Sharif Industrial Engineering and Management Journal (2026), **41**(2), 45-61

Sharif University of Technology

https://sjie.journals.sharif.edu/

Research Article

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

Razieh Rahimi[1] and Reza Ghasemy Yaghin[*2]

[1]Department of Textile Engineering, Amirkabir University of Technology, Tehran, Iran.
[2]Department of Industrial Engineering and Management Systems, Amirkabir University of Technology.
* corresponding author: (yaghin@aut.ac.ir)

Article Info

Article history:

Received:15 October 2024

Revised:15 January 2025

Accepted:10 March 2025

Keywords:

Sentiment analysis,
machine learning,
natural language processing,
text mining,
customer behavior.

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 the 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 punctuation marks, stop words and words with fewer than three letters, converting the first letter of the words to lowercase and lemmatization, data cleaning was done. After that, topic modeling was done and each of the topics and words related to them were examined. In the next step, the topics were examined in different clusters. These analyses provide a comprehensive understanding of the key themes and concerns customers have when considering womenswear items in each of the four topics.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflicts of Interest: The authors declare no conflict of interest.

Author Contributions: All authors contributed to the reported work for conceptualization, methodology, validation, writing-review and editing, and supervision.

To Cite this article:
Rahimi, R. and Ghasemy Yaghin, R. 2026. Topic modeling and sentiment analysis of customers using natural language processing and machine learning techniques, Sharif Industrial Engineering and Management Journal, **41**(2), pp.45-61. https://doi.org/10.24200/j65.2025.65318.2418

E-ISSN: 2676-475X

مجله علمی مهندسی صنایع و مدیریت شریف

https://sjie.journals.sharif.edu

مقاله‌ی پژوهشی

مدل‌سازی موضوعی و تحلیل احساس‌های مشتریان با استفاده از روش‌های پردازش زبان طبیعی و یادگیری ماشین

راضیه رحیمی[۱] و رضا قاسمی‌یقین[۲*]

[۱] دانشکده مهندسی نساجی، دانشگاه صنعتی امیرکبیر، تهران، ایران

[۲] دانشکده‌ی مهندسی صنایع و سیستم‌های مدیریت، دانشگاه صنعتی امیرکبیر، تهران، ایران.

[*] نویسنده مسئول (yaghin@aut.ac.ir)

اطلاعات مقاله

تاریخ دریافت ۱۴۰۳/۰۷/۲۴

تاریخ اصلاحیه : ۱۴۰۳/۱۰/۲۶

تاریخ پذیرش : ۱۴۰۳/۱۲/۲۰

واژگان کلیدی:

تجزیه‌وتحلیل احساس‌ها،

یادگیری ماشین،

پردازش زبان طبیعی،

متن‌کاوی،

رفتار مشتری،

مدل‌سازی موضوعی.

چکیده

در نوشتارحاضر، به بررسی و تجزیه‌وتحلیل نظرهای مشتریان پوشاک زنان با استفاده از روش‌های یادگیری ماشین و پردازش زبان طبیعی پرداخته شده است. مدل‌های یادگیری ماشین استفاده‌شده، شامل: ماشین بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده‌ی چندجمله‌ای، بیز ساده‌ی مکمل، XGBoost ، و LightGBM بوده‌اند. به‌منظور برداری‌کردن و استخراج ویژگی متن نظرها از الگوریتم‌های TF-IDF و Word2vec استفاده شده است. مدل-سازی موضوعی با استفاده از روش تخصیص دیریکله‌ی پنهان و خوشه‌بندی K-Means انجام شده است. مجموعه‌ی داده‌ی استفاده‌شده مربوط به نظرهای پوشاک زنان و متغیر هدف، امتیازهای مشتریان در نظرها بوده است. مطالعه‌ی حاضر، در حالت‌های: 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 شده است.

۱. مقدمه

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

به همین منظور در پژوهش حاضر، رویکردی مبتنی بر روش‌های پردازش زبان طبیعی و مدل‌های یادگیری ماشین براساس نظرها و رتبه‌بندی‌های مشتریان ارائه شده است، که به کمک آن می‌توان با آگاهی از احساس‌های مشتریان خود، خدمات و راهبردهای بازاریابی مناسب را به‌کار گرفت. در نوشتار حاضر، با استفاده از مجموعه‌ی داده‌ی نظرهای پوشاک زنان در خریدهای الکترونیکی به بررسی رضایت یا نارضایتی از خرید مشتریان، شناسایی عوامل مؤثر در رفتار آن‌ها از طریق شناسایی احساس‌های متن نظرها، مدل‌سازی موضوعی براساس دسته‌بندی محصول‌ها به‌منظور شناسایی موضوع‌های کلیدی به همراه خوشه‌بندی مشتریان و درنهایت، ارائه‌ی مدل‌های کلاس‌بندی احساس‌ها با استفاده از روش‌های پردازش زبان طبیعی و یادگیری ماشین پرداخته شده است. درک احساس‌های نهفته در نظرهای مشتریان اهمیت فراوانی دارد و اطلاعات ارزشمندی در مورد رضایت مشتریان نسبت به محصول‌ها یا خدمات شرکت ارائه می‌دهد. احساس‌های مثبت، نشان‌دهنده‌ی خوشحالی و رضایت مشتریان و احساس‌های منفی، نشان‌دهنده‌ی زمینه‌های نیازمند بهبود هستند. با شناسایی احساس‌های مذکور، کسب‌وکارها می‌توانند راهبردهای خود را برای رسیدگی به نیازها و نگرانی‌های خاص مشتریان خود تنظیم کنند. به همین منظور هدف کلی نوشتار حاضر، ارائه‌ی روش‌ها و مدل‌هایی است که با ترکیب پردازش زبان طبیعی و یادگیری ماشین، اطلاعات و داده‌های شرکت‌های

استناد به این مقاله:

رحیمی، راضیه و قاسمی یقین، رضا، ۱۴۰۴. مدل‌سازی موضوعی و تحلیل احساس‌های مشتریان با استفاده از روش‌های پردازش زبان طبیعی و یادگیری ماشین. مهندسی صنایع و مدیریت شریف ، ۴۱(۲)، صص.۴۵-۶۱. https://doi.org/10.24200/j65.2025.65318.2418

شاپا الکترونیکی: ۲۶۷۶-۴۷۵X

پوشاک را به بینش‌های عمیق و عملی برای رشد و موفقیت آن‌ها تبدیل و از قدرت بازخوردهای مشتریان برای بهبود محصول‌ها، خدمات، و تجربه‌ی کلی مشتری از خرید خود استفاده کنند.

ساختار ادامه‌ی نوشتار به این صورت است، که در بخش ۲، به مرور ادبیات موضوع و بیان کارهای مرتبط؛ در بخش ۳، به ارائه‌ی روش‌شناسی و مبانی نظری؛ در بخش ۴، به ارزیابی مدل‌های پیاده‌سازی‌شده و ارائه‌ی نتایج به‌دست‌آمده؛ و در بخش ۵، به ارائه‌ی نتیجه‌گیری و پیشنهادهای آینده پرداخته شده است.

۲. مرور ادبیات موضوع

خلاصه‌ی کارهای مرتبط در جدول ۱ ارائه شده است؛ که مطابق آن، مطالعات پیشین به‌منظور بررسی نظرهای مشتریان پوشاک در خریدهای الکترونیک با مجموعه داده‌های متفاوت، متغیر هدف، و مدل‌های یادگیری ماشین متفاوت انجام شده‌اند. از آنجایی که مجموعه داده‌ی استفاده‌شده در نوشتار حاضر با عنوان نظرهای تجارت الکترونیک پوشاک زنانه با برخی نوشتارهای پیشین،[۳ و ۵-۱۴] یکسان بوده است؛ در ادامه، به بررسی بیشتر آن‌ها پرداخته شده است. مجموعه داده‌ی مذکور در مطالعات متعددی استفاده شده است، که بیشتر بر تحلیل احساس‌های مشتریان و پیش‌بینی رفتارهای آن‌ها متمرکز بوده‌اند. برخی از آن‌ها، شامل تحلیل اکتشافی داده‌ها و پیش‌بینی احساس‌ها با استفاده از مدل‌های یادگیری ماشین، مانند ماشین بردار پشتیبان و درخت تصمیم، همچنین مقایسه‌ی مدل‌های مختلف یادگیری ماشین برای تحلیل نظرهای مشتریان بوده‌اند. در مقایسه با پژوهش‌های ذکرشده، پژوهش حاضر با استفاده از مدل‌های مختلف یادگیری ماشین تلاش کرده است تا دقت مدل‌ها را در تحلیل و طبقه‌بندی نظرهای مشتریان بهبود بخشد. علاوه‌بر این، استفاده از روش‌های پردازش زبان طبیعی مانند TF-IDF و Word2Vec در استخراج ویژگی‌ها و مدل‌سازی موضوعی با روش تخصیص دیریکله‌ی پنهان، از ویژگی‌های خاص پژوهش حاضر است، که به دقت بیشتر در پیش‌بینی احساس‌ها و دسته‌بندی نظرها کمک می‌کند. لذا، با استفاده از ترکیب روش‌های مختلف و ارزیابی دقیق‌تر مدل‌ها، به نتایج متفاوتی نسبت به سایر مطالعات مشابه دست یافته است. در برخی مطالعات،[۳، ۹-۵، و ۱۱] متغیر هدف در نظر گرفته‌شده به‌منظور پیاده‌سازی مدل‌های یادگیری ماشین، ستون ویژگی شاخص پیشنهاد محصول در مجموعه‌ی داده و در حالت دوکلاسه (۰ برای پیشنهاد نکردن محصول توسط مشتری و ۱ برای پیشنهاد محصول توسط مشتری) انجام شده است. همچنین، در مطالعه‌های دیگری،[۱۲ و ۱۳] متغیر هدف در نظر گرفته‌شده، ستون ویژگی امتیاز محصول‌ها و در حالت دوکلاسه با مرز امتیاز ۳ (امتیازهای برابر ۳ و بالاتر به‌عنوان کلاس ۱ و امتیازهای کمتر از ۳ به‌عنوان کلاس ۰) صورت گرفته‌اند. موضوع قابل‌توجه این است که مجموعه داده‌ی استفاده‌شده در همه‌ی مطالعات یکسان بوده و همان‌طور که در ادامه با تجزیه‌وتحلیل داده‌های اکتشافی نشان داده شده است، مجموعه داده‌ی مذکور برای متغیرهای هدف شاخص، پیشنهاد محصول، و امتیاز مشتریان متعادل نبوده است. از آنجایی که 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 غیرقابل‌اعتمادبودن نتایج و معیارهای ارزیابی مدل‌های پیاده‌سازی‌شده، موضوع متعادل‌کردن مجموعه‌ی داده مهم است و باید به آن توجه کرد.

چهار مطالعه در میان کارهای مرتبط انجام‌شده، بحث متعادل‌کردن مجموعه‌ی داده را مطرح و بررسی کرده‌اند. کوبروسلی[۱] و همکاران (۲۰۲۲)،[۶] به‌منظور انجام این کار به‌صورت تصادفی، مجموعه‌های متعادل آموزش و آزمایش را انتخاب کرده و توضیح دقیق‌تری از روش انجام کار ارائه نداده‌اند. سونیل و شیرازی (۲۰۲۳)،[۸] به‌منظور متعادل‌کردن مجموعه‌ی داده از روش افزایش تعداد داده‌ها در کلاس اقلیت استفاده کرده‌اند، اما به روش انجام آن به‌صورت صریح اشاره‌ای نکرده‌اند. لوکیلی[۲] (۲۰۲۳)،[۹] در بخش تجزیه‌وتحلیل اکتشافی، به موضوع متعادل‌نبودن مجموعه‌ی داده با متغیر هدف شاخص پیشنهاد محصول پی برده است. با این حال در ادامه‌ی نوشتارش، توضیحی مبنی بر استفاده از روش‌هایی برای رفع نامتعادل‌بودن مجموعه‌ی داده در نظر گرفته نشده است. شتی[۳] و همکاران (۲۰۲۳)،[۱۱] به نامتعادل بودن مجموعه‌ی داده اشاره کرده‌اند؛ اما راهکاری به‌منظور برطرف‌کردن آن ارائه نکرده‌اند. با این حال، ایشان بیان کرده‌اند که به‌دلیل نامتعادل‌بودن مجموعه‌ی داده، معیار درستی قابل‌اعتماد نیست و بهتر است از معیارهای دقت و پوشش برای ارزیابی عملکرد مجموعه‌ی داده‌های نامتعادل استفاده شود.

در مطالعه‌ی حاضر، با درنظرگرفتن امتیاز محصولات به‌عنوان متغیر هدف و به‌کارگیری روش SMOTE برای رفع مشکل نامتعادل‌بودن مجموعه‌ی داده و به کمک روش‌های یادگیری ماشین و پردازش زبان طبیعی، به تجزیه‌وتحلیل احساس‌ها در حالت‌های دوکلاسه، سه‌کلاسه، و پنج‌کلاسه پرداخته شده است. روش SMOTE، با استفاده از الگوریتم k- نزدیکترین همسایه به تولید داده‌های مصنوعی جدید می‌پردازد،[۱۵] و در مقایسه با روش‌های متداول، افزایش تعداد داده‌ها در کلاس اقلیت مانند OverSampling، که ممکن است داده‌های تکراری زیادی به مجموعه اضافه شوند و نتایج را غیرقابل اعتماد کنند، انتخاب بهتری است. همچنین انجام نوشتار حاضر در حالت‌های دو کلاسه، سه‌کلاسه، و پنج‌کلاسه به بررسی دقیق‌تر و با جزئیات بیشتری به رفتار و احساس‌های مشتریان پرداخته است. علاوه‌بر موضوعاتی که اشاره شد، در نوشتار کنونی، مدل‌سازی موضوعی و برآورد موضوع‌های نهفته در متن نظرها با استفاده از روش تخصیص دیریکله‌ی پنهان نیز انجام شده است. نتایج حاصل از همه‌ی موارد ذکرشده، اطلاعات مهمی در رابطه با شناخت عوامل مؤثر در رفتار مشتریان و تحلیل احساس‌های آن‌ها در اختیار پژوهشگران قرار می‌دهد.

۳. روش‌شناسی

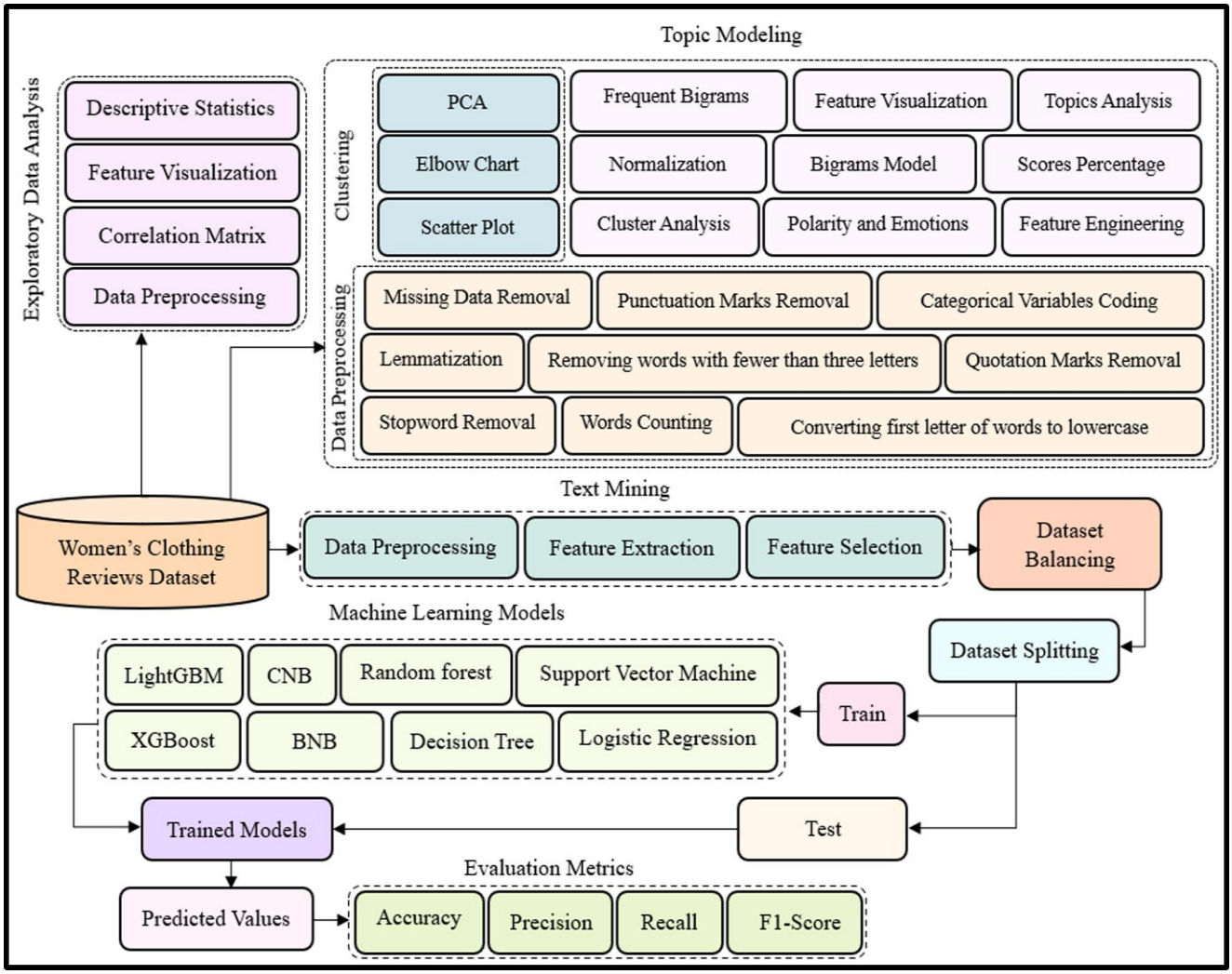
در شکل ۱، مراحل مطالعه‌ی حاضر مشاهده می‌شود؛ که مطابق آن، پس از انتخاب مجموعه‌ی داده‌ی مربوط به نظرهای پوشاک زنان، به انجام تمیزکردن داده‌ها و پیش‌پردازش‌هایی چون مدیریت مقادیر گمشده و حذف ستون مربوط به شماره‌ی ردیف داده‌ها پرداخته شده است.

در مرحله‌ی بعد، ابتدا به‌منظور شناخت و درک مجموعه‌ی داده به تجزیه‌وتحلیل اکتشافی پرداخته شده است، که در ادامه در مورد آن بحث شده است. در بخش بعد، پس از انجام پیش‌پردازش به انتخاب ویژگی[۴]، استخراج[۵]، و مهندسی ویژگی[۶] توسط الگوریتم TF-IDF، متعادل‌کردن مجموعه‌ی داده با استفاده از

[1] Kubrusly
[2] Loukili
[3] Shetty
[4] Feature Selection
[5] Feature Extraction
[6] Feature Engineering

جدول ۱. خلاصه‌ی کارهای مرتبط.

Title and publication year	Dataset	Target variable	Dataset balancing	ML models
Sentiment analysis of e-commerce customer reviews based on natural language processing 2020 [5]	Women's e-commerce clothing reviews	Recommended Index	-	**LR, SVM, RF, XGBoost, LightGBM**
A statistical analysis of textual e-commerce reviews using tree-based methods 2022 [6]	Women's e-commerce clothing reviews	Recommended Index	Random selection of balanced training (5,840) and test (2,504) Sets	**RF, XGBoost, DT, GB**
Analyzing online reviews of customers using machine learning techniques 2022 [7]	Women's e-commerce clothing reviews	Recommended Index	-	**NB, LR, SVM, NN**
Customer review classification using machine learning and deep learning techniques 2023 [8]	Women's e-commerce clothing reviews	Recommended Index	Enhancing the minority class sample size	**LR, Linear SVM, DT, RF, AdaBoost**
Explainable sentiment analysis for textile personalized marketing 2023 [3]	Women's e-commerce clothing reviews	Recommended Index	-	**LR**
Sentiment analysis of product reviews for e-commerce recommendation based on machine learning 2023 [9]	Women's e-commerce clothing reviews	Recommended Index	-	**LR, RF, KNN, CatBoost**
Exploring commonly used terms from online reviews in the fashion field to predict review helpfulness 2023 [16]	Customer reviews on amazon’s fashion platform	Positive feedback count	-	**RF, SVM, LR, SGB, TM, LDA**
It is about inclusion! Mining online reviews to understand the needs of adaptive clothing customers 2023 [17]	Customer reviews from Amazon, Silverts, IZ-Adaptive, and ResellerRatings	-	-	**TM, LDA**
Using Topic Modeling for Extracting Customers’ Expectations: a case of women apparel 2023 [10]	Women's e-commerce clothing reviews	-	-	**TM, LDA**
Sentiment forecasting in women’s fashion e-Commerce: a machine learning perspective 2024 [18]	Reviews on Women’s Apparel	Sentiment	-	**NB, SVM, NN, LR**
Hyperparameter optimization of machine learning models using grid search for amazon review sentiment analysis 2024 [11]	Women's e-commerce clothing reviews	Recommended Index	-	**DT, LR, SVM, NB, XGBoost, RF**
Sentiment exploring on feedback of e-commerce data using machine learning algorithms 2024 [12]	Women's e-commerce clothing reviews	Customer Ratings (Binarized with a Threshold of 3)	-	**LR, NB, MNB, BNB, SVM, RF, AdaBoost**
Factors influencing recommendations for women's clothing satisfaction: a latent dirichlet allocation approach using online reviews 2024 [14]	Women's e-commerce clothing reviews	-	-	**TP, LDA**
This study: Topic Modeling and Sentiment Analysis of Customers Using Natural Language Processing and Machine Learning Techniques	Women's e-commerce clothing reviews	Customer Ratings in Binary, Three-class, and Five-class Settings	SMOTE	**LR, SVM, DT, RF, CNB, MNB, XGBoost, LightGBM, TM, LDA, NLP**



شکل ۱. روش‌شناسی مطالعه‌ی حاضر.

روش SMOTE، و تفکیک آن به دو دسته‌ی مجموعه‌ی داده‌های آموزش و آزمایش صورت گرفته است. درنهایت، مدل‌های یادگیری ماشین موردنظر، پیاده‌سازی، و عملکرد آن‌ها ارزیابی شده است.

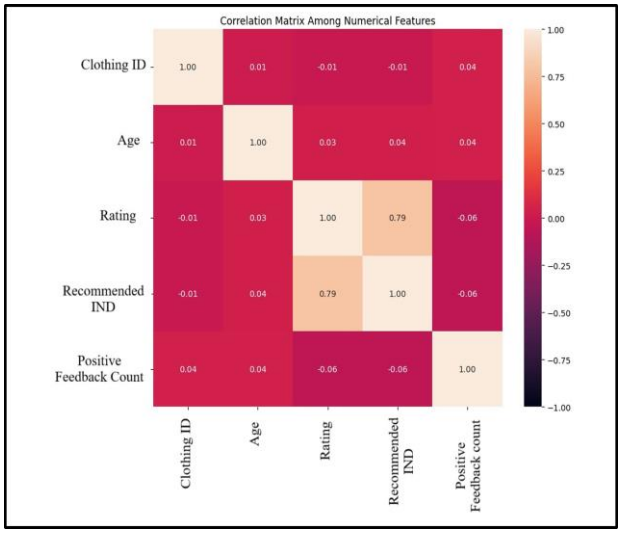
۳.۱. جمع‌آوری داده

مجموعه‌ی داده‌ی استفاده‌شده در مطالعه‌ی حاضر، شامل داده‌های تجاری واقعی و با عنوان نظرهای تجارت الکترونیک پوشاک زنانه[۷]، از وب‌سایت Kaggle دانلود شده است.[۱۹] مجموعه‌ی داده شامل ۲۳۴۸۶ ردیف و ۱۰ ستون ویژگی بوده است و هر ردیف آن بیان‌کننده‌ی شناسه‌ی قطعه‌ی در حال بررسی، سن مشتری صاحب‌نظر، 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[۸]، نام بخش محصول، و نام کلاس محصول است. ستون‌های ویژگی مجموعه‌ی داده‌های اخیر، شامل: متغیرهای عددی[۹]، رشته‌ای[۱۰]، عددی ترتیبی[۱۱]، عددی دودویی[۱۲]، و طبقه‌ای[۱۳] می‌شود.

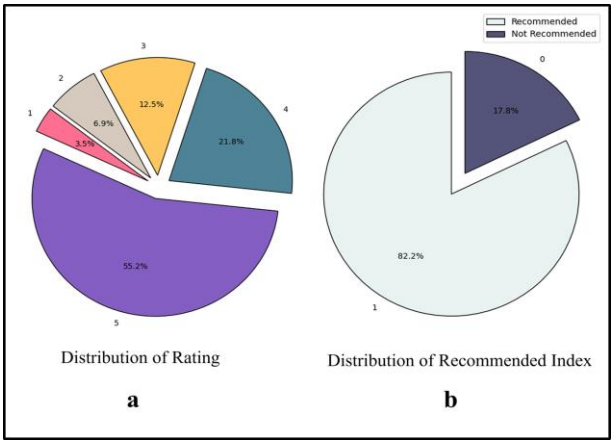
۳.۲. تجزیه‌وتحلیل داده‌های اکتشافی

به‌منظور انجام تجزیه‌وتحلیل داده‌های اکتشافی (EDA)[۱۴]، پس از بررسی و حذف مقادیر داده‌ی گمشده[۱۵]، ابتدا محاسبه‌ی پارامترهای آمار توصیفی برای ویژگی‌های عددی انجام و سپس ماتریس همبستگی ویژگی‌های مذکور ترسیم شده است. در ادامه، به تحلیل ویژگی‌ها پرداخته شده است. پس از حذف مقادیر داده‌های گمشده، تعداد ردیف‌های موردبررسی مجموعه‌ی داده برابر با ۱۹۶۶۲ در نظر گرفته شده است. میانگین گروه سنی افرادی که نظرهای خود را در رابطه با پوشاک خریداری‌شده‌ی زنان به اشتراک گذاشته‌اند، حدود ۴۳ سال و در بازه‌ی ۱۸ تا ۹۹ سال برآورد شده است. 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 داشته‌اند. تعداد مشتریانی که نظرهای ثبت‌شده را مفید و کمک‌کننده دانسته‌اند، نیز از ۰ تا ۱۲۲ عدد متغیر بوده است. در شکل ۲، ماتریس همبستگی ویژگی‌های عددی مشاهده می‌شود؛ که مطابق آن، بیشترین مقدار ضریب همبستگی مربوط به ویژگی‌های امتیاز و شاخص پیشنهاد محصول و برابر با ۰٫۷۹ برآورد شده است. شکل ۳ (a)، نحوه‌ی

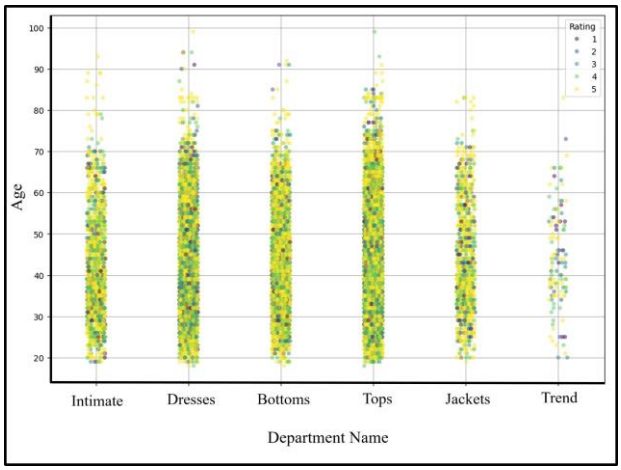
[7] Women's clothing e-commerce reviews
[8] Product high level division
[9] Numerical
[10] String
[11] Ordinal variable
[12] Binary variable
[13] Categorical
[14] Exploratory Data Analysis
[15] Missing values



شکل ۲. ماتریس همبستگی ویژگی‌های عددی.



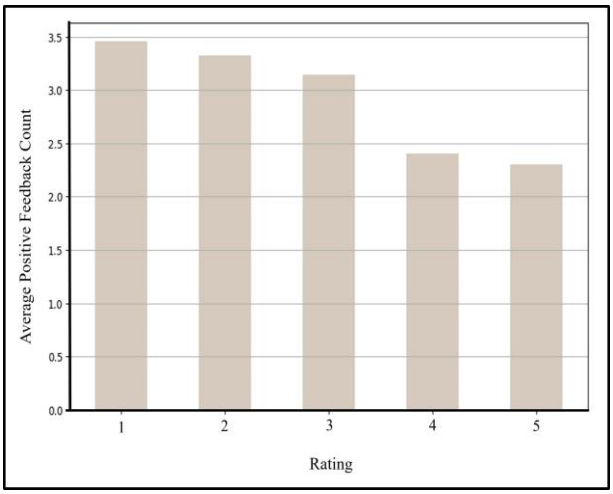
شکل ۳. (a) نحوه‌ی توزیع امتیازها، (b) توزیع شاخص پیشنهاد محصول.



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

توزیع امتیازهای موجود در مجموعه‌ی داده، شکل ۳ (b)، نحوه‌ی توزیع شاخص پیشنهاد محصول توسط مشتریان، و شکل ۴، امتیازهای مشتریان را با توجه به گروه سنی و نام بخش نشان می‌دهند.

همان‌طور که مشاهده می‌شود، ۸۲٫۲٪ از مشتریان، محصول خریداری‌شده را به دیگران پیشنهاد می‌کنند. امتیاز ۵ با مقدار ۵۵٫۲٪، بیشترین سهم را در



شکل ۵. میانگین تعداد مفیدبودن نظرها برای هر امتیاز.

تعداد کل مجموعه‌ی داده داشته و پس از آن به ترتیب امتیازهای ۴، ۳، ۲، و ۱ قرار گرفته‌اند. بیشتر امتیازها توسط افراد در محدوده‌ی سنی ۲۰ تا ۷۵ سال ثبت شده و بیشترین نظرها، امتیاز ۳ و بالاتر داشته‌اند. در شکل ۵، میانگین تعداد مفیدبودن نظرها برای هر امتیاز مشاهده می‌شود؛ که مطابق آن، بیشترین مقدار میانگین تعداد مفیدبودن نظرها، مربوط به امتیازهای ۱ تا ۳ بوده است.

۳.۳. مدل‌سازی موضوعی

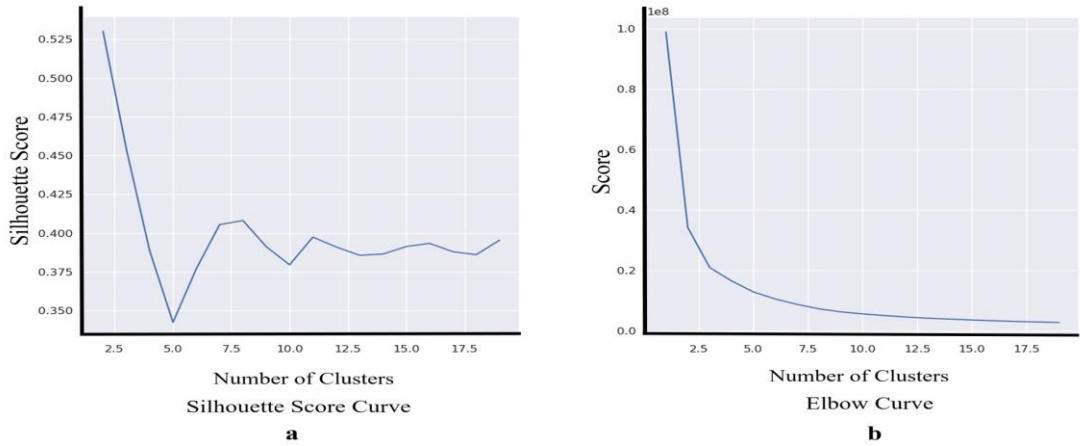
در بخش کنونی، ابتدا به بیان توضیح مختصری از روش تخصیص پنهان دیریکله (LDA)[۱۶] در مدل‌سازی موضوعی[۱۷]، خوشه‌بندی[۱۸]، و چرایی استفاده از دو روش مذکور در نوشتار حاضر پرداخته شده است. در ادامه، روش‌شناسی بخش حاضر، بیان و نتایج برآوردشده گزارش شده است.

- **تخصیص پنهان دیریکله**، شکلی از یادگیری ماشین بدون نظارت است، که برای کشف موضوع‌های پنهان از متن نظرها استفاده می‌شود و هدف آن، یافتن گروهی از موضوع‌هاست که کلمات مشاهده‌شده را در تمام اسناد به بهترین شکل توصیف کنند. اطلاعات تولیدشده از مدل ذکرشده، شامل واژگان کلیدی مرتبط با هر یک از موضوع‌ها و احتمال مرتبط‌شدن هر یک از نظرهای متنی با هر موضوع است.[۲۰]

- **خوشه‌بندی K-means**، از الگوریتم‌های یادگیری بدون نظارت است، که برای طبقه‌بندی مجموعه‌ای از داده‌ها به کلاس‌هایی از داده‌های مشابه استفاده می‌شود. طبقه‌بندی مجموعه‌ی داده‌ی N مورد براساس ویژگی‌ها به k زیرمجموعه‌ی مجزا با کمینه‌سازی فاصله بین آیتم داده و مرکز خوشه‌ی مرتبط انجام می‌شود.[۲۱]

در نوشتار حاضر، برای انجام مدل‌سازی موضوعی از روش تخصیص دیریکله‌ی پنهان و خوشه‌بندی K-Means به‌طور ترکیبی استفاده شده است. تخصیص دیریکله‌ی پنهان، به‌عنوان یک روش مبتنی بر احتمال، توانایی شبیه‌سازی توزیع موضوع‌های پنهان در اسناد مختلف را دارد و می‌تواند به‌طورمؤثری، موضوع‌های غالب در متن‌های مشتریان را شناسایی کند. روش تخصیص دیریکله‌ی پنهان به تحلیل ساختار معنایی داده‌ها کمک می‌کند و امکان استخراج ویژگی‌های دقیق موضوعی از متون نظرهای مشتریان را فراهم

[16] Latent Dirichlet Allocation
[17] Topic Modeling
[18] Clustering



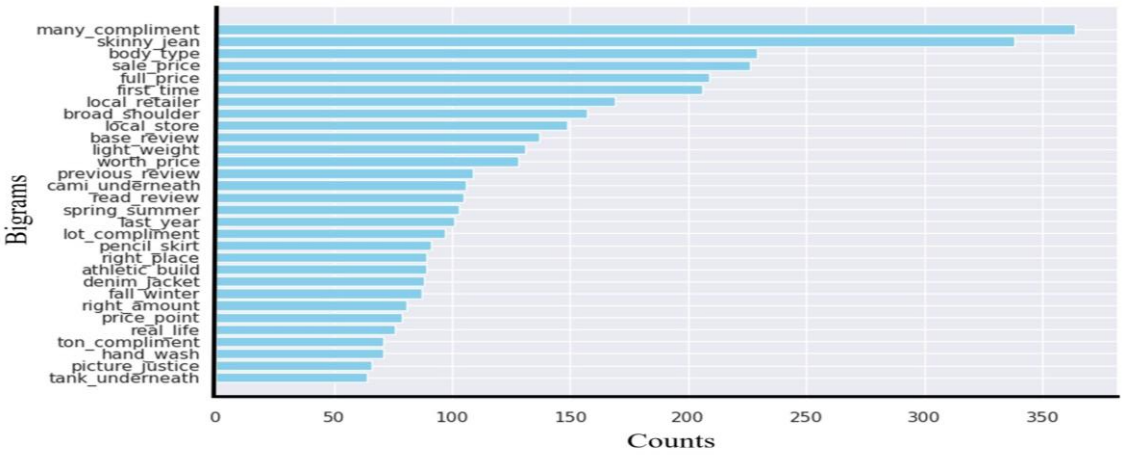
شکل ۶. (a) نمودار میانگین سیلوئت، (b) نمودار Elbow.

می‌آورد. به‌علاوه، به‌دلیل 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 موضوع‌های 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، براساس توزیع‌های موضوعی به‌دست‌آمده از تخصیص پنهان دیریکله استفاده شده است. ترکیب مذکور، داده‌ها را براساس شباهت‌های موضوعی به‌طورمؤثری گروه‌بندی می‌کند و تحلیل‌های دقیق‌تری را از نظرهای مشتریان امکان‌پذیر می‌سازد. علاوه‌بر این، در مطالعه‌ی حاضر، ارتباط موضوع‌های غالب در هر خوشه با سن مشتریان و تعداد کلمات در متن نظرهای مشتریان بررسی شده است. تحلیل‌های انجام‌شده نشان می‌دهند که چگونه سن و طول متن نظرها می‌تواند بر تمایل به صحبت در مورد موضوع‌های خاص تأثیر بگذارد و به درک بهتر الگوهای رفتاری و اولویت‌های مشتریان کمک کند.

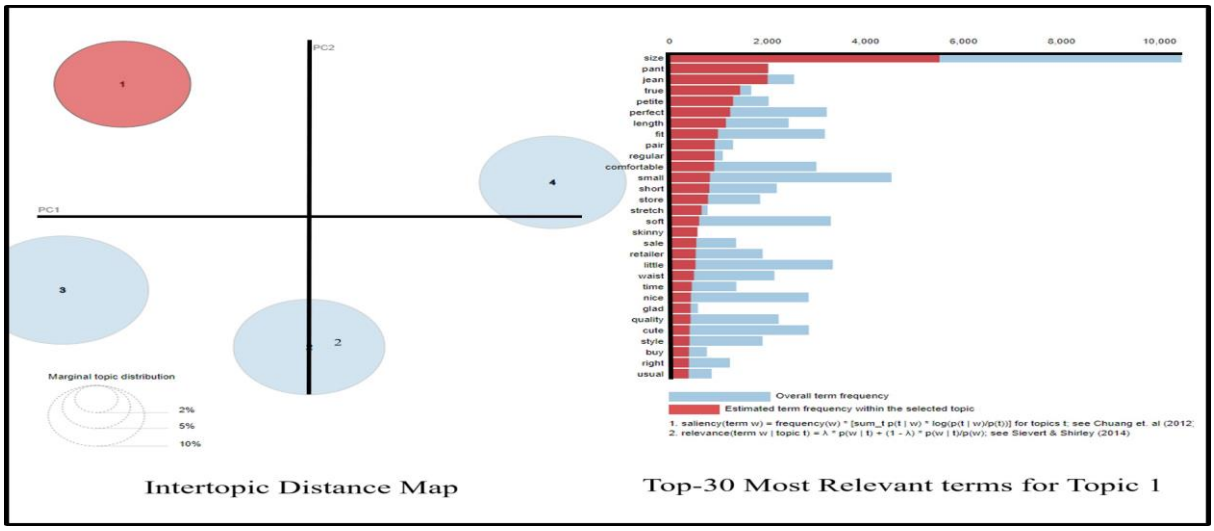
بررسی هر خوشه به‌طور ویژه مزایای زیادی از منظر راهبردهای بازاریابی و توسعه‌ی کسب و کار در بر دارد. با شناسایی ویژگی‌های خاص هر خوشه، کسب‌وکارها می‌توانند راهبردهای بازاریابی متناسب با نیازها و خواسته‌های خاص مشتریان در هر گروه را طراحی کنند. برای مثال، یک خوشه ممکن است تمایل زیادی به ویژگی‌های خاص محصول 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 تجربه‌ی مشتری علاقه‌مند باشد. به‌علاوه، تحلیل موضوع‌های موجود در هر خوشه می‌تواند به کسب‌وکارها کمک کند تا پیشنهادهای محصولات و خدمات را براساس اولویت‌های خاص مشتریان هر خوشه سفارشی کند و تجربه‌ی مشتری را بهبود بخشد. در نتیجه، روش مذکور نه فقط به تحلیل دقیق‌تر رفتار مشتریان کمک می‌کند، بلکه می‌تواند راهکارهای مؤثری برای رشد کسب‌وکار 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 با فراوانی کمتر از ۱٪ تعداد کل، شناسایی و شناسه‌ی یکسانی به آن‌ها نسبت داده شده است. سپس تعداد تکرار هر یک از شناسه‌های لباس، نرمال و در ستونی با نام شناسه‌ی لباس جدید به مجموعه‌ی داده اضافه شده است. در مرحله‌ی بعد، میزان قطبیت و احساس‌های متن نظرها به همراه عنوان آن‌ها، برآورد و در ستونی با عنوان امتیاز احساس‌های کل به مجموعه‌ی داده افزوده شده است. به‌منظور بهبود عملکرد الگوریتم خوشه‌بندی K-Means، کاهش ابعاد و رسم مجموعه‌ی داده، داده‌ها با استفاده از تجزیه‌وتحلیل مؤلفه‌های اصلی (PCA)[۱۹] و سپس تحلیل T-SNE[۲۰] به ابعاد پایین‌تر کاهش یافته و به‌عنوان ورودی الگوریتم K-Means در نظر گرفته شده‌اند. T-SNE یک روش کاهش ابعاد غیرخطی است، که برای نمایش ساختارهای پیچیده‌ی داده‌ها در فضای دو یا سه بُعدی بسیار مؤثر است و امکان خوشه‌بندی مؤثرتر داده‌های پیچیده را با حفظ روابط غیرخطی میان نقاط داده فراهم می‌کند.[۲۳] پس از آن، با استفاده از نمودار Elbow و معیار سیلوئت، تعداد بهینه‌ی خوشه‌ها برآورد شده است. روش elbow برای تعیین تعداد بهینه‌ی خوشه‌ها در الگوریتم K-Means براساس معیار مجموع مربعات خطا (SSE)[۲۱] عمل می‌کند. معیار SSE، 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 کرده است، به‌عنوان مقدار بهینه‌ی تعداد خوشه‌ها انتخاب شده است. کاهش SSE به این معناست که داده‌ها به خوشه‌های مجزا و منسجم‌تری تقسیم می‌شوند.[۲۴] میانگین سیلوئت نیز به‌منظور ارزیابی کیفیت خوشه‌بندی داده‌ها استفاده شده است، که نشانگر انسجام درون‌خوشه‌ای و جداشدگی بین خوشه‌هاست؛ که برای محاسبه‌ی آن، فاصله‌ی هر نقطه از نقاط دیگر در خوشه‌ی خود و نزدیک‌ترین خوشه محاسبه می‌شود.[۲۵] در مطالعه‌ی حاضر، مقدار سیلوئت برای تعداد دو خوشه برابر با ۰٫۵۳ و برای تعداد سه خوشه برابر با ۰٫۴۵ به‌دست آمده است. این تفاوت نشان می‌دهد که هر دو تعداد خوشه (دو و سه خوشه) کیفیت مشابهی در خوشه‌بندی داده‌ها داشته‌اند. از طرفی، با توجه به نمودار elbow، تعداد سه خوشه به‌عنوان تعداد بهینه‌ی خوشه‌ها با کمترین مجموع مربعات خطا و تقسیم‌بندی بهتر داده‌ها شناسایی شده است. با توجه به اینکه نتایج روش elbow و امتیاز سیلوئت به‌طورکلی هم‌راستا هستند و تعداد سه خوشه، هم‌زمان کیفیت بالاتری در خوشه‌بندی و انسجام بهتر داده‌ها ارائه می‌دهد، تعداد سه خوشه به‌عنوان تعداد بهینه‌ی خوشه‌ها در نظر گرفته شده است. درنهایت، عملیات خوشه‌بندی با استفاده از الگوریتم K-means انجام شده است. در شکل ۶ (a)، نمودار میانگین سیلوئت و در شکل

[19] Principal component analysis
[20] T-distributed Stochastic Neighbor Embedding
[21] Sum of Squared Errors



شکل ۷. سی بایگرام پرتکرار.



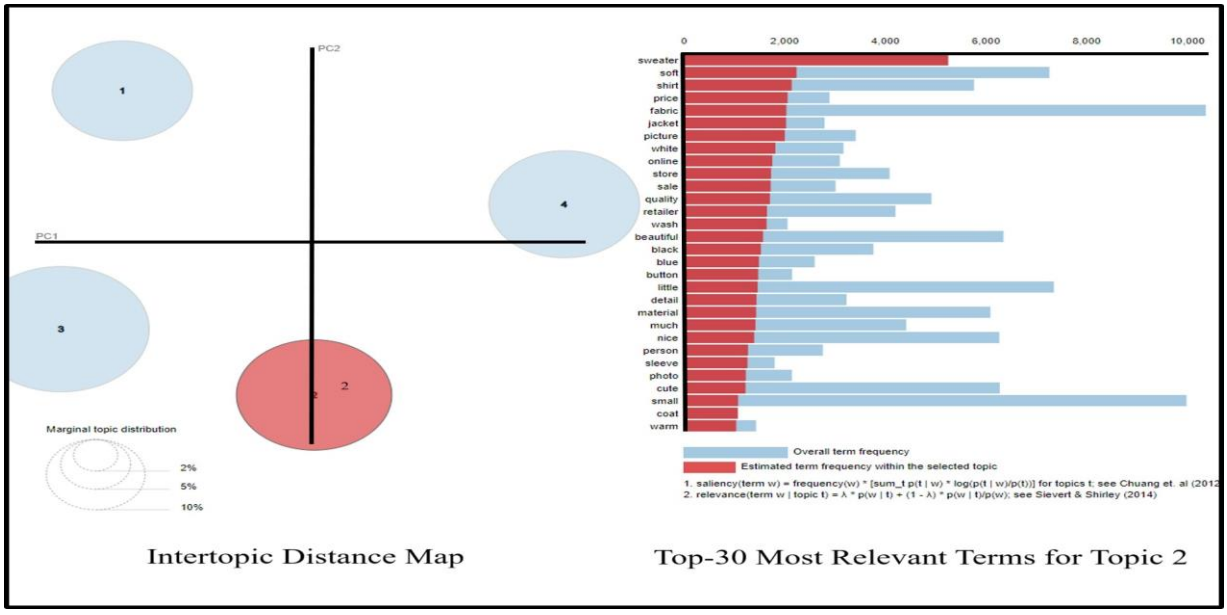
شکل ۸. نمودار فاصله‌ی بین موضوعی و کلمات پرتکرار موضوع یک.

۶ (b)، نمودار Elbow برای محاسبه‌ی تعداد خوشه‌های بهینه مشاهده می‌شوند.

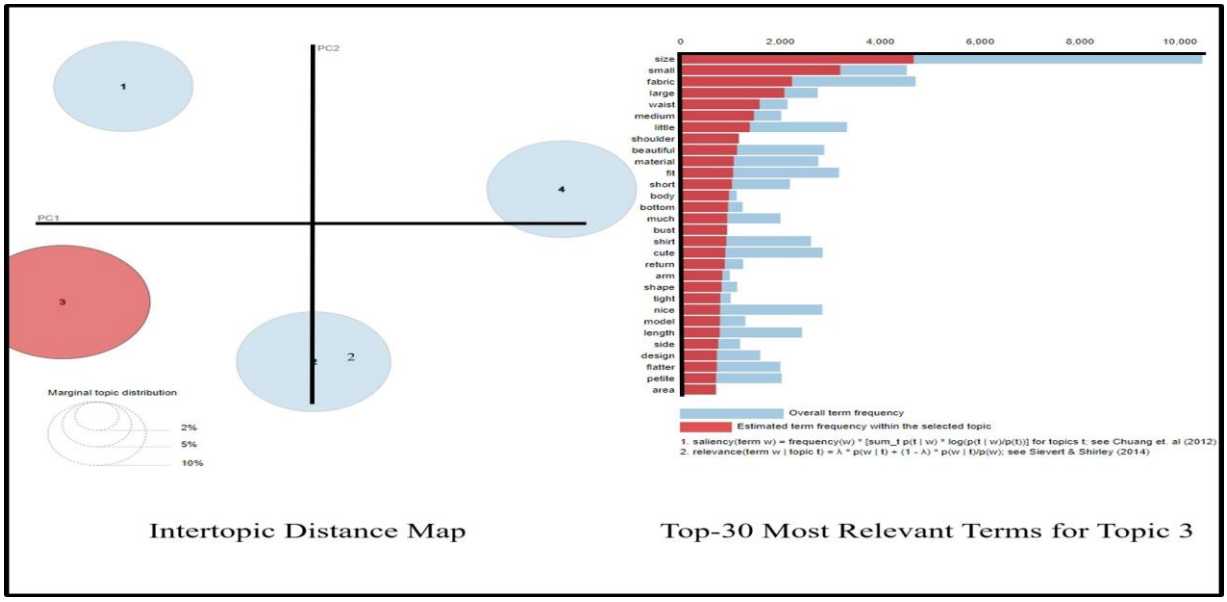
در مرحله‌ی بعد، برای اطمینان حاصل‌کردن از اینکه کلمات به‌درستی براساس بخش گفتارشان[۲۲] به‌صورت کلمه بیان می‌شوند، با استفاده از یک تابع، اسم، صفت، قید، و یا فعل بودنشان مشخص شده است. پس از آن با حذف علامت‌های نگارشی، کلمات توقف و کلمات دارای کمتر از سه حرف، تبدیل حرف اول کلمات به حرف کوچک و ریشه‌یابی[۲۳] تمیزسازی داده‌ها انجام شده است. سپس با استفاده از کلمات اسم و صفت به ساخت مدل بایگرام[۲۴] به کمک روش Word2vec (برای استخراج ویژگی) و برآورد ۳۰ بایگرام پرتکرار پرداخته شده است. بایگرام‌های مذکور در شکل ۷ مشاهده می‌شوند؛ که مطابق آن، عبارت‌هایی مانند تمجید فراوان نشان می‌دهد 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 نحوه‌ی شست‌وشو از عبارت‌های پرتکرار و در نتیجه از موضوعاتی هستند که باید به آن‌ها توجه کرد. دسته‌بندی لباس‌ها، مانند: شلوار جین، دامن، ژاکت، و غیره، تناسب و اندازه‌ی لباس‌ها و همچنین فصل‌های سال نیز از موضوعات مهم در نظر گرفته شده‌اند. در شکل ۸، نمودار فاصله بین موضوعی و کلمات پرتکرار مربوط به موضوع یک؛ در شکل ۹، نمودار بین موضوعی و کلمات پرتکرار مربوط به موضوع دو؛ در شکل ۱۰، نمودار فاصله بین موضوعی و کلمات پرتکرار موضوع سه؛ و در شکل ۱۱، نمودار فاصله بین موضوعی و کلمات پرتکرار مربوط به موضوع چهار مشاهده می‌شوند. همچنین، در جدول ۲، بررسی هر یک از موضوع‌ها و کلمات مرتبط با آن‌ها به صورت مختصر ارائه شده است. تحلیل‌های صورت‌گرفته، درک جامعی از موضوع‌های 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 جدول اخیر، ترتیب دسته‌بندی هر یک از موضوع‌ها براساس 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 هستند. پس از آن، تخصیص دیریکله‌ی پنهان برای حالت‌های ۲ تا ۲۳ موضوعی آزمایش و با توجه به معیار coherence، بهترین تعداد موضوع‌ها، چهار در نظر گرفته شده است.

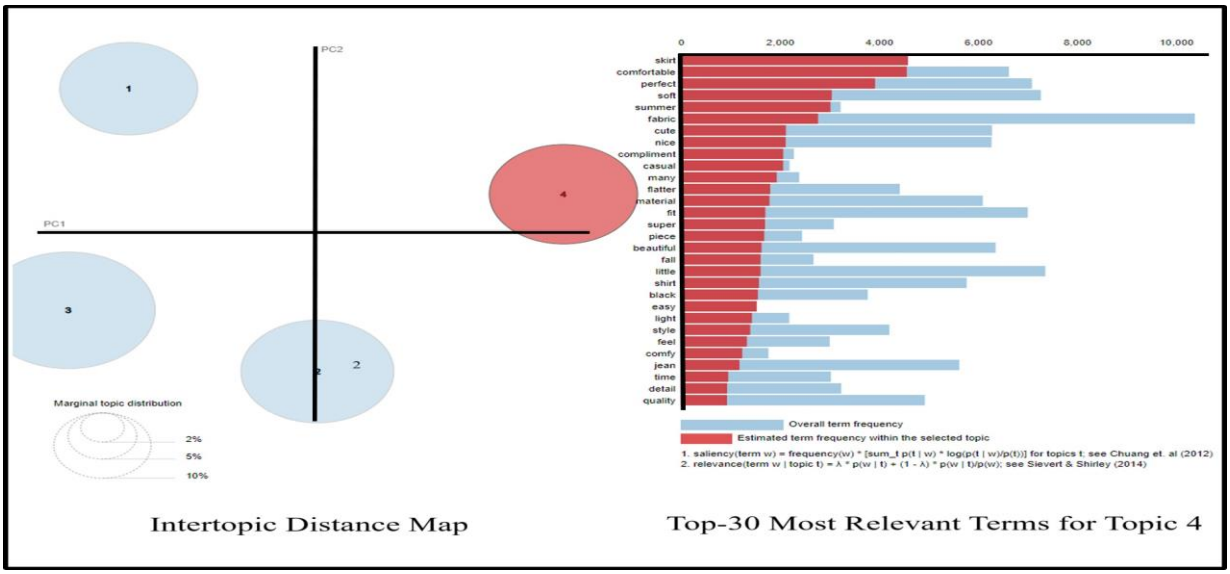
[22] Part-of-speech
[23] Lemmatization
[24] Bigram



شکل ۹. نمودار فاصله‌ی بین موضوعی و کلمات پرتکرار موضوع دو.



شکل ۱۰. نمودار فاصله‌ی بین موضوعی و کلمات پرتکرار موضوع سه.



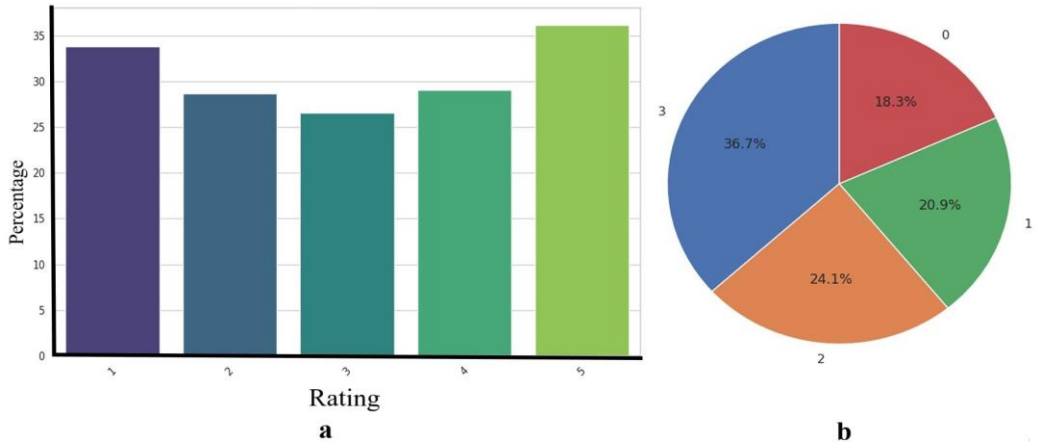
شکل ۱۱. نمودار فاصله‌ی بین موضوعی و کلمات پرتکرار موضوع چهار.

جدول ۲. تجزیه‌وتحلیل نتایج مدل‌سازی موضوعی.

	Items		Frequent words	Analysis
Topics	Topic one (focusing on pants and jeans)	Fit	"size," "petite," "regular," "small," "waist," "short," "usual"	Phrases such as size, petite, regular, small, waist, short, and usual emphasize the importance of fit. Buyers frequently mention the fit of pants or jeans, the need for various sizes (e.g., small), and achieving a perfect fit.
		Comfort	"comfortable," "stretch","soft", "skinny"	Words such as comfortable, stretch, and soft indicate that buyers prioritize the feel of the clothing.
		Style	"pair," "perfect," "nice," "cute," "style," "quality"	Words such as pair, perfect, nice, cute, style, and quality highlight the importance of the appearance and overall style of the pants and jeans.
		Shopping experience	"store," "sale," and "retailer"	Phrases such as stores, sale, and retailer indicate that the shopping experience, including the location and method of purchasing items, plays a significant role in customer satisfaction.
		Occasions and applications	"time," "buy," "right,", "glad"	Words such as time, buy, right, and glad indicate that customers evaluate these items based on their suitability for various occasions and their overall satisfaction with their purchase.
	Topic two (focusing on jacket and shirt)	Fabric and comfort	"soft," "fabric," "material," "warm," "nice"	Words such as soft, fabric, material, warm, and nice indicate that customers value comfortable and high-quality materials in jackets and shirts.
		Style	"beautiful," "cute," "black," "blue," "button," and "detail"	Phrases such as beautiful, cute, black, blue, button, and details indicate that visual appeal and design details are very important to customers.
		Price and value	"price," "sale," and "retailer"	Words such as price, sale, and retailer indicate that customers consider the cost and value of money when evaluating these items. Finding good deals and feeling that the item is worth the price are also important factors.
		Fit	"small" and "little"	The use of words such as small and little indicates that fit and size are important considerations. Customers often discuss whether items meet their expectations and whether the sizing is accurate.
		Shopping experience	"store," "online," "photo," and "picture"	Phrases such as store, online, photo, and picture highlight the importance of the shopping experience, both in-store and online. Customers discuss how items appear in photos versus in person and their overall shopping experience.
	Topic three (focusing on shirts and tops)	Fit	"size," "small," "large," "medium," "shoulder," "waist," "body," "bust," and "arm"	Words such as size, small, large, medium, shoulder, waist, body, bust, and arm indicate a strong emphasis on how shirts fit different parts of the body and concerns regarding overall sizing.
		Fabric and material	"fabric," "material," and "soft"	Phrases such as fabric, material, and soft highlight the importance of fabric quality and comfort. Customers often comment on their feelings about the materials and their overall quality.
		Style	"beautiful," "cute," "nice," "design," "flatter," "shape," and "model"	Words such as beautiful, cute, nice, design, flatter, shape, and model indicate the importance of appearance and style. Customers consider how shirts look and whether they are flattering to their body shape.
		Occasions and applications	"return", "design"	The word return indicates that some reviews might discuss the return process due to fit or quality issues. Words such as design suggest that customers consider shirts for various occasions.
		Details	"short," "bottom," "tight," "length," "side," "area"	Terms like short, bottom, tight, length, side, and area highlight the significance of specific details regarding the fit and design of shirts to customers. These terms frequently appear in customer feedback, underscoring the importance of particular shirt attributes in influencing customer satisfaction.
	Topic four (focusing on skirts and casual	Fit and comfort	"comfortable," "soft," "fit," "comfy," "perfect"	Terms like comfortable, soft, fit, comfy, and perfect suggest that customers prioritize the feel and fit of skirts. The focus is on selecting items that offer both comfort and a proper fit.
		Style	“cute,” “nice,” “beautiful,” “flatter,” “style,” “detail,” and “quality”	Words such as cute, nice, beautiful, flatter, details, and quality underscore the significance of appearance and style. Customers value skirts that are visually appealing and feature attractive design elements.
		Occasions and applications	"casual," "summer," "fall," "piece," "easy,"	Phrases such as casual, summer, fall, piece, and easy suggest that customers seek versatile skirts suitable for different seasons and occasions. The ability to pair these skirts with various tops, such as shirts and blouses, is also a key consideration.
		Fabric and material	"fabric," "material," "light," and "feel"	Words such as fabric, material, light, and feel highlight the importance of fabric quality and the tactile experience of skirts. Customers frequently remark on the suitability of materials for various weather conditions and their overall quality.
		Shopping experience and satisfication	"compliment," "time," and "glad"	Words such as compliments, time, and glad suggest that customer satisfaction is significantly influenced by the feedback they receive when wearing skirts. Positive experiences and compliments from others enhance their overall satisfaction with the purchase.

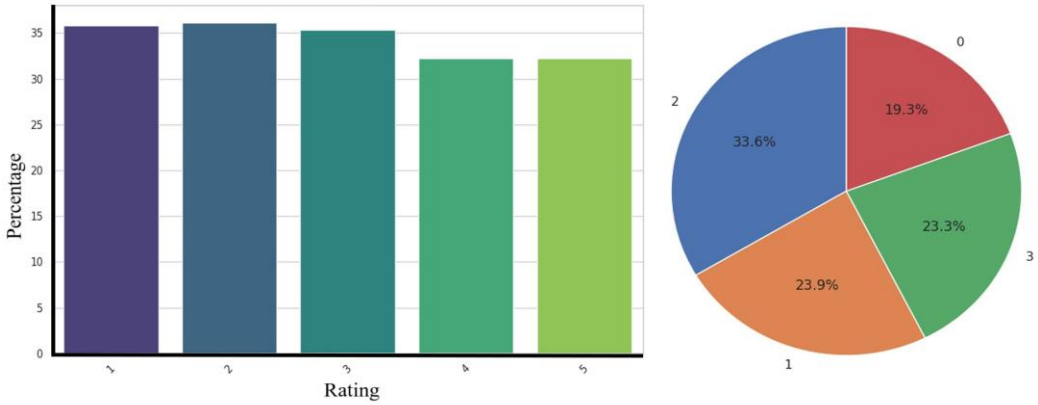
درنهایت، تجزیه‌وتحلیل مربوط به هر خوشه با توجه به درصد امتیازها، مصورسازی ویژگی‌های عددی، و موضوع‌های اصلی در آن خوشه انجام شده است. در شکل‌های ۱۲ الی ۱۴، به ترتیب درصد موضوع‌ها و امتیازها در خوشه‌های یک، دو، و سه مشاهده می‌شوند.

در شکل ۱۵، نحوه‌ی توزیع برخی از ویژگی‌های عددی در هر سه خوشه مشاهده می‌شود. با توجه به نمودارهای گزارش‌شده، خوشه‌ی یک، شامل زنان جوان و میانسال با گروه سنی ۲۵-۴۰ است، که نظرهای نسبتاً دقیقی (۱۰-۵۰ کلمه) نوشته‌اند و حدود ۴۷٪ از نظرهای آن‌ها مربوط به موضوع ۳ و متمرکز بر تناسب و راحتی، پارچه و جنس، ظاهر و استایل، و جزئیات پیراهن و تاپ است. خوشه‌ی دو، دربرگیرنده‌ی نظرهای زنان میانسال و سالمند با گروه سنی ۶۰-۳۵ سال است؛ که نظرهای آن‌ها به‌صورت توصیفی با نزدیک به ۴۰-۹۰ کلمه و تقریباً ۳۳٪ 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. خوشه‌ی سه، شامل نظرهای همه‌ی گروه‌های سنی از ۲۵ تا ۶۰ سال است، که نظرهای بسیار دقیق با تعداد ۸۰-۱۱۰ کلمه نوشته‌اند. نزدیک به ۴۲٪ از نظرهای اخیر مربوط به موضوع ۲ است.



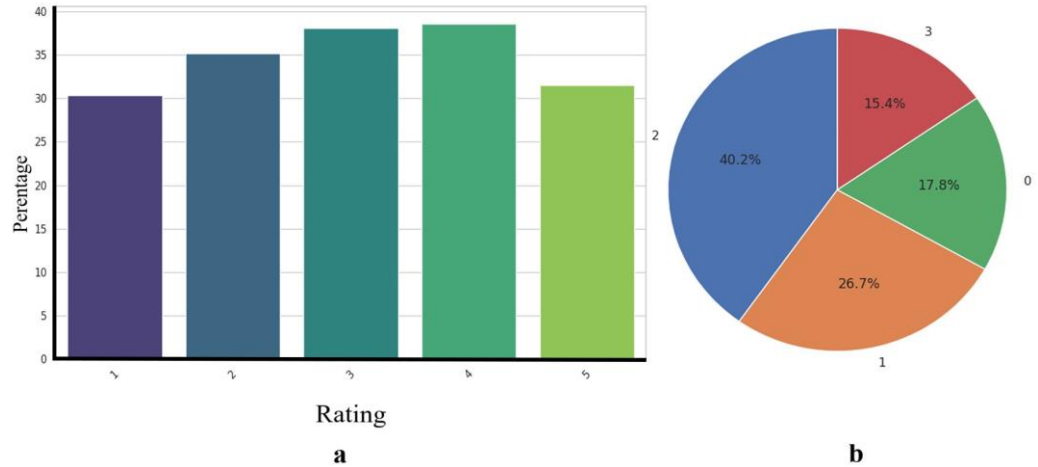
a b

شکل ۱۲. (a) درصد امتیازها، (b) درصد موضوع‌ها (در خوشه‌ی یک).



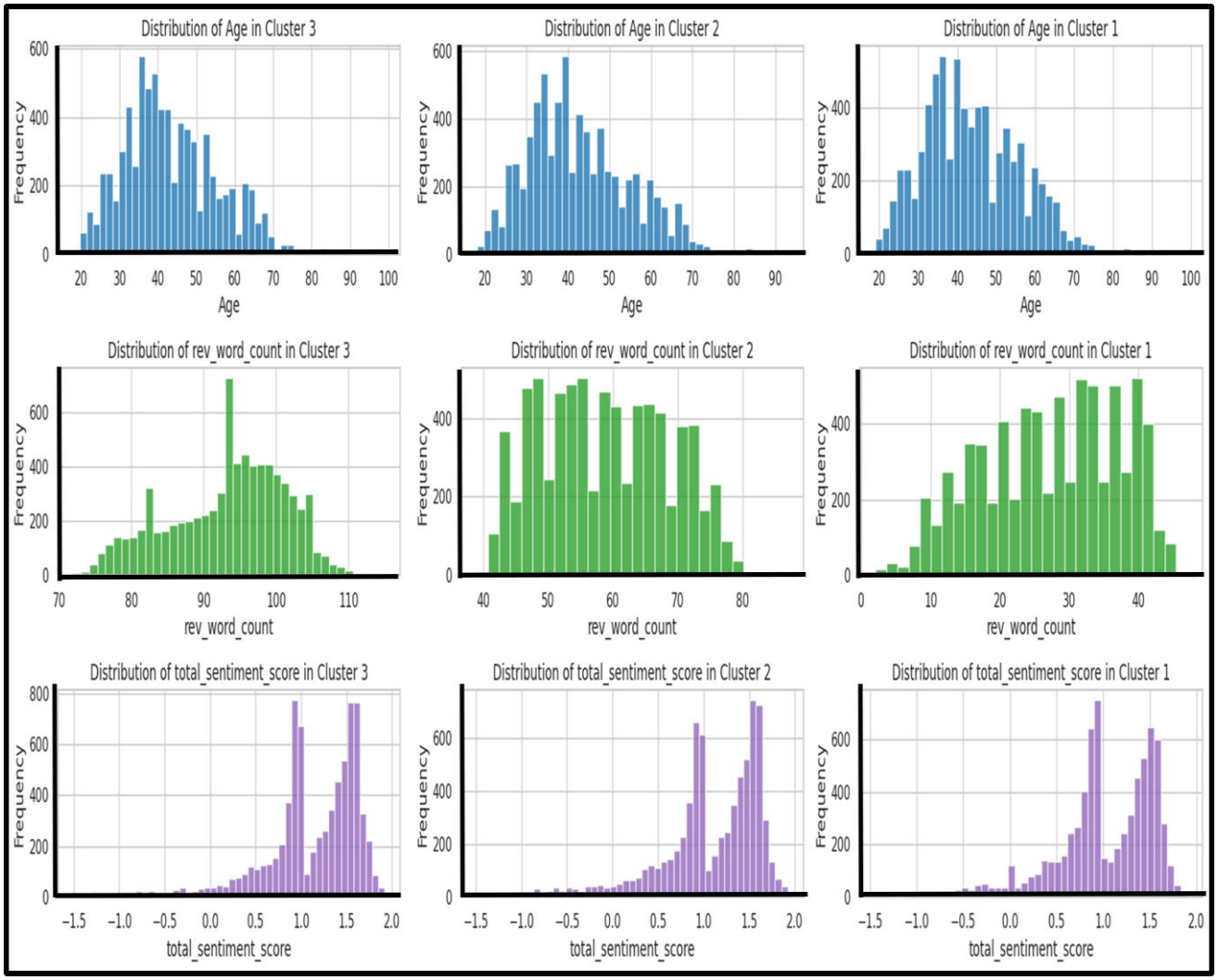
a b

شکل ۱۳. (a) درصد امتیازها، (b) درصد موضوع‌ها (در خوشه‌ی دو).



a b

شکل ۱۴. (a) درصد موضوع‌ها، (b) درصد امتیازها (در خوشه‌ی سه).



شکل ۱۵. نحوه‌ی 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 درستی به‌دست‌آمده از اعمال مدل‌های یادگیری ماشین بر مجموعه‌ی داده، تصمیم گرفته شد که به‌منظور انجام مطالعه‌ی حاضر از ویژگی متن نظرها و امتیاز استفاده شود.

۳.۵. پیش‌پردازش داده‌ها و استخراج ویژگی

در مطالعه‌ی حاضر، به‌منظور انجام پیش‌پردازش و استخراج ویژگی از الگوریتم TFIDF به همراه حذف کلمات توقف استفاده شده است؛ که در ادامه، به بیان توضیح مختصری از آن‌ها پرداخته شده است.

- **کلمات توقف:** رایج‌ترین اصطلاح‌ها در یک زبان که ارزش معنایی ندارند و فقط به درک کلی متن کمک می‌کنند، را کلمات توقف می‌گویند؛ 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 محاسبه‌ی وزن نسبت داده‌شده به هر عبارت موجود در سند، ابزار TfidfVetorizer از کتابخانه‌ی sklearn استفاده می‌شود، که برای انجام عملیات اخیر از الگوریتم TF-IDF استفاده می‌شود؛ و درنهایت، کلمات موجود در متن نظر را به یک ماتریس برداری تبدیل می‌کند.[۸ و ۱۱]

۳.۶. متعادل‌کردن مجموعه‌ی داده

همان‌طور که در قسمت تجزیه‌وتحلیل داده‌های اکتشافی در بخش بررسی ویژگی امتیاز مشاهده شد، مجموعه‌ی داده دارای توزیع نامتعادل کلاس امتیازهاست. بنابراین، به‌منظور اعمال مدل‌های یادگیری ماشین، ابتدا باید مجموعه‌ی داده متعادل شود. از آنجایی که در روش‌های معمول، افزایش داده‌های اقلیت، مانند OverSampling، ممکن است داده‌های تکراری زیادی به مجموعه اضافه و نتایج را غیرقابل‌اعتماد کنند، روش انتخابی برای متعادل‌کردن مجموعه‌ی داده در مطالعه‌ی حاضر، روش SMOTE[۲۵] انتخاب شده است؛ که با استفاده از الگوریتم k- نزدیک‌ترین همسایه به‌منظور متعادل‌کردن مجموعه‌ی داده، نمونه‌های مصنوعی جدید تولید می‌کند.[۱۵] از آنجایی که متغیر هدف امتیاز معمولاً پنج کلاس (امتیاز ۱ تا ۵) دارد، به‌منظور انجام مطالعه و اجرای مدل‌ها در حالت‌های دوکلاسه و سه‌کلاسه، امتیازها دسته‌بندی شده‌اند. در حالت دوکلاسه، امتیازهای قبل از ۳ به‌عنوان کلاس ۰ و امتیاز ۳ و بالاتر به‌عنوان کلاس ۱ تعریف شده‌اند. در حالت سه‌کلاسه، امتیازهای قبل از

[25] Synthetic Minority Oversampling Technique

۳ به‌عنوان کلاس ۰، امتیاز ۳ به‌عنوان کلاس ۱، و امتیازهای بالاتر از آن به‌عنوان کلاس ۲ در نظر گرفته شده‌اند. در مرحله‌ی بعد، به اجرای مدل‌های یادگیری ماشین پرداخته شده است. پس از بررسی کارهای مرتبط و مطالعات پیشین و همچنین اعمال برخی از مدل‌های یادگیری ماشین و با توجه به درستی حاصل، مدل‌های یادگیری ماشین در نظر گرفته‌شده برای مطالعه‌ی حاضر عبارت از: رگرسیون لجستیک، ماشین بردار پشتیبان، جنگل تصادفی، LightGBM، درخت تصمیم، XGBoost، بیز ساده‌ی چندجمله‌ای، و بیز ساده‌ی مکمل (CNB)[۲۶] بوده‌اند، که در ادامه، به بیان توضیح مختصری از آن‌ها پرداخته شده است.

۳.۷. مدل‌های یادگیری ماشین

- **رگرسیون لجستیک،** از مدل‌های یادگیری ماشین تحت‌نظارت است، که در کلاس‌بندی، تحلیل، و پیش‌بینی با استفاده از مشاهده‌های قبلی یک مجموعه‌ی داده استفاده می‌شود. رگرسیون لجستیک، از یک تابع لجستیک به نام تابع سیگموئید که احتمال رخ‌دادن متغیر وابسته را بین ۰ و ۱ محدود می‌کند، برای تخمین احتمال‌ها استفاده می‌کند.[۱۲ و۲۶]
- **ماشین بردار پشتیبان،** از الگوریتم‌های یادگیری ماشین تحت‌نظارت به‌منظور انجام وظایف کلاس‌بندی است و هدف آن یافتن ابرصفحه‌ای است که بیشترین حاشیه بین کلاس‌های داده‌های آموزشی را داشته باشد.[۵ و ۲۶]
- **درخت تصمیم،** روش یادگیری ماشین تحت‌نظارت است، که در مسائل رگرسیون و همچنین کلاس‌بندی استفاده می‌شود و با خلاصه‌کردن ویژگی-های داده از داده‌های آموزشی به پیش‌بینی مقدار متغیر هدف کمک می-کند.[۲۶ و ۲۷]
- **جنگل تصادفی،** از روش‌های کلاس‌بندی مبتنی بر یادگیری گروهی[۲۷] است، که با استفاده از روش گروه‌بندی موازی[۲۸] و ترکیب چندین درخت تصمیم و انتخاب تصادفی داده‌ها و ویژگی‌ها، عملکرد مدل را بهبود می‌بخشد. بنابراین، مشکل بیش‌برازش را به میزان کمینه می‌رساند و درستی پیش‌بینی را افزایش می‌دهد.[۱۲ و ۲۶]
- **LightGBM،**[29] الگوریتمی مبتنی بر یادگیری گروهی و درخت‌های تصمیم است، که با استفاده از چارچوب GBDT[30] طراحی شده است. این مدل با استفاده از روش‌های نمونه‌برداری یک‌طرفه‌ی مبتنی بر گرادیان (GOSS)[۳۱] و دسته‌بندی ویژگی‌های انحصاری (EFB)[۳۲]، در حالی که زمان آموزش و استفاده از حافظه را کاهش می‌دهد، عملکرد مدل را بهبود می‌بخشد.[۵ و ۲۶]
- **بیز ساده‌ی چندجمله‌ای و مکمل:** بیز ساده از الگوریتم‌های یادگیری ماشین تحت‌نظارت مبتنی بر احتمال و قضیه‌ی 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.[۲۶ و ۲۸] بیز ساده‌ی چندجمله‌ای براساس تعداد دفعاتی که یک کلمه در یک سند ظاهر می‌شود، عمل می‌کند. الگوریتم بیز ساده‌ی مکمل، نیز اقتباسی از بیز ساده‌ی چندجمله‌ای است، که به جای محاسبه‌ی احتمال یک نمونه متعلق به یک کلاس خاص، احتمال تعلق نمونه به همه‌ی کلاس‌ها را محاسبه می‌کند.[۱۲ و ۲۹]
- **XGBoost** [۳۳] (تقویت گرادیان شدید)، 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 براساس یک سری مدل‌های جداگانه، معمولاً درخت‌های تصمیم، تولید می‌کند.[۲۶ و ۲۷]

۳.۸. معیارهای ارزیابی

معیارهای ارزیابی استفاده‌شده در مطالعه‌ی حاضر عبارت‌اند از: درستی، دقت، پوشش، و امتیاز F1، که خلاصه‌ای از نحوه‌ی محاسبه و توضیحات آن‌ها در جدول ۳ ارائه شده است.

۴. ارزیابی و نتایج

در بخش کنونی، به بیان معیارهای ارزیابی برآوردشده توسط مدل‌های یادگیری

جدول ۳. معیارهای ارزیابی و نحوه‌ی محاسبه.[۶، ۹، و ۱۱]

Name	Formula	Definition
Recall	$TPR = \frac{TP + TN}{FN + TP}$	**The proportion of true positive samples that are correctly classified as positive**
Accuracy	$A = \frac{TP + TN}{TN + FP + FN + TP}$	**The proportion of correct predictions among the total number of predictions**
Precision	$P = \frac{TP}{FP + TP}$	**The proportion of predicted positive instances that are actually positive**
F1-Score	$F1 = 2\frac{P \times TPR}{P + TPR}$	**To measure the harmonic mean of precision and recall, providing a balanced evaluation by considering both false positives and false negatives**

[26] Complement Naive Bayes
[27] Ensemble learning
[28] Parallel ensembling
[29] Light Gradient Boosting Machine
[30] Gradient Boosted Decision Trees
[31] Gradient-based One-Side Sampling
[32] Exclusive Feature Bundling
[33] Extreme Gradient Boosting (XGBoost)

جدول ۴. مقایسه‌ی نتایج حالت دوکلاسه با کارهای مشابه.

		shetty, et.al (2023)				This study			
		Evaluation Metrics							
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Machine Learning Models	Support Vector Machine	0.94	0.94	0.99	0.96	0.95	0.96	0.95	0.95
	Logistic regression	0.93	0.94	0.99	0.97	0.90	0.91	0.92	0.91
	Decision Tree	-	-	-	-	0.91	0.92	0.92	0.92
	Random forest	0.90	0.90	1.00	0.95	0.98	0.96	0.99	0.98
	AdaBoosting	0.91	0.93	0.97	0.95	-	-	-	-
	Bernoulli Naive Bayes	0.92	0.95	0.97	0.96	-	-	-	-
	Multinomial Naive Bayes	0.89	0.89	1.00	0.94	0.87	0.86	0.94	0.89
	Complement Naive Bayes	-	-	-	-	0.87	0.92	0.86	0.89
	XGBoost	-	-	-	-	0.91	0.91	0.94	0.93
	LightGBM	-	-	-	-	0.90	0.91	0.92	0.91

جدول ۵. نتایج حالت‌های سه‌کلاسه و پنج‌کلاسه.

		Three-class				Five-class			
		Evaluation Metrics							
		Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
Machine Learning Models	Support Vector Machine	0.90	0.90	0.88	0.89	0.83	0.83	0.83	0.83
	Logistic regression	0.84	0.83	0.78	0.79	0.75	0.74	0.74	0.75
	Decision Tree	0.83	0.81	0.81	0.81	0.72	0.72	0.72	0.72
	Random forest	0.95	0.96	0.93	0.94	0.91	0.92	0.91	0.91
	Multinomial Naive Bayes	0.78	0.83	0.68	0.67	0.73	0.73	0.73	0.73
	Complement Naive Bayes	0.81	0.81	0.75	0.76	0.72	0.73	0.72	0.72
	XGBoost	0.86	0.87	0.82	0.83	0.79	0.80	0.79	0.79
	LightGBM	0.85	0.85	0.81	0.82	0.81	0.81	0.81	0.80

ماشین اجراشده در حالت‌های دوکلاسه، سه‌کلاسه، و پنج‌کلاسه پرداخته شده است. در بین کارهای مرتبط بیان‌شده در بخش‌های قبل، مطالعه‌ی شتی و همکاران (۲۰۲۳)،[۱۲] با عنوان کاوش احساس‌های بازخورد داده‌های تجارت الکترونیک با استفاده از الگوریتم‌های یادگیری ماشین، دارای متغیر هدف یکسان با مطالعه‌ی حاضر بوده است، اما فقط در حالت دوکلاسه صورت گرفته است. در جدول ۴، به‌منظور مقایسه‌ی معیارهای ارزیابی، دو مطالعه‌ی صورت-گرفته طراحی شده‌اند. معیارهای ارزیابی بیان‌شده در جدول اخیر، هر دو بر روی مجموعه‌ی داده‌ی یکسان و با استفاده از الگوریتم استخراج ویژگی TF-IDF در نظر گرفته شده‌اند. همان‌طور که مشاهده می‌شود، بالاترین درستی در مطالعه‌ی شتی و همکاران، مربوط به مدل ماشین بردار پشتیبان با درستی ۰٫۹۴ و بهترین عملکرد در مطالعه‌ی حاضر مربوط به مدل جنگل تصادفی با درستی ۰٫۹۸، و پس از آن درستی مدل ماشین بردار پشتیبان برابر با ۰٫۹۵ در نظر گرفته شده است. موضوع قابل‌توجه این است که در مطالعه‌ی شتی و همکاران (۲۰۲۳)، به‌منظور رفع نامتعادل‌بودن مجموعه‌ی داده، راهکاری ارائه نشده است.

مقادیر درستی، دقت، پوشش، و امتیاز F1 برای مدل‌های یادگیری ماشین اجراشده در حالت‌های سه‌کلاسه و پنج‌کلاسه در جدول ۵ مشاهده می‌شود. در حالت سه‌کلاسه، بهترین عملکرد مربوط به مدل جنگل تصادفی با درستی ۰٫۹۵، دقت ۰٫۹۶، پوشش ۰٫۹۳، و امتیاز F1 ۰٫۹۴ بوده است. پس از آن مدل ماشین بردار پشتیبان با داشتن مقادیر درستی ۰٫۹۰، دقت ۰٫۹۰، پوشش ۰٫۸۸، و امتیاز F1 ۰٫۸۹ عملکرد خوبی داشته است. در حالت پنج‌کلاسه نیز بهترین عملکرد مربوط به مدل جنگل تصادفی با درستی ۰٫۹۱، دقت ۰٫۹۲، پوشش

۰٫۹۱ ، و امتیاز F1 ۰٫۹۱ بوده است. پس از آن مدل ماشین بردار پشتیبان با داشتن مقادیر درستی ۰٫۸۳، دقت ۰٫۸۳، پوشش ۰٫۸۳، و امتیاز F1 ۰٫۸۳ عملکرد خوبی داشته است.

۵. نتیجه‌گیری و پیشنهادهای آینده

با توجه به اهمیت شناخت رفتار مشتریان به‌منظور ادامه‌ی فعالیت در دنیای امروز، در مطالعه‌ی حاضر به بررسی احساس‌های نظرهای تجارت الکترونیک پوشاک زنان با استفاده از روش‌های پردازش زبان طبیعی و یادگیری ماشین پرداخته شده است. مدل‌های یادگیری ماشین استفاده‌شده، شامل ماشین بردار پشتیبان، رگرسیون لجستیک، درخت تصمیم، جنگل تصادفی، بیز ساده‌ی چندجمله‌ای، بیز ساده‌ی مکمل، XGBoost، و LightGBM هستند. به‌منظور برداری‌کردن و استخراج ویژگی متن نظرها از الگوریتم‌های TF-IDF و Word2vec استفاده شده است. مدل‌سازی موضوعی با استفاده از روش تخصیص دیریکله‌ی پنهان و خوشه‌بندی k-means انجام شده است. معیارهای ارزیابی استفاده‌شده در مطالعه‌ی حاضر، شامل درستی، دقت، پوشش، و امتیاز F1 بوده‌اند. از آنجایی که متغیر هدف امتیاز در حالت معمول پنج کلاس (امتیاز ۱ تا ۵) دارد، به‌منظور انجام مطالعه و اجرای مدل‌ها در حالت‌های دوکلاسه و سه‌کلاسه، امتیازها دسته‌بندی شده‌اند. در حالت دوکلاسه، امتیازهای قبل از ۳ به‌عنوان کلاس ۰ و امتیاز ۳ و بالاتر به‌عنوان کلاس ۱ تعریف شده‌اند. در حالت سه‌کلاسه، امتیازهای قبل از ۳ به‌عنوان کلاس ۰، 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 درستی ۰٫۹۸ در حالت دوکلاسه، ۰٫۹۵ در حالت سه‌کلاسه، و ۰٫۹۱ در حالت پنج‌کلاسه برآورد شده‌اند.

مطالعه‌ی حاضر، ترکیبی از متن‌کاوی و پردازش زبان طبیعی است، اما براساس اهداف و روش‌شناسی، بیشتر بر روی حوزه‌ی متن‌کاوی متمرکز بوده است. هدف اصلی پژوهش، تحلیل و طبقه‌بندی نظرهای مشتریان با استفاده از الگوریتم‌های یادگیری ماشین بوده است. روش‌های پردازش زبان طبیعی، مانند: TF-IDF و Word2Vec برای برداری‌سازی و استخراج ویژگی‌های متنی استفاده شده و نقش پشتیبانی‌کننده‌ای در پیش‌پردازش داده‌ها و آماده‌سازی آن‌ها برای مدل‌سازی ایفا کرده‌اند. علاوه‌بر این، فرآیندهایی مانند دسته‌بندی مجدد متغیر هدف، متعادل‌سازی داده‌ها با SMOTE و استفاده از الگوریتم‌های مختلف یادگیری ماشین برای پیش‌بینی و تحلیل احساس‌ها، نشان‌دهنده‌ی تمرکز پژوهش بر کشف الگوها و پیش‌بینی‌ها از داده‌های متنی هستند. در حالی که روش‌های پردازش زبان طبیعی در بخش‌هایی، مانند مدل‌سازی موضوعی و استخراج ویژگی به‌کار رفته‌اند، نقش اصلی آن‌ها پشتیبانی از تحلیل داده‌های متنی و مدل‌سازی بوده است. بنابراین، پژوهش حاضر، بیشتر در حوزه‌ی متن‌کاوی قرار می‌گیرد و روش‌های پردازش زبان طبیعی به‌عنوان ابزارهایی برای تسهیل فرآیند متن‌کاوی استفاده می‌شوند. به‌منظور انجام پژوهش در کارهای آینده می‌توان علاوه‌بر مدل‌های یادگیری ماشین، از مدل‌های یادگیری عمیق نیز به‌منظور تجزیه‌وتحلیل احساس‌ها استفاده کرد. روش‌های استخراج ویژگی در پردازش زبان طبیعی نیز الگوریتم‌های متفاوتی دارند، که می‌توان به‌منظور بررسی نظرها و رفتار مشتریان آن‌ها را به‌کار گرفت.

References - منابع

1. Mabrouk, A., Redondo, R.P.D. and Kayed, M., 2021. Seopinion: summarization and exploration of opinion from e-commerce websites. *Sensors*, *21*(2), pp.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. 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. 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. 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. 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. 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. 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. 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*, pp.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), pp.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. Shobha Rani, V., Deepthi, K., Ramu, V. and Shirisha, G., 2024, January. Sentiment forecasting in women's fashion e-Commerce: a machine learning perspective. In *International Conference on Multi-Strategy Learning Environment*, pp.597-610. https://doi.org/10.1007/978-981-97-1488-9_44.

19. NICAPOTATO; Women's e-commerce clothing reviews, 2018. https://www.kaggle.com/datasets/nicapotato/womens-ecommerce-clothing-reviews.

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), pp.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*, pp.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. 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. https://doi.org/10.1109/ACCESS.2023.3299853.