتم TFIDF به همراه حذف کلمات توقف استفاده می‌شود که در ادامه به بیان توضیح مختصری از آن‌ها خواهیم پرداخت.

- **کلمات توقف:** رایج‌ترین اصطلاحات در یک زبان که ارزش معنایی ندارند و فقط به درک کلی متن کمک می‌کنند را کلمات توقف گویند که با حروف تعریف، حروف اضافه، علائم نگارشی، حروف ربط و ضمایر مشخص می‌شوند. حذف آن‌ها مقدار توکن‌ها را کاهش می‌دهد و تحلیل را بهبود می‌بخشد [۶].
- **الگوریتم TFIDF برای استخراج ویژگی:** به‌منظور تبدیل متن به بردار و همچنین امکان محاسبه وزن نسبت-داده‌شده به هر عبارت موجود در سند، ابزار TfidfVetorizer از کتابخانه sklearn استفاده می‌شود که برای انجام این عملیات از الگوریتم TF-IDF استفاده و درنهایت کلمات موجود در متن نظر را به یک ماتریس برداری تبدیل می‌کند [۸, ۱۱].

۳-۶- متعادل کردن مجموعه داده

همان‌طور که در قسمت تجزیه‌وتحلیل داده‌های اکتشافی در بخش بررسی ویژگی امتیاز مشاهده شد، مجموعه داده دارای توزیع نامتعادل کلاس امتیازها است. بنابراین به‌منظور اعمال مدل‌های یادگیری ماشین ابتدا باید مجموعه داده متعادل شود. ازآنجایی‌که در تکنیک‌های معمول افزایش داده‌های اقلیت مانند OverSampling ممکن است داده‌های تکراری زیادی به مجموعه اضافه و نتایج را غیرقابل‌اعتماد کنند، روش انتخابی برای متعادل کردن مجموعه داده در مطالعه پیش‌رو روش SMOTE[1] است. این روش با استفاده از الگوریتم k-نزدیک‌ترین همسایه به‌منظور متعادل کردن مجموعه داده نمونه‌های مصنوعی جدید تولید می‌کند [۱۵]. ازآنجایی‌که متغیر هدف امتیاز در حالت معمول دارای پنج کلاس (امتیاز ۱ تا ۵) است به‌منظور انجام مطالعه و اجرای مدل‌ها در حالت دوکلاسه و سه‌کلاسه امتیازها دسته‌بندی می‌شوند. در حالت دوکلاسه امتیازهای قبل از ۳ به‌عنوان کلاس صفر و امتیاز ۳ و بالاتر به‌عنوان کلاس یک تعریف می‌شوند. در حالت سه‌کلاسه امتیازهای قبل از ۳ به‌عنوان

[1] Synthetic Minority Oversampling Technique

کلاس صفر، امتیاز ۳ به‌عنوان کلاس ۱ و امتیازهای بالاتر از آن به‌عنوان کلاس ۲ در نظر گرفته می‌شوند. در مرحله‌ی بعد به اجرا مدل‌های یادگیری ماشین پرداخته خواهد شد. پس از بررسی کارهای مرتبط و مطالعات گذشته و همچنین اعمال برخی از مدل‌های یادگیری ماشین و با توجه به صحت حاصل، مدل‌های یادگیری ماشین در نظر گرفته شده برای مطالعه پیش‌رو عبارت هستند از رگرسیون لجستیک، ماشین بردار پشتیبان، جنگل تصادفی، LightGBM، درخت تصمیم، XGBoost، بیز ساده چندجمله‌ای و بیز ساده مکمل[1] که در ادامه به بیان توضیح مختصری از آن‌ها پرداخته خواهد شد.

۳-۷- مدل‌های یادگیری ماشین

- **رگرسیون لجستیک:** از مدل‌های یادگیری ماشین تحت‌نظارت است که در کلاس‌بندی، تحلیل و پیش‌بینی با استفاده از مشاهدات قبلی یک مجموعه داده به کار گرفته می‌شود. رگرسیون لجستیک از یک تابع لجستیک به نام تابع سیگموئید که احتمال رخ دادن متغیر وابسته را بین صفر و یک محدود می‌کند، برای تخمین احتمالات استفاده می‌کند [۲۶, ۱۲].
- **ماشین بردار پشتیبان:** از الگوریتم‌های یادگیری ماشین تحت‌نظارت به‌منظور انجام وظایف کلاس‌بندی است و هدف آن یافتن ابر صفحه‌ای است که دارای بیشترین حاشیه بین کلاس‌های داده‌های آموزشی باشد [۵, ۲۶].
- **درخت تصمیم:** روش یادگیری ماشین تحت‌نظارت است که در مسائل رگرسیون و همچنین کلاس‌بندی استفاده می‌شود و با خلاصه کردن ویژگی‌های داده از داده‌های آموزشی به پیش‌بینی مقدار متغیر هدف کمک می‌کند [۲۷, ۲۶].
- **جنگل تصادفی:** از روش‌های کلاس‌بندی مبتنی بر یادگیری گروهی[2] است که با استفاده از روش گروه‌بندی موازی[3] و ترکیب چندین درخت تصمیم و انتخاب تصادفی داده‌ها و ویژگی‌ها عملکرد مدل را بهبود می‌دهد. بنابراین مشکل بیش‌برازش را به حداقل می‌رساند و صحت پیش‌بینی را افزایش می‌دهد [۱۲, ۲۶].
- **LightGBM:** LightGBM[4] الگوریتمی مبتنی بر یادگیری گروهی و درختان تصمیم است که با استفاده از چارچوب GBDT[5] طراحی شده است. این مدل با استفاده از تکنیک‌های نمونه‌برداری یک‌طرفه مبتنی بر گرادیان (GOSS)[6] و دسته‌بندی ویژگی‌های انحصاری (EFB)[7] درحالی‌که زمان آموزش و استفاده از حافظه را کاهش می‌دهد عملکرد مدل را بهبود می‌بخشد [۵, ۲۶].
- **بیز ساده چندجمله‌ای و مکمل:** بیز ساده از الگوریتم‌های یادگیری ماشین تحت‌نظارت مبتنی بر احتمال و قضیه بیز با فرض استقلال بین هر جفت ویژگی است و هدف آن بهبود احتمال پسین با استفاده از داده‌های آموزشی است که درنهایت منجر به یک قانون تصمیم‌گیری برای داده‌های جدید می‌شود [۲۸, ۲۶]. بیز ساده چندجمله‌ای بر اساس تعداد دفعاتی که یک کلمه در یک سند ظاهر می‌شود عمل می‌کند. الگوریتم بیزساده مکمل اقتباسی از بیزساده چندجمله‌ای است که به جای محاسبه احتمال یک نمونه متعلق به یک کلاس خاص، احتمال تعلق نمونه به همه کلاس‌ها را محاسبه می‌کند [۱۲, ۲۹].
- **XGBoost** : تقویت گرادیان شدید (XGBoost)[8] شکلی از الگوریتم‌های تقویت گرادیان است که تقریب‌های دقیق‌تری را هنگام تعیین بهترین مدل در نظر می‌گیرد و مانند جنگل‌های تصادفی یک الگوریتم یادگیری مجموعه‌ای است که یک مدل نهایی را بر اساس یک سری مدل‌های جداگانه، معمولا درخت‌های تصمیم، تولید می‌کند [۲۷, ۲۶].

۳-۸- معیارهای ارزیابی

معیارهای ارزیابی مورداستفاده در این مطالعه عبارت هستند از صحت، دقت، پوشش و امتیاز F1 که خلاصه‌ای از نحوه محاسبه و توضیحات این معیارها در جدول ۳ نشان داده‌شده است.

[1] Complement naive Bayes (CNB)
[2] Ensemble learning
[3] Parallel ensembling
[4] Light Gradient Boosting Machine
[5] Gradient Boosted Decision Trees
[6] Gradient-based One-Side Sampling (GOSS)
[7] Exclusive feature bundling (EFB)
[8] Extreme gradient boosting (XGBoost)

جدول ۳ معیارهای ارزیابی و نحوه محاسبه [۶, ۱۱, ۹]

نام	نحوه محاسبه	توضیحات
پوشش	$TPR=\frac{TP+TN}{FN+TP}$	نسبت مثبت‌های واقعی که به درستی به‌عنوان یک کلاس مثبت شناسایی می‌شوند.
صحت	$A=\frac{TP+TN}{TN+FP+FN+TP}$	نسبت پیش‌بینی‌های صحیح از تعداد کل پیش‌بینی‌های یک مدل
دقت	$P=\frac{TP}{FP+TP}$	نسبت مثبت‌های پیش‌بینی‌شده که درواقع در یک کلاس مثبت هستند.
امتیاز F1	$F1=2\frac{P\times TPR}{P+TPR}$	میانگین وزنی صحت و پوشش است که با در نظر گرفتن مثبت کاذب و منفی کاذب ارزیابی دقیق‌تری از توانایی یک مدل برای کلاس‌بندی صحیح نمونه‌ها ارائه می‌دهد.

۴- ارزیابی و نتایج

در این بخش به بیان معیارهای ارزیابی برآوردشده توسط مدل‌های یادگیری ماشین اجراشده در حالت‌های دوکلاسه، سه‌کلاسه و پنج‌کلاسه خواهیم پرداخت. در بین کارهای مرتبط بیان‌شده در بخش‌های قبل مطالعه شتی و همکاران با عنوان کاوش احساسات بازخورد داده‌های تجارت الکترونیک با استفاده از الگوریتم‌های یادگیری ماشین دارای متغیر هدف یکسان با مطالعه‌ی انجام شده است و تنها در حالت دوکلاسه صورت گرفته است. جدول ۴ به‌منظور مقایسه معیارهای ارزیابی دو مطالعه صورت‌گرفته طراحی شده است. معیارهای ارزیابی بیان‌شده در جدول هر دو بر روی مجموعه داده یکسان و با استفاده از الگوریتم استخراج ویژگی TF-IDF در نظر گرفته شده‌اند. همان‌طور که مشاهده می‌شود، بالاترین صحت در مطالعه شتی و همکاران مربوط به مدل ماشین بردار پشتیبان با صحت ۰/۹۴ و بهترین عملکرد در این مطالعه مربوط به مدل جنگل تصادفی با صحت ۰/۹۸ و پس از آن صحت مدل ماشین بردار پشتیبان برابر با ۰/۹۵ در نظرگرفته می‌شود. موضوع قابل‌توجه این است که در مطالعه شتی و همکاران (۲۰۲۴) به‌منظور رفع نامتعادل بودن مجموعه داده راهکاری ارائه نشده است.

مقادیر صحت، دقت، پوشش و امتیاز F1 برای مدل‌های یادگیری ماشین اجراشده در حالت سه‌کلاسه و پنج‌کلاسه در جدول ۵ مشاهده می‌شود. در حالت سه‌کلاسه، بهترین عملکرد مربوط به مدل جنگل تصادفی با صحت ۰/۹۵، دقت ۰/۹۶، پوشش ۰/۹۳ و امتیاز F1 ۰/۹۴ است. پس از آن مدل ماشین بردار پشتیبان با دارا بودن مقادیر صحت ۰/۹۰، دقت ۰/۹۰، پوشش ۰/۸۸ و امتیاز F1 ۰/۸۹ عملکرد خوبی دارد. در حالت پنج‌کلاسه نیز بهترین عملکرد مربوط به مدل جنگل تصادفی با صحت ۰/۹۱، دقت ۰/۹۲، پوشش ۰/۹۱ و امتیاز F1 ۰/۹۱ است. پس از آن مدل ماشین بردار پشتیبان با دارا بودن مقادیر صحت ۰/۸۳، دقت ۰/۸۳، پوشش ۰/۸۳ و امتیاز F1 ۰/۸۳ عملکرد خوبی دارد.

۵- نتیجه‌گیری و پیشنهادهای آینده

با توجه به اهمیت شناخت رفتار مشتریان به‌منظور ادامه فعالیت در دنیای امروز، در این مطالعه به بررسی احساسات نظرات تجارت الکترونیک پوشاک زنان با استفاده از تکنیک‌های پردازش زبان طبیعی و یادگیری ماشین پرداخته شد. مدل‌های یادگیری ماشین به کار گرفته‌شده شامل ماشین بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده چندجمله‌ای، بیز ساده مکمل، XGBoost و LightGBM هستند. به‌منظور برداری کردن و استخراج ویژگی متن نظرات از الگوریتم TF-IDF و Word2vec استفاده شد. مدل‌سازی موضوعی با استفاده از روش تخصیص دیریکله پنهان و خوشه‌بندی k-means انجام شد. معیارهای ارزیابی استفاده‌شده در این مطالعه شامل صحت، دقت، پوشش و امتیاز F1 است. ازآنجایی‌که متغیر هدف امتیاز در حالت معمول دارای پنج کلاس (امتیاز ۱ تا ۵) است به‌منظور انجام مطالعه و اجرای مدل‌ها در حالت دوکلاسه و سه‌کلاسه امتیازها دسته‌بندی شده‌اند. در حالت دوکلاسه امتیازهای قبل از ۳ به‌عنوان کلاس صفر و امتیاز ۳ و بالاتر به‌عنوان کلاس یک تعریف شده‌اند. در حالت سه‌کلاسه امتیازهای قبل از ۳ به‌عنوان کلاس صفر، امتیاز ۳ به‌عنوان کلاس ۱ و امتیازهای بالاتر از آن به‌عنوان کلاس ۲ در نظر گرفته شده‌اند. پس از اعمال الگوریتم‌های یادگیری ماشین و پردازش زبان طبیعی در هر سه حالت بهترین عملکرد مربوط به مدل جنگل تصادفی با صحت ۰/۹۸ در حالت دوکلاسه، ۰/۹۵ در حالت سه‌کلاسه و ۰/۹۱ در حالت پنج‌کلاسه برآورد شد.

این مطالعه ترکیبی از متن‌کاوی و پردازش زبان طبیعی است، اما بر اساس اهداف و روش‌شناسی، بیشتر به حوزه متن‌کاوی متمرکز است. هدف اصلی تحقیق، تحلیل و طبقه‌بندی نظرات مشتریان با استفاده از الگوریتم‌های یادگیری ماشین است. تکنیک‌های پردازش زبان طبیعی مانند TF-IDF و Word2Vec برای برداری‌سازی و استخراج ویژگی‌های متنی به کار گرفته شده‌اند و نقش پشتیبانی‌کننده‌ای در پیش‌پردازش داده‌ها و آماده‌سازی آنها برای مدل‌سازی ایفا کرده‌اند. علاوه بر این، فرآیندهایی مانند دسته‌بندی مجدد متغیر هدف، متعادل‌سازی داده‌ها با SMOTE و استفاده از الگوریتم‌های مختلف یادگیری ماشین برای پیش‌بینی و تحلیل احساسات، نشان‌دهنده تمرکز تحقیق بر کشف الگوها و پیش‌بینی‌ها از داده‌های متنی است. درحالی‌که تکنیک‌های پردازش زبان طبیعی در بخش‌هایی مانند مدل‌سازی موضوعی و استخراج ویژگی به‌کار رفته‌اند، نقش اصلی آن‌ها پشتیبانی از تحلیل داده‌های متنی و مدل‌سازی است. بنابراین، این تحقیق بیشتر در حوزه متن‌کاوی قرار می‌گیرد و تکنیک‌های پردازش زبان طبیعی به‌عنوان ابزارهایی برای تسهیل فرآیند متن‌کاوی استفاده شده‌اند. به‌منظور انجام پژوهش در کارهای آینده می‌توان علاوه بر مدل‌های یادگیری ماشین، از مدل‌های یادگیری عمیق نیز به‌منظور تجزیه‌وتحلیل احساسات استفاده کرد. تکنیک‌های استخراج ویژگی در پردازش زبان طبیعی نیز دارای الگوریتم‌های متفاوتی هستند که می‌توان به‌منظور بررسی نظرات و رفتار مشتریان آن‌ها را به کار گرفت.

جدول ۴ مقایسه نتایج حالت دوکلاسه با کارهای مشابه

		شتی و همکاران (۲۰۲۴)				این مطالعه			
		معیارهای ارزیابی							
		صحت	دقت	پوشش	امتیاز F1	صحت	دقت	پوشش	امتیاز F1
مدل‌های یادگیری ماشین	ماشین بردار پشتیبان	۰٫۹۴	۰٫۹۴	۰٫۹۹	۰٫۹۶	**۰٫۹۵**	**۰٫۹۶**	۰٫۹۵	۰٫۹۵
	رگرسیون لجستیک	۰٫۹۳	۰٫۹۴	۰٫۹۹	۰٫۹۷	۰٫۹۰	۰٫۹۱	۰٫۹۲	۰٫۹۱
	درخت تصمیم	–	–	–	–	۰٫۹۱	۰٫۹۲	۰٫۹۲	۰٫۹۲
	جنگل تصادفی	۰٫۹۰	۰٫۹۰	۱٫۰۰	۰٫۹۵	**۰٫۹۸**	**۰٫۹۶**	۰٫۹۹	**۰٫۹۸**
	AdaBoosting	۰٫۹۱	۰٫۹۳	۰٫۹۷	۰٫۹۵	–	–	–	–
	بیز ساده برنولی	۰٫۹۲	۰٫۹۵	۰٫۹۷	۰٫۹۶	–	–	–	–
	بیز ساده چندجمله‌ای	۰٫۸۹	۰٫۸۹	۱٫۰۰	۰٫۹۴	۰٫۸۷	۰٫۸۶	۰٫۹۴	۰٫۸۹
	بیز ساده مکمل	–	–	–	–	۰٫۸۷	۰٫۹۲	۰٫۸۶	۰٫۸۹
	XGBoost	–	–	–	–	۰٫۹۱	۰٫۹۱	۰٫۹۴	۰٫۹۳
	LightGBM	–	–	–	–	۰٫۹۰	۰٫۹۱	۰٫۹۲	۰٫۹۱

جدول ۵ نتایج حالت سه‌کلاسه و پنج‌کلاسه

		سه‌کلاسه				پنج‌کلاسه			
		معیارهای ارزیابی							
		صحت	دقت	پوشش	امتیاز F1	صحت	دقت	پوشش	امتیاز F1
مدل‌های یادگیری ماشین	ماشین بردار پشتیبان	۰٫۹۰	۰٫۹۰	۰٫۸۸	۰٫۸۹	۰٫۸۳	۰٫۸۳	۰٫۸۳	۰٫۸۳
	رگرسیون لجستیک	۰٫۸۴	۰٫۸۳	۰٫۷۸	۰٫۷۹	۰٫۷۵	۰٫۷۴	۰٫۷۵	۰٫۷۴
	درخت تصمیم	۰٫۸۳	۰٫۸۱	۰٫۸۱	۰٫۸۱	۰٫۷۲	۰٫۷۲	۰٫۷۲	۰٫۷۲
	جنگل تصادفی	**۰٫۹۵**	۰٫۹۶	۰٫۹۳	۰٫۹۴	**۰٫۹۱**	۰٫۹۲	۰٫۹۱	۰٫۹۱
	بیز ساده چندجمله‌ای	۰٫۷۸	۰٫۸۳	۰٫۶۸	۰٫۶۷	۰٫۷۳	۰٫۷۳	۰٫۷۳	۰٫۷۳
	بیز ساده مکمل	۰٫۸۱	۰٫۸۱	۰٫۷۵	۰٫۷۶	۰٫۷۲	۰٫۷۳	۰٫۷۲	۰٫۷۲
	XGBoost	۰٫۸۶	۰٫۸۷	۰٫۸۲	۰٫۸۳	۰٫۷۹	۰٫۸۰	۰٫۷۹	۰٫۷۹
	LightGBM	۰٫۸۵	۰٫۸۵	۰٫۸۱	۰٫۸۲	۰٫۸۱	۰٫۸۱	۰٫۸۱	۰٫۸۰

۶- مراجع

[1] Mabrouk, A., Redondo, R.P.D. and Kayed, M., 2021. Seopinion: summarization and exploration of opinion from e-commerce websites. *Sensors*, *21*(2), p.636. https://doi.org/10.3390/s21020636.

[2] Huang, H., Asemi, A. and Mustafa, M.B., 2023, July. Sentiment Analysis Application in E-Commerce: Current Models and Future Directions. In *International Conference on Electronic Government and the Information Systems Perspective* (pp. 67-72). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-39841-4_5.

[3] Kasimu, M., Hellen, N. and Marvin, G., 2023. Explainable Sentiment Analysis for Textile Personalized Marketing. In *The Fourth Industrial Revolution and Beyond: Select Proceedings of IC4IR+* (pp. 473-488). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-8032-9_33.

[4] Zainal, S.R.M. and Xiang, C., 2023. Examining Women's E-Commerce Clothing Reviews. *International Journal of Business and Technology Management*, *5*(1), pp.22-38. Available at: <https://myjms.mohe.gov.my/index.php/ijbtm/article/view/21637>. Date accessed: 04 aug. 2024.

[5] Lin, X., 2020, April. Sentiment analysis of e-commerce customer reviews based on natural language processing. In *Proceedings of the 2020 2nd international conference on big data and artificial intelligence* (pp.32-36). https://doi.org/10.1145/3436286.3436293.

[6] Kubrusly, J., Neves, A.L. and Marques, T.L., 2022. A statistical analysis of textual e-commerce reviews using tree-based methods. *Open Journal of Statistics*, *12*(3), pp.357-372. https://doi.org/10.4236/ojs.2022.123023.

[7] Muniasamy, A. and Bhatnagar, R., 2022. Analyzing online reviews of customers using machine learning techniques. In *Rising Threats in Expert Applications and Solutions: Proceedings of FICR-TEAS 2022* (pp. 485-493). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-19-1122-4_51.

[8] Sunil, N. and Shirazi, F., 2023, July. Customer review classification using machine learning and deep learning techniques. In *International Conference on Human-Computer Interaction* (pp. 581-597). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-35915-6_42.

[9] Loukili, M., Messaoudi, F. and El Ghazi, M., 2023. Sentiment analysis of product reviews for e-commerce recommendation based on machine learning. *International Journal of Advances in Soft Computing & Its Applications*, *15*(1). https://doi.org/10.15849/IJASCA.230320.01.

[10] Sharma, H., 2023. Using Topic Modeling for Extracting Customers' Expectations: A Case of Women Apparel. *Business Perspectives and Research*, p.22785337221150831. https://doi.org/10.1177/22785337221150831.

[11] Shetty, A.M., Aljunid, M.F., Manjaiah, D.H. and Shaik Afzal, A.M., 2023, July. Hyperparameter Optimization of Machine Learning Models Using Grid Search for Amazon Review Sentiment Analysis. In *International Conference on Data Science and Applications* (pp. 451-474). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7814-4_36.

[12] Shetty, A.M., Aljunid, M.F. and Manjaiah, D.H., 2023, February. Sentiment Exploring on Feedback of E-commerce Data Using Machine Learning Algorithms. In *International Conference on Emerging Research in Computing, Information, Communication and Applications* (pp. 107-129). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-7622-5_8.

[13] Yabes, I., Aryanti, N., Pradana, R.J., Setiawan, K.E. and Hasani, M.F., 2023, October. Classifying and Predicting The Rating Sentiment of Women's E-commerce Clothing Reviews: A Comparative Study Using SVM, ANN, and BERT Models. In *2023 5th International Conference on Cybernetics and Intelligent System (ICORIS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICORIS60118.2023.10352189.

[14] Shashank, S. and Behera, R.K., 2024. Factors influencing recommendations for women's clothing satisfaction: A latent dirichlet allocation approach using online reviews. *Journal of Retailing and*

Consumer Services, *81*, p.104011. https://doi.org/10.1016/j.jretconser.2024.104011.

[15] Padurariu, C. and Breaban, M.E., 2019. Dealing with data imbalance in text classification. *Procedia Computer Science*, *159*, pp.736-745. https://doi.org/10.1016/j.procs.2019.09.229.

[16] Mahdikhani, M., 2023. Exploring commonly used terms from online reviews in the fashion field to predict review helpfulness. *International Journal of Information Management Data Insights*, *3*(1), p.100172. https://doi.org/10.1016/j.jjimei.2023.100172.

[17] Li, M., Zhao, L. and Srinivas, S., 2023. It is about inclusion! Mining online reviews to understand the needs of adaptive clothing customers. *International Journal of Consumer Studies*, *47*(3), pp.1157-1172. https://doi.org/10.1111/ijcs.12895.

[18] Mahmud, F.A.M., Mullick, S.B.R.A. and Anas, T.C.M., Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach.

[19] NICAPOTATO; Women's E-Commerce Clothing Reviews, https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews, 2018.

[20] Heng, Y., Gao, Z., Jiang, Y. and Chen, X., 2018. Exploring hidden factors behind online food shopping from Amazon reviews: A topic mining approach. *Journal of Retailing and Consumer Services*, *42*, pp.161-168. https://doi.org/10.1016/j.jretconser.2018.02.006.

[21] Kim, S.W. and Gil, J.M., 2019. Research paper classification systems based on TF-IDF and LDA schemes. *Human-centric Computing and Information Sciences*, *9*, pp.1-21. https://doi.org/10.1186/s13673-019-0192-7.

[22] Zhong, K., Jackson, T., West, A. and Cosma, G., 2024. Natural Language Processing Approaches in Industrial Maintenance: A Systematic Literature Review. *Procedia Computer Science*, *232*, pp.2082-2097. https://doi.org/10.1016/j.procs.2024.02.029.

[23] Vadivel, A., Meena, K., Sumathy, P., Selvaraj, H., Shanmugavadivu, P., & S. G., S., Eds., 2024. *Interactive and Dynamic Dashboard: Design Principles*,1st Edn., CRC Press. https://doi.org/10.1201/9781003542735.

[24] Cui, M., 2020. Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, *1*(1), pp.5-8. https://dx.doi.org/10.23977/accaf.2020.010102.

[25] Januzaj, Y., Beqiri, E. and Luma, A., 2023. Determining the Optimal Number of Clusters using Silhouette Score as a Data Mining Technique. *International Journal of Online & Biomedical Engineering*, *19*(4). https://doi.org/10.3991/ijoe.v19i04.37059.

[26] Sarker, I.H., 2021. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, *2*(3), p.160. https://doi.org/10.1007/s42979-021-00592-x.

[27] Pedregosa, F., 2011. Scikit-learn: Machine learning in python Fabian. *Journal of machine learning research*, *12*, p.2825. https://doi.org/10.1002/hbm.25822.

[28] Noor, A. and Islam, M., 2019, July. Sentiment Analysis for Women's E-commerce Reviews using Machine Learning Algorithms. In *2019 10th International conference on computing, communication and networking technologies (ICCCNT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICCCNT45670.2019.89444.36.

[29] Chrismanto, A.R., Sari, A.K. and Suyanto, Y., 2023. Enhancing Spam Comment Detection on Social Media With Emoji Feature and Post-Comment Pairs Approach Using Ensemble Methods of Machine Learning. *IEEE Access*. https://doi.org/10.1109/ACCESS.2023.3299853.