Lecture Notes in Networks and Systems 1090

Cengiz Kahraman · Sezi Cevik Onar ·
Selcuk Cebi · Basar Oztaysi ·
A. Cagrı Tolga · Irem Ucal Sari *Editors*

Intelligent and Fuzzy Systems

Intelligent Industrial Informatics and Efficient Networks Proceedings of the INFUS 2024 Conference, Volume 3

Springer

Lecture Notes in Networks and Systems 1090

The series "Lecture Notes in Networks and Systems" publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the worldwide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Cengiz Kahraman · Sezi Cevik Onar ·
Selcuk Cebi · Basar Oztaysi · A. Cagrı Tolga ·
Irem Ucal Sari
Editors

Intelligent and Fuzzy Systems

Intelligent Industrial Informatics and Efficient Networks Proceedings of the INFUS 2024 Conference, Volume 3

Springer

Editors
Cengiz Kahraman
Department of Industrial Engineering
Istanbul Technical University
Istanbul, Türkiye

Sezi Cevik Onar
Department of Industrial Engineering
Istanbul Technical University
Istanbul, Türkiye

Selcuk Cebi
Department of Industrial Engineering
Yildiz Technical University
Besiktas, Istanbul, Türkiye

Basar Oztaysi
Department of Industrial Engineering
Istanbul Technical University
Besiktas, Istanbul, Türkiye

A. Cagrı Tolga
Department of Industrial Engineering
Galatasaray University
Istanbul, Türkiye

Irem Ucal Sari
Department of Industrial Engineering
Istanbul Technical University
Besiktas, Istanbul, Türkiye

ISSN 2367-3370 ISSN 2367-3389 (electronic)
Lecture Notes in Networks and Systems
ISBN 978-3-031-67191-3 ISBN 978-3-031-67192-0 (eBook)
https://doi.org/10.1007/978-3-031-67192-0

Preface

INFUS is an acronym for intelligent and fuzzy systems. INFUS 2019 was an on-site conference organized in Istanbul, Turkey. INFUS 2020 and INFUS 2021 conferences were organized as online conferences because of pandemic conditions. INFUS 2022 conference was organized as both online and on-site conference in Izmir with the cooperation of Yasar University and Izmir Bakircay University. INFUS 2023 was the fifth conference of this series organized by Istanbul Technical University in Istanbul. INFUS 2024 is the sixth conference that is organized in Canakkale with the cooperation of Istanbul Technical University and Canakkale Onsekiz Mart University.

The theme of INFUS 2024 conference this year is intelligent industrial informatics and efficient networks which are at the forefront of modern industrial systems. This cutting-edge field integrates advanced technologies, such as artificial intelligence, machine learning, and data analytics, into industrial processes, revolutionizing the way industries operate. By harnessing the power of intelligent informatics, businesses can optimize their operations, enhance productivity, and improve decision-making processes. One key aspect of this paradigm shift is the implementation of smart sensors and IoT devices, which facilitate real-time data collection and communication. These sensors continuously monitor various parameters, providing a wealth of information on factors like temperature, humidity, pressure, and machine performance. These data are then processed by intelligent algorithms, enabling predictive maintenance strategies, and reducing downtime. Moreover, through the application of machine learning, these systems can adapt and learn from historical data, allowing for more accurate and efficient operations over time. Efficient networks play a pivotal role in this ecosystem, providing the infrastructure for seamless communication and data exchange. High-speed, low-latency networks ensure that information flows effortlessly between devices, enabling timely responses and enabling the coordination of complex manufacturing processes. This network architecture supports the integration of edge computing, where data processing occurs closer to the source, reducing latency and enabling faster decision-making. Furthermore, the concept of digital twins, which involves creating virtual replicas of physical systems, is a cornerstone of intelligent industrial informatics. These digital counterparts allow for simulation, testing, and optimization of processes in a virtual environment before implementation in the real world. This not only minimizes the risk of errors but also enables continuous improvement and innovation.

Researchers from nearly 40 countries such as Turkey, Russia, China, Iran, Poland, India, Azerbaijan, Bulgaria, Spain, Ukraine, Pakistan, South Korea, UK, Indonesia, USA, Vietnam, Finland, Romania, France, Uzbekistan, Italy, and Austria contributed to INFUS 2024. Our invited speakers this year are Prof. Krassimir Atanassov, Prof. Vicenc Torra, Prof. Janusz Kacprzyk, Prof. Ahmet Fahri Özok, and Prof. Ajith Abraham, and Prof. Irina Perfilieva. It is an honor to include their invaluable speeches in our conference program. We appreciate their voluntary contributions to INFUS 2024, and we hope to see them at INFUS conferences for many years. This year, the number of submitted papers

became 394. After the review process, about 36% of these papers have been rejected. More than 60% of the accepted papers are from other countries outside Turkey.

We again thank all the representatives of their countries for selecting INFUS 2024 as an international scientific arena to present their valuable research results. We are honored and aware of our responsibility that our participants have chosen us in a highly competitive environment with hundreds of conferences in the same field and organized in close dates to each other. INFUS conferences manage high-cost international conference participation processes for the benefit of the participants, with lower registration fees but more well-known expert invitations and rich social activities.

We also thank the anonymous reviewers for their hard works in selecting high-quality papers of INFUS 2024. Each of the organizing committee members provided invaluable contributions to INFUS 2024. INFUS conferences would be impossible without their efforts. We hope meeting all of our participants next year in Turkey one more time with a new research theme at a new city and new social activities.

We would like to thank our publisher Springer Publishing Company, Series editor Prof. Janusz Kacprzyk, Interdisciplinary and Applied Sciences and Engineering and Editorial Director Thomas Ditzinger for their supportive, patient, and helpful roles during the preparation of this book.

Book Editors: Prof. Cengiz Kahraman, Istanbul Technical University, Prof. Sezi Cevik Onar, Istanbul Technical University, Prof. Selcuk Cebi, Yildiz Technical University, Prof. Basar Oztaysi, Istanbul Technical University, Prof. A. Cagri Tolga, Galatasaray University, Assoc. Prof. Irem Ucal Sari, Istanbul Technical University.

Cengiz Kahraman
Sezi Cevik Onar
Selcuk Cebi
Basar Oztaysi
A. Cagri Tolga
Irem Ucal Sari

Organization

Program Committee Chairs

Kahraman, Cengiz	ITU, Industrial Engineering Department, Istanbul, Turkey
Cevik Onar, Sezi	ITU, Industrial Engineering Department, Istanbul, Turkey
Cebi, Selcuk	YTU, Industrial Engineering Department, Istanbul, Turkey
Tolga, A. Cagri	GSU, Industrial Engineering Department, Istanbul, Turkey
Oztaysi, Basar	ITU, Industrial Engineering Department, Istanbul, Turkey
Ucal Sari, Irem	ITU, Industrial Engineering Department, Istanbul, Turkey

Program Committee Members

Bolturk, Eda	ITU, Industrial Engineering Department, Turkey
Dogan, Onur	Bakircay University, Industrial Engineering Department, Turkey
Haktanir Aktas, Elif	BAU, Industrial Engineering Department, Turkey
Gulbay, Murat	Gaziantep University, Vocational School of Naci Topcuoglu, Turkey
Tekin, Ahmet Tezcan	Soft Towel Games Ltd., Malta
Cebi, Selcuk	YTU, Industrial Engineering Department, Istanbul, Turkey
Cevik Onar, Sezi	ITU, Industrial Engineering Department, Istanbul, Turkey
Findikoglu, Fuat	YTU, Foreign Languages Education Department, Istanbul, Turkey
Gül, Alize Yaprak	ITU, Industrial Engineering Department, Turkey
Sahin, Gulsah	YTU, Industrial Engineering Department, Turkey
Caliskan Demir, Melisa	Istanbul Aydin University, Industrial Engineering Department, Turkey
Alkan, Nursah	ITU, Industrial Engineering Department, Turkey
Ilhan Findikoglu, Dilek	YTU, Educational Sciences Department, Turkey
Kadikoylu, Seren	ITU, Industrial Engineering Department, Turkey

Kahraman, Cengiz	ITU, Industrial Engineering Department, Istanbul, Turkey
Kizilkan, Zeynep Burcu	ITU, Industrial Engineering Department, Turkey
Otay, Irem	Istanbul Bilgi University, Industrial Engineering Department, Istanbul, Turkey
Oztaysi, Basar	ITU, Industrial Engineering Department, Istanbul, Turkey
Tolga, A. Cagri	GSU, Industrial Engineering Department, Istanbul, Turkey
Ucal Sari, Irem	ITU, Industrial Engineering Department, Istanbul, Turkey

Contents

Machine Learning

Type-n Fuzzy Sets

Pythagorean Fuzzy Sets

Machine Learning

Uncovering Business Accounts Among Retail Shoppers: Insights from Supermarket Transactions Using an Intelligent Decision Support System

Eslem Güler(✉) and Ömer Zeybek

R&D Department, Migros Ticaret A.S., Atasehir, Istanbul 34735, Turkey
{eslemg,omerze}@migros.com.tr

Abstract. Customer segmentation is crucial to retail operations, pivotal in marketing strategies, and improving operational efficiency. This study proposes an innovative approach to optimizing customer segmentation and enhancing business efficiency. The primary objective is to classify customers currently shopping in supermarkets but exhibit behaviors indicative of potential interest in preferring the wholesale channel. Although the propensity of retail customers to buy specific products and their probability of churn have been well documented by previous marketing analytics research, less is known about service-oriented customer segmentation. By leveraging customer behavior and transactional sales-related data from a retail grocer serving both retail and wholesale channels, this research aims to identify retail customers exhibiting wholesale behavior patterns. Various unsupervised classification algorithms were employed during the estimation process. Among them, the KMeans algorithm, known for its usage in segmentation models, was chosen as the primary classification model. Implementing this strategy has the potential to streamline customer targeting efforts, ultimately maximizing the effectiveness of directing individuals toward the most relevant retail outlet. By discovering these patterns, we aim to direct customers toward the retailer's most efficient service channel, streamlining customer targeting efforts and enhancing operational efficiency.

Keywords: Machine Learning · Customer Analytics

1 Introduction

In the dynamic environment of the retail industry, understanding and finding solutions to the diverse needs of customers is essential for sustainable long-term profitability. Customer segmentation, the process of categorizing consumers based on their demographics, shopping behaviors, or channel/product preferences, has a massive impact on the efficiency of retail operations. In this regard, segmentation is a way to know the customer and find solutions to problems.

C. Kahraman et al. (Eds.): INFUS 2024, LNNS 1090, pp. 3–10, 2024.
https://doi.org/10.1007/978-3-031-67192-0_1

When companies have the essence of the question of "what?" it is always easier to reach guidelines on "how?". This strategic practice shapes the marketing efforts' goals and is fundamental in optimizing procurement, investment, and sales efficiency.

In the fiercely competitive retail environment, where consumers have many choices, the ability to tailor marketing strategies to specific segments of customers is a distinguishing factor. Customer segmentation allows retailers to point to distinct groups within their customer base, enabling them to personalize promotional campaigns, product offerings, and communication channels.

Furthermore, the impact of customer segmentation extends beyond marketing considerations, influencing operational efficiency at every level of the retail supply chain. By understanding different customer segments' diverse needs and preferences, retailers can streamline inventory management, optimize pricing strategies, and improve the overall customer experience. This results in enhanced customer loyalty and ensures that resources are allocated more efficiently throughout the retail operations ecosystem.

In this context, our study delves into customer segmentation with a specific focus on its role in retail operations by studying different channels of the same brand and directing its customers to those channels based on the segmentation results. By proposing an innovative approach to identifying and classifying customers exhibiting wholesale customer behavior, we aim to guide these particular shoppers to business-to-business (B2B) service channels. Serving customers through a specific service channel that corresponds to their needs and shopping behavior is critical to increasing loyalty and improving customer experience. Our modeling approach is a foundation for an intelligent decision system and ties potential small business owners like shopkeepers, artisans, or craftspeople to retailers' business-type customer segments. This strategy contributes to the ongoing discourse on enhancing marketing strategies and operational efficiency in the retail sector and offers suitable marketing campaigns for customers.

The rest of the study is organized as follows: Section 2 offers a brief literature review, Sect. 3 presents methodology, Sect. 4 explains data preparation and processing, Sect. 5 addresses the conceptual model, and Sect. 6 discusses empirical findings.

2 Literature Review

Our research objective is a combination of customer segmentation and profiling. Since the early days of marketing, companies have gradually recognized that not all customers are the same; this phenomenon leads them to distinguish people from each other using an analytical methodology. Customer segmentation is a way to achieve this goal as a decision support system that identifies the best cohorts for customized marketing campaigns [5].

Thanks to rapid improvements in computerized data management systems from the 1970s, customer segmentation has become a promising study area in the marketing analytics domain. Early studies on service segmentation of customer

characteristics and behaviors began in the banking sector. One of the first studies on profiling, financial service usage, and impact on profitability [4] employed data from a retail bank to understand the relationship between financial service usage patterns and customer segments. Customer segmentation is a project that can be done using multiple techniques and various forms of data. It can be achieved using demographic data to find patterns [3], or product-specific variables and transactional history can also be used to discover customer relationships [10].

In retail, it is crucial to understand customer behavior based on the recency, frequency, and monetary (RFM) value of shopping. Understanding the shopping mission brings unique customer information and should be used in conjunction with traditional methods [9].

3 Conceptual Model

3.1 Overview

Companies must direct their customers to the proper service format based on their needs. This paper will use data from one of the national-level leader retailers' data to investigate its formats and group customers to them based on their shopping behaviors. One of the target formats is the wholesale stores, which sell products in large quantities. The other is the supermarket, which serves individual customers and is mainly served for household consumption. Based on the definitions of the segments given above, customers should be served in channels (formats) that suit their characteristics and needs the best.

If a wholesale customer purchases from regular supermarkets, such as in the wholesale stores, they would create a shortage of supplies, leading to less customer satisfaction for regular supermarkets. Besides, a wholesale customer would find better price offers in stores focused on commercial consumption due to varying frequency patterns and sales volume. Directing wholesale customers to wholesale stores will make it easy for them to purchase many supplies and offer beneficial prices. However, the benefits of promotional products will increase customer satisfaction for regular supermarket customers. To solve this problem, we used customers who only shopped in districts where the company has a direct-to-retail supermarket and a wholesale-focused outlet. We chose to use unsupervised learning algorithms to identify customers' patterns and find meaningful cohorts among retail customers that correspond to the essential characteristics of wholesale customers.

3.2 Data Collection and Preprocessing

The data we used to analyze and find patterns consists of customer data from those who shopped in the 98 districts where wholesale stores are located and all stores in neighborhoods close to these stores. We derive customer-level features, the main KPIs distinguishing a wholesaler from a retail customer. Our other constraint was only including active customers in our analysis, so we only included

customers who shopped in the last six months. We collected and summarised data for both wholesalers and retail customers. All features are calculated for 2023.

The features are collected based on individuals and calculated for each customer. For instance, some features include item counts for each category, the count of times a product is bought more than twelve times in one transaction, and the amount of money spent on online and offline channels. While collecting the features, we assume that a customer will most likely be a wholesaler if they purchase a large quantity of a specific product or heavily promoted SKU in one shopping trip.

Other feature examples that will demonstrate wholesaler behavior are how often an individual shops from a wholesale store and the count of VAT percentage if it is one and between particular dates. This is because, during the pandemic, only wholesalers could shop with 1% VAT.

All data processing and estimation processes are carried out in a Python environment [11] with scikit-learn libraries for modeling [8], pandas for data preprocessing [7] and matplotlib package for visualizations [6].

4 Methodology

4.1 Method and Empirical Findings

This research is shaped based on a business request to solve stock problems and misguidance in the retail sector in Turkey. In our analysis, we used data from a national supermarket chain. This supermarket chain serves both individual customers and wholesalers (commercial customers). The data contains transactional data on customers who shopped in any format where wholesale stores of the retailer are located and all stores in the neighborhoods close to these stores for the last six months.

Initially, we checked for missing and outlier values in the data. Missing data are interpolated according to data properties with averages, whereas outliers are left still. However, our data were too complex, with 3.9 million observations and 28 columns. Although personal data on customers from wholesalers and individual customers were anonymous, we had a flag indicating whether a customer was a wholesaler. However, in our data, the ratio between wholesalers and all customers was minimal at the 0.0006 level.

In their study [12] stated that data sets with rare target events would suffer bias. However, they discuss solutions like increasing sample size, using synthetic data-dependent models like SMOTE, or using evolution criteria other than accuracy. In our study, we follow a completely different procedure. We follow studies of [1,2] in which semi-supervised methods were employed instead of classification methods in the presence of unbalanced data. However, unlike them, we used the supervised clustering method to label a cluster with the most robust characteristics related to wholesaler data.

We decided to use factor analysis to simplify complex data and reduce dimension, and we used derived scores in the clustering algorithm. Additionally, we

treated factor scores as more reliable features to find hidden patterns and set the stage for a deeper and more focused analysis. To determine the number of factors needed to run factor analysis, we looked at eigenvalues, which measure how much of the variance of the original variables a factor explains. An eigenvalue of more than one means that the factor explains more variance than a unique variable. We chose eight factors as an optimal number of variables in the clustering study. Following dimension reduction, these eight variables for all observations were used as inputs to the k-means clustering algorithm.

4.2 K-Means Algorithm

K-means clustering aims to partition a dataset into k clusters where each data point belongs to the cluster with the nearest mean. The algorithm works by randomly initializing k centroids. Then, assign each data point to the closest centroid. Recalculate the centroids as the mean of all data points assigned to them. Repeating the assignment and centroid update steps until convergence (i.e., centroids no longer change significantly or a maximum number of iterations is reached).

a. Distance Calculation (Euclidean distance is commonly used):

$$\text{Euclidean distance}(x_i, c_j) = \sqrt{\sum_{d=1}^{D}(x_{id} - c_{jd})^2}$$

where x_i is a data point, c_j is a centroid, D is the number of dimensions/features, and d ranges from 1 to D.

b. Objective Function (Squared Euclidean distance is often minimized):

$$J = \sum_{i=1}^{N}\sum_{j=1}^{k} r_{ij}\|x_i - c_j\|^2$$

where N is the number of data points, k is the number of clusters, r_{ij} is a binary indicator (1 if x_i is assigned to cluster j, 0 otherwise), x_i is a data point, c_j is a centroid.

c. Assignments:

$$r_{ij} = \begin{cases} 1 & \text{if } j = \text{argmin}_k\|x_i - c_k\|^2 \\ 0 & \text{otherwise} \end{cases}$$

d. Centroid Update:

$$c_j = \frac{1}{\sum_{i=1}^{N} r_{ij}} \sum_{i=1}^{N} r_{ij} x_i$$

This equation computes the new centroid c_j by averaging the data points assigned to cluster j.

This is a graphical method to find the optimal K value in a k-means clustering algorithm. The elbow graph shows the values of the sum of squares within the cluster on the y-axis corresponding to the different values of K. The optimal K value is the point at which the graph forms an elbow. Based on Fig. 1, the cluster size is selected as four, and the data obtained from factor analysis with eight features is used after scaling.

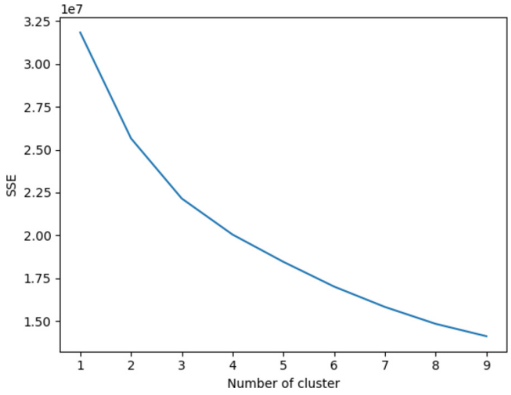

Fig. 1. The elbow method graph.

With this information, we employed the K-means algorithm [13], one of the most frequently used algorithms for clustering and segmentation. Although our research is not based on segmentation, finding similar patterns in our customers is essential.

4.3 Model Evaluation and Validation

Using factor analysis to reduce the dimension of the features to eight, we profile and name eight newly created features using correlations between factor loadings and original variables (See Fig. 2).

Those columns explain customer behavior, such as prime customers, customers who shop online and use promotions frequently, customers who use promotions all the time, shoppers who shop online the most, buy alcohol at night, shop mainly in the mornings, and offline shoppers. A newly created eight-column factor score data frame is used to perform a k-means analysis with four clusters, determined using the elbow method.

Table 1 represents the clustering results based on the percentage of customers. One of the most important results is that cluster zero is fully clustered as regular customers, not wholesalers. On the other hand, cluster one has the most wholesalers and regular customers in its cluster. Even though the model could not perfectly cluster wholesalers into one cluster, we present an approach to

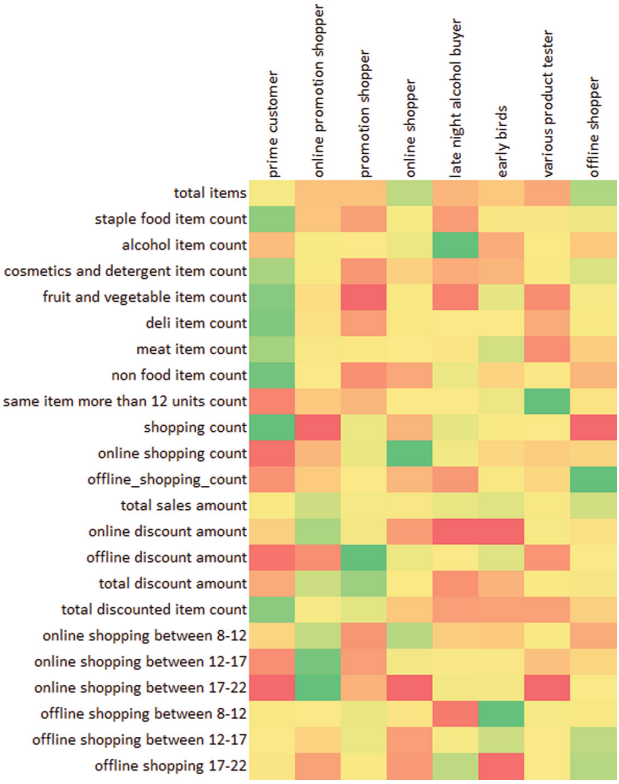

Fig. 2. Factor analysis loadings with reformed columns.

solve this problem. The supervised model was overfitting because of the lack of wholesaler customers, and we approached the issue with unsupervised methods.

The cluster with the most wholesalers can give us insights about similar behavior in other customers not classified as wholesalers.

Table 1. Results of K-means model based on the counts of customers.

Cluster Label	Ratio of Wholesaler Cust.
Cluster 0	0
Cluster 1	0.0003
Cluster 3	0.02
Cluster 2	0.003

5 Conclusion

In conclusion, this paper can be used as a guide to using unsupervised clustering methods on big data to find similar customer behavior. Although the group we suggest for possible wholesaler customers only includes 2% of registered segments, the hit ratio is nearly ten times higher than the average compared to other segments. The results can be improved by using different algorithms and getting more smoothly divided clusters. Marketers can use the information delivered to offer special promotions to regular customers who are found to be wholesalers and direct them to the wholesale stores so regular stores do not have to experience inventory shortages; thus, regular customers can benefit.

References

1. Arshad, A., Riaz, S., Jiao, L.: Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification. IEEE Access **7**, 28100–28112 (2019). https://doi.org/10.1109/ACCESS.2019.2901860
2. Chakraborty, S., Phukan, J., Roy, M., Chaudhuri, B.: Handling the class imbalance in land-cover classification using bagging-based semisupervised neural approach. IEEE Geosci. Remote Sens. Lett. **17**, 1493–1497 (2020). https://doi.org/10.1109/LGRS.2019.2949248
3. Gulluoglu, S.: Segmenting customers with data mining techniques. In: 2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC), pp. 154–159 (2015). https://doi.org/10.1109/dinwc.2015.7054234
4. Harrison, T.: Mapping customer segments for personal financial services. Int. J. Bank Market. **12**, 17–25 (1994). https://doi.org/10.1108/02652329410069010
5. The Evolution of Customer Segmentation: From Mass Marketing to Individualized Approaches (2023)
6. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput Sci. Eng. **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
7. McKinney, W., et al.: Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, Austin, TX, vol. 445, pp. 51–56 (2010)
8. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
9. Sokol, O., Hol'y, V.: The role of shopping mission in retail customer segmentation. Int. J. Mark. Res. **63**, 454–470 (2019). https://doi.org/10.1177/1470785320921011
10. Tsai, C., Chiu, C.C.: A purchase-based market segmentation methodology. Expert Syst. Appl. **27**, 265–276 (2004). https://doi.org/10.1016/j.eswa.2004.02.005
11. Van Rossum, G., Drake, F.L.: Python 3 Reference Manual. CreateSpace, Scotts Valley, CA (2009)
12. Wang, L., Han, M., Li, X., Zhang, N., Cheng, H.: Review of classification methods on unbalanced data sets. IEEE Access **9**, 64606–64628 (2021). https://doi.org/10.1109/ACCESS.2021.3074243
13. Wu, J.: Advances in K-Means Clustering: A Data Mining Thinking. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-29807-3

Beyond the Scent: A Holistic NLP Study of the Fragrance World

Simay Turan(✉), Fatih Kerem Doğan, and Tolga Kaya

Department of Management Engineering, Istanbul Technical University, 34367 Istanbul, Turkey
{turans20,doganfa20}@itu.edu.tr

Abstract. This paper presents a holistic NLP-based approach to the fragrance industry. Using a methodology based on online reviews, this study aimed to gain insights into customer sentiment and product preferences towards fragrance brands. In this study, analyses were conducted by creating a unique dataset by randomly selecting 13,828 reviews of 36 perfumes from a website specializing in the global perfume industry. By sentiment analysis, it is investigated how well user comments' sentiment compound scores and overall perfume evaluations correlate. Using Latent Dirichlet Allocation method, commonly used terms and topics are extracted from comments to provide insights into user sentiments. Moreover, Term Frequency-Inverse Document Frequency analysis helped to draw out keywords and word patterns from evaluations that are distinctive to a given brand. The most commonly used words in these comments were identified by examining user reviews. By bridging perfumes and language within user-generated content, we believe this study would contribute valuable insights to management practices in the fragrance industry.

Keywords: Natural Language Processing · Latent Dirichlet Allocation · Fragrance · Machine Learning · Sentiment Analysis · Topic Modelling · Perfume

1 Introduction

As the digital age progresses and user-generated content becomes more prevalent on the Internet, consumers are becoming more engaged in expressing their opinions about a variety of products and services. Since fragrances are a very subjective and individualized product category that frequently evokes a wide range of feelings and views in customers, customer feedback regarding the product is crucial. In addition to perfume manufacturers, marketers, retailers, and researchers looking to improve product presentment and customer experiences could all benefit from an understanding of the feelings and preferences of fragrance enthusiasts. This study presents three different natural language processing (NLP) techniques for extracting insights from user-generated content related to reviews written by perfume enthusiasts. These methods include TF-IDF analysis, sentiment analysis, and Latent Dirichlet Allocation (LDA). The study begins by examining the correlation between the sentiment compound scores obtained from the related user comments and the overall ratings given to scents on a website. The top 100

C. Kahraman et al. (Eds.): INFUS 2024, LNNS 1090, pp. 11–19, 2024.
https://doi.org/10.1007/978-3-031-67192-0_2

most often occurring words in these comments are then subjected to LDA modeling to examine their frequency distributions. On top of that, TD-IDF methods are used to perform an extensive brand analysis. Keyword extraction and recurrent word patterns are identified by the study using this analysis of reviews. The rest of this paper is organized as follows: Section 3 provides an overview of methodologies of related work in the field of sentiment analysis and topic modeling. Section 4 presents the data preparation methodology and process. Section 5 discusses the findings and visualizations of the sentiment analysis and LDA topic modeling. Section 6 offers the conclusions according to our analysis for the fragrance industry, followed by further research.

2 Literature Review

The computer-based technique known as Natural Language Processing (NLP) makes it easier for researchers to derive useful inferences from textual information [1]. To explore customer attitudes and preferences, NLP is widely utilized in research in the fields of information systems and marketing. In this context, text preprocessing, text representation, classification, topic modeling, sentiment analysis and deep learning are among the methods applied as the main natural language processing (NLP) techniques in studies [2]. As an illustration, a study investigated the Amazon product review section, where users assessed the product overall rating and their experience about buying it. While evaluating customer reviews might give a clear perspective about the product, it is advised to use advanced NLP tool kits to analyze and summarize these unstructured evaluations, as it can take a long time in some circumstances when hundreds of reviews are displayed for the product [3].

Sentiment Analysis is a branch of Natural Language Processing that focuses on determining the attitude or viewpoint of a text's subjective aspects [4]. The proliferation of the Internet and its applications has led to an increase in textual data from a variety of sources. Users are posting content and providing copious amounts of information from social networks, blogs, Internet forums, and websites that review products [5]. Sentiment analysis, in the context of product reviews, is the act of examining these evaluations to ascertain the general viewpoint or sentiment of a product [6]. The fundamental objective of sentiment analysis is polarity classification, which divides product reviews into positive and negative categories [5]. Bhadane et al. indicate that lexical and machine learning techniques are the two primary topics of sentiment classification research. Using the lexical approach, polarity values for words are stored in a dictionary. The polarity score of each word in a text is added up if it appears in a dictionary to get a "overall polarity score," which indicates the sentiment of the text [4]. On the other hand, a technique for locating word clusters (or topics) in a text corpus was defined as Topic Modeling. Accordingly, topic modeling algorithms have been created for text mining since manual topic identification is not only difficult but also neither scalable nor efficient due to the large volume of data [7]. LDA which is a method of Topic Modelling, was developed from the Probabilistic Latent Semantic Indexing model in the field of machine learning [8]. One major benefit of LDA compared to pLSI is its assumption of multiple latent topics within a corpus, with each topic being a distribution of words found throughout the documents [9]. Lastly, two methods that are widely employed in text categorization

for information retrieval are TF-IDF and NLP [10]. To be more specific, only the most related terms remain in the corpus after TF-IDF removes the most frequently occurring terms [11]. In the following sections, sentiment analysis, LDA modeling and TF-IDF technique findings obtained from perfumery data will be discussed.

3 Methodology

3.1 Concepts

Tokenization: Tokenization is the process of semantically breaking data into small units. Comments were tokenized with NLTK to set up LDA Modeling smoothly.

Latent Dirichlet Allocation (LDA): Natural language processing (NLP) and machine learning both use the statistical model known as LDA (Latent Dirichlet Allocation), which has several uses in both domains, including text mining, text classification, text clustering, and text representation. Finding hidden subjects, or themes, in documents and establishing word associations with them is the main objective of LDA.

Term Frequency-Inverse Document Frequency (TF-IDF): Is a metric used to quantify a word's significance to a document inside a corpus or collection, taking into account the fact that certain terms are used more frequently than others. This method identified the most important words based on the brand, and as shown in Fig. 4, an example for 5 brands has been demonstrated.

Sentiment Analysis: Using a natural language processing (NLP) technique called sentiment analysis, one can automatically ascertain whether a document or set of data expresses positive, negative, or neutral emotions or feelings.

Topic Modelling: A machine learning method called "Topic Modeling" is used to automatically identify hidden themes or subjects in large-scale of text collections.

3.2 Data Methodology

Data Access: For this purpose, comments on perfumes were manually and randomly collected from the publicly available https://www.fragrantica.com/ website.

Data Cleaning: In this stage, various Python libraries were used. Firstly, non-English comments were detected using the 'langdetect' library. Then, using NLTK (Natural Language Toolkit), stopwords, punctuation marks (for LDA Modeling), and emojis were removed.

Data Analysis: In this part, Sentiment Analysis was initially conducted using the SentimentAnalyzer algorithm from the NLTK library to generate 'Compound' scores. In the LDA Modeling section, the gensim library was employed for Topic Modeling, and comments were clustered and weighted. Successful results were obtained by comparing the sentiment scores created using the NLTK library with the overall perfume ratings found on the website for each perfume. Key terms for each perfume were derived based on the TD-IDF score, but this process was reiterated after excluding non-adjective words such as "like," "love," and "fragrance." The objective was to omit adjectives, eliminate verbs, and discard unnecessary words.