Intelligent Feature Design in Mass-market Retail Campaigns

Leonardo Augusto Fiorelli Luzio

Master Dissertation

FEUP Supervisor: Prof. Eduardo Luís de Meireles e Oliveira

FEUP Co-supervisor: Prof. António Manuel Ferreira Mendes Lopes



Master in Mechanical Engineering

September 2023

Resumo

Com a crescente competição por consumidores, as empresas foram forçadas a adaptar as suas estratégias operacionais. A implementação de táticas de marketing tornou-se crucial para o sucesso de qualquer negócio, pois estas não só aumentam a visibilidade da marca, como também atraem novos clientes e retêm os já existentes. Adicionalmente, estas estratégias podem ser empregadas para lançar novos produtos ou serviços, impulsionar as vendas e melhorar a reputação da empresa. Através de campanhas publicitárias, eventos e outras iniciativas, as empresas têm a oportunidade de comunicar aos consumidores os seus valores e diferenciar-se da concorrência. Esta comunicação pode ajudar a estabelecer uma conexão com a marca, potencialmente resultando numa maior lealdade do cliente e frequência de compra.

Uma destas ações de marketing amplamente adotadas pelas empresas é a implementação de campanhas promocionais. Estas ferramentas podem ser usadas para aumentar as vendas, atrair novos clientes e incentivá-los a comprar mais produtos ou serviços da empresa. A evolução tecnológica permitiu que as empresas adotassem estratégias centradas no consumidor, através da personalização dos produtos e serviços comercializados, o que tornou possível a criação de extensas bases de dados com informações sobre a base de clientes e o seu comportamento de compra.

A partir da recolha de dados comportamentais dos clientes, e utilizando técnicas de análise de dados, tornou-se possível tomar decisões focadas nas suas necessidades, desejos e preferências. No que diz respeito às ações de marketing, estas tornaram-se mais bem planeadas e direcionadas para consumidores específicos. Ao direcionar campanhas e comunicações apenas para clientes relevantes, resultados excelentes podem ser obtidos tanto para a empresa como para o cliente.

Este trabalho apresenta um projeto desenvolvido para a classificação e avaliação de campanhas promocionais massivas realizadas por empresas do setor retalhista. O objetivo deste trabalho foi o desenvolvimento de uma ferramenta fiável e eficaz para classificação de campanhas através da aplicação de técnicas como web scraping e text mining. A partir de informações históricas sobre o comportamento de compra da base de clientes, foi calculado o score de afinidade entre a campanha e grupos de clientes.

O uso destas técnicas permite que as empresas obtenham informações valiosas sobre os seus clientes e as suas preferências, permitindo-lhes criar campanhas mais eficazes e direcionadas. Isto pode levar ao aumento das vendas e à fidelização do cliente. Além disso, o uso destas técnicas também pode ajudar as empresas a identificar tendências emergentes no mercado e adaptar as suas estratégias em conformidade.

Intelligent Feature Design in Mass-market Retail Campaigns

Abstract

With the current competition for consumers, companies have had to adapt their operational strategies. The adoption of marketing actions is fundamental to the success of any company. They help increase brand visibility, attract new customers, and retain existing ones. In addition, these actions can also be used to launch new products or services, increase sales, and improve the company's image. Through advertising campaigns, events, and other initiatives, companies can communicate their values and differences to consumers. This can help create a connection with the brand, which can lead to greater customer loyalty and purchase recurrence.

One of these widely adopted marketing actions by companies is the implementation of promotional campaigns. These tools can be used to increase sales, attract new customers, and encourage customers to buy more products or services from the company. Technological evolution has allowed businesses to adopt consumer-centric strategies, through the personalization of marketed products and services, enabling extensive databases with information on the customer base and their purchasing behavior to be created.

From the collection of behavioral data from customers, using data analysis techniques, it became possible to make decisions focused on their needs, desires, and preferences. Regarding marketing actions, these have become better planned and directed towards specific consumers. By directing campaigns and communications only to relevant customers, excellent results can be obtained for both the company and the customer.

This work presents a project developed for the classification and evaluation of massive promotional campaigns carried out by companies in the retail sector. The objective of this work was the development of a reliable and effective tool would be obtained for campaign classification through the application of techniques such as web scraping and text mining. With the classified campaign, using historical information on the purchasing behavior of the customer base, the affinity score between the campaign and groups of customers was calculated.

The use of these techniques allows companies to obtain valuable information about their customers and their preferences, enabling them to create more effective and targeted campaigns. This can lead to increased sales and customer loyalty. In addition, the use of these techniques can also help companies identify emerging trends in the market and adapt their strategies accordingly.

Ackowledgements

I would like to express my deepest appreciation to all those who provided me with the possibility to complete this master thesis. A special gratitude I give to my supervisor, Professor Eduardo Luís de Meireles e Oliveira, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project, especially in writing this thesis.

Furthermore, I would also like to acknowledge with much appreciation the crucial role of the staff of Sonae MC, who gave the permission to use all required equipment and the necessary materials to complete the task. A special thanks goes to my team leader, Leandro Miranda Sousa Pinto, for his understanding and assistance.

Finally, I must express my very profound gratitude to my family and friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Contents

1 Introduction 1
1.1 General context 1
1.2 MC Sonae presentation 2
1.3 Cartão Continente 3
1.4 Dissertation objectives 3
1.5 Document Structure 4
2 State of the art 6
2.1 Data analysis 6
2.2 Data selection and sampling 9
2.2.1 Web scraping 9
2.2.2 APIs 10
2.3 Text Mining 10
2.3.1 Data Pre-processing 11
2.3.2 Data transformation and reduction 12
2.3.3 Data evaluation and visualization 13
3 Problem descripiton 16
3.1 Promotional campaigns in MC 16
3.2 Customer segmentation in MC 18
3.3 The project 19
3.4 Objectives 20
4 Methodology 21
4.1 Campaign definition 22
4.1.1 Data selection and sampling 22
4.1.2 Data cleaning and pre-processing 22
4.1.3 Data transformation and reduction 23
4.1.4 Data evaluation and visualization 25
4.2 Customer assignment 26
4.2.1 Customer segments affinity model 26
4.2.2 Affinity score 27
5 Results 30
5.1 KitchenAid 30
5.1.1 Campaign definition 32
5.1.2 Customer assignment 34
5.2 Smeg 35
5.3 Hey Clay 38
6 Discussion 42
7 Conclusions and future work 47
7.1 Conclusions 47
7.2 Limitations and future work 48
References 51

List of figures

Figure 1: An overview of the steps comprising the KDD process (Fayyad, 2005).....8
Figure 2: Example results for stemming and lemmatization (Khyani et al., 2021).....12
Figure 3: Relationships on word vectors (Sarwan, 2017). 13
Figure 4: Different clustering approaches (Taskesen, 2023b).....14
Figure 5: A visual curve with an explicit elbow point (Shi et al., 2021).....15
Figure 6: KitchenAid stamp campaign.....17
Figure 7: KDD process adapted to the problem. 21
Figure 8: Methodology for calculating affinity between campaigns and customer segments. 20
Figure 9: Comparison between agglomerative technique and other techniques. 25
Figure 10: Comparison of the cosine function (green), similarity function (blue), and linear function (red).....28
Figure 11: KitchenAid (a) Distribution by segment of the customer base; (b) Promotional participation.....31
Figure 12: Resulting clusters for the brand KitchenAid.....34
Figure 13: KitchenAid affinity score.....35
Figure 14: Smeg stamp campaign. 36
Figure 15: Smeg (a) Distribution by segment of the customer base; (b) Promotional participation.....37
Figure 16: Resulting clusters for the brand Smeg. 38
Figure 17: Smeg affinity score. 38
Figure 18: Hey Clay stamp campaign. 39
Figure 19: Hey Clay (a) Distribution by segment of the customer base; (b) Promotional participation.....40
Figure 20: Resulting clusters for the brand Hey Clay. 41
Figure 21: Hey Clay affinity score. 41

List of tables

Table 1: Words and relative frequencies. 24
Table 2: Similar words. 24
Table 3: Distribution of active customers in the promotional period for KitchenAid campaign. 32
Table 4: First 20 terms resulting from the search for the term "KitchenAid". 33
Table 5: First 20 terms ordered by importance. 33
Table 6: Distribution of active customers in the promotional period for Smeg campaign. 37
Table 7: Distribution of active customers in the promotional period for Hey Clay campaign. 40
Table 8: Top 20 terms for KitchenAid resulting cluster. 42
Table 9: Top 20 terms for Smeg resulting cluster. 43
Table 10: Segments ordering for KitchenAid campaign. 44
Table 11: Segments ordering for Smeg campaign. 44
Table 12: Top 20 terms for Hey Clay resulting cluster. 44
Table 13: Segments ordering for Hey Clay campaign. 45

1 Introduction

1.1 General context

The retail sector is a vital sector of the economy, as it is the industry responsible for selling goods and services to the end consumer. This sector includes all stores, from kiosks and small grocery stores to large supermarket chains and shopping malls that offer products for personal and household use. Physical retail is the dominant channel of this industry, however non-physical channels, such as e-commerce, have become increasingly popular (Hameli, 2018).

In 2022, the global retail market moved more than 27 trillion US dollars, with a forecast to exceed 30 trillion US dollars by 2024 (Sabanoglu, 2023). In Portugal, according to the Portuguese Association of Distribution Companies (APED), the distribution and retail sector represented 12.4% of the country's gross domestic product (GDP). The retail industry is vital to the global economy, promoting large-scale employability for skilled and unskilled workers. According to Pordata, a database with official statistics on Portugal, in 2022, 14,61% of the employed Portuguese population was in the trade sector, which includes retail (Lourinho, 2023).

Despite retail having accompanied human history, challenges such as high taxation, economic instability and global uncertainty have been challenging companies in this sector to remain competitive in this market. Concerns about humanized service, intelligent logistics processes, integration of sales processes and optimized shopping journey are increasingly common routines for businesses in this sector. Companies which are more connected, more receptive to new technologies, and working on personalization for consumers tend to stand out (Smith, 2011).

Due to the high competitiveness of this sector, companies are always looking for new marketing strategies to attract attention. Some of the strategies consolidated by companies to remain competitive include the use of marketing campaigns through various channels, which offer advantages for business communication. In addition, promotional campaigns are also used with the aim of attracting new customers, creating recurrence for existing customers and consequently increasing total sales. Promotional campaigns exist in several different formats depending on the objectives expected to be achieved with their implementation. Among the types of promotional campaigns, these can offer monetary discounts on products, offers on groups of products when buying others, loyalty points or collaborative campaigns with other companies (Shyan Fam, 2003).

However, despite being one of the main strategies of companies in the retail sector, carrying out promotional campaigns as a marketing and customer relationship tool needs to be well defined and planned since these are based on reductions in profit margins.

To create a successful promotional campaign, it is important to define the objectives of the campaign, plan the campaign, organize logistics, and actively promote it. It is essential to have clear objectives to structure the campaign in the best way. Having well-defined goals will also

be essential when analyzing the results at the end of the promotional campaign. Some examples of objectives can be to encourage new consumers to try and buy the product, influence current customers by encouraging loyalty to the brand, increase market share, introduce product innovations, differentiate the brand from its competitors or eliminate or drastically reduce stock. It is no use having a perfectly defined plan and not having a structure that can support it. If the promotional campaign is directed at a physical store, it is necessary to ensure sufficient stock to satisfy sales. If it is an online store, make sure that the site is prepared to receive a large volume of simultaneous access without compromising the user experience (Lee & Trim, 2006).

With the high competitiveness of the sector, companies have had to adapt their strategy. With changes in economic and social conditions, consumers have become more demanding of the products and services they consume, generating great effort on the part of businesses to attract new customers and retain recurring customers. Understanding consumers, their needs, desires, and interests has become powerful and essential tools in the retail sector. These changes have led companies in this sector to evolve from product-centric to consumer-focused strategies, through the establishment of a loyal relationship with their customer base (Anshari et al., 2019).

To achieve the goal of having customer relationship as a strategic pillar, many companies in the sector have opted to create extensive databases with information about customers. This stored data ranges from transactional information to demographic, social, and economic data of consumers. With this knowledge about the customer base, it is possible to perform various analyses to obtain patterns that can assist in decision-making, whether at the strategic planning level or in evaluating results. This information gives the business a strategic advantage, as it allows them to know who their consumer is to direct personalized campaigns and communications based on their purchasing behavior. Customizing offers and communications for consumers has a huge impact on customer relationships, as they receive a better shopping experience (Anshari et al., 2019).

In the context of carrying out promotional campaigns, knowledge about the consumer base is of high importance. Knowing the purchasing behavior of their customers, it is possible to determine products of greater interest for specific or generalist groups of customers. It is then possible to create promotional campaigns on target products for the audience that is desired to be reached or even exclusive offers for these customers. However, it may be necessary or commercially interesting to carry out massive campaigns available to all consumers. These advantageous campaigns can have various objectives, from reducing stock of an excess product to attracting new customers who are interested in the offer.

Due to the high impact that carrying out campaigns has on the business, careful planning is necessary; otherwise, the results may be opposite to those expected, leading to loss for the business. Planning must go through ensuring necessary stock during the promotional period and correctly defining margins and discounts so that the campaign does not bring loss. In the case of massive campaigns, this can be quite a challenging task since it is necessary to estimate the sales result of this campaign for recurring and new customers, both for the product participating in the campaign and for the resulting impact on sales patterns for all other products.

1.2 MC Sonae presentation

Sonae is a Portuguese multinational company founded in 1959 with the aim of producing decorative thermolaminates. Over the years, the company expanded and diversified, acquiring other companies, and expanding into new business areas. Today, Sonae is a company that manages a diverse portfolio of businesses in retail, financial services, technology, real estate, and telecommunications (Smeg, 2023).

MC Sonae (MC) is the Sonae Group company responsible for retail. It began its activity in 1985 with the opening of Continente Matosinhos, the first hypermarket in Portugal. Currently, MC Sonae has more than 1342 stores, with more than 36 thousand employees and more than 4 million families as customers. The company is made up of a multi-format business portfolio that offers its customers a wide range of high-quality products and services at competitive prices. Some of the brands managed by MC include:

- Food retail business: Continente, Continente Modelo, Continente Bom Dia, Continente Online, Meu Super;
- Health, beauty and well-being: Wells, Arenal, Dr. Wells, Go Natural;
- New growing businesses: Bagga, Note!, ZU, Home Story.

MC is an innovative and constantly growing company, always seeking to offer its customers the best in products and services. It is an important company for the Portuguese economy and a successful example in the retail sector. One of MC's biggest differentials compared to competitors is its successful loyalty program, Cartão Continente (CC).

1.3 Cartão Continente

Loyalty programs are strategies created by companies to retain their customers by offering them rewards, discounts, and other benefits in exchange for frequent purchases. These programs are an effective way to keep current consumers, increasing engagement with the brand and allowing greater bargaining power resulting from the trust that the consumer places in the brand. In addition, loyalty programs allow companies to better understand their customers' habits, building a database with extensive knowledge about them.

An example of a loyalty program is the Cartão Continente, created by MC Sonae. With it, customers have access to personalized discounts and can accumulate and use balance in more than 2,000 stores of the 20 brands associated with the CC. In addition, the CC offers exclusive promotions and other advantages for its members. CC has more than 4.5 million families in its customer base (Continente, 2023).

To effectively manage a loyalty program like the CC, companies can use Customer Relationship Management (CRM) strategies. CRM is a technology for managing all your company's relationships and interactions with customers and potential customers. The goal is simple: Improve business relationships to grow your business. A CRM system helps companies stay connected to customers, streamline processes, and improve profitability (Beketova, 2023).

Data analysis plays a key role in CRM strategy. CRM systems typically store large amounts of customer data, including demographic information, purchase history, and communication preferences. Analyzing this data can provide valuable insights into customer behavior, preferences, and needs. This allows companies to provide more personalized service to their customers, increasing customer satisfaction and consequently revenues.

1.4 Dissertation objectives

Given the complexity associated with massive promotional campaigns, for the development of this work, it was decided to study these campaigns in isolation, that is, independent of their impact on other products. In addition, the current customer base of the business was used as a study group, ignoring the impact of new customers during the campaign.

The problem this work seeks to address is the lack of a data-driven methodology for defining and analyzing massive promotional campaigns. Currently, many retail sector companies struggle to effectively plan and implement promotional campaigns due to a lack of understanding of their customer base and the impact of these campaigns on customer behavior and sales patterns.

Thus, among the objectives of the work, two main points stand out. Firstly, it was expected to obtain the characterization of the campaign from information available on web pages. This data source was chosen because it allows access to large amounts of updated information, allowing for the collection of information from different sources in a simplified and fast way.

With the campaign characterization carried out, the next step is to relate it to the customer base for which it is desired to study. At this stage, the main challenge is defining the customer base in a comparative way with the campaign, for calculating affinity between promotional activity and customer groups. Defining affinity as the similarity between participating products in the campaign and recurring products of customers.

To achieve these objectives, a model based on data analysis techniques was developed. The process is divided into two subprocesses: campaign characterization and calculation of affinity with customer segments. Campaign characterization defines the process of selecting databases, collecting information, and processing at the campaign level. With the expected result of classifying the campaign from clusters of words. In the affinity calculation phase, challenges associated with relating campaigns and customers are presented. A model for characterizing pre-established customer segments in business is presented, as well as the method used to calculate affinity score with campaigns.

This project aims to investigate several key research questions in the context of promotional campaigns in the retail sector. The first question is how information available on web pages can be utilized to characterize promotional campaigns. This involves exploring methods for collecting and processing information from web pages to gain insights into the nature and effectiveness of promotional campaigns.

The second research question concerns the calculation of affinity between promotional campaigns and customer segments. This involves developing and evaluating methods for measuring the similarity between the products participating in a campaign and the recurring products of customers, to determine the potential impact of the campaign on different customer segments.

The third research question focuses on the impact of promotional campaigns on customer behavior and sales patterns. This involves analyzing data on customer behavior and sales patterns before, during, and after a promotional campaign, to assess the effectiveness of the campaign and its impact on customer behavior.

By addressing these research questions, this dissertation seeks to contribute to the understanding of how data analysis techniques can be applied to the retail sector, specifically in the context of promotional campaigns. The results of this research could provide valuable insights for businesses in the retail sector, helping them to make more informed decisions when planning and implementing promotional campaigns.

1.5 Document Structure

Chapter 2 is dedicated to the analysis of the state of the art. It presents data collection techniques using web scraping, as well as text processing techniques and methods for evaluating and visualizing results for textual data.

The following chapters explore the problem in greater detail and present the methodology used in the process for campaign characterization and calculation of its affinity with the customer base.

The results obtained, for example campaigns, used in the development and validation of the project, are then presented in Chapter 5. These examples are real campaigns carried out by MC, allowing clear objectives to be defined based on historical data for comparison with the results found.

The results are then discussed in a Chapter 6, followed by general conclusions of the work, limitations, and future work in Chapter 7.

2 State of the art

In this chapter, two major themes will be explored. Firstly, to contextualize the area in which the project was developed, some concepts related to data analysis and its application in retail sector companies will be presented. Subsequently, web scraping resources and the use of APIs are explored, covering data acquisition and processing.

2.1 Data analysis

With the evolution and adoption of personal electronic devices such as smartphones and laptops, an enormous amount of data is being produced, which increases exponentially every day (Seetharaman et al., 2016). The collection of this data has become relevant as companies have begun to see it as a potentially useful source of information that, from a strategic point of view, can provide competitive advantages (Chen et al., 2012).

To utilize this information, selecting truly relevant data has become an indispensable task. As a result, various data analysis techniques have been developed, allowing for the extraction of valuable insights that help understand trends, behaviors, habits, and consumer needs. Additionally, these tools can help optimize processes, reduce costs, identify business opportunities, and forecast future demands and profits. Therefore, data analysis can create actionable ideas for delivering sustained value by providing smart data and establishing competitive advantages (Wamba et al., 2015).

As the volume and variety of data sources continue to grow, the data sets are composed of large amounts of heterogeneous and unstructured information that cannot be analyzed using traditional approaches (Waller & Fawcett, 2013). This has led to the emergence of dedicated resources for data infrastructure, as traditional approaches to data analysis are often inadequate. The challenges posed by these heterogeneous and unstructured datasets require specialized techniques and tools to effectively extract insights from them (Santoro & Usai, 2018).

Data analysis tools are used to find hidden or unknown information from large databases using different criteria, allowing the discovery of patterns and relationships between data. These criteria define how the database will be analyzed, as different conclusions can be drawn from the same data sample depending on the objective being sought. According to Sallam et al. (2011), the information obtained from these analyses can be used to support decisions, predictions, and estimations.

To make these better-informed decisions, retail companies have used data mining technologies to obtain information that allows them to offer more personalized and customized services that meet their specific customer needs, as opposed to mass campaigns. By studying their customers' buying behavior, companies can make promotional and commercial campaigns addressed to specific customer profiles, resulting in benefits for both the company and the consumer. These actions can result in a considerable reduction in costs for the company, as well as the benefit for customers of receiving only offers that really matter, leading to a higher level of satisfaction with the services offered (Ziafat & Shakeri, 2014).

Companies need to know their customers' needs and desires to ensure that they send the right message to the right customer in the best possible way (Anderson & Kerr, 2002). This process involves the use of data mining techniques to evaluate their customers, understand and predict their behavior, analyzing patterns that allow for the improvement of the relationship with the consumer (Ziafat & Shakeri, 2014).

The experience and knowledge of business experts are indispensable, as the results obtained from the use of data mining can be even more significant when associated with business expertise. These professionals are responsible for validating the results obtained by the data mining model or even participating in its development (Berry & Linoff, 2000). Despite the great help and improvement to the process obtained by business knowledge, some patterns found by data mining models would go unnoticed even by the most experienced experts, so, because of these factors, companies gain a competitive advantage in managing their customer bases (Ziafat & Shakeri, 2014).

The two main categories of data mining: predictive and descriptive. Predictive models seek to deduce values from the available data, that is, they aim to forecast unknown values, future trends, or behaviors based on historical information present in the database. On the other hand, descriptive models aim to find patterns that define and characterize the values present in the database (Han et al., 2022).

The selection of data mining techniques to be used in each problem depends on various factors such as the way the data is structured and the patterns that are expected to be obtained as a result. There is no single appropriate method for any type of problem, and a universal solution is a utopian idea when it comes to data mining. However, there are techniques commonly used to perform basic data mining tasks, such as association rules, sequential patterns, clustering, and classification (Ziafat & Shakeri, 2014).

In the retail sector, various techniques are used to analyze consumer behavior, based on clusters or association rules. The main principle underlying these techniques is that past data contains useful information for the future. These associations are possible because customer behavior is not random, but rather reflects their needs, preferences, and desires. Although the generated data is confusing and noisy, identifying and using useful information can extract patterns in historical data. As a result of this process, it is possible to obtain customer segments, which facilitates the development of personalized offers that meet these customers' specific priorities. It can also allow for personalized interactions and communications with customers, increasing satisfaction and benefiting the customer relationship (Ziafat & Shakeri, 2014).

To maximize the chances of finding valid patterns with potential impact in data, Fayyad (2005) described a process for Knowledge Discovery in Databases (KDD). The term KDD was first used by Piateski and Frawley (1991) and is still widely used today as a generalized model capable of solving complex or repetitive problems involving data.

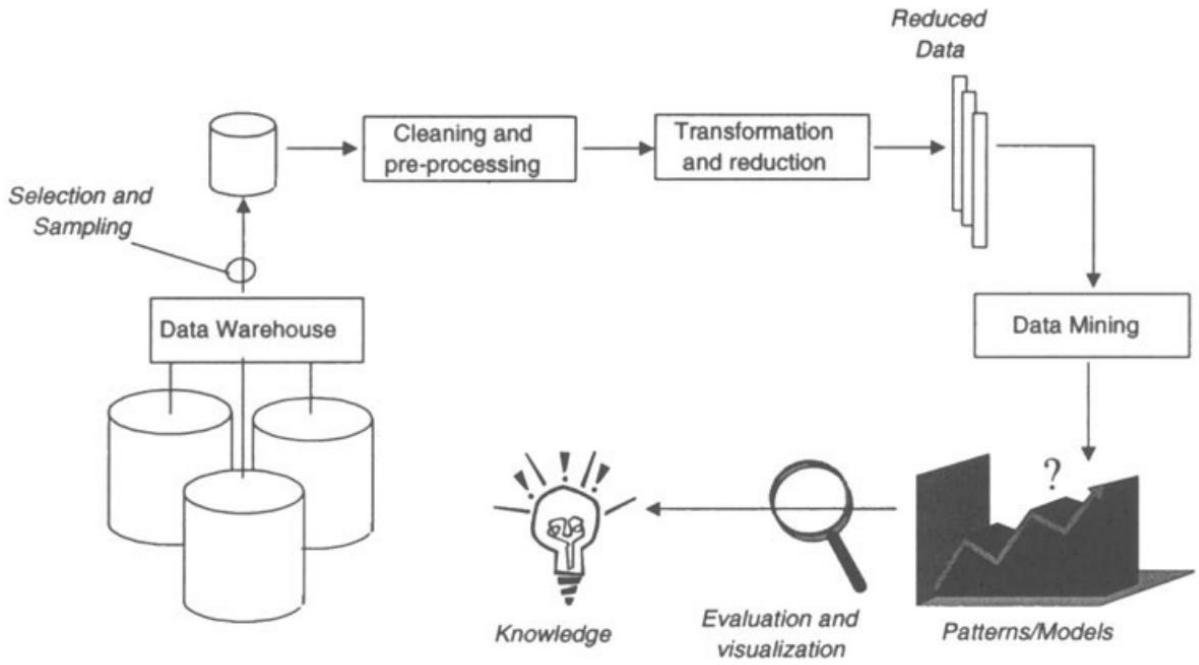


Figure 1: An overview of the steps comprising the KDD process (Fayyad, 2005).

The KDD process, as well as the distinction of its stages, is presented in Figure 1. The stages that make it up are data selection, data pre-processing and transformation, data mining and evaluation. Despite the model being presented sequentially, a great advantage of using KDD is that it is fully iterative, which means that even after the end of the process, intermediate stages can be repeated adaptively to further refine the result obtained. Below is a brief description of each of these stages:

- Data selection and sampling: after selecting the Data Warehouse, the source of the data, the data to be used in the analysis must be decided. Knowledge of the context and objective of the analysis is extremely important for selecting data and attributes since if an inadequate sample of data is used valid patterns may be found but without application.
- Data cleaning and pre-processing: this is where basic data processing operations are performed, such as removing noisy or duplicate information and formatting data types. It is also at this stage that the approach with empty fields is determined.
- Data transformation and reduction: to ensure that the data is prepared for the data mining process, it is often necessary to do treatment in its fields. The most common are normalization and standardization responsible for standardizing the scale of quantitative data. For qualitative data conversion to quantitative may be necessary since this type of data offers facilitated analysis for the machine.
- Data mining: stage where computing and statistical techniques are applied to the dataset to obtain patterns.
- Data evaluation and visualization: Since potentially infinite patterns can be found in a single dataset validation of found patterns must be validated to ensure that obtained knowledge is well-founded on relevant data. As mentioned earlier at this stage it may be concluded that too broad or too specific parameters were used in previous stages so one can resort to this process's iteratively for correction.

2.2 Data selection and sampling

Before diving into the specifics of data collection methods such as web scraping and APIs, it's important to understand the broader context. Data selection and sampling are crucial steps in any data analysis process. The goal is to identify and collect data that is relevant to the research question or objective. This involves determining what type of data is needed, where it can be found, and how it can be accessed. The chosen data must be representative of the population or phenomenon under study to ensure the validity of the analysis results. The methods for data collection can vary widely depending on the nature of the project, available resources, and ethical considerations.

2.2.1 Web scraping

Data collection is one of the most critical aspects of any data mining process, as the quality of the data can directly affect the results of the analysis and insights that will be generated from them. It is essential that data collection is carried out efficiently and effectively to ensure that the data is accurate, complete, and relevant to the objective of the analysis.

This step can be carried out in different ways, such as through questionnaires, interviews, administrative records, among others. However, with the advancement of technology, data collection through the internet has become increasingly common. This technique is known as web scraping.

Web scraping is the name given to techniques for extracting data from web pages, either manually or automatically. As a huge amount of heterogeneous data is constantly generated online, web scraping is widely recognized as an efficient method for data collection. Due to its efficiency and adaptability to a wide variety of scenarios, the technique of web scraping has evolved from small projects with human assistance to fully automated systems capable of converting entire web pages into organized datasets (Bar-Ilan, 2001). As such, for the process depicted in Figure 1, the practice of web scraping utilizes the internet as the Data Warehouse.

The process of collecting data from the internet usually begins with the creation of an HTTP request to obtain data from a target website. When this request is received and processed by the website, it is returned to the web scraping program with the collected data from the web. With the heterogeneous data obtained, the process continues with the formatting and organization of the data in a structured manner. Thus, the two essential modules in a web scraping program are: a module to compose the HTTP request, such as Urllib3 or Selenium, and another to extract information from the HTML code, such as Beautiful Soup or Pyquery (Zhao, 2017).

The Urllib3 module defines a series of functions to handle HTTP requests, such as authentication, redirection, cookies, among others (Petrov, 2023). On the other hand, Selenium allows the automation of the search process on a website through programming (Huggins & Hammant, 2023). In data extraction, Beautiful Soup was developed to extract data from HTML and XML documents using a simple and intuitive interface (Richardson, 2023). Similarly, Pyquery was also developed for the same purpose, with the main difference being that Pyquery only supports lxml for fast XML processing (Lauzanne, 2022).

Although it is a powerful tool in the acquisition stage of large amounts of data, the use of web scraping raises some legal issues related to copyright and violation of others' intellectual property. Data collection through web scraping is allowed since only snippets are obtained, but for use in large volumes of information, attention must be paid to the existence of copyrights in the material. In addition, if the data collection application sends requests with high frequency, it may be denied access, as it can be mistaken for an attack and has the potential to cause problems on the server where the website is hosted. An ethical web scraping tool pays attention to the frequency at which requests are made, keeping it at a reasonable level (Zhao, 2017).

2.2.2 APIs

Even though the use of web scraping raises ethical and legal questions, it is undoubtedly one of the most efficient ways to obtain data nowadays. A solution found by many companies to avoid the use of these tools on their websites was the creation of Application programming interfaces (APIs), which allow data owners to control how they are shared.

APIs are interfaces for computer programs that allow software to communicate with other software. They can also be used to share data and services with external developers, allowing them to create applications or integrate features into their own products. This can lead to a vibrant ecosystem of applications and services, with benefits for both developers and end users (Lomborg & Bechmann, 2014).

Unlike web scraping techniques that only allow access to public web data, the use of APIs allows access to non-public data, such as data that requires authentication by login and password, since communication is done directly with the service to which the information belongs. Despite representing a simplified form of information access, there are issues related to the quality and accuracy of information, as data collection presents an intermediary (Lomborg & Bechmann, 2014).

An important application for APIs is their use for web search. Search engines compile an extensive range of information, with the advantage of effective ranking to ensure that the most useful and relevant results are found in the vast library of data that is the internet (Stylos & Myers, 2006). Thus, to avoid scraping of its page, Google provides a search API based on the company's established search engine.

Both in qualitative and quantitative analyses, there are several challenges associated with the use of APIs in data collection, structuring, and analysis. Initially, computational capacity is required for data collection with APIs, and knowledge of different types of software, their capabilities, and limitations, allows the selection of relevant tools. The choice of specific software can influence the results that can be obtained from the data (Lomborg & Bechmann, 2014).

2.3 Text Mining

Text mining is a variation of data mining that aims to find useful patterns in large databases containing unstructured information in text format. To achieve this, text indexing techniques have been developed to deal with unstructured data. Text analysis involves a semantic analysis of information to obtain quality data results. Depending on what needs to be extracted from the text, different analysis techniques can be combined or repeated to obtain relevant insights for the user.

The main challenge of text mining is the complexity of natural language. The same word can be ambiguous and have different meanings depending on the accompanying words or even the context in which the sentence is inserted. While it is undesirable to completely remove ambiguity, as it provides flexibility and is intrinsic to human language, it is desirable to control this challenge or be aware of this factor in interpreting the results. There are various text mining techniques, depending on the type of information that needs to be obtained, the methods can be classified into four groups: Term-Based Method (TBM), Phrase-Based Method (PBM), Concept-Based Method (CBM), and Pattern Taxonomy Method (PTM) (Gaikwad et al., 2014).

In the TBM, the document is analyzed, considering each word with semantic meaning, or term. One of the main problems of this method is the ambiguity and synonyms of the words, so the result may not always meet the user's needs (Feldman et al., 1998). As phrases present greater semantic information, with reduced ambiguity, they are used in the PBM. In contrast, phrases

have less statistical information and lower occurrence frequency than terms, leading to many noisy phrases being included in the analysis (Ahonen et al., 1998).

Text mining techniques are typically based on statistical analyses of words or phrases in the document. This analysis records the importance of each term in the document. However, terms with the same frequency in the same document can contribute in different ways to understanding. Terms that best capture the semantics of the text should receive more importance. To achieve this, terms should be analyzed at the level of the sentence and the document, a technique introduced by the CBM. This model can efficiently discriminate unimportant terms from essential terms in understanding a document (Shehata et al., 2009).

Text documents can also be analyzed based on patterns. Although discovering patterns is difficult and often inefficient, it can offer excellent advantages in the text mining process. The difficulty associated with this method concerns patterns with potential value for analysis that often have too low a frequency to be detected, or the excess of short patterns, mostly noisy, leading to interpretation problems. The PTM is an effective technique for discovering patterns in text documents. It was initially proposed to deal with interpretation problems or low frequency in text mining (Wu et al., 2006).

Regardless of the final model to be used, the process of analyzing text documents goes through several intermediate stages to obtain interesting insights. The stages typically present in a text mining process are information extraction, summarization, categorization, clustering, and information visualization. Each of these stages applies different techniques and technologies to deal with the information it receives and to ensure it returns the appropriate output for the next phase.

In the initial stage of text document processing, sentences are semantically interpreted. This process of analyzing unstructured data is called information extraction, from identifying the main phrases and how they relate to the rest of the text. For this task, the model searches for predefined patterns in the text. Following this process, the obtained information is categorized. Categorization is a supervised learning type, as it uses known examples to automatically categorize the new data sample. Statistical classification techniques are used in this phase, such as Naïve Bayesian classifier, Nearest Neighbor classifier, and Decision Tree (Gaikwad et al., 2014).

Another technique for grouping data is called clustering, where documents with similar content are in the same group (called cluster). Unlike categorization, this is an unsupervised learning process, where groups are created based on similarities between the documents and not pre-defined in the training. The expected ideal outcome is that documents with greater similarity are within the same cluster, and between clusters, documents with different characteristics (Allahyari et al., 2017).

The final stages of the generic text mining process consist of data visualization and evaluation. Visualization techniques can be used to improve and simplify the discovery of relevant results. Flags are typically used to determine the categories of text documents, and the frequency of each term or phrase is also analyzed. In summarization, the amount of information in the document is reduced, so that only the most important points for understanding the text are retained (Gaikwad et al., 2014).

2.3.1 Data Pre-processing

Data pre-processing is the first step in a text mining process and has a huge impact on the results that can be obtained from a dataset. This section will present the main techniques for text pre-processing: removing stop-words, removing numbers and special characters, and stemming.

When analyzing text documents, the most common words that make them up are articles, prepositions, and pronouns, which do not add much value in understanding the document. These terms are called stop-words. Removing these words is a common practice to reduce the size of the dataset and improve the efficiency of text analysis algorithms (Vijayarani et al., 2015).

Another factor that can have a great influence on the results obtained from text mining is numbers or special characters in the text. Thus, it is of interest to remove these characters in the pre-processing stage. In addition, for standardization of the document, words can be converted entirely to uppercase or lowercase letters.

To avoid related words being considered separately in the analysis of the text document, an established practice is to remove suffixes from words. This technique aims to reduce the inflected or derived form of words to their basic form (root), this process is called stemming. With this tool it is possible to reduce the dimension of the dataset by grouping words into their base words. Lemmatizing is a similar technique but uses analysis of vocabulary and morphology of the word to return them to dictionary form (lemma). An advantage of this approach is the possibility of grouping synonyms allowing for even greater cleaning of the document before analysis (Khyani et al., 2021). Figure 2 presents the different results obtained using these methods on the same set of words.

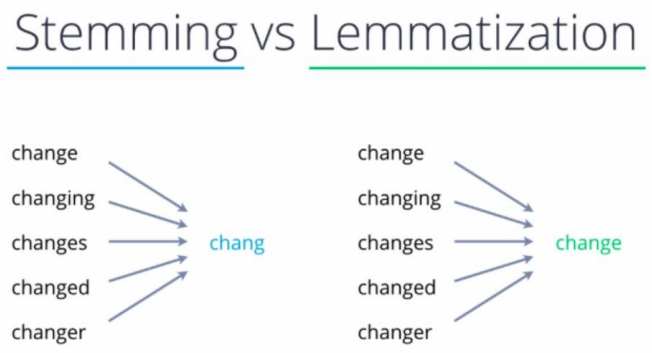


Figure 2: Example results for stemming and lemmatization (Khyani et al., 2021).

2.3.2 Data transformation and reduction

Since machines cannot understand natural language like humans and deal better with quantitative data than qualitative data, transforming text documents into numerical data is a step that ensures better use of the dataset. There are several consolidated techniques for this purpose, the main ones being Bag of Words and word embeddings.

In the classification of text documents, Bag of Words is a vector representing the number of occurrences of each word, this information is also called the document's histogram. Thus, this technique converts the textual information of the document into an attribute vector where each dimension is the absolute frequency of each of the words that make it up.

One such approach that has gained significant attention in text mining research is the use of word embeddings. These embeddings are generated by mapping words to high-dimensional vectors, allowing for the capture of semantic and syntactic relationships between words.

The word embedding models are typically created using neural networks, which are trained on enormous amounts of text data. By analyzing the co-occurrence patterns of words in each text corpus, the neural network can learn to associate words that are often used together and assign them similar vector representations. This enables the computer model to capture the nuanced relationships between words, even when they are used in different contexts. With the help of word embedding, machines can perform a variety of natural language processing tasks, such as sentiment analysis, text classification, and machine translation, with increasing accuracy and

efficiency. As such, word embeddings have become a valuable tool in the field of computational linguistics and are being used in a wide range of applications (Onan, 2019).

Using word representation as fixed-length vectors as done by word embeddings it is possible to find hidden similarities in vectors at both a syntactic and semantic level. These similarities between vectors allow finding common relationships between words that correspond to arithmetic operations between vectors (Almeida & Xexéo, 2019). Figure 3 presents in a simplified way the relationship between words and the corresponding relationship between vectors.

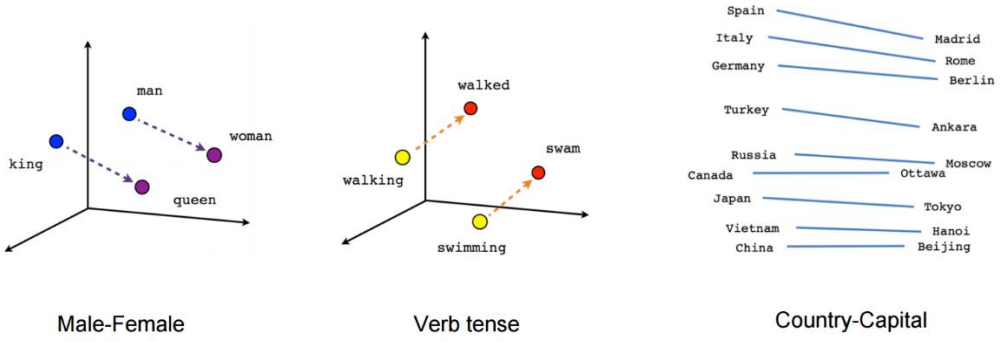


Figure 3: Relationships on word vectors (Sarwan, 2017).

2.3.3 Data evaluation and visualization

In the final stage of the text mining process is when the actual data classification and pattern discovery occurs. One of the most widely used techniques for evaluating text documents is clustering. This is an unsupervised data grouping technique where the goal is to group data into groups called clusters that have similar characteristics to each other. The use of evaluation methods allows for more accurate determination of clustering tendency, its quality and selecting the best number of clusters for each case.

With unsupervised clustering, the goal is to determine “natural” groups in the dataset. Thus, different techniques can result in different groups since the way data is analyzed is modified. To evaluate different methods, a good clustering is generally defined as one in which elements are as similar as possible within the same cluster and as different as possible between clusters. According to Omran et al. (2007), the most widely used clustering methods are divided based on two main techniques used, hierarchical and partitional.

The hierarchical clustering method uses merging or splitting techniques on the dataset to determine the final clusters. The so-called divisive algorithms start with the entire dataset in a single group and progressively use splitting techniques to create and restrict clusters based on patterns in the data. In the case of agglomerative algorithms, merging techniques are used on an initial dataset where each item starts associated with a cluster by similarity clusters are joined until a result is obtained (Omran et al., 2007).

Several agglomerative methods have been proposed differing in how the most similar clusters are defined. Single link algorithms join clusters where the distance between their closest elements is the smallest. In complete link algorithms it is the distance between the furthest elements of clusters that is considered to select similar clusters and join them. In general, complete link algorithms tend to generate more compact clusters compared to single link ones. Another less popular method of evaluating similarity between clusters in agglomerative technique is from centroid distance. A disadvantage of this method is that cluster size relationship is lost in evaluation. Another method used is Ward linkage, in which the distance between two hierarchical clusters is calculated as the increase in the sum of squared errors, that is, the grouping steps are sequentially defined to minimize this value. In this method, the error

is defined by the distance between each element and the centroid of the cluster to which it belongs. (Omran et al., 2007).

Partitional methods are used to divide the dataset into a specific number of clusters by minimizing certain criteria. The technique used for this purpose is to randomly determine centroid positions then move on to an iterative process where elements belonging to each cluster by proximity are calculated and centroid position recalculated. The process repeats until a stopping criterion is reached. This technique is known as K-means and is widely used due to its ease of implementation and satisfactory results. The disadvantages of this process are that all elements have equal importance in characterizing clusters and these, as they are calculated from distance to centroid, assume spherical shape which may not be sufficient for some datasets (Omran et al., 2007).

One of the most important steps in the clustering process is determining the number of clusters into which one wants to divide the dataset, which can be extremely challenging when dealing with unstructured or unknown data. To make this decision analytically there are several clustering evaluation techniques capable of determining the ideal number of clusters and analyzing data tendency. Figure 4 presents different datasets and their ground-truth clusters to illustrate how different approaches are needed for different datasets.

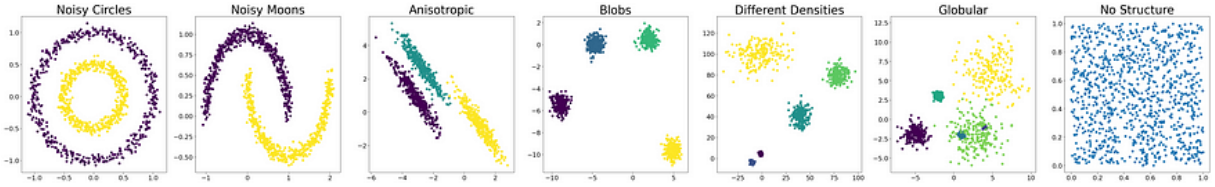


Figure 4: Different clustering approaches (Taskesen, 2023b).

The silhouette coefficient is a widely used metric for evaluating the quality of clusters in clustering analyses. It measures how well an object fits into its assigned cluster compared to other clusters. The coefficient value ranges from -1 to 1, where values close to 1 indicate that the object is well assigned to the cluster and values close to -1 indicate that the object may be better in another cluster. This metric is particularly useful in choosing the ideal number of clusters in a dataset and can be applied to various clustering methods including k-means and hierarchical (Aranganayagi & Thangavel, 2007).

Another metric applied to evaluate the quality of resulting clusters is the Davies-Bouldin index (DBindex), which can be described as the ratio between within-cluster distances and between-cluster distances. A score between 0 and 1 is assigned where low values represent better assigned clusters. A disadvantage of DBindex is that it frequently overshoots as lower values are more common with a growing number of clusters (Xiao et al., 2017).

It is also possible to evaluate clustering quality from the distance of elements in a group to their respective centroid. Variance (SSE) can be used, sum of distances between each observation and centroid squared, to consider elements closer to centroid as better assigned than those on the periphery. One way to determine the ideal number of clusters is to calculate the average value of this for several numbers of clusters and what has become popular as the elbow method can be used. This visual method consists of graphical analysis of variance average by corresponding number of clusters. Figure 5 presents an ideal situation for using this method where an inflection point can be easily identified. Mathematically determination of inflection point can be obtained from derivative method. This method consists of calculating the second derivative of curve where optimal point is where second derivative is maximum (Shi et al., 2021).

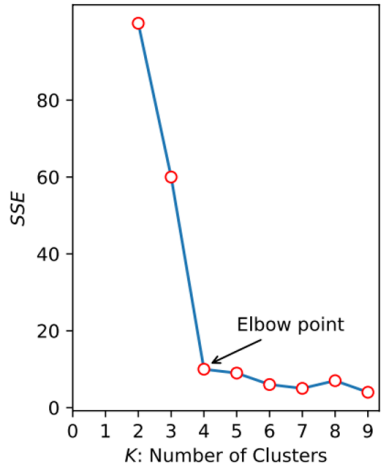


Figure 5: A visual curve with an explicit elbow point (Shi et al., 2021).

3 Problem descripiton

In the past, companies had homogeneous products in their portfolios with a limited range of differentiation. However, as more companies consolidated in the market, the supply of products and brands grew. As a result, retail companies had to adapt, and understanding their customers and their buying behavior became essential to ensure the availability of products that meet their needs and desires.

The acquisition of knowledge pertaining to customer behavior has become a crucial practice in various sectors of the economy, particularly in retail. Data analysis has emerged as a key technique for extracting meaningful insights from large volumes of data, thereby enabling businesses to make informed decisions (Chen et al., 2012). The retail industry has increasingly relied on data analysis to understand consumer behavior and develop more effective marketing strategies (Bradlow et al., 2017).

With the rapid evolution of technology and the increasing volume of data generated every day, data analysis has become increasingly complex. Therefore, it is essential to have data analysis tools capable of dealing with large volumes of information (Chen et al., 2012).

Promotional campaigns are important tools adopted by retail companies to attract and retain customers, increase the average purchase ticket, that is, the total amount spent, and boost sales of specific products (Mendez et al., 2015). However, often business knowledge alone is not enough to plan and evaluate these strategic decisions. Through data analysis, it is possible to identify consumption patterns, understand the preferences of the target audience, define the best time to carry out promotions, and evaluate the performance of campaigns. Thus, data analysis contributes to the success of promotional campaigns, increasing the efficiency of investment and improving the customer experience.

3.1 Promotional campaigns in MC

An important marketing strategy adopted by retail companies is the adoption of promotional campaigns, to promote the relationship with the customer base and maximize business profits (Mendez et al., 2015). In MC, there are two types of promotional activities: segmented and non-segmented. Segmented campaigns are created based on the definition of a product and a target audience, and only customers with pre-determined buying behavior will receive these offers. Thus, knowing the consumption pattern of the target customers for the selected products, it is possible to estimate with relative precision the expected sales impact with this action and consequently estimate the necessary stock in the stores during the campaign period.

On the other hand, there are non-segmented campaigns, which are carried out for all customers, regardless of any conditions. The non-segmented promotional campaigns in MC are carried out jointly with a brand or product, usually a new product being introduced to the catalog or limited products associated with special campaigns from their manufacturers. This type of campaign is the focus of this work.

In massive actions, loyalty stamps campaigns are highlighted. In these promotional activities, a consumption value is stipulated for which the customer receives a stamp, for example 20 euros. Thus, the higher the customer's expenditure, the more seals he/she receives. In the example used, a customer with an expenditure of 70 euros would receive 3 participation stamps. These stamps can be used to exchange for participating products in the campaign. Figure 7 shows the available products and how many stamps were required to obtain them, in the last campaign of this type, in which the product catalog was from the brand KitchenAid.



Figure 6: KitchenAid stamp campaign.

Despite promotional campaigns being a valuable tool for establishing a relationship with the customer base and encouraging their recurrence, they can pose risks to the business if poorly planned. With reduced profit margins on participating products, the ultimate impact of these campaigns is expected to be positive, avoiding losses for the company.

In general, a good promotional campaign is one that can attract many customers and increase their overall consumption in the store. To ensure the participation of many customers, it is necessary to understand their purchasing behavior, needs, and desires. With this knowledge, it is possible to direct campaigns to the right customers and ensure that they are relevant to them. Increasing overall customer consumption is associated with an indicator of incremental sales, meaning that the impact of the campaign on the customer's average ticket value is studied.

So, the expected result of carrying out a promotional campaign of this type is that customers, such as the ones in the previous example who made a purchase of 70 euros and would receive 3 promotional stamps, due to their level of affinity with the campaign, may choose to spend an additional 10 euros to ensure an additional stamp. In simplified terms, the maximum value that a customer spends on incremental sales can be related to their affinity with the campaign.

The sales forecasting model used at MC runs at the beginning of each year, predicting total sales for the current year. To adapt the results obtained to specific situations and trends, corrections to the initial forecast are made every two weeks. This model considers various factors, such as the time of year, temperature, holidays, and vacations, as well as the expected purchasing behavior for each customer segment. Promotional activities have a huge impact on the business, and knowledge of the campaigns is necessary to study their impact on total sales.

Currently, the sales forecasting automation considers the existence of promotional campaigns in a generic way, signaling the periods in which a promotional campaign will take place. Promotional activities are usually divided into phases with different impacts, including advertisement, start, and end of the campaign. The distinct phases of a promotional campaign present distinct impacts on the business in terms of customer engagement and consequent

increase in sales. Thus, the campaign phase is a crucial factor influencing the sales forecasting model.

Although promotional campaigns have very similar structures and proposals, the participating products and target audience result in non-standardized impact. Therefore, based on the history of past campaigns and business knowledge, generic and informal categories have been created for the campaigns, defining the expected impact as mild, moderate, or high.

In summary, the parameters related to promotional campaigns used in the sales forecasting automation are associated with the existence of promotional activities during the period, the phase of the campaign occurring during that period, and the expected level of impact. In segmented campaigns, defining the level of impact is a simpler task since the participating products are already consolidated in the business catalog, meaning that the purchasing behavior of the customer base is known for the products selected. In contrast, non-segmented campaigns are associated with new or exclusive products, making it difficult to predict consumer reactions. Therefore, the solution for formalizing these campaigns, predicting sales, and inventory is to find a previous campaign with similar characteristics.

A strategy used by companies to facilitate this process is through customer segmentation, which consists of grouping the customer base by common characteristics (Marcus, 1998). From this strategy, the business can better understand its customers, being able to create actions for specific groups, which offers several advantages in terms of planning than carrying out actions for each customer individually.

3.2 Customer segmentation in MC

Given the wide range of advantages resulting from the use of customer segmentation, this is a widely used strategy at MC. At MC, not only a single segment is used to characterize the customer base, but rather a set of several segments that define the characteristics and behaviors of a customer. The segments used group customers based on different criteria, so that when grouped together, they generate several possible combinations, based on demographic characteristics, frequency, volume of purchases, among others. The more segments are used, the more personalized the characterization of a group of customers will be, since the probability of a set of segments repeating decreases.

The division of the consumer base into segments presents several advantages for the business during the planning and execution of promotional campaigns. By determining the groups of customers with the greatest affinity to the promotion, better-founded decisions can be made, such as contacting customers for whom the campaign is relevant, as well as evaluating the feasibility and impact of carrying out this campaign in advance.

Several different factors determine a customer’s buying behavior during the period of promotional campaigns. Therefore, it is extremely challenging to estimate their reaction. Customers with purchases in the same product niche may have very different tendencies for promotional campaigns, even if the product in question is extremely relevant to their purchasing pattern.

To simplify the problem, a single segmentation was selected for analysis. The chosen segmentation is one of the most generalist, grouping customers into 8 groups by their usual shopping basket. Although these groups are not defined by unique characteristics of the customer, it is possible to identify niches with growth potential among consumer interests, offering guidance for business strategies. The niches that can be identified from the products normally purchased by each customer are related to their lifestyle, creating groups that allow identifying customers with a tendency to buy healthy or premium products, customers with children, dietary restrictions, among others.

To assist in calculating the affinity of segments with the promotional campaign, some data analysis techniques will be used. These tools allow you to obtain valuable insights that often go unnoticed in superficial analyses of historical data. In addition, these techniques can help identify hidden patterns and trends in data that can be useful in guiding strategic business decisions.

However, despite the segments used being well established in the business structure, studying the behavior of these customers in relation to new products can be a challenging task. As promotional campaigns are usually carried out with temporary or new products in the catalog, reliably relating them to customer segments is a task of great value for the business.

Since campaign definition comes from using web scraping techniques, word embeddings and clustering it is necessary to find a way to relate customer segments with the result of this process.

3.3 The project

With the high competition in the retail sector, digitalization and technological evolution have become a necessity to maintain competitive advantages (Bradlow et al., 2017). Thus, this project aims to obtain greater knowledge about non-segmented campaigns carried out by MC, with the objective of determining more accurately the influence of these actions on the company's total sales.

The proposal is to automate the process of categorizing campaigns, nowadays based solely on business expertise. Defining the level of impact of an action by comparing it only with previous actions is an overly simplistic way of treating a tool of such enormous influence on the business. More accurate predictions can be achieved if categorization is expanded beyond mild, moderate, or high impact. Furthermore, by incorporating data from multiple sources, you can gain a more comprehensive understanding of the factors that influence your business and make more informed decisions. The diversification of data sources is also a way to reduce the bias of an analysis, leading to more accurate and reliable results.

The similarity sought between campaigns is relative to their type, target audience, and participating products. However, the manual nature of this stage offers certain uncertainties, such as subjectivity in categorization, and, as defended by Ziafat (2014), some patterns found using data mining models can go unnoticed even by business experts.

As participating products in massive promotional campaigns are new or exclusive products, it is unknown how consumers will behave with this action, since there is no sales history. Although it can be expected that customer behavior with related products will be repeated in a comparable way, determining which products are those can be a challenging task. One way to find related products is to categorize them into defining keywords, where products that share the same categories can be considered related.

Determining the category of a product can be done in several ways, such as through analysis of the product description on its official website and keywords used in its commercial communications. However, these methods can be imprecise or require a lot of manual effort, especially on sites with enormous amounts of products. In this context, web scraping can be an interesting solution, as it allows data about the product to be collected from various sources in an automated and fast way. Thus, this methodology can be an initial step in obtaining information about the product for subsequent analysis and categorization.

With an extensive database extracted from different web pages, by processing the information with text mining techniques and tools, it is possible to find patterns in terms or phrases characterizing that product, which can be used as categorizers. An advantage of this method is the possibility of expanding the research to the brand, since not all campaigns are for a single

product, but rather for a catalog of products from a particular brand, as in the example of KitchenAid presented in Figure 6. With the campaign characterized, we proceed to calculate its affinity with the customer segments existing in the business.

3.4 Objectives

With the proposal to improve the planning stage of promotional campaigns, the main objective is to calculate the affinity of promotional campaigns with existing customer segments in the business. The validation of the results obtained will be done from the participation of the customer base in the campaign, for the selected examples. A customer is defined as a participant in the campaign if they exchanged stamps for one of the promotional products. Each customer is considered only once, regardless of the number of exchanges they make.

To achieve the main objective, there is an intermediate objective of defining the promotional campaign from participating products using web scraping and text mining techniques. The quality of this stage is crucial for the overall success of the project, as this information will be used to calculate affinity with the customer base.

In addition to the main objective of calculating the affinity of promotional campaigns with existing customer segments, there is another intermediate goal which is to characterize the customer base in a way that it is comparable with the result obtained for campaign characterization. This means defining it based on products. Figure 7 presents the simplified methodology for calculating affinity. On the left, the goal of characterizing the segments is represented, and on the right, the characterization of the campaign. The referred KDD process will be explored in greater detail in the following chapter. The figure also represents, in a color scale, the affinity found between the presented segments and the campaign.

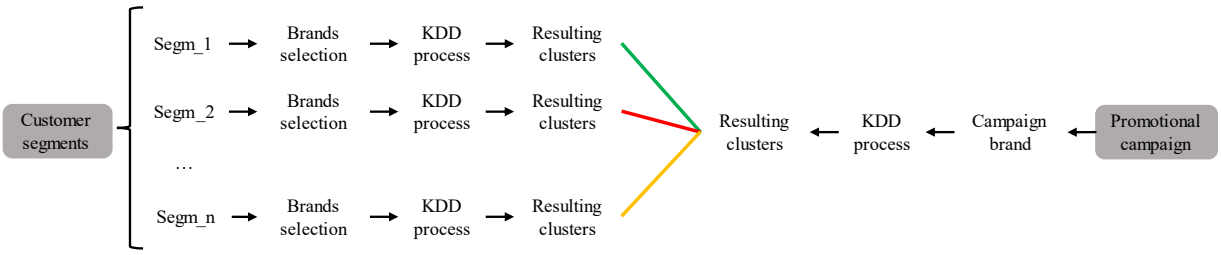


Figure 7: Methodology for calculating affinity between campaigns and customer segments.

4 Methodology

The solution to the problem was developed using the Python programming language. This language was chosen due to its simple syntax and readability, which facilitate understanding and development. Additionally, the vast number of libraries and frameworks available for Python, built to simplify established techniques, are key factors that have made Python widely adopted and popular.

Thus, from adapting the model proposed in Figure 1 to the specified problem. The structure of the proposed solution consists of using the Google search API to return the most relevant web pages for an input term. Next, all textual information from the most relevant results is extracted, this data is pre-processed and then goes through the clustering process. Figure 8 presents a simplified representation of the process using the search term “KitchenAid” as an example. This solution will be explored in the next section.

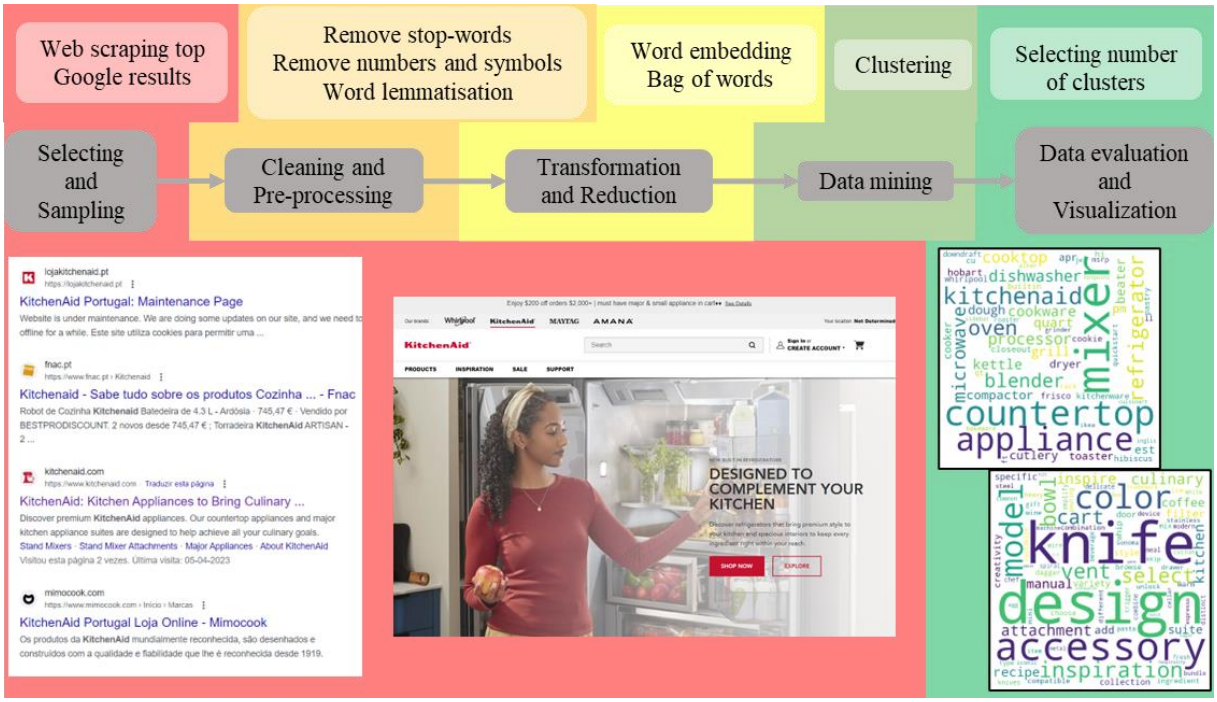


Figure 8: KDD process adapted to the problem.

4.1 Campaign definition

After defining the campaign, the next step is to select and sample the data. This is a crucial step in the methodology as it determines the quality and relevance of the information that will be used in the subsequent stages. The data selection process involves identifying and choosing the most relevant sources of information, while sampling involves determining how much data to collect from these sources. The goal is to gather a comprehensive and representative sample of data that accurately reflects the characteristics of the search term in question. This data will then be used for extraction and analysis in the following stages of the methodology.

4.1.1 Data selection and sampling

Due to the proposal of the project, the described problem is of a completely unsupervised nature, as labeled historical data will be used only for validation of the results. Therefore, it is necessary to collect as much information as possible that characterizes the search term in question. As previously discussed, the internet offers a vast collection of data from various sources, some of which are potential characterizers of the search term and others are noise in characterization. Therefore, a rigorous selection of sources to be used is necessary.

In the presented solution, the consolidated Google search engine is used to obtain the most relevant results for the search. Thus, by using the company's search API it is possible to have simplified access to these results. The Google Search API works through HTTP requests where developers can send specific search queries to the API server. These queries can include keywords filters and other custom parameters to refine search results. The API processes these queries and returns results in structured format such as JSON or XML so developers can manipulate and display information as needed (Google, 2023). The most relevant results returned will be used as input for the next web scraping process.

To standardize the number of sources used for the next phases, the 10 most relevant search results where data collection by web scraping is possible will be used since many sites block this tool or only allow data collection through registration such as social networks. With this, the objective is to select sufficient data sources with relevant information about the input term. We then proceed to extract textual information from selected web pages using web scraping techniques.

To perform web scraping techniques Urllib and Beautiful Soup libraries were used in python. The Urllib library is a highly efficient and robust tool for handling URL manipulation and HTTP requests. Its wide range of features allows access to relevant data from various online sources such as web pages and Google search API (Zhao, 2017). The Beautiful Soup library is useful in extracting and analyzing HTML and XML data, being able to extract text information from HTML documents. Another differential of this library is its ability to handle complex and malformed web pages (Richardson, 2023).

4.1.2 Data cleaning and pre-processing

Cleaning and pre-processing of text plays a key role in preparing data for analysis. During the cleaning stage, it is common to remove unwanted elements such as stop-words, symbols, and numbers. As stop-words are high-frequency words but with low attached meaning, removing these terms helps reduce data dimensionality and focus on the essence of the text. In addition, symbols and numbers may not contribute to textual analysis and their elimination helps simplify text and reduce noise.

Another important process carried out during pre-processing is lemmatization. The lemmatizer seeks to identify the basic or root form of words known as lemma which helps treat different inflected forms of a word as a single unit facilitating text categorization and information

retrieval. Lemmatization also contributes to reducing text dimensionality since several variations of a word are simplified to a single term. Thus, cleaning and pre-processing of text helps prepare data for efficient and more accurate textual analysis.

To perform these tasks, the spaCy library cwas usesd. It is a free, open-source library for Natural Language Processing in Python. It is designed to be easy to use and to build systems for information extraction or general-purpose natural language processing. Its lemmatization component can help identify the basic or root form of words known as lemma which helps treat different inflected forms of a word as a single unit facilitating text categorization and information retrieval. Furthermore, it has a built-in list of stop-words for several languages (Smeg, 2023).

4.1.3 Data transformation and reduction

The data reduction and transformation stage in text mining plays a key role in the numerical representation of text documents enabling statistical analyses and the application of categorization algorithms. A common approach to performing this transformation is the Bag-of-Words technique where each document is represented by a vector containing the count of the words that make it up.

However, when the goal is to identify keywords that define a dataset it is necessary to consider that some words may influence more than others in defining a document. Some words that appear only once in each document although they may be important for overall understanding of content, are not necessarily good keywords. Therefore, as these words add little value to analyses for development of this project they will be eliminated.

With definition of characterizing words for each of URLs then proceed to union of this information from all sources used in analysis. With all information gathered it is possible to select words that best define the initial search term. Union of data from different URLs is done by creating a general characterizing vector for term using Bag-of-Words technique. At this stage words with total absolute frequency equal to 2 are eliminated since these originated from only one URL that is a word of little relevance in characterizing site and even less search term. Table 1 presents data structure used to relate words found as characterizing search term and respective relative frequency.

One problem with characterizing the search term from Bag-of-Words is the lack of sensitivity to synonyms and related words. Thus, several words may present little value to analysis when analyzed individually but in fact because they are related, they have much greater importance. In addition, words with the same absolute frequency are considered equally relevant as characterizers which is very difficult to be true especially for low frequency words which are usually the most common. To circumvent this issue a metric was created for defining word importance from word embedding and Bag-of-Words.

Word embeddings are a technique that allows words to be represented as dense vectors in a continuous dimensional space. These vectors capture the semantic meaning of words allowing us to calculate relationships between them. In other words, semantically similar words are represented by vectors that are close to each other in this space (Almeida & Xexéo, 2019). For this purpose, Gemsim library was used in Python.

The Gensim library in Python is a powerful tool for natural language processing and topic modeling. One of Gensim’s most popular features is its efficient implementation of the Word2Vec algorithm which allows converting words to their equivalent vector. These models capture semantic and syntactic relationships between words based on their context. The Gensim library provides a series of trained models to facilitate use (Řehůřek & Sojka, 2023).

Table 1: Words and relative frequencies.

Word	Relative Frequency
W_1	$Freq_1$
W_2	$Freq_2$
W_3	$Freq_3$
W_i	$Freq_i$

Available models differ in data source used as basis and number of dimensions used in final vector. Model selection is usually associated with objectives to be achieved and degree of project complexity. For this project dataset "glove-wiki-gigaword-300" was used created from Wikipedia page information providing representation of words in 300-dimensional vector space where each of these presents a semantic or syntactic aspect of word in question (Pennington et al., 2014).

Another concept used for defining importance metric is based on "similar_word" function present in Gensim library. This function is used to find words like a reference word in a word model. Thus, it is possible to find related words based on their vector proximity in word embedding space. In addition to related words a similarity score is also assigned calculated based on vector distance between words in vector space (Řehůřek & Sojka, 2023). For calculation of importance metric this function is used on all words found as characterizing search term also relative occurrence frequency calculated from absolute frequency is used. Table 2 presents the final structure of similar words resulting from the use of Gensim library.

Table 2: Similar words.

Word	Similar Word	Similarity Score
W_1	SW_1	$Score_{11}$
W_1	SW_2	$Score_{12}$
W_1	SW_3	$Score_{13}$
W_2	SW_1	$Score_{21}$
W_2	SW_3	$Score_{23}$
W_i	SW_j	$Score_{ij}$

Thus, the process of finding related characterizing words uses similar words resulting from the Gensim function to relate them. To keep words with higher frequency as more representative the process of calculating the importance metric uses words with lower relative frequency as "boosters" for words with higher relative frequency. Therefore, the method of calculating the importance metric can be defined from the equation:

$$Importance_i = Freq_i + \sum_{j=1}^{n}\sum_{k=2}^{m} Score_{ij} * Score_{kj} * Freq_k, \quad Freq_i \geq Freq_k \wedge \ j \neq k \tag{1}$$

With the calculation of the importance metric for all characterizing words, it is expected as a result that words with low relative frequency and a high number of related words gain greater importance in the analysis. In addition, a decrease in the number of words with equal relevance is expected.

4.1.4 Data evaluation and visualization

In the final stages of the text mining process, the collected and properly processed information is used in the evaluation process. In this stage various models and statistical techniques are applied for discovery of quality insights in dataset. Data visualization techniques are also used for a format that facilitates understanding and interpretation of results.

In text mining applications, clustering is one of the most widely used and established techniques for evaluating text documents. It is an unsupervised data grouping technique where the goal is to group data into clusters that share similar characteristics. By using evaluation methods, we can more accurately determine the clustering tendency, assess its quality, and select the optimal number of clusters for each case.

Among the various existing clustering techniques an agglomerative model was used in solution. These models belong to a group of clustering algorithms that build nested clusters by successively merging data points. This hierarchy of clusters can be represented as a tree diagram known as a dendrogram. The top of tree is a single cluster with all data points while bottom contains individual points.

The reason for choosing this technique in this work was the capability of this method to separate the data set under analysis with a smaller number of clusters. The other clustering techniques evaluated, when forcing one more cluster than the result, the new group contained elements from more than one previous cluster, while with the agglomerative model, the new cluster, in most of the usual cases tested, contains values from only one previous cluster. Figure 9 illustrates fictitious results in a random data set and clusters that show the behavior that led to the choice of the agglomerative technique.

Agglomerative clustering essentially starts from an individual cluster (each data point is considered as an individual cluster also called leaf) then each cluster calculates its distance with others. The two clusters with the shortest distance between them merge creating what we call a node. Newly formed clusters again calculate distance of members of their cluster with another cluster outside their cluster. The process repeats until all data points are assigned to a cluster called root (Omran et al., 2007).

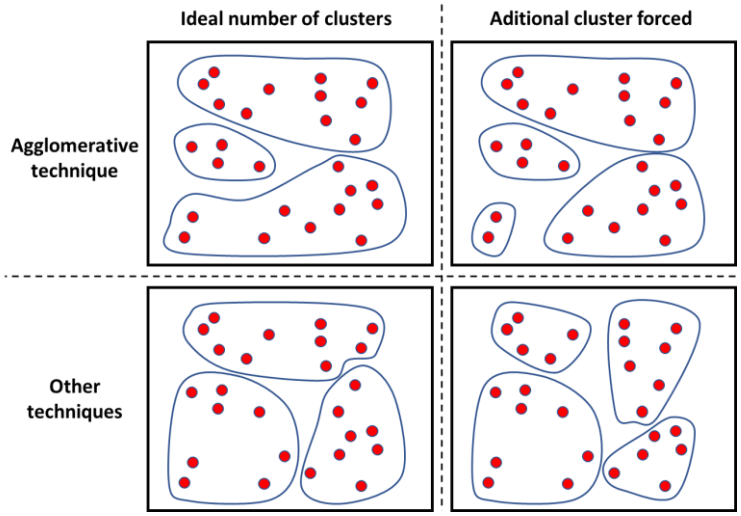


Figure 9: Comparison between agglomerative technique and other techniques.

The Clusteval library is a Python package developed to determine and evaluate clusters in an optimized way. Widely used for allowing selection of pre-defined clustering models and evaluation techniques in a simplified way (Taskesen, 2023a). Due to its application to evaluate the quality of the clustering process and determine the ideal number of clusters this library was used in development.

To determine the ideal number of clusters to be used in analysis the quality of clustering process was evaluated for a limit of 10 groups using the silhouette coefficient. The silhouette coefficient measures how well each object is grouped with its own cluster compared to other clusters. It ranges from -1 to 1 where a high value indicates that the object is well grouped with its own cluster and poorly grouped with neighboring clusters. The selected configuration is where most objects have a high value (Aranganayagi & Thangavel, 2007).

The dataset used for clustering is composed of vectors resulting from use of Word2Vec algorithm provided by Gensim library. With clusters defined, a visualization technique implemented for evaluation and presentation of clusters with the most common words in characterizing search term is WordCloud. This method displays cluster elements so that font size is adjusted to visually represent the importance metric defined for each term.

From the importance metric, Equation 1, it is also possible to calculate cluster centroids to reflect importance of words that make it up in characterizing document. This value can be calculated from a weighted average of vectors that make up a cluster by importance metric. Resulting in a centroid located closer to more significant words.

Thus, from this method it is possible to reduce dataset originating from internet composed of hundreds of unique items and transform it into at most 10 vectors capable of characterizing initial search term.

4.2 Customer assignment

After the characterization process of the promotional campaign using the webscraping process described in the previous chapter, an analysis of the affinity of this campaign with the customer base is conducted. In this way, it is possible to obtain relevant insights for the business, facilitating the forecast of the necessary stock during the campaign period and evaluating its success both economically and in ensuring that it is interesting for its consumers.

A widely used strategy by retailers to characterize their customers is through segmentation. Customer segmentation is the process of dividing customers into groups that share characteristics. This can include demographic, geographic, psychological or behavioral factors. By creating segments in their consumer base, the business is able to better understand its customers, being able to create targeted actions for specific groups, based on their needs, interests or preferences. These actions can range from offering products desired by customers to sending communications efficiently and using the customer's preferred medium, resulting in greater affinity and customer loyalty with the business (Marcus, 1998).

Efficient communication can be achieved by sending it to those who will really be interested and convert into purchases, focusing on the right customers avoids dissatisfaction with the business due to excessive irrelevant communications. As a result of these actions, company resources are optimized.

4.2.1 Customer segments affinity model

The application of advanced data analysis techniques on customer segment information can play a key role in predicting participation in promotional campaigns, based on customer segment information. However, to increase the accuracy of forecasts, it is necessary to use comparable data. Therefore, it is essentialp to develop a methodology to convert the customer

group determined by segmentation into values comparable to the vectors equivalent to the centroids of the word clusters that characterize the campaign.

Considering that promotional campaigns are carried out in partnership with specific brands, whose names are used as input in the campaign classification model, one possible approach is to classify customer segments also based on brands. In this way, the characterization of segments is carried out from the most consumed brands by them in the last year. The five most purchased brands for each segment are then selected, which represent this group.

It is important to note that basic needs or recurring consumption products tend to be acquired regardless of the customer's segment. Therefore, brands associated with these products should be avoided when creating distinct groups. To define the segments based on brands, preferably these should not be repeated from one segment to another. Thus, to select the representative brands of each segment, they are sorted by total consumption and among the first n brands of each segment, the five most consumed that do not repeat between segments are selected, thus creating unique identifiers for the segments.

In this proposed solution, the relationship between customer segments and the promotional campaign is established from the relationship between the brands that identify a segment, and the brands present in the campaign. With this, it is expected to obtain an affinity index of the customer group with the campaign, which represents the similarity between campaign products and group characterizers of customers, thus associating participation rate with customer's usual shopping basket.

Despite campaign and segment characterization being at brand level, more information beyond names is needed. If a supervised analysis of the problem were carried out, it could be done from the market structure of brands. In this hypothesis, customer segments' purchasing categories would be studied, creating association rules, or forecasting models for defining affinity degree between campaign and customer segments. As it is an unsupervised analysis, affinity score was defined using mathematical techniques.

To obtain comparable data in defining segments, characterizing brands can be defined using the same methodology used in defining campaign. Using the same technique ensures similar data for both representative brands of segments and promotional campaign.

The process of characterizing segments then goes through stages described in previous chapter where brands are classified using web scrapping and word embeddings techniques. The result is brand definition in characterizing vectors; these vectors are defined from centroids of characterizing word clusters found online. Thus, starting from customer base, it can be transformed into groups defined from most consumed unique brands by this segment.

With all brands represented by word vectors, customer segments are now defined as set of vectors from brands identified as characterizing this group. In this way, conversion of customer segments into data related to obtained campaign definition is carried out.

4.2.2 Affinity score

After the operations performed, the data to be compared are vectors representative of the characterizing terms of the brands, so the calculation of affinity between campaign and segments can be interpreted as the calculation of affinity between vectors. There are several ways to calculate the affinity score between vectors. One technique that can be used is the calculation of the angle between them.

The cosine of the angle between two vectors can be calculated through the dot product of the vectors divided by the product of their magnitudes. The equation for calculating the cosine between two angles is represented below, where A and B are vectors of the same dimension. It

assumes a value of 1 when the angle between the vectors is zero, corresponding to the maximum affinity score.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} \tag{2}$$

The similarity calculation by cosine method offers excellent advantages when comparing high-dimensionality vectors, as is the case with mathematical representation of textual data.

To adapt use of this evaluation method to needs of this project, affinity score value must be a value between 0 and 1, for angles between 0 and 180º. Normalizing the result between 0 and 1 facilitates interpretation, with the resulting value being the estimate of the percentage of participation for each group of customers, since negative affinity values do not have a real meaning. In addition, to value vectors with smaller angle between them, resulting score can be raised to an integer. Thus, final formula for calculating affinity score is:

$$affinity(A,B) = \left(\frac{\cos(\theta)+1}{2}\right)^n = \left(\frac{\frac{A \cdot B}{\|A\|\|B\|}+1}{2}\right)^n \tag{3}$$

With the changes made to the function, the period of the cosine function was doubled, and it varies between 0 and 1. In addition, the constant n allows to control how angles are interpreted in the affinity score. Based on experimental tests, it was determined that $n = 2$ for calculating affinity between vectors.

Figure 10 illustrates the effect of the exponent on the function, as well as compares it with a linear variation function, in the angular variation interval between vectors of at most 180º. The affinity function (3) is represented in blue, in green is represented the function

$$f(\theta) = \frac{\cos(\theta)+1}{2} \tag{4}$$

and in red the linear function

$$g(\theta) = 1 - \frac{\theta}{\pi} \quad . \tag{5}$$

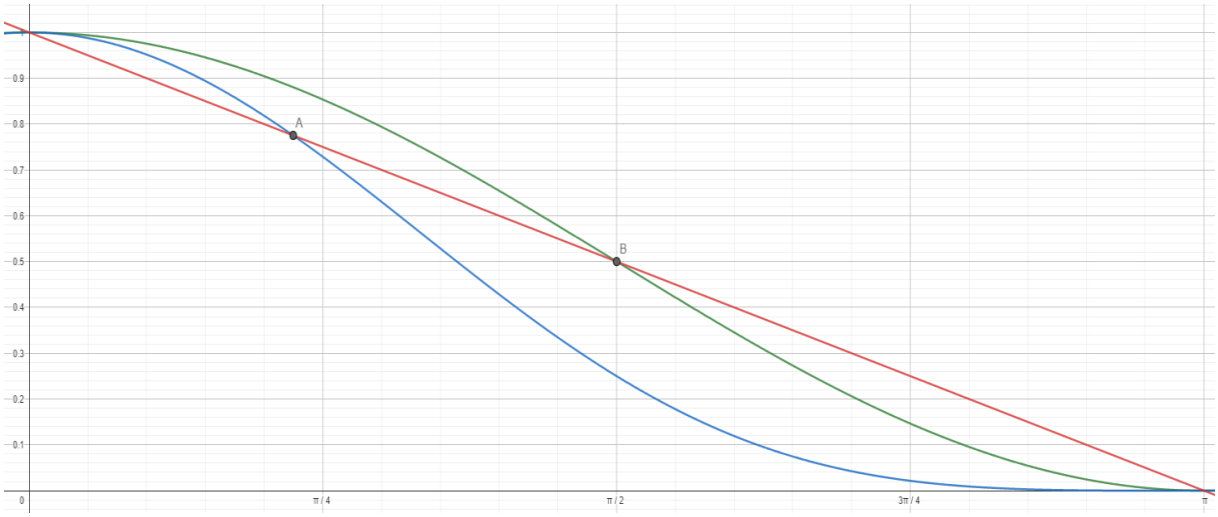


Figure 10: Comparison of the cosine function (green), similarity function (blue), and linear function (red).

Equation 4 assigns a higher affinity score than the linear function (5) for angles smaller than the intersection point $B = (90°; 0{,}5)$. In the case of the affinity function, point $C = (\sim 40{,}45°; \sim 0{,}78)$ defines the intersection. Using the exponent in the function allows to more easily isolate pairs of dissimilar vectors, assigning smaller score values.

With the definition of the formula for calculating affinity, the next phase is to calculate affinities between customer segments and campaigns. At this stage, as described in previous sections, both segments and campaigns are defined by representative vectors of their characterizing clusters. As between 2 and 10 groups are created in clustering process, there are between 10 and 50 vectors that define each segment for set of 5 distinct brands that characterize it. Similarly, campaign is defined by a set of 2 to 10 vectors.

Thus, to calculate affinity between a customer segment and a campaign, affinity score is calculated for each pair composed of a campaign vector and a segment vector. The average of these values is then used as final affinity value between analyzed campaign and segment. Expressing calculation of affinity score mathematically:

$$score = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} affinity(A_i, B_j)}{n * m} \tag{6}$$

In this equation, A_i represents a campaign vector and n is the number of vectors that characterize it, m represents the number of vectors that characterize the segment and B_i is the vector of this group being analyzed. This method is like ward linkage in how calculation is done. Using this technique as an analogy, calculation of affinity score is an average of compatibility of campaign vectors with clusters already defined for segments.

5 Results

To validate and assist in the development of the project, two past campaigns were used. These campaigns were chosen because they are the stamp campaigns described earlier as the focus of this work. In addition, these campaigns had a high participation rate, which provides good data for comparison and validation. The brands of the selected campaigns are KitchenAid and Smeg.

In addition to the KitchenAid and Smeg campaigns, the Hey Clay campaign was also used for result validation. This is a stamp campaign, but with a shorter duration and lower promotional participation than the previous ones.

5.1 KitchenAid

KitchenAid is an American home appliance brand owned by Whirlpool Corporation. The KitchenAid brand was first introduced as a brand name for an electric mixer developed by Hobart Manufacturing Company in 1919. In the early 1900s, Hobart began producing the first electrically powered machines for grinding food, including coffee beans, peanuts, and hamburgers. In addition to mixers, KitchenAid also offers a wide range of other home appliances, including food processors, blenders, coffee makers and juicers, kettles, countertop ovens, electric cookers and multi-cookers, toasters, and commercial products (Smeg, 2023).

Despite being a brand focused on producing kitchen appliances, the stamp campaign carried out by MC did not offer this type of product. The products available in the campaign were the kitchen equipment shown in Figure 6. This catalog is a special selection of reduced cost products to be offered as gifts.

This promotional campaign at MC took place between December 21, 2022 and March 12, 2023 at Continente Online and all Continente, Continente Bom Dia and Continente Modelo stores. Following MC's traditional stamp campaign model, for every €20 spent by the customer they were offered a stamp. From 5 accumulated stamps, the customer can make an exchange for participating products of the campaign. The exchange, in addition to using the stamp as currency, can be free or require addition of a value between €0.99 and €99.99, depending on product value and number of stamps used in exchange. Thus, a single campaign has several repercussions on business.

The first impact of a campaign is on the choice of business against competitors; customers may choose to buy at MC stores because they see this campaign as a differential benefit. Once a customer buys at one of stores, if the campaign attracts them, they may decide to spend more to reach a multiple value of €20 and increase number of stamps received, increasing incremental sales indicator. There is also microeconomy created from stamps as currency, feeding formal economy of product redemption for stamps and money, in addition to informal exchanges of stamps between customers.

Even when analyzing historical data, it is difficult to determine the total impact that carrying out a campaign has on business due to various repercussions presented. Therefore, for the development of this project, the metric used as basis for validation was customer participation

rate in product exchange. Participation rate is the percentage of customers with purchases during campaign period who made a stamp exchange.

Figure 11 and Table 3 present distribution of customers with purchases during campaign period divided by groups from segment selected for analysis. From the graph it is possible to verify the representativeness of each segment in the customer base. Segm_6 had the highest participation rate, at 22.51%, followed by Segm_7. At a lower level are segments Segm_2, Segm_3, Segm_4, and Segm_5. With a much lower participation rate than the other segments are segments Segm_1 and Segm_8.

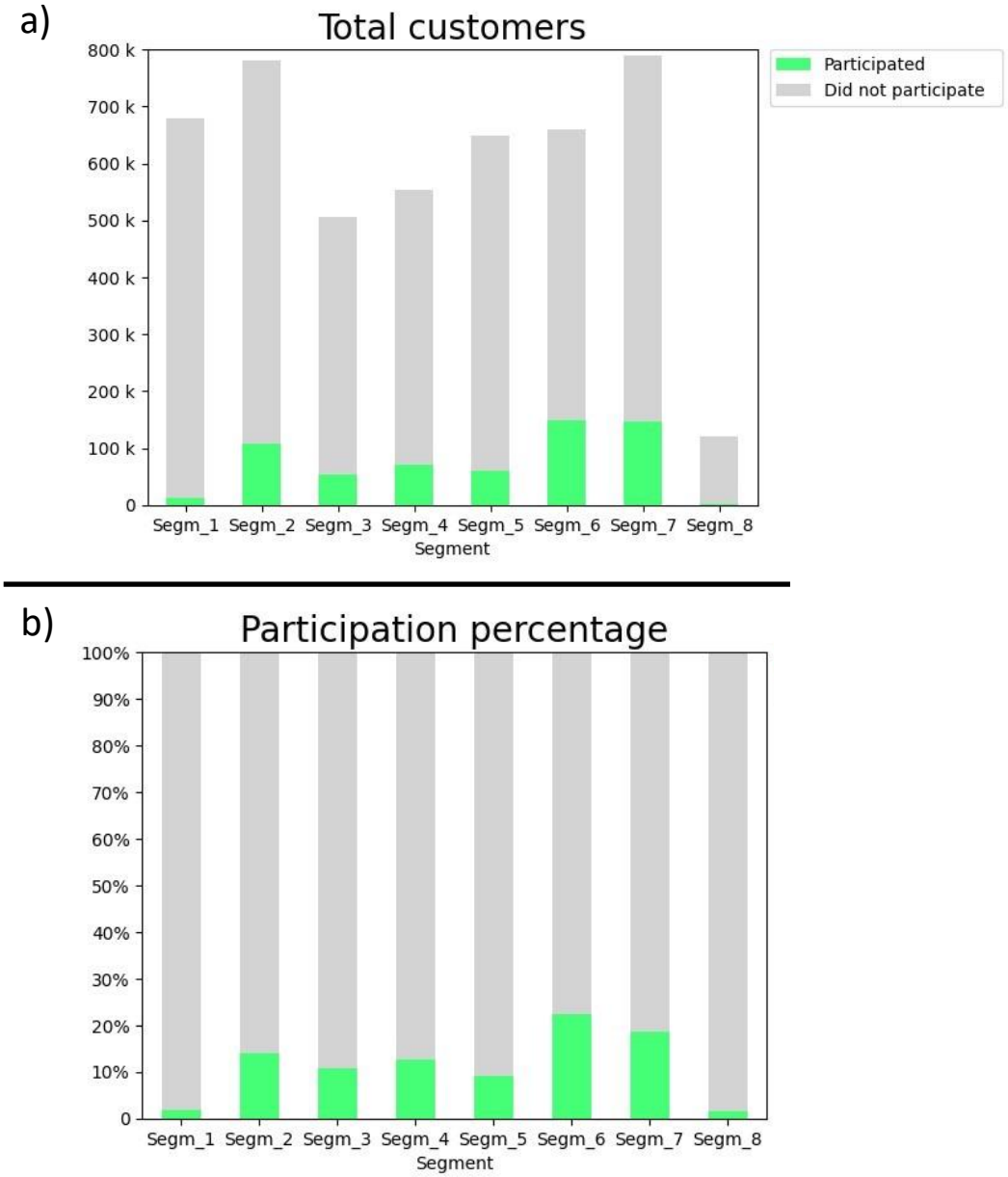


Figure 11: KitchenAid (a) Distribution by segment of the customer base; (b) Promotional participation.

Table 3: Distribution of active customers in the promotional period for KitchenAid campaign.

Segment	Total	Did not participate	Participated	Participation percentage
Segm_1	679578	667176	12402	1,82%
Segm_2	780457	672040	108417	13,89%
Segm_3	505295	451402	53893	10,67%
Segm_4	552588	482466	70122	12,69%
Segm_5	648047	589110	58937	9,09%
Segm_6	659180	510831	148349	22,51%
Segm_7	788620	642249	146371	18,56%
Segm_8	120680	118773	1907	1,58%
Total	4734445	4134047	600398	12,68%

5.1.1 Campaign definition

This campaign was the main sample used to create the promotional campaign characterization model. Being a brand related to home appliances and kitchen utensils, the expected result for characterizing this campaign by webscraping from the brand name is to define it from terms associated with this niche.

Table 4 presents the terms obtained from webscraping of the 10 most relevant web pages resulting from the search for the “KitchenAid” brand using the Google Search Engine. The search for this brand returns 602 distinct terms. The resulting terms undergo text preprocessing, and their absolute frequency is calculated.

At this stage, how related a term is to the brand is defined solely by its relative frequency on web pages from which data was extracted. As expected, many terms are related to each other and have the same relative frequency, which makes it difficult to analyze the real relevance of each term in characterization. The formula for calculating the importance score described earlier is then applied. At this stage, noisy words with no real meaning are also eliminated, resulting in a total of 582 terms. Table 5 presents resulting terms after this treatment. Although it is not necessary to calculate the importance score of each term in the next clustering process, this stage proved to be of enormous importance for calculating centroid of final clusters.

The resulting terms from data collection process by web scraping are then converted into representative vectors using pre-trained models. With information converted into quantitative values, clustering process can be carried out to identify relevant groups in these terms. For KitchenAid brand, model obtains 2 clusters as result, which are presented in Figure 12. In this figure, resulting clusters are represented as WordClouds with terms that compose them, and font size related to importance score.

Table 4: First 20 terms resulting from the search for the term "KitchenAid".

Rank	Term	Relative frequency	Rank	Term	Relative frequency
1	kitchenaid	2,70%	11	accessory	0,82%
2	appliance	2,08%	12	knife	0,81%
3	mixer	1,94%	13	shop	0,75%
4	product	1,58%	14	buy	0,75%
5	stand	1,21%	15	company	0,73%
6	PS	1,17%	16	select	0,73%
7	brand	1,11%	17	kitchen	0,72%
8	new	1,03%	18	countertop	0,71%
9	major	0,85%	19	close	0,68%
10	sale	0,83%	20	whirlpool	0,68%

Table 5: First 20 terms ordered by importance.

Rank	Term	Importance	Rank	Term	Importance
1	kitchenaid	0,024676	11	major	0,007381
2	appliance	0,018118	12	company	0,007108
3	product	0,017417	13	select	0,006796
4	mixer	0,016099	14	design	0,006778
5	stand	0,010293	15	knife	0,006747
6	brand	0,010028	16	offer	0,006674
7	sale	0,009421	17	accessory	0,006667
8	new	0,009265	18	kitchen	0,006643
9	buy	0,008877	19	whirlpool	0,006472
10	shop	0,007893	20	food	0,006204



Figure 12: Resulting clusters for the brand KitchenAid.

It is possible to observe that the results obtained are satisfactory for the expected values, obtaining a cluster with terms related to the kitchen and another mostly composed of noise. With two resulting clusters, the main cluster can be easily identified and isolated, but with a higher number of clusters, this task becomes extremely delicate. From tests carried out, it was observed that the clusters with the highest volume of words are the least relevant, since the relevant words are a minority in the data set. As noise is always expected to be present in the data set, if the least significant cluster is not eliminated, the final affinity may be affected by the similarity between the noisy clusters of the brands. Thus, to select the final clusters characterizing the campaign, the cluster with the most elements is eliminated.

5.1.2 Customer assignment

From the clusters selected as characterizers, in the case of KitchenAid only one (Cluster 1 in Figure 12), the affinity scores with the segments are calculated. At this stage, the segments are already defined in the same way as the campaign, from the vectors resulting from the characterizing clusters of the representative brands.

With campaign and segments in comparable data formats, it is enough to apply the calculation of the affinity score described in Equation (6. The results obtained, presented in Figure 13, represent the affinity of the campaign with each customer segment. From the results presented, it is possible to see that Segm_7 has the highest calculated affinity score, followed by Segm_8 and Segm_3. The lowest calculated affinity scores are segments Segm_2 and Segm_5.

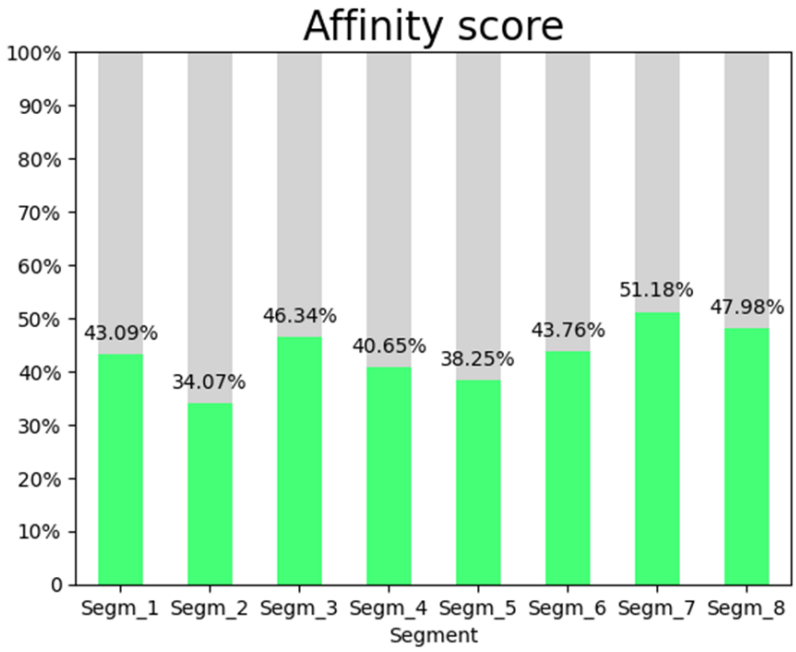


Figure 13: KitchenAid affinity score.

5.2 Smeg

Smeg is an Italian home appliance brand owned by Smeg S.p.A. The Smeg brand was first introduced as a brand name for a cooker developed by Vittorio Bertazzoni Sr in 1948. In the early 1900s, Smeg began producing the first electrically powered machines for grinding food, including coffee beans, peanuts, and hamburgers. In addition to cookers, Smeg also offers a wide range of other home appliances, including food processors, blenders, coffee makers and juicers, kettles, countertop ovens, electric cookers and multi-cookers, toasters, and commercial products (Smeg, 2023).

Despite being a brand focused on producing home appliances, the stamp campaign carried out by MC did not offer this type of product. The products available in the campaign were the kitchen equipment shown in Figure 14. This catalog is a special selection of reduced cost products to be offered as gifts.

This promotional campaign at MC took place between October 18, 2021 and February 20, 2022 at Continente Online and all Continente, Continente Bom Dia and Continente Modelo stores. Following MC's traditional stamp campaign model, for every €20 spent by the customer they were offered a stamp. From 5 accumulated stamps, the customer can make an exchange for participating products of the campaign. The exchange, in addition to using the stamp as currency, can be free or require addition of a value between €0.99 and €79.99, depending on product value and number of stamps used in exchange.

Figure 15 and Table 6 present the distribution of customers with purchases during the campaign period divided by groups from the segment selected for analysis. Segm_6 had the highest participation rate, at 22.51%, followed by Segm_7. At a lower level are segments Segm_2, Segm_3, Segm_4, and Segm_5. With a much lower participation rate than the other segments are segments Segm_1 and Segm_8.

Figure 16 shows the clusters resulting from the characterization process of the Smeg campaign using web scraping and text mining. With the elimination of the noisy cluster, using the most voluminous cluster as a criterion, the resulting cluster for characterizing the Smeg campaign is Cluster 2.



Figure 14: Smeg stamp campaign.

Using the resulting cluster in the process of calculating the affinity of the campaign with customer segments, from Equation 6, the resulting scores are obtained and presented in Figure 17. From the results presented, it is possible to see that Segm_7 has the highest calculated affinity score, followed by Segm_3 and Segm_8. The lowest calculated affinity scores are segments Segm_2 and Segm_5.

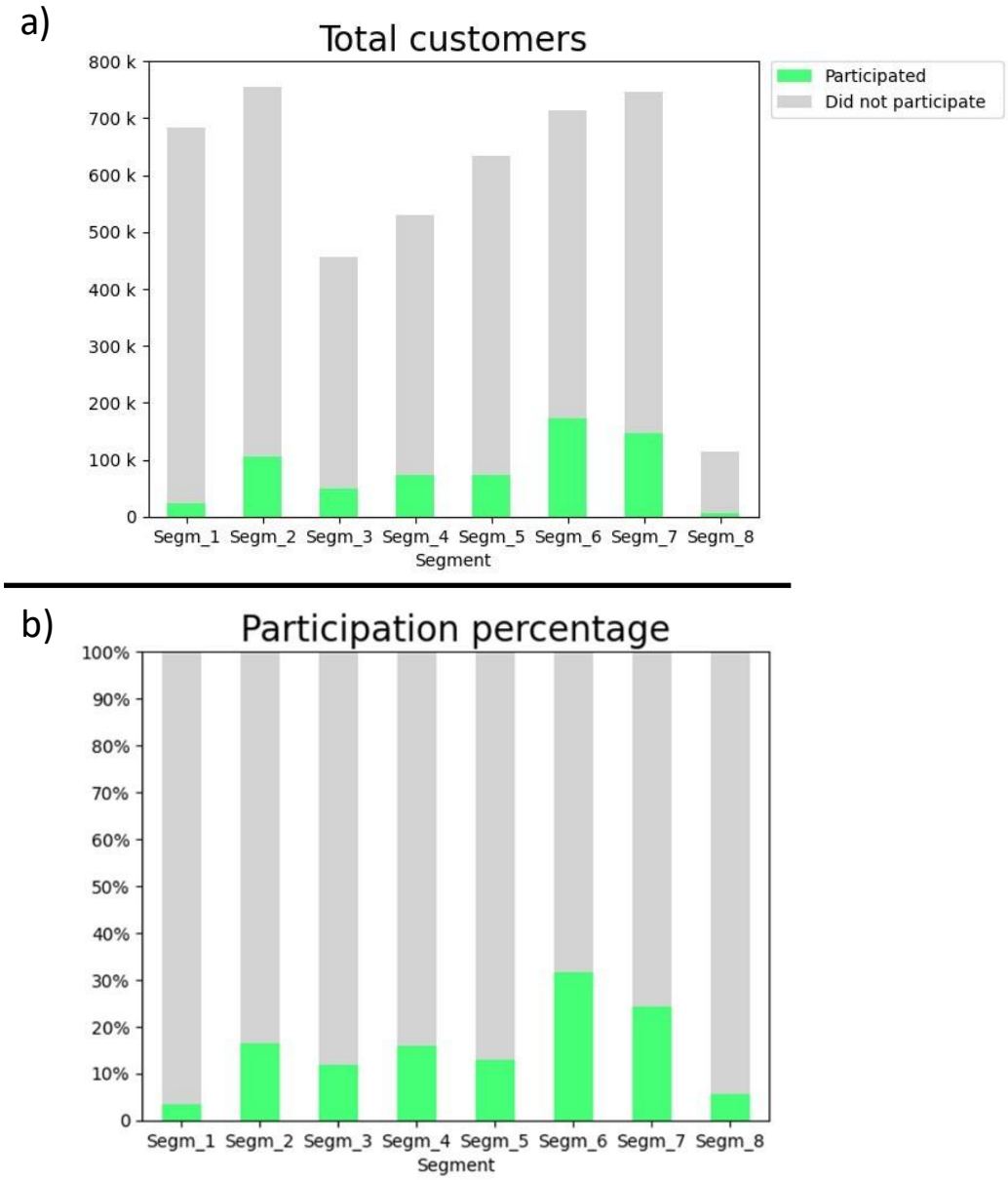


Figure 15: Smeg (a) Distribution by segment of the customer base; (b) Promotional participation.

Table 6: Distribution of active customers in the promotional period for Smeg campaign.

Segment	Total	Did not participate	Participated	Participation percentage
Segm_1	684215	660799	23416	3,54%
Segm_2	754485	648395	106090	16,36%
Segm_3	456689	408327	48362	11,84%
Segm_4	528811	456775	72036	15,77%
Segm_5	632906	560759	72147	12,87%
Segm_6	714537	542773	171764	31,65%
Segm_7	746782	600246	146536	24,41%
Segm_8	113979	108036	5943	5,50%
Total	4632404	3986110	646294	16,21%



Figure 16: Resulting clusters for the brand Smeg.

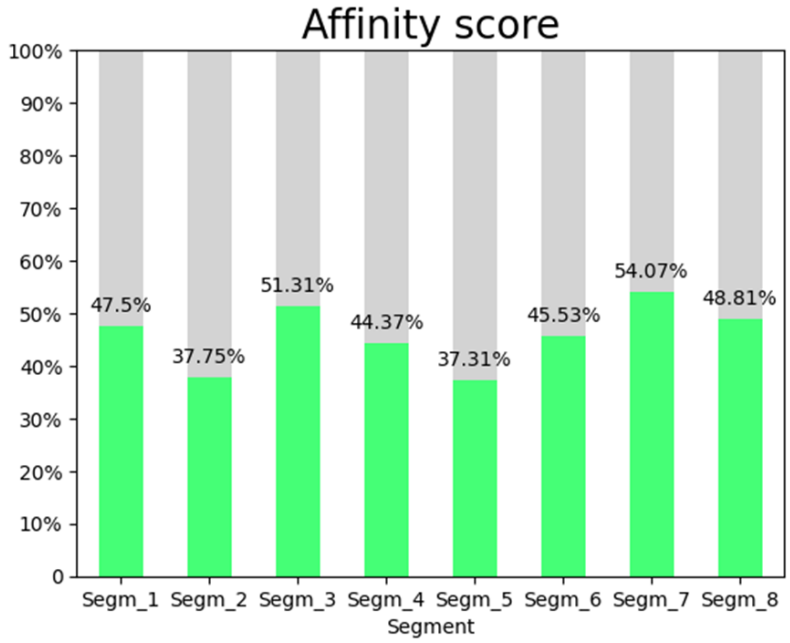


Figure 17: Smeg affinity score.

5.3 Hey Clay

Hey Clay is a brand that offers a wide range of kids' art clay & dough. Since its launch in the UK market in May 2022, Hey Clay owes its recognition to the mix of artistic properties of the clay, digital modeling experiences, and beautifully designed characters. Although this is a brand of modeling clay for children, the products available in the promotional campaign are plush toys. The available models are presented in Figure 18.



Figure 18: Hey Clay stamp campaign.

This promotional campaign at MC took place between June 12, 2023 and August 27, 2023 at Continente Online and all Continente, Continente Bom Dia and Continente Modelo stores. Following MC's traditional stamp campaign model, for every €20 spent by the customer they were offered a stamp. In addition, for every €4.99 spent on the brand's products, the customer receives 5 extra stamps. With 10 stamps plus €3.99, the customer can exchange for one of the available plush toy models.

The participation results obtained for this campaign are presented in Figure 19 and Table 7. From the graph, it is possible to observe the distribution of customer segments that made stamp exchanges in this campaign. For Hey Clay promotional campaign, Segm_4 had the highest participation rate, at 22.51%, followed by Segm_6. At a lower level are segments Segm_3 and Segm_7. With a much lower participation rate than the other segments are segments Segm_1 and Segm_8. Figure 20 illustrates the resulting clusters in the campaign characterization process.

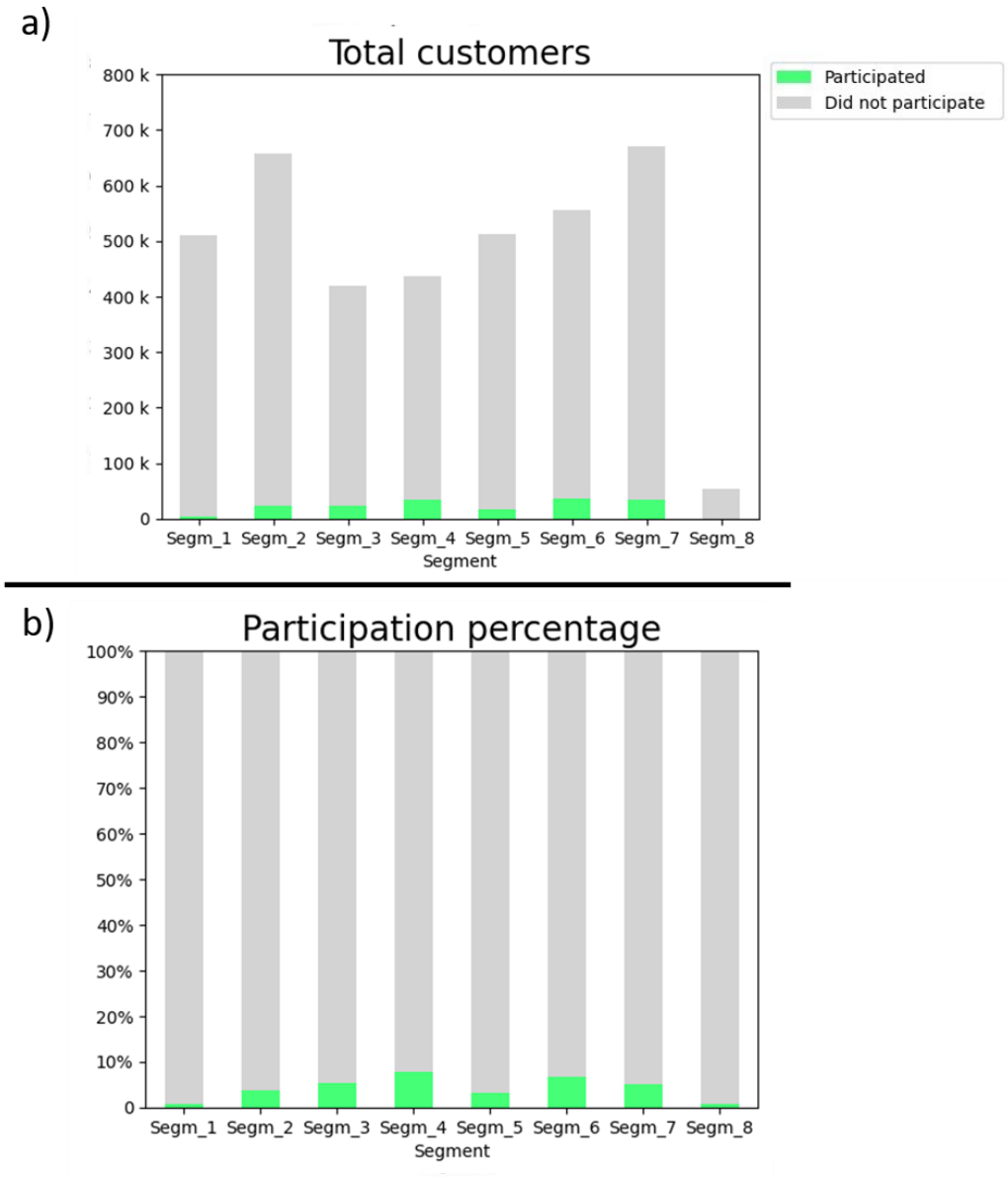


Figure 19: Hey Clay (a) Distribution by segment of the customer base; (b) Promotional participation.

Table 7: Distribution of active customers in the promotional period for Hey Clay campaign.

Segment	Total	Did not participate	Participated	Participation percentage
Segm_1	511170	507233	3937	0,77%
Segm_2	656557	632712	23845	3,63%
Segm_3	418603	396047	22556	5,39%
Segm_4	436403	401965	34438	7,89%
Segm_5	512544	496831	15713	3,07%
Segm_6	554728	517787	36941	6,66%
Segm_7	670520	636342	34178	5,10%
Segm_8	53711	53356	355	0,66%
Total	3814236	3642273	171963	4,51%



Figure 20: Resulting clusters for the brand Hey Clay.

After eliminating the noisy cluster, with the criterion of the largest volume cluster, the final cluster of campaign characterization is Cluster 1. Figure 21 presents calculated affinity scores for each segment using resulting cluster from campaign characterization process to calculate affinity with customer segments. From the results presented, it is possible to see that Segm_3 has the highest calculated affinity score, followed by Segm_8 and Segm_4. The lowest calculated affinity scores are segments Segm_1 and Segm_2.

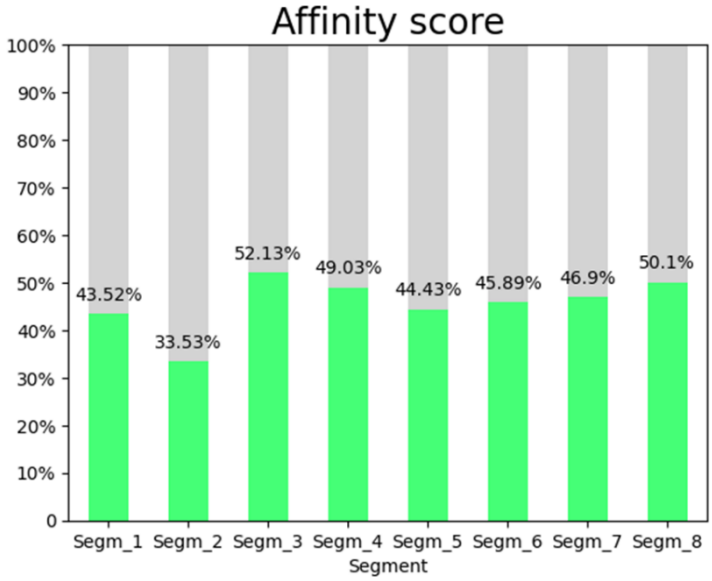


Figure 21: Hey Clay affinity score.

6 Discussion

The KitchenAid and Smeg campaigns are brands with similar product catalogs, consisting of appliances and kitchen accessories (pots and pans for KitchenAid (Figure 6), knives for Smeg (Figure 14)). As the campaign characterization process using web scraping and text mining techniques has the brand name as input and not the participating products, the expected result for characterizing these campaigns is similar, being clusters related to kitchen products.

For KitchenAid, from the resulting clusters of the process, presented in Figure 12, and used as a criterion for eliminating the noisiest cluster the most voluminous one, the resulting cluster is Cluster 1. Table 8 presents the 20 most significant terms present in this cluster.

Table 8: Top 20 terms for KitchenAid resulting cluster.

Rank	Term	Rank	Term
1	mixer	11	artisan
2	kitchenaid	12	recipe
3	knife	13	cook
4	dough	14	bowl
5	kitchen	15	add
6	whirlpool	16	chef
7	food	17	dagger
8	attachment	18	processor
9	mix	19	toaster
10	cookie	20	culinary

In the case of Smeg, campaign characterization also results in a single cluster after excluding the least significant group, Cluster 2 presented in Figure 16. Table 9 presents the most significant terms that make up the resulting cluster of the campaign.

From the results presented for the resulting clusters of KitchenAid and Smeg campaigns' characterization, it is possible to verify that these correspond to expected results, being possible from final terms obtained to conclude that they are brands related to kitchen utensils. In addition, by applying affinity calculation (Equation 3) for centroids of campaign characterizing clusters, an affinity of approximately 70,5% is obtained, which highlights similarity between brands.

Table 9: Top 20 terms for Smeg resulting cluster.

Rank	Term	Rank	Term
1	smeg	11	italy
2	oven	12	cook
3	italian	13	fab
4	fridge	14	dolce
5	caffe	15	guastalla
6	dishwasher	16	toaster
7	cooking	17	washing
8	coffee	18	pan
9	refrigerator	19	wikipedia
10	cookie	20	tray

Knowing that brands and consequently characterizations of KitchenAid and Smeg campaigns are similar, it is expected that results obtained for calculating affinity of these campaigns with customer segments will be similar to each other. From analyzing historical data from campaigns presented in Figure 11 and Figure 15, it can be observed that for both campaigns the total number of customers is approximately equal; KitchenAid has a volume 2% higher. As KitchenAid's campaign was carried out about 10 months after Smeg's product campaign ended, this difference can be justified by new business customers during this additional period.

Another point that can be observed with analysis of historical data is low promotional participation of Segm_1 and Segm_8 segments when compared to remaining segments in both campaigns. Segm_1 consists of customers associated with higher financial availability, that is, customers who tend to prefer more expensive products valuing brand or quality resulting in low promotional participation. Segm_8 is with lowest volume of customers consisting of customers without segment that is new customers or without sufficient purchase history for analysis. Due to criteria used in this segmentation these customer segments are not good comparative bases for results of this project either due to low promotional participation or lack of information about the customer.

With elimination of Segm_1 and Segm_8 segments as they are not good comparative bases it is possible to compare ordering of segments by calculated affinity and real participation in campaign. Table 10 presents this ordering for segments for KitchenAid campaign. In both cases Segm_6 and Segm_7 segments showed themselves in positions of greatest relevance among segments same occurs for Segm_4 and Segm_5 segments among least relevant.

Similarly, Table 11 presents the ordered segments for the Smeg campaign. As can be observed, both for calculated and real values, the ordering is very similar to that of KitchenAid, only changing the position of the least significant segments. However, Segm_6 and Segm_7 segments remain among the most significant and Segm_4 and Segm_5 segments among the least relevant.

Unlike the KitchenAid and Smeg campaigns in which brands are associated with kitchen products, the Hey Clay campaign is inserted in the niche of children's products, more specifically plasticine and in the case of promotional products plush toys. The difference in characterizing this campaign compared to those previously presented can be verified by applying affinity calculation (Equation 3) for centroids of campaign characterizing clusters, resulting in an affinity of 49.7% with Smeg campaign and 54.5% with KitchenAid campaign.

Table 10: Segments ordering for KitchenAid campaign.

Rank	Calculated affinity score	Participation
1	Segm_7	Segm_6
2	Segm_3	Segm_7
3	Segm_6	Segm_2
4	Segm_4	Segm_4
5	Segm_5	Segm_3
6	Segm_2	Segm_5

Table 11: Segments ordering for Smeg campaign.

Rank	Calculated affinity score	Participation
1	Segm_7	Segm_6
2	Segm_3	Segm_7
3	Segm_6	Segm_2
4	Segm_4	Segm_4
5	Segm_2	Segm_5
6	Segm_5	Segm_3

Thus, the expected result for characterizing the campaign is terms associated with the child audience. The resulting cluster of the characterization process is Cluster 1, defined in Figure 20. Table 12 presents the 20 most significant terms of this cluster.

Table 12: Top 20 terms for Hey Clay resulting cluster.

Rank	Term	Rank	Term
1	clay	11	white
2	toy	12	g
3	craft	13	soft
4	add	14	jigsaw
5	basket	15	stationery
6	item	16	color
7	eur	17	kg
8	dry	18	bake
9	colour	19	kit
10	supply	20	learn

Despite the results obtained for characterizing Hey Clay campaign not being as straightforward as in case of KitchenAid and Smeg campaigns, it can be verified that there are terms related to

child audience and home crafting associated with modeling clay main product marketed by brand.

Table 13 presents this ordering for segments for Hey Clay campaign. It is possible to note that Segm_3 and Segm_4 both for calculated scores and real participation in campaign were among the most relevant. Similarly, Segm_2 and Segm_5 segments were in both analyses among least significant.

Table 13: Segments ordering for Hey Clay campaign.

Rank	Calculated affinity score	Participation
1	Segm_3	Segm_4
2	Segm_4	Segm_6
3	Segm_7	Segm_3
4	Segm_6	Segm_7
5	Segm_5	Segm_2
6	Segm_2	Segm_5

From results presented for three example promotional campaigns it can be noted similarity between results of KitchenAid and Smeg campaigns given similarities between campaigns and results for Hey Clay campaign that distinguishes itself from others in niche of offered products.

In addition to comparing similarities and differences in the results, it is possible to validate the affinity calculation from business knowledge. As the segmentation used in the analysis is consolidated in the business, observing the purchase behavior of these segments over time allows for greater knowledge about the customers that make it up. Segm_3 is described as composed of parents who tend to change their buying behavior if their children value the product in question. Segm_4 is composed of customers with low experimentation, who take advantage of promotions on items they already consume. The business describes Segm_6 as the segment of customers with low purchasing power and low promotional participation. Segm_7 is composed of competitors' customers, who seek the business for convenience or promotional interest.

Thus, it can be concluded that KitchenAid and Smeg campaigns have participating products of high interest to the customer base, with the main participating customer segments being those with low usual promotional participation, Segm_6 and Segm_7. In the case of Hey Clay campaign, associated with children's audience, it is possible to notice a higher promotional participation of Segm_3 and Segm_4 segments when compared to other campaigns.However, comparing quantitative values of affinities calculated for each segment against their effective participation rate large differences are noted despite similarities in ordering. Several factors can cause these differences such as bias of pre-existing segmentation or even data incompatibility.

Segmentation bias is clear in the case of Segm_1 and Segm_8 segments due to expected low promotional participation or insufficient transactional data as previously explained. However, since only one customer segmentation is being used among several existing ones in business it may be too simplistic a way to characterize the customer base and more subtle biases may go unnoticed.

In case of data incompatibility, it may not be accurate to compare affinity between products and promotional participation although this is the best existing metric. Characterizing promotional campaigns from participating brands and subsequently segments in same way makes calculated affinity based on similarity between brands. Although affinity between usually purchased

products and those participating in campaign is an important factor in customer participation it is not only one; in case of stamp campaigns focused on this project customers with higher average spending may tend to greater participation by accumulating stamps more naturally.

7 Conclusions and future work

7.1 Conclusions

From the analysis of the evolution of the retail sector, it can be concluded that improving customer relations and adopting consumer-focused strategies are of utmost importance in the current context of increasing competition. From strategies focused on the consumer, their needs, desires, and preferences, companies in the sector gain a competitive advantage and consequently an increase in sales volume.

With these changes in the strategic planning of companies in the retail sector, they had to adapt consolidated business tools and create strategies to remain competitive. A widely used strategy that contributes to improving relationships with the current customer base and attracting new customers is carrying out marketing actions. These actions can range from communications to discounts on products and promotional campaigns.

From a personalization perspective, marketing actions have become primarily focused on customers. Business communications have become directed towards truly relevant customers, to improve relationships with the customer base by reducing unwanted messages. Another approach used is the selection of personalized offers, where a customer receives promotions and discounts on selected products based on their needs and desires, generating greater affinity and loyalty with the business.

An important strategy adopted by companies in the sector is carrying out massive promotional campaigns, that is, available to the entire customer base. Although these campaigns are not personalized, from transactional data from consumers it is possible to plan them with greater assertiveness.

In the context of these promotional actions, this work presented a methodology for modernizing the planning process for massive promotional campaigns. The model for automating the process had as its objectives the characterization of the campaign and calculation of its affinity with the customer base.

Web scraping and text mining techniques were used for campaign characterization, resulting in a campaign defined from the most relevant terms extracted from web pages related to the participating brand of the campaign. With campaign characterization carried out in this way, an extremely diverse and up-to-date data source is obtained, allowing for more precise and effective analysis of the target audience and market trends associated with the campaign.

For calculating campaign affinity with consumers, concepts of customer base segmentation were presented. A technique that consists of creating groups of customers with common characteristics, whether demographic, social or behavioral, facilitating analysis. To compare with the campaign, customer segments were defined based on the brands they consume most, allowing for application of the methodology used for campaign characterization, resulting in customer groups defined from relevant terms associated with selected brands. From using word

embedding techniques, qualitative results are converted into quantitative values that can be clustered; with resulting clusters affinities between segments and campaigns are calculated.

From the results obtained for example campaigns used, it is possible to observe that the campaign characterization process had satisfactory results from terms related to kitchen equipment for KitchenAid and Smeg campaigns and terms associated with children's products and home crafting for Hey Clay campaign.

Regarding the calculation of the affinity of campaigns with segments, there was some similarity in the ordering of the most relevant segments when compared to the historical promotional participation of the campaign. It is possible to identify similarities and differences between campaigns. When compared individually, the results are not so conclusive, however there are doubts about the interpretation of comparing affinity results with participation history.

The methodology applied for calculating affinity between campaign and segments uses information about brands to characterize both the promotional action and the customer base. Thus, the calculated affinity score can be interpreted as how related the campaign brand is to the brands that define the segments. Although one of the determining factors of promotional participation is affinity with participating products in the campaign, other variables can lead a customer with high affinity not to participate in the campaign.

Although there are no historical metrics that can be used to validate these results, it is believed that the results obtained can be applied in the disclosure phase of promotional campaigns. Using calculated affinity scores, it is possible to predict which segments for which the campaign will have greater relevance, creating specific strategies for each segment. In the case of more relevant segments, the message may have a tone that reinforces customer affinity with the campaign and for those with lower affinity, a message that encourages their participation.

In a current scenario where business strategies of companies in the retail sector are focused on customers through personalization of offered products and services, this project's contributions were significant. A methodology was created that allows for a simplified evaluation of the impact of carrying out a promotional campaign, which can be used comparatively with previous campaigns. In addition, the model can be used to personalize communications related to promotional campaigns, allowing for more attractive and efficient messages. Therefore, with eventual validation, the methodology can be used as a competitive advantage in planning marketing communications in promotional campaigns.

Relating the results obtained to the research questions defined in the introduction of this document, this project proved to be efficient in using web scraping for characterizing promotional campaigns. This allows for a basis of comparison for campaigns with little prior information.

Regarding the second question defined, the results obtained can be used for a primary analysis of the impact of carrying out promotional campaigns. However, it allows for planning campaign communication actions to reinforce customer base participation.

To understand the impacts of using this project in planning promotional campaigns, as proposed by the third research question defined, it is necessary to carry out tests for validation. With smaller scale and controlled tests, it is possible to validate the impact of creating personalized communications on customer buying behavior during the promotional period.

7.2 Limitations and future work

Despite the significant impact of the results obtained in the communication phase of the promotional campaign, there were several limitations that affected the developed methodology.

The first limiting factor of the quality of the results is the use of only one customer segmentation for calculating affinity with the campaign. In business, several segmentations are used, each using a criterion to group the customer base. Thus, each customer is inserted into a combination of the various existing segments.

Using only one segmentation in the analysis may oversimplify the problem. However, the methodology used to characterize segments makes sense for the segment used in which customers are grouped by the products they usually buy, but not so much for a segmentation in which the criterion used is average spending or purchase frequency.

As previously pointed out, promotional participation is associated with other factors besides affinity with participating products. In the case of stamp campaigns used as examples, average spending and purchase frequency are relevant factors for campaign participation, as customers with higher average spending and higher purchase frequency tend to accumulate stamps more naturally.

In addition, the developed methodology depends on several highly variable factors, which makes the process less robust. As campaign characterization information is obtained from web pages that undergo constant updating, different results can be obtained in executions for the same campaign since different web pages can be used for extraction. The quality of characterization is directly related to the information found in the search for the brand. Therefore, better results tend to be obtained for larger and consolidated brands due to a greater amount of online information.

The stage of characterizing segments from brands is also subject to this variability of results since it uses the same methodology applied in defining the campaign. However, an additional factor that compromises the robustness of the model at this stage is in selecting representative brands for each segment. As transactions from the last year are used to determine the most relevant brands, the five brands selected as representative may change.

Thus, result variability is related to external factors' influence at all stages of applied methodology.

Another limiting factor in developing this work was using pre-existing word banks to convert terms into vectors. The problem with using these tools is a lack of knowledge about their structure. They are often generalist models trained for use in analyzing relationships between words. However, when carrying out this project, word similarity is more relevant than their relationship; syntax matters little. Therefore, using these word banks can bring bias to similarity analyses.

As a final limitation, there is low compatibility for comparing obtained results and historical data for validation. Comparing data of different nature can lead to hasty conclusions. Thus, with currently existing information, validating tool use in planning communications for promotional campaigns becomes a difficult task.

As future work, it would be interesting to create a word bank specific to the business, created from the products in the catalog and inserted into the market structure. The methodology used for campaign characterization can be used on brands present in the business, allowing the characterizing terms found to be related to the market structure. Thus, for new brands, information found on web pages would be used to estimate the category of this brand in an already consolidated structure. With this, campaign affinity with the brand could be calculated from segment affinity with the estimated department and category of the campaign, making the analysis better grounded than comparing with the most consumed brands by segments.

In addition, test communications can be created for smaller campaigns, where the hypothesis of using the methodology for personalizing communications related to the campaign would be tested. If the results are positive, that is, customers who received personalized communications

have greater participation when compared to a control group, it can be studied to apply the methodology for sending general communications, related or not to promotional campaigns. In this case, the project would be a kind of search engine capable of offering insights in planning general communications with the customer base.

References

Ahonen, H., Heinonen, O., Klemettinen, M., & Verkamo, A. I. (1998). Applying data mining techniques for descriptive phrase extraction in digital document collections. Proceedings IEEE International Forum on Research and Technology Advances in Digital Libraries-ADL'98-,

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). A brief survey of text mining: Classification, clustering and extraction techniques. *arXiv preprint arXiv:1707.02919*.

Almeida, F., & Xexéo, G. (2019). Word embeddings: A survey. *arXiv preprint arXiv:1901.09069*.

Anderson, K., & Kerr, C. (2002). *Customer relationship management*. McGraw-Hill.

Anshari, M., Almunawar, M. N., Lim, S. A., & Al-Mudimigh, A. (2019). Customer relationship management and big data enabled: Personalization & customization of services. *Applied Computing and Informatics*, *15*(2), 94-101.

Aranganayagi, S., & Thangavel, K. (2007). Clustering categorical data using silhouette coefficient as a relocating measure. International conference on computational intelligence and multimedia applications (ICCIMA 2007),

Bar-Ilan, J. (2001). Data collection methods on the Web for infometric purposes—A review and analysis. *Scientometrics*, *50*, 7-32.

Beketova, S. (2023). CRM and Data Analysis in a Modern Marketing Approach.

Berry, M. A., & Linoff, G. S. (2000). Mastering data mining: The art and science of customer relationship management. *Industrial Management & Data Systems*, *100*(5), 245-246.

Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of retailing*, *93*(1), 79-95.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 1165-1188.

Continente, C. (2023). *Homepage*. https://www.cartaocontinente.pt/

Fayyad, U. (2005). Knowledge discovery in databases: An overview. Inductive Logic Programming: 7th International Workshop, ILP-97 Prague, Czech Republic September 17–20, 1997 Proceedings,

Feldman, R., Fresko, M., Kinar, Y., Lindell, Y., Liphstat, O., Rajman, M., Schler, Y., & Zamir, O. (1998). Text mining at the term level. Principles of Data Mining and Knowledge Discovery: Second European Symposium, PKDD'98 Nantes, France, September 23–26, 1998 Proceedings 2,

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, *85*(17).

Google. (2023). *Google for developers documentation*. https://developers.google.com/

Hameli, K. (2018). A literature review of retailing sector and business retailing types. *ILIRIA International Review*, *8*(1), 67-87.

Han, J., Pei, J., & Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.

Huggins, J., & Hammant, P. (2023). *Selenium 4.11.2 documentation*. https://selenium-python.readthedocs.io/

Khyani, D., Siddhartha, B., Niveditha, N., & Divya, B. (2021). An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, *22*(10), 350-357.

Lauzanne, O. (2022). *Pyquery 2.0.0 documentation*. https://pyquery.readthedocs.io/en/latest/api.html

Lee, Y. I., & Trim, P. R. (2006). Retail marketing strategy: The role of marketing intelligence, relationship marketing and trust. *Marketing Intelligence & Planning*, *24*(7), 730-745.

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, *30*(4), 256-265.

Lourinho, J. C. (2023). Distribuição e Retalho representou 12,4% do PIB português em 2022. *O Jornal Económico*. https://jornaleconomico.pt/noticias/distribuicao-e-retalho-representou-124-do-pib-portugues-em-2022/

Marcus, C. (1998). A practical yet meaningful approach to customer segmentation. *Journal of consumer marketing*, *15*(5), 494-504.

Mendez, M., Bendixen, M., Abratt, R., Yurova, Y., & O'Leary, B. (2015). Sales promotion and brand loyalty: some new insights. *International Journal of Education and Social Science*, *2*(1), 103-117.

Omran, M. G., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. *Intelligent Data Analysis*, *11*(6), 583-605.

Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, *7*, 145614-145633.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP),

Petrov, A. (2023). *Urllib3 2.0.4 documentation*. https://urllib3.readthedocs.io/en/stable/

Piateski, G., & Frawley, W. (1991). *Knowledge discovery in databases*. MIT press.

Řehůřek, R., & Sojka, P. (2023). *Gensim 4.3.1 documentation*. https://radimrehurek.com/gensim/#

Richardson, L. (2023). *Beautiful Soup 4.12.2 documentation*. https://www.crummy.com/software/BeautifulSoup/

Sabanoglu, T. (2023). *Retail market worldwide - Statistics & Facts*. https://www.statista.com/topics/5922/retail-market-worldwide/#topicOverview

Sallam, R. L., Richardson, J., Hagerty, J., & Hostmann, B. (2011). Magic quadrant for business intelligence platforms. *Gartner Group, Stamford, CT*.

Santoro, G., & Usai, A. (2018). Knowledge exploration and ICT knowledge exploitation through human resource management: A study of Italian firms. *Management Research Review*, *41*(6), 701-715.

Sarwan, N. (2017). *An intuitive understanding of word embeddings: From count vectors to word2vec*. https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Seetharaman, A., Niranjan, I., Tandon, V., & Saravanan, A. (2016). Impact of big data on the retail industry. *Journal of Corporate Ownership & Control*, *14*(1), 506-518.

Shehata, S., Karray, F., & Kamel, M. (2009). An efficient concept-based mining model for enhancing text clustering. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1360-1371.

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP Journal on Wireless Communications and Networking*, *2021*(1), 1-16.

Shyan Fam, K. (2003). Managing an effective promotion campaign: lessons from retail advertisers in New Zealand, Portugal and Hungary. *Marketing Intelligence & Planning*, *21*(5), 282-291.

Smeg. (2023). *History*. https://www.smeg.com/company/history).

Smith, A. D. (2011). Retail managers' perspectives on the effectiveness of corporate e-personalisation initiatives. *International Journal of Business Information Systems*, *8*(3), 211-246.

Stylos, J., & Myers, B. A. (2006). Mica: A web-search tool for finding api components and examples. Visual Languages and Human-Centric Computing (VL/HCC'06),

Taskesen, E. (2023a). *Clusteval 2.2.1 documentation*. https://erdogant.github.io/clusteval/pages/html/index.html

Taskesen, E. (2023b). *From Data to Clusters: When is Your Clustering Good Enough?* https://towardsdatascience.com/from-data-to-clusters-when-is-your-clustering-good-enough-5895440a978a

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, *5*(1), 7-16.

Waller, M. A., & Fawcett, S. E. (2013). Data science, predictive analytics, and big data: a revolution that will transform supply chain design and management. In (Vol. 34, pp. 77-84): Wiley Online Library.

Wamba, S. F., Akter, S., Edwards, A., Chopin, G., & Gnanzou, D. (2015). How 'big data'can make big impact: Findings from a systematic review and a longitudinal case study. *International journal of production economics*, *165*, 234-246.

Wu, S.-T., Li, Y., & Xu, Y. (2006). Deploying approaches for pattern refinement in text mining. Sixth International Conference on Data Mining (ICDM'06),

Xiao, J., Lu, J., & Li, X. (2017). Davies Bouldin Index based hierarchical initialization K-means. *Intelligent Data Analysis*, *21*(6), 1327-1338.

Zhao, B. (2017). Web scraping. *Encyclopedia of big data*, 1-3.

Ziafat, H., & Shakeri, M. (2014). Using data mining techniques in customer segmentation. *Journal of Engineering Research and Applications*, *4*(9), 70-79.