

NOVA

IMS

Information
Management
School

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

**The application of Anna Karenina Principle on Portuguese
Restaurants Online Reviews: A BERTopic Approach**

Dinis Cortilho Melo

Master Thesis

presented as partial requirement for obtaining a Master's Degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

**The application of Anna Karenina Principle on Portuguese Restaurants Online Reviews: A
BERTopic Approach**

by

Dinis Cortilho Melo

Master Thesis presented as partial requirement for obtaining the Master's degree in Data
Science and Advanced Analytics, with a specialization in Data Science

Supervised by

Flávio Pinheiro, PhD, Nova Information Management School

November, 2024

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Rules of Conduct and Code of Honor from the NOVA Information Management School.

Lisbon, November 2024

ABSTRACT

Online reviews play a significant role in shaping consumer decisions, influencing perceptions, and serving as key reference points for various choices. The Anna Karenina Principle, derived from Leo Tolstoy's classic novel, provides valuable insights into success and failure in complex systems. Recent studies have explored the application of this principle to online reviews in the tourism and restaurant sectors, focusing on challenges in identifying areas of dissatisfaction within predominantly positive feedback. By employing state-of-the-art methodologies, such as BERTopic for topic modeling, this research provides a more transparent view of the relevance of the Anna Karenina Principle in the context of Portuguese restaurant online reviews. Building on these insights, this research delves deeper into the success and failure factors specific to the Portuguese restaurant sector, providing a nuanced understanding of consumer sentiment and its alignment with the Anna Karenina Principle, particularly in identifying key factors that contribute to the success and failure of Portuguese restaurants.

KEYWORDS

Anna Karenina Principle; Portuguese Restaurants; Topic Modelling; Sentiment Analysis;
BERTopic

TABLE OF CONTENTS

1. Introduction.....	1
2. Literature review	2
2.1. Anna Karenina Concept.....	2
2.2 Impact of online Reviews in tourism success	2
2.3 Application of Anna Karenina Principle on Tourism Online Reviews.....	3
2.4 Restaurants’ Success Factors.....	3
2.5 Understanding the Sentiment of online Reviews.....	4
2.6 Automated Topic Modeling State-of-Art.....	5
3. Methodology	7
3.1 Data	7
3.2 Sentiment Analysis	8
3.3 Text Embeddings	9
3.4 Dimensionality Reduction	9
3.5 Clustering.....	10
3.6 Clusters’ Evaluation	11
3.7 Success and Dissatisfaction Factors Analysis	12
4. Results and discussion	13
4.1 Results	13
4.1.2 Internal Clusters’ Variability Analysis	15
4.1.2 Cluster Distinctiveness Analysis	21
4.2 Discussion	22
5. Conclusions and future works	24
Bibliographical References	26

LIST OF FIGURES

Figure 1 - Methodology Layout	7
Figure 2 – Example of Reviews Classified by Vader	9
Figure 3 - Spatial Distribution of Topics in the Dimensional Embedding Space	13
Figure 4 - Number of Reviews By Sentiment Compound Scores	13
Figure 5 - Number of Topics by Review Count Range	14
Figure 6 - Word Clouds Representing Positive and Negative Topics	14
Figure 7 - MICCD vs. Percentage of Positive Reviews by Cluster (Entire Dataset)	15
Figure 8 - MICCD vs. Percentage of Positive Reviews by Cluster (Clusters >90% Positive Reviews)	17
Figure 9 - β 1s and p-value vs. Cluster Top Size Selection % (Positive Clusters).....	19
Figure 10 - β 1s and p-value vs. Cluster Top Size Selection % (All Clusters)	20
Figure 11 - Pairwise Inter-Cluster Cosine Distances by Sentiment and Rating Clusters.....	21

LIST OF TABLES

Table 1 - Data Dictionary.....	8
Table 2 - Results of OLS Analysis for the Entire Dataset	16
Table 3 - Results of Spearman Correlation Analysis for the Entire Dataset.....	17
Table 4 - Results of OLS Analysis for Clusters with >90% Positive Reviews.....	18
Table 5 - Results of Spearman Analysis for Clusters with >90% Positive Reviews	18

LIST OF ABBREVIATIONS AND ACRONYMS

AKP	Anna Karenina Principle
MICCD	Mean Intra-Cluster Cosine Distance
VADER	Valence Aware Dictionary and Sentiment Reasoning
BERT	Bidirectional Encoder Representations from Transformer
UMAP	Uniform Manifold Approximation and Projection
eWOM	Electronic Word Of Mouth
LDA	Latent Dirichlet Allocation
OLS	Ordinary Least Squares
IQR	Interquartile Range

1. INTRODUCTION

The influence of online reviews on consumer decisions has been investigated in recent years, with such reviews becoming critical references across various industries, especially in the tourism and restaurant sectors. Consumers rely heavily on online feedback to make informed choices, and these reviews hold significant potential to either enhance or harm business reputations (Zhang et al., 2010). This study draws on the Anna Karenina Principle (AKP), a concept from Tolstoy's literature that has been adapted to explore factors contributing to the success or failure of complex systems. According to AKP, success arises from satisfying a full set of criteria, while failure can stem from the breakdown of any single component (Bornmann & Marx, 2011).

The Anna Karenina Principle (AKP) provides a valuable framework for examining the intricate factors that drive restaurant success or failure, particularly by highlighting the contrasting dynamics between positive and negative customer reviews. Although previous studies have explored this phenomenon within tourism attractions (Kirilenko et al., 2021) and, more specifically, Portuguese restaurants, (Mukhacheva, 2023) they have primarily relied on traditional methodologies such as Latent Dirichlet Allocation (LDA) for topic modeling. As a result, these findings are often constrained by the limitations of such methods, leading to tentative conclusions that lack the robustness of more advanced techniques.

This study aims to address these gaps within the context of Portuguese restaurant online reviews by employing state-of-the-art approaches, such as BERTopic for topic modeling. The objective is to achieve clearer results by providing a more transparent understanding of the clustering process while considering the contextual nuances of the online reviews. This enhanced methodology enables more precise, cluster-based analyses, allowing for a more thorough investigation into the assumptions made in previous studies. Specifically, this study aims to address the following questions:

- When analyzed with state-of-the-art methodologies, does the Anna Karenina Principle hold in the context of Portuguese restaurant reviews?
- What are the main success and failure factors for Portuguese restaurants?

The following section will begin with exploring the Anna Karenina Principle (AKP), examining its applicability in the analysis of online reviews. This will be followed by a discussion on methods for extracting sentiment from online reviews, alongside an analysis of the state-of-the-art topic modeling approaches. The methodology section will then provide a detailed account of the analytical process employed in this study. Finally, the results section will present findings that shed light on the applicability of the AKP in the context of Portuguese restaurants online reviews and on the factors driving customer satisfaction.

2. LITERATURE REVIEW

The following literature review encapsulates the theoretical framework pertinent to the thesis's inquiries and outcomes. This review explores the Anna Karenina concept and its correlation with online reviews in the tourism industry, this review traverses through the domains of Sentiment Analysis and Topic Modeling. It culminates by delving into optimal strategies and fundamental principles essential for the effective implementation of the methodology.

2.1. ANNA KARENINA CONCEPT

Introduced by Leo Tolstoy's in the novel Anna Karenina and extend by Diamond, the Anna Karenina Principle states that "all happy families are alike; each unhappy family is unhappy in its own unique way." (Bornmann & Marx, 2011). This principle suggests that for something to be successful it has always to satisfy a specific set of factors, while the absence of one of those factors leads to failure. "Favorable outcomes require every detail to be right, whereas unfavorable outcomes only require one wrong detail" (Bornmann & Marx, 2011). The Anna Karenina Principle helps identify critical factors for success and failure in complex systems, informing decision-making and problem-solving. The AKP has found applications in various fields, from evolutionary biology to economics and even to the study of technological systems, revealing itself to be a useful mechanism to study the success and failures in science (Bornmann & Marx, 2011).

2.2 IMPACT OF ONLINE REVIEWS IN TOURISM SUCCESS

Online reviews play a crucial role in the success of businesses, particularly in the tourism industry. The positive electronic word of mouth (eWOM) can significantly enhance the reputation and credibility of a tourism establishment, attracting more customers and ultimately leading to increased bookings and revenue (Zhang et al., 2010). This can be especially impactful for smaller brands looking to establish themselves in the market compared to larger competitors (Y. Wang et al., 2021) . According to the paper "The Economic Value of Online Reviews" online reviews have a positive impact on both costumers and restaurants when it comes to decision-making and overall business performance (Wu et al., 2015). Approximately seventy percent of consumers takes online reviews as part of the decision-making process (Y. Wang et al., 2021) and the contextual characteristics of the reviews has a much higher significance on consumer behavior as compared to the overall rating (Schuckert et al., 2015),(Wu et al., 2015). The quantity of reviews usually also affect the online popularity of the business, as potential customers perceive a higher number of reviews as a sign of credibility and trustworthiness (Zhang et al., 2010). While positive electronic word of mouth (eWOM) can enhance the reputation of a tourism establishment, it's important to note that negative reviews can have a detrimental impact. One negative review can potentially outweigh the impact of multiple positive ones. In fact, research has shown that consumers tend to give more weight to negative reviews than positive ones when making purchasing

decisions (Kirilenko et al., 2021). In conclusion, while online reviews can certainly impact the success of tourism businesses, it's crucial to consider the potential drawbacks and limitations associated with them. Either way analyzing the sentiments of online reviews can provide valuable insights into the success or failures of restaurants.

2.3 APPLICATION OF ANNA KARENINA PRINCIPLE ON TOURISM ONLINE REVIEWS

In recent years, some studies have been conducted on the application of the Anna Karenina principle in the field of tourism. Until now, there are two relevant studies that serve as a basis for our research: "Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply?" (Kirilenko et al., 2021) and "The application of Anna Karenina principle on restaurant online reviews: a Portuguese case study" (Mukhacheva, 2023).

The first paper indicates that the Anna Karenina principle could be applied to tourism attractions. The researchers conducted an analysis of a large dataset comprising online reviews from various touristic attractions worldwide. It was observed that identifying dissatisfaction-related topics posed greater challenges compared to satisfaction-related topics when employing automated topic modeling using Latent Dirichlet Allocation (LDA). This suggests that the larger variability of complaints in negative reviews results in limitations for automated topic modeling, confirming the AKP. The authors acknowledge this limitation and suggest exploring dictionary-based methods or supervised learning approaches in future studies. It was also recommending manual verification of results, especially when tokens related to a specific topic are challenging to interpret.

The second study inspired by the first one, focused specifically on the application of the Anna Karenina principle to restaurant online reviews in Portugal. Using a similar methodology as the first study (application of LDA for topic identification), the researcher analyzed a dataset of 307,000 online reviews of Portuguese restaurants extracted from TripAdvisor. This time applying a sentiment analysis on the reviews and using the sentiments along with the ratings to classify between positive and negative reviews. The study concluded that the Anna Karenina principle can indeed be applied to restaurant online reviews in Portugal in these conditions. The thesis outlines limitations such as the need to specify the number of topics in advance when using the LDA topic modeling method, potential online review bias due to retrieving only one part of available TripAdvisor reviews, and the fact that the bag-of-words model used by Latent Dirichlet Allocation disregards word order and document structure. This results in not capturing semantic meaning. Additionally, splitting the review dataset into positive and negative reviews presented limitations, along with a lack of research on the effect of AKP on online reviews and its relation to sentiment analysis.

2.4 RESTAURANTS' SUCCESS FACTORS

Restaurants across the globe encompass a diverse range of types, each designed to offer distinct experiences to customers. While these experiences may vary significantly, it is important to investigate whether the factors contributing to a restaurant's success differ

across these various types, or if they remain consistent throughout the spectrum of restaurants. Understanding these dynamics can shed light on whether universal success factors exist or if tailored strategies are required for different restaurant categories.

In a study conducted by Hwang and Zhao in 2008, surveying 390 customers in Miami, Florida, it was found that customer satisfaction in restaurants is primarily influenced by factors such as food quality, price, and cleanliness (Hwang & Zhao, 2010). These findings are still consistent with recent research analyzing 111 academic papers, which identified food, service, atmosphere, and price as the primary factors contributing to restaurant success, with food being the most significant determinant (Zanetta et al., 2024).

Another study conducted by Anil Bilgihan, Soobin Seo, and Jihee Choi utilized online reviews to investigate the factors contributing to restaurant success. Their study employed various data visualization techniques, including tag clouds (word clouds) for qualitative analysis, and Multivariate Analysis of Variance (MANOVA) for quantitative analysis. The findings revealed that food, menu offerings, ambiance, and service are key factors driving restaurant success, aligning also with the previous studies (Bilgihan et al., 2018).

Closer to our research focus, Luis Miguel Pacheco examined online reviews of upscale Iberian (Portugal and Spain) restaurants. Through statistical tests and regression analysis, the study concluded that food and service are the primary determinants of success for these establishments (Pacheco, 2018).

The outcomes of these studies consistently highlight food as the most critical factor for restaurant success, supported by other key elements such as service, ambiance, price, and cleanliness. These findings offer a general understanding of the factors contributing to restaurant success and provide valuable insights into developing effective strategies to enhance customer satisfaction and ensure sustained business growth.

2.5 UNDERSTANDING THE SENTIMENT OF ONLINE REVIEWS

In the realm of online tourism reviews classification, there can be discrepancies between the star ratings and the true sentiment expressed in the review text. To effectively study the opinions and tones conveyed in the digital landscape, it is crucial to have an accurate classification of the review's overall tone. Based on research, identifying the sentiment expressed within the review text could prove beneficial, either complementing or providing an alternative to the existing rating systems (Al-Natour & Turetken, 2020).

For this purpose, different sentiment analysis techniques have been developed to capture emotions in reviews. Over time, these methods have been improved, resulting in advanced approaches with their own strengths and weaknesses.

Rule-based sentiment analysis models are algorithms designed to determine the sentiment of text by applying predefined linguistic rules and sentiment lexicons. The Valence Aware Dictionary and sEntiment Reasoning (VADER) model is part of that group. It was specifically

developed for the context of social media text. It integrates a validated sentiment lexicon with five key grammatical and syntactical rules to accurately measure sentiment intensity. This model exhibits a domain-agnostic classification approach, which enables it to achieve satisfactory performance across diverse domains without requiring extensive prior training. Additionally, its efficient performance is an advantage when working with large datasets, granting flexibility to quickly test and experiment during research endeavors. VADER's transparency, in terms of its lexicon and rules, facilitates easy interpretation and understanding, without requiring extensive domain-specific expertise. This allows the model to be accessible and applicable across a diverse range of research fields and study areas. However, besides its strengths, VADER may not fully grasp the intricacies of sentiment in more nuanced or complex contexts. As a result, its performance can be limited when compared to more advanced machine learning models that are capable of learning from large, diverse datasets (Hutto & Gilbert, 2014).

In recent years, large language models (LLMs) have been increasingly utilized for sentiment analysis, leveraging the power of deep learning to extract complex semantic meanings from text and classify them more accurately. Examples such as BERT, RoBERTa, and GPT have demonstrated superior performance in complex scenarios compared to traditional methods. This superiority arises from their attention mechanisms, which allow the models to weigh the importance of words and understand relationships across long distances within text. Additionally, LLMs generate dynamic word embeddings based on context, enabling them to capture nuances in meaning, such as sarcasm and irony. Fine-tuning on domain-specific datasets further enhances their ability to recognize sentiment in specialized language. However, LLMs also have disadvantages, including their high computational requirements, which may limit accessibility for some users (Tabinda Kokab et al., 2022).

2.6 AUTOMATED TOPIC MODELING STATE-OF-ART

Automated topic modeling has experienced significant advancements in recent years, enabling its application in various domains and tasks. There is a variety of different types of Topic Modeling techniques with their unique strengths and applications.

Introduced by Blei, Latent Dirichlet Allocation (LDA) is a foundational topic modeling technique that frames each document as a mixture of multiple latent topics, with each topic represented as a distribution over words (Blei et al., 2003). LDA operates under a probabilistic framework, assuming that words within a document are generated from a set of hidden topics. This allows LDA to reveal overarching themes without requiring prior knowledge of the data's structure, making it highly versatile and broadly applicable across various fields (Egger & Yu, 2022). A significant advantage of LDA is its computational efficiency, as it can handle large, sparse datasets by effectively representing topics in lower-dimensional space. However, LDA assumes topics are independent, which can limit the model's effectiveness when relationships between topics are nuanced or when data requires a more sophisticated representation of correlations (Blei, 2012).

Building upon the foundations of LDA, newer advancements in topic modeling, such as Top2Vec, addressed some of these limitations by providing a more flexible and efficient approach for handling large, unstructured datasets (Egger & Yu, 2022; Grootendorst, 2022). Top2Vec generates joint document and word embeddings, allowing for topic identification without requiring predefined topic numbers or extensive preprocessing. By leveraging embeddings from models like Word2Vec and Doc2Vec, Top2Vec maps words and documents to a high-dimensional space, capturing both context-independent word meanings (Mikolov et al., 2013) and the contextual relationships of larger text segments (Le & Mikolov, 2014). Topics are then derived through clustering in this embedding space, with the centroids representing the topics and the nearest words reflecting their primary content (Angelov, 2020).

BERTopic takes these advancements further by incorporating transformer-based embeddings, clustering, dimensionality reduction, and class-based TF-IDF to enhance the topic extraction process (Grootendorst, 2022). Using models like BERT, which produce context-aware word representations (Devlin et al., 2019; Vaswani et al., 2023), BERTopic creates embeddings that capture rich semantic information. To manage high-dimensional embeddings, BERTopic applies UMAP for dimensionality reduction, preparing the data for clustering (Angelov, 2020). The clustering process is performed using HDBSCAN, a density-based algorithm that identifies clusters based on variations in data density (Grootendorst, 2022). Finally, class-based TF-IDF (c-TF-IDF) is employed to highlight key terms within each cluster, making the resulting topics both more interpretable and contextually relevant (Grootendorst, 2022).

There is considerable evidence in the literature suggesting that BERTopic generally outperforms other topic modeling methods across various contexts. For example, in a comparative study evaluating the performance of different algorithms for analyzing Twitter posts, BERTopic demonstrated superior capabilities in generating more meaningful and interpretable topics when compared to Latent Dirichlet Allocation (LDA) and Top2Vec. This finding highlights BERTopic's effectiveness in extracting meaningful and interpretable topics from short and informal text data, often characterized by sparse and noisy content (Egger & Yu, 2022). Additionally, when applied to Consumer Financial Protection Bureau (CFPB) data and Arabic text data, BERTopic was found to outperform LDA, producing more diverse and semantically coherent topics (Abuzayed & Al-Khalifa, 2021; Sangaraju et al., 2022). These findings collectively highlight BERTopic's effectiveness in extracting high-quality topics across diverse domains, including both short-form and highly structured text data.

3. METHODOLOGY

This section outlines the data and methodological approaches employed in the present research. It describes the nature of the data, the process of its acquisition, as well as the steps performed to respond to our research question. Figure 1 illustrates the methodological steps in a flowchart, providing a visual representation of the research process. The diagram demonstrates that, following data collection, two distinct analyses were conducted using VADER and BERTopic. These analyses diverged into separate pathways to address different research questions. One pathway focused on cluster evaluation to determine the applicability of the Anna Karenina Principle (AKP) within this context. The other pathway aimed to identify the factors contributing to success and failure in the Portuguese restaurant sector.

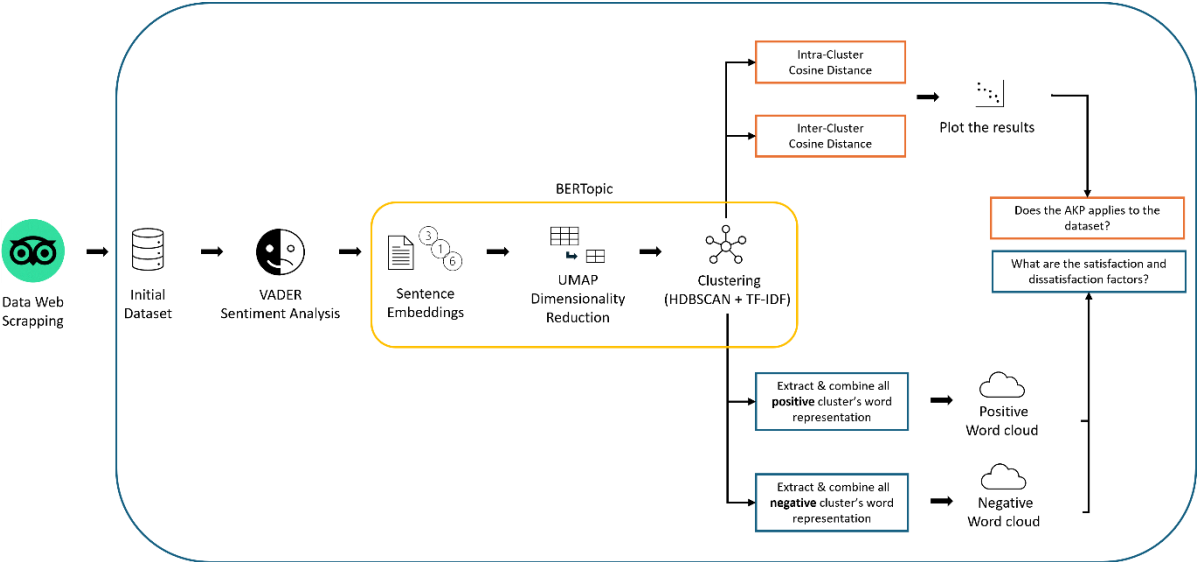


Figure 1 - Methodology Layout

3.1 DATA

The dataset is composed by approximately 307,000 online reviews from 24,000 Portuguese restaurants on the TripAdvisor platform, covering the period from 2007 to 2022. Using web scraping trough a Python script, a maximum of 30 English reviews were collected per restaurant, selected in descendent order based on their date. The following table presents the data dictionary.

Table 1 - Data Dictionary

Attributes	Data Type	Description
Bubbles	Integer	Star Rating (from 1 to 5)
Review	String	Complete review's text
Quotes	String	Quote from the review
Len	Integer	Length of the review
RatingDt	Date	Rating date
Region_URL	String	Region URL
Restaurnat_URL	String	Restaurant URL
Address	String	Restaurant's Address
Gastronomy	String	Restaurant's Gastronomy

To understand how well this data represents the Portuguese restaurant scenario, some exploration was made. The review ratings revealed that 60% of the reviews awarded 5 stars, 21% awarded 4 stars, and the remainder received 3 stars or lower. On average, 83% of the restaurants had ratings of 4 stars or higher, making the dataset an imbalanced one. The geographic distribution showed that Lisbon and the Northern region accounted for the largest share of restaurants, while Algarve and Lisbon received the highest number of reviews, likely due to their popularity as tourist destinations. From this analysis we can conclude that, majority of our reviews are positive having in consideration the star ratio classification, and that the Northern region, Lisbon and Algarve are most predominant zones in terms of restaurants and reviews.

3.2 SENTIMENT ANALYSIS

Aiming to use the sentiment of the reviews in the future, as a way of understanding the tone behind the text, the reviews were first evaluated using the VADER (Valence Aware Dictionary and Sentiment Reasoner) model. Due to its computational efficiency and proved satisfactory performance in social media text, this model allowed for a higher flexibility while doing the research (Hutto & Gilbert, 2014).

The sentiment analysis enabled the comparison of the review star ratings with the corresponding sentiment scores in subsequent analyses. Specifically, the compound scores obtained from the VADER were used to categorize the reviews into positive, and negative classifications. Reviews with compound scores exceeding 0.00 were designated as 'Positive', and reviews with compound scores less than 0.00 were labeled as 'Negative'.

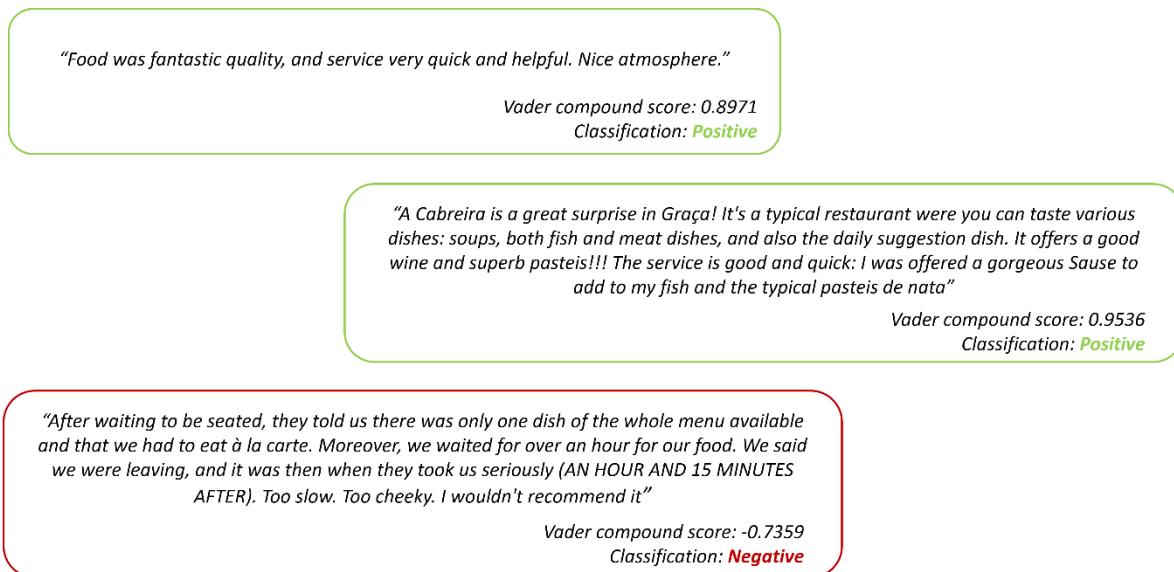


Figure 2 – Example of Reviews Classified by Vader

3.3 TEXT EMBEDDINGS

After the sentiment analysis was conducted, the objective was to group the reviews into clusters. But prior to clustering, it was necessary to transform the text data into embeddings. For this purpose, we utilized the "SentenceTransformer" model, especially the "all-MiniLM-l6-v2" version. The Sentence Transformer, or SBERT, is a pre-trained deep learning model which utilizes a Transformer architecture as its embedding method. Transformers are distinguished by a self-attention mechanism that captures dependencies between words in a sentence, irrespective of their distance by employing the self-attention mechanism. This allows SBERT to generate contextualized embeddings, where a word can have varying representations depending on the context in which it appears (Devlin et al., 2019). The all-MiniLM-L6-v2 model exhibits high efficiency due to its 384-dimensional embedding size and 6-layer architecture, enabling rapid processing and low memory requirements, making it well-suited for large-scale or real-time applications. It also demonstrates strong performance across a variety of tasks, leveraging deep learning and attention mechanisms to comprehend complex textual data. Research indicates that this model maintains a balance between accuracy and efficiency, often performing comparably to larger models. (W. Wang et al., 2020).

3.4 DIMENSIONALITY REDUCTION

In an effort to enhance the quality of the clustering, we proceeded with dimensionality reduction, specifically utilizing the UMAP (Uniform Manifold Approximation and Projection) technique. UMAP is a nonlinear dimensionality reduction technique that has gained popularity for reducing high-dimensional data to a manageable size while preserving its structure.

Developed by McInnes (McInnes et al., 2020), UMAP efficiently captures local and global data structures, making it particularly valuable for high-dimensional embeddings generated by deep learning models (Grootendorst, 2022). Unlike PCA, which focuses on linear reduction, or t-SNE, which is computationally intensive, UMAP offers a balance of speed and structural preservation. Its effectiveness at maintaining semantic relationships in reduced dimensions has made UMAP a standard choice for topic modeling frameworks such as BERTopic and Top2Vec (Angelov, 2020). Applying UMAP prior to doing the clustering helps improve both the performance and efficiency of the topic modeling process. High-dimensional embeddings are computationally intensive and susceptible to the "curse of dimensionality" where distance metrics become less meaningful. UMAP reduces the dimensionality of these embeddings, enabling clustering algorithms like HDBSCAN to function more effectively by focusing on the most significant features. Furthermore, UMAP mitigates noise and redundancy in the data, which enhances the separation between clusters and results in more coherent topics. Consequently, UMAP augments both the computational efficiency and the quality of topics extracted by BERTopic (Allaoui et al., 2020).

3.5 CLUSTERING

Following the creation of the text embeddings, and the dimensionality reduction it was possible to proceed with clustering the reviews using the HDBSCAN algorithm plus the c-TF-IDF coming from BERTopic.

HDBSCAN extends the DBSCAN clustering algorithm by incorporating hierarchical clustering principles. It begins by constructing a hierarchy of clusters based on density variations. The algorithm then applies a stability criterion to extract a flat clustering, favoring clusters that persist across multiple density thresholds. This approach is particularly advantageous for datasets with varying densities, as it robustly identifies clusters and isolates noise (Campello et al., 2013). To enhance interpretability, c-TF-IDF was employed for topic modeling. Unlike traditional TF-IDF, which evaluates term importance at the document level, c-TF-IDF considers each cluster as a distinct class. By computing term relevance within each cluster relative to the overall corpus, c-TF-IDF emphasizes words unique to each topic. In the BERTopic framework, this results in clear and contextually significant topic representations, highlighting vocabulary that distinguishes one topic from another (Grootendorst, 2022).

After clustering, the model generated topic representations each associated with a set of keywords and distribution percentages. These topics were analyzed for sentiment and rating distribution. Topics were classified as "Positive" if over 50% of their sentiment or rating reviews were positive; otherwise, they were categorized as "Negative."

3.6 CLUSTERS' EVALUATION

To explore how inner cluster variability and compactness could prove the presence of the Anna Karenina Principle (AKP) in the dataset, we measured the distance between each review and its assigned cluster centroid. By calculating the mean intra-cluster cosine distance for each cluster (MICCD), we could assess how varied the reviews within each topic were. The goal was to show that topics with more positive reviews had a lower average cosine distance to the cluster centroid, compared to those with negative reviews. A smaller distance indicates that the reviews within a cluster are more similar to each other, while clusters with larger distances contain more diverse themes, supporting the AKP. Cosine distance was used for the purpose of this research because it is generally more effective than measures like Euclidean distance when dealing with high-dimensional data (Liu et al., 2019).

$$\text{MICCD} = \frac{1}{N_k} \sum_{i=1}^{N_k} 1 - \frac{v_i \cdot c_k}{\|v_i\| \cdot \|c_k\|} \quad (1)$$

where N_k is the number of reviews in cluster k , v_i the selected review vector and c_k the corresponding centroid vector.

Another metric for evaluating the presence of AKP was cluster distinctiveness/separability. To assess this, the Inter-Cluster Cosine Distance was calculated by constructing a distance matrix, measuring the cosine distance between each cluster centroid and all others. Consistent with the MICCD, cosine distance was selected due to its effectiveness in capturing similarity between vectors. This analysis was conducted separately for clusters based on Positive Sentiment, Positive Rating, Negative Sentiment, and Negative Rating. The resulting distance matrixes were then visualized using a box plot, enabling a clearer interpretation of cluster distinctiveness across these four categories.

$$\text{Parwise Inter Cluster Cosine Distance} = 1 - \frac{c_i \cdot c_j}{\|c_i\| \cdot \|c_j\|} \quad (2)$$

where c_i and c_j are the centroids of clusters i and j

To examine potential correlations and relationships, some statistical analyses were conducted, including Ordinary Least Squares (OLS) regression and the Spearman correlation test. The OLS regression model was employed to assess whether a linear relationship exists between MICCD and the percentage of positive reviews (both sentiment and rating) within each cluster. This model inherently involves Pearson correlation, which evaluates the strength and direction of linear associations. In parallel, the Spearman correlation test was applied to detect any monotonic relationships between the same variables. Following these analyses, hypothesis testing was performed, and conclusions were drawn based on the results of both tests. If the findings indicate a negative impact of the percentage of positive reviews on

MICCD, whether through a linear relationship (as demonstrated by the OLS and Pearson analyses) or a monotonic relationship (as indicated by the Spearman test), this would support the Anna Karenina Principle. In the other hand, the absence of such a negative relationship would suggest insufficient evidence to confirm this principle.

3.7 SUCCESS AND DISSATISFACTION FACTORS ANALYSIS

To identify the success and dissatisfaction factors of Portuguese restaurants, we first utilized BERTopic to extract the most relevant words associated with each set of topics (Positive and Negative). For each set, we compiled all relevant words into a unique corpus, one for positive clusters and another for negative clusters. Using these corpora as a foundation, we conducted a word cloud analysis to gain an overview of the topics and themes characterizing each group of clusters. This approach was inspired by the study development by Bilgihan (Bilgihan et al., 2018), where tag clouds were used for a qualitative analysis when studying success factors. In this way was possible to visualize the predominant factors influencing customer perceptions of satisfaction and dissatisfaction within Portuguese restaurants.

4. RESULTS AND DISCUSSION

This section presents and analyzes the results obtained from the implemented methodology. The findings are interpreted and discussed in the context of the relevant insights and in relation to the previous literature review.

4.1 RESULTS

To examine the overall distribution and behavior of the clusters, a general analysis was performed by plotting the two groups of clusters (positive and negative) independently and in combination in two dimensions, using BERTopic and UMAP. This approach provided an overview of data behavior with the default model settings.

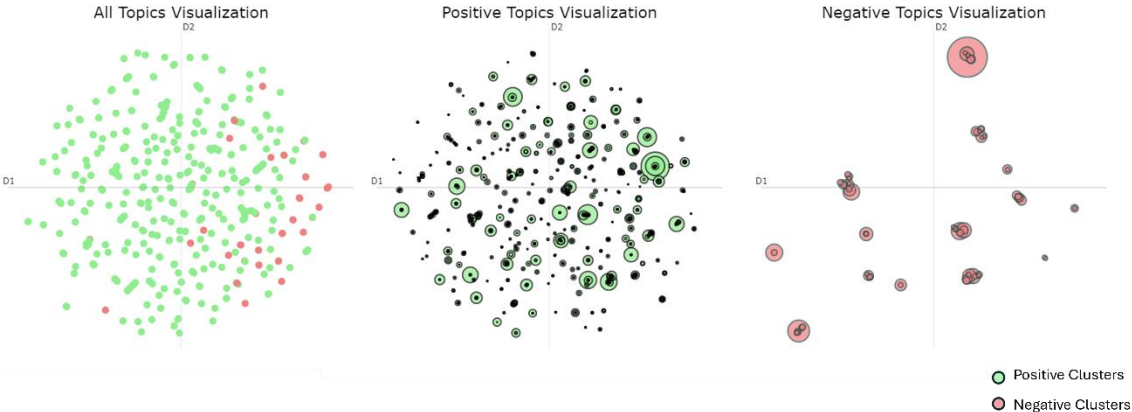


Figure 3 - Spatial Distribution of Topics in the Dimensional Embedding Space

The plots in the Figure 3 were organized to present three perspectives: All Topics, Positive Topics, and Negative Topics. A total of 1218 clusters were identified, with 1132 positive clusters and 86 negative clusters. Figure 3 reveals a significant disparity in the number of positive versus negative topics, mirroring the imbalance in positive and negative review volumes noted during data exploration.

Number of Reviews by Sentiment Compound Scores

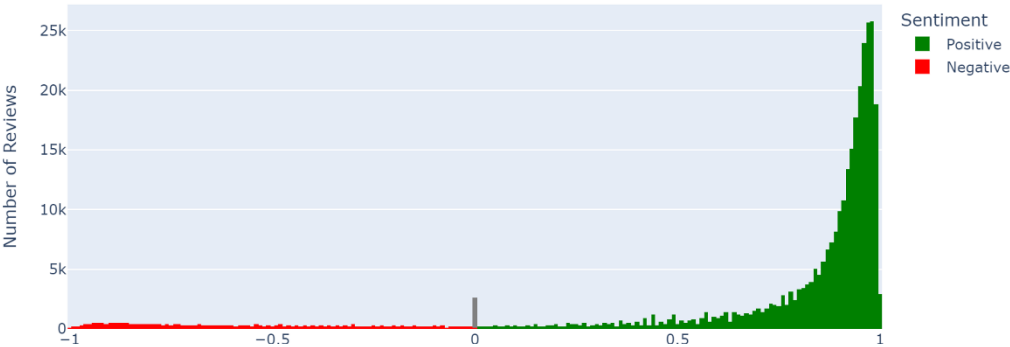


Figure 4 - Number of Reviews By Sentiment Compound Scores

4.1.2 Internal Clusters' Variability Analysis

To pursue the investigation about the AKP applicability, it was examined whether there were differences in the internal variability of negative versus positive clusters. Specifically, we analyzed whether the mean intra-cluster cosine distance of clusters varied according to the percentage of positive reviews within each cluster.

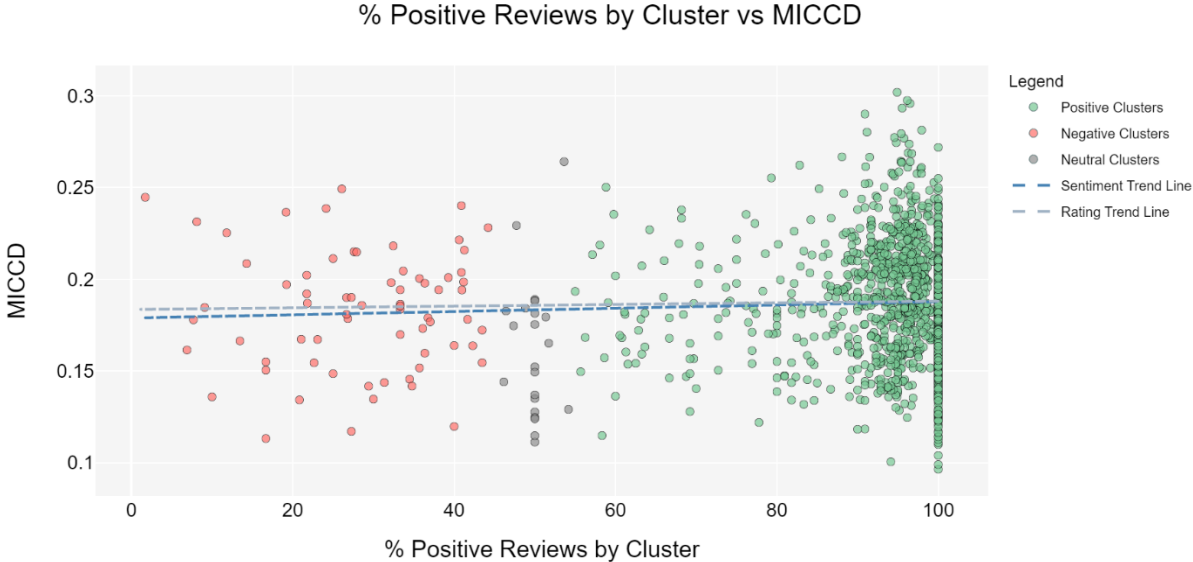


Figure 7 - MICCD vs. Percentage of Positive Reviews by Cluster (Entire Dataset)

This approach aimed to identify a potential relationship between these two variables. A negative relationship would suggest that clusters with higher proportions of positive reviews are gradually more tightly grouped, indicating decreased variability and a focus around specific themes. The analysis began by plotting the mean intra-cluster cosine distance of each cluster against the percentage of positive sentiment and rating reviews for each cluster. This step aimed to investigate the existence of a linear correlation between these variables.

The resulting scatter plot indicated that positive clusters displayed a broader range of MICCD values, with a notable concentration of clusters where over 90% of the reviews were positive. However, despite these observations, the MICCD exhibited minimal variation across the spectrum of positive sentiment and rating percentages. This lack of variation suggests an absence of correlation between the two variables, indicating that sentiment and rating positivity do not significantly linearly affect the internal cohesion of the clusters.

Sentiment Trend Line Equation

$$\text{MICCD} = \beta_{0s} + \beta_1 \cdot \% \text{Positive Sentiments} + \epsilon \quad (3)$$

Rating Trend Line Equation

$$\text{MICCD} = \beta_{0r} + \beta_1 \cdot \% \text{Positive Ratings} + \epsilon \quad (4)$$

Table 2 presents some relevant values from the Ordinary Least Squares (OLS) regression analysis. The p-values for the % Positive Sentiment Coefficient and the % Positive Rating Coefficient were 0.095 and 0.278, respectively. For this analysis, a significance level of $\alpha = 0.05$ was adopted.

Table 2 - Results of OLS Analysis for the Entire Dataset

Trend Line	Variable	Coefficient	Standard Error	P-Value
Sentiment	β_{0s}	0.179	0.005	0.101
Sentiment	β_{1s}	8.907×10^{-5}	0.000	0.095
Rating	β_{0r}	0.183	0.004	0.000
Rating	β_{1r}	4.510×10^{-5}	4.16×10^{-5}	0.278

The results show that both values exceed the 0.05 significance threshold (based on a 95% confidence interval). This indicates that the null hypothesis cannot be rejected, supporting the conclusion that the β_{1s} and the β_{1r} do not have a statistically significant effect. This means that there is no evidence of a significant linear correlation between the MICCD (Mean Intra-Cluster Cosine Distance) and the percentage of positive sentiment and rating reviews.

Although a linear relationship between the two variables could not be confirmed, a Spearman correlation analysis was conducted to explore the possibility of a monotonic relationship. The analysis examined the correlation between MICCD and the percentage of positive sentiment and rating reviews by cluster. The Spearman correlation coefficients were found to be -0.091 (p-value = 0.001) for positive sentiment and -0.057 (p-value = 0.046) for positive ratings. For this test we also considered $\alpha=0.05$.

Table 3 - Results of Spearman Correlation Analysis for the Entire Dataset

Type of positive reviews	Spearman Coefficient	Correlation	P-Value
Sentiment	-0.091		0.001
Ratings	-0.057		0.046

Looking to the Table 3 we can verify that both p-values are smaller than 0.05, demonstrating statistical significance at the 95% confidence level confirming a weak negative monotonic relationship.

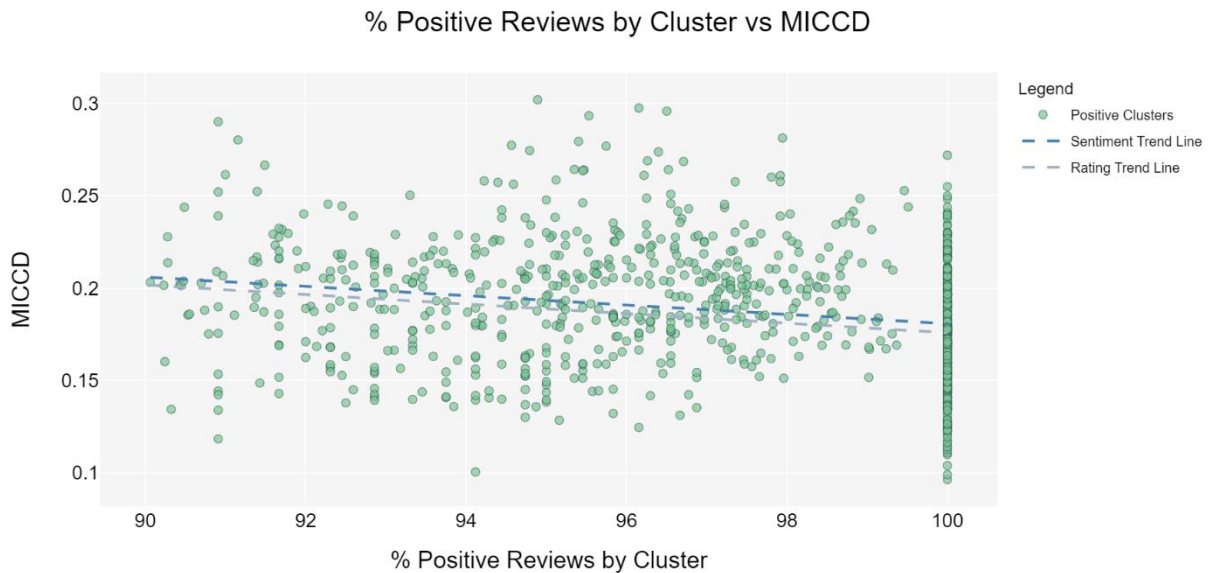


Figure 8 - MICCD vs. Percentage of Positive Reviews by Cluster (Clusters >90% Positive Reviews)

In an additional attempt to gain insights into the potential applicability of the Anna Karenina Principle (AKP), the same Ordinary Least Squares (OLS) and Spearman analyses were conducted specifically in areas where the clusters exhibited a higher concentration, defined as instances where the percentage of positive reviews within a cluster exceeded ninety percent. This analysis is significant as it may provide insights how the potential outcomes would look like if there was a more homogeneous dataset regarding the positive and negative proportions of reviews.

Table 4 - Results of OLS Analysis for Clusters with >90% Positive Reviews

Trend Line	Variable	Coefficient	Standard Error	P-Value (Rounded)
Sentiment	β_{0s}	0.433	0.036	0.000
Sentiment	β_{1s}	-2.5×10^{-3}	0.000	0.000
Rating	β_{0r}	0.435	0.035	0.000
Rating	β_{1r}	-2.6×10^{-3}	0.000	0.000

The analysis indicates that the slopes for sentiment ($\beta_{1s} = -0.0025$) and rating ($\beta_{1r} = -0.0026$) suggest a statistically significant negative linear relationship with the Mean Intra-Cluster Cosine Distance (MICCD), with rounded p-values of 0.000. Specifically, for each one-unit increase in the percentage of sentiment or rating reviews by cluster, the MICCD is expected to decrease by approximately 0.025 and 0.026 units, respectively. Furthermore, the intercepts for sentiment ($\beta_{0s} = 0.433$) and rating ($\beta_{0r} = 0.435$) are also statistically significant, both yielding rounded p-values of 0.000. The statistical significance of the slopes, reflected in the p-values of 0.000, allows us to reject the null hypothesis. Based on the null hypothesis test conducted, we conclude that there is substantial evidence to support a linear negative relationship between the percentages of both sentiment and rating reviews by cluster and the MICCD.

Table 5 - Results of Spearman Analysis for Clusters with >90% Positive Reviews

Type of positive reviews	Spearman Correlation Coefficient	P-Value (Rounded)
Sentiment	-0.242	0.000
Ratings	-0.264	0.000

Based on the results from Table 5, we can conclude that both p-values are less than 0.05, the findings indicate statistical significance at the 95% confidence level. This suggests that there is evidence of a weak negative monotonic relationship between the Mean Intra-Cluster Cosine Distance (MICCD) and the percentage of positive reviews.

Another analysis was conducted to examine how the β_1 s coefficient and the corresponding p-value change when considering all negative clusters and only the top X% of the largest positive clusters. The intent behind this analysis was to understand if removing the smaller positive clusters would function similarly to reducing noise in the dataset. In each iteration, the positive and negative clusters were separated. The positive clusters were filtered to include only the top X% clusters based on quantity of reviews. Afterwards, the negative clusters and the selected positive clusters were recombined, creating a new dataset containing all the negative clusters along with the top X% of positive clusters based on quantity of reviews. An Ordinary Least Squares (OLS) regression was then performed on this dataset. This process was repeated for each iteration to observe changes in β_1 s and its p-value across different filtering levels.

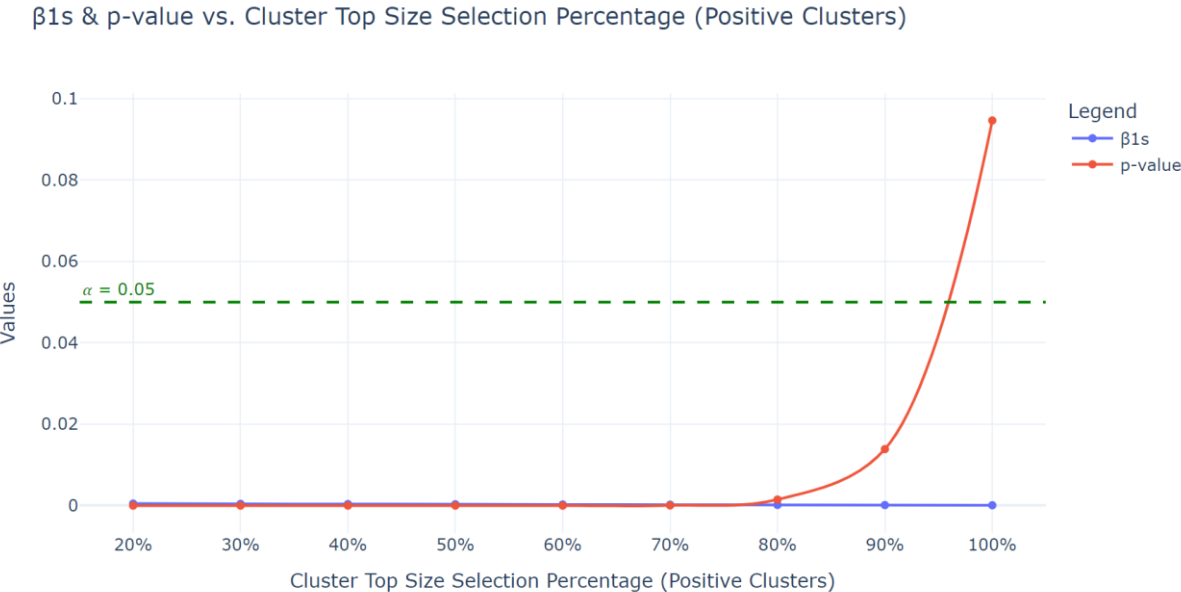


Figure 9 - β_1 s and p-value vs. Cluster Top Size Selection % (Positive Clusters)

Based on Figure 9, we can conclude that decreasing the number of positive clusters based on their size does not significantly impact the value of β_1 s. However, the statistical significance of β_1 s increases as the number of positive clusters decreases, which can be attributed to this noise reduction effect. In the figure, it is possible to see that all the p-values below 90% are below the 0.05 threshold, proving the β_1 s significance. The results support the idea that there is no negative correlation between MICCD and the percentage of positive reviews within each cluster.

A further investigation was performed by applying the cluster reduction process to the entire set of clusters, including both positive and negative clusters. The top X% criterion was applied uniformly across the entire dataset. The Figure 10 indicates that, under this approach, the p-value fluctuates when the top X% decreases, moving β 1s further from statistical significance. This outcome may be attributed to the reduced number of negative clusters, which generally contain fewer reviews. Since the reduction process affects all clusters, negative clusters are one of the firsts to be removed. This removal disrupts the balance between the positive and negative classes, potentially contributing to the observed high variation in p-values.



Figure 10 - β 1s and p-value vs. Cluster Top Size Selection % (All Clusters)

4.1.2 Cluster Distinctiveness Analysis

To assess the distinctiveness of the clusters, it was essential to analyze the distances between each cluster and all the others. For this analysis, we measured the cosine distance between the centroids of each cluster and plotted the resulting matrix in the form of a box plot, which facilitates a clearer understanding of the distribution of values. The objective was to determine whether positive and negative clusters exhibited different distributions and values, thereby drawing conclusions regarding these differences in the context of the Anna Karenina Principle (AKP).

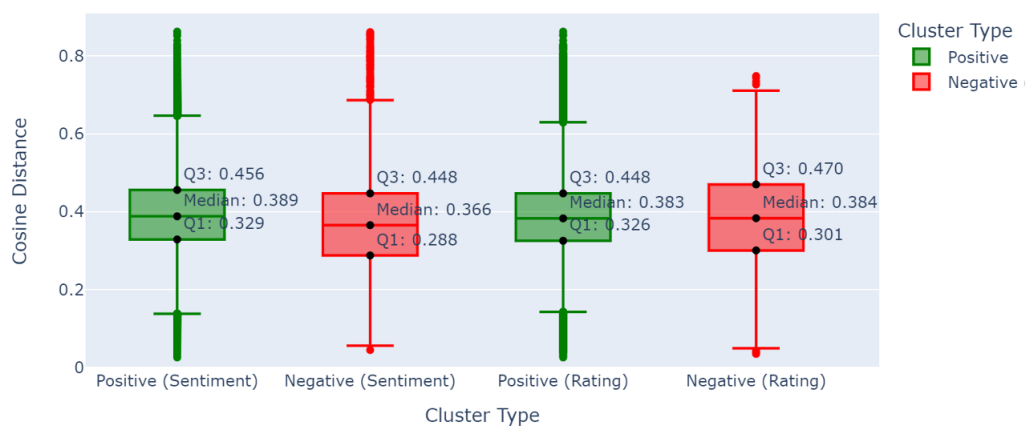


Figure 11 - Pairwise Inter-Cluster Cosine Distances by Sentiment and Rating Clusters

The boxplot displays the inter-topic cosine distance distribution for positive and negative topics concerning both Sentiment and Rating scores. For Positive Sentiment clusters, the median cosine distance is 0.389, with an interquartile range (IQR) of 0.127. In contrast, Negative Sentiment clusters have a median cosine distance of 0.366 and a wider IQR of 0.160. Regarding the positive rating clusters the median cosine distance is 0.383, with an interquartile range (IQR) of 0.122. For negative rating clusters the median cosine distance is 0.384, with an interquartile range (IQR) of 0.169.

The results indicate that Positive Sentiment clusters have a higher median inter-cluster cosine distance but a smaller interquartile range (IQR) than negative sentiment clusters. This suggests that while positive sentiment clusters are slightly more spaced out, they maintain a consistent distance pattern, implying a focus around core themes. In contrast, negative sentiment clusters display a greater IQR, likely reflecting a wider range of reasons for dissatisfaction and thus more diverse thematic clusters. For Positive and Negative Rating clusters, both have similar median distances, but positive rating clusters exhibit a smaller IQR, indicating a more stable clustering pattern around specific theme. Negative rating clusters, with a larger IQR,

suggest more variation in clustering, reflecting a broader diversity in negative review content. The pattern between sentiment clusters and rating cluster is similar. This observation aligns with the notion that positive feedback may focus on specific aspects of satisfaction, while negative feedback has in consideration a broader spectrum of dissatisfaction sources.

4.2 DISCUSSION

The results do not provide clear evidence to support the presence of the AKP principle in this dataset, despite insights suggesting its potential existence:

- Across the entire dataset, a weak monotonic negative relationship appears to exist between the Mean Intra-Cluster Cosine Distance and the percentage of positive sentiment and rating reviews per cluster.
- Focusing on clusters with over 90 percent of positive (sentiment and rating) reviews, a linear and monotonic negative relationship between the two variables is observed..
- The analysis of inter-cluster cosine distances further corroborates the idea of AKP, with positive clusters exhibiting greater homogeneity and negative clusters displaying more heterogeneity.
- The qualitative analysis about the themes of the clusters reveals a greater variety of negative topics than positive ones.

While these findings indicate the potential presence of AKP, no strong correlation and significant values have been identified to prove the existence of this principle within this study's context. Positioning this study within the broader literature enables further refinement of conclusions. Previous research has noted that negative reviews are often more distinct and varied, reflecting the complexities in interpreting them.

In the context of Portuguese restaurants online reviews, this study suggests that negative clusters tend to be more distinct than positive ones, as evidenced by the larger IQR in inter-cluster cosine distances. Positive reviews cover a broader array of topics focused on the same success factors, with variations such as different food types or restaurant styles. Negative topics, while fewer, focus on distinct and diverse issues, as confirmed by cluster distinctiveness analysis showing a larger IQR. However, these findings may be influenced by the discrepancy in the volume of positive and negative reviews, as a greater number of reviews increases the possibility of identifying new themes within clusters.

Regarding variability, although minor evidence suggests a difference between negative and positive clusters, the results do not confirm that the variability of negative reviews was a primary factor in the ineffectiveness of topic identification through Latent Dirichlet Allocation (LDA) in previous studies. Although this research suggests that negative clusters may be

slightly more dispersed than positive ones, attributing topic identification challenges solely to "greater variability" in negative clusters may be misleading, as this variability appears minimal. Topic modeling methods like LDA are subject to certain limitations, and the disparity in the quantity of positive versus negative reviews, coupled with the nuanced nature of negative reviews, may contribute to these limitations in LDA, rather than cluster variability alone.

With regard to success and failure factors for the Portuguese restaurants, this research identifies primary determinants of restaurant success as food quality, food type, and location, while failure factors are more diverse, encompassing aspects such as waiting times, pricing, service quality, and food quality. Specifically, it can be concluded that customers primarily express positive sentiments regarding food quality. While other factors such as price, service, and ambiance also influence customer satisfaction (as described in the literature review), these are more frequently mentioned when they fail to meet expectations. A noteworthy observation is that customers generally expect a certain standard of satisfaction in terms of price, ambiance, and service, which is why these aspects tend to be highlighted in negative reviews when they fall short. In contrast, food quality stands out as a factor that has the potential to positively impact customer experience, encouraging repeat visits to the restaurant. Still regarding the success factor analysis, the methodology used in this study is considered more consistent and reliable. It employs BERTopic to extract the most relevant words within each cluster, enhancing the credibility of the base corpus for word cloud analysis. While a bag-of-words approach is also used in word cloud analysis, this corpus is notably more robust. By deriving top words from clusters using BERTopic and leveraging c-TF-IDF to identify contextually significant terms, this approach improves interpretability and relevance compared to a general preprocessed corpus, allowing deeper insights when analyzing through word frequency.

5. CONCLUSIONS AND FUTURE WORKS

This study aimed to investigate further the presence of the Anna Karenina Principle (AKP) using state-of-the-art tools within the context of Portuguese restaurants, and identify the main factors driving their success and failure. A novel methodology, distinct from previous literature approaches, was implemented, leveraging restaurant reviews as the foundation for analysis. Large Language Models, such as SentenceTransformer for semantic understanding, UMAP for dimensionality reduction, and BERTopic for clustering, were utilized to address the research questions. This approach proved to be more robust than previous methods, enabling the extraction of meaning and context from textual data and clarifying some empirical statements made in the previous studies.

The applicability of the Anna Karenina Principle within this context cannot be substantiated due to the limited strength and significance of the observed relationships. Evidence indicates that negative topics are slightly more varied and distinct than positive ones. Across the dataset, a weak negative monotonic relationship was observed between the mean intra-cluster cosine distance of reviews in each topic and the percentage of positive reviews by topic, suggesting that negative topics exhibit more significant variability. Furthermore, negative topics appear more dispersed within the dimensionality space, while positive topics tend to cluster more closely. This pattern suggests that negative topics are marginally more distinct than positive ones. The qualitative analysis of success and failure factors in Portuguese restaurants supports these findings. Negative topics were more distinctive, covering a range of issues such as price, waiting times, food quality, and service. In contrast, positive topics were more concentrated, focusing primarily on food (quality and type) and restaurant location. These findings align with the AKP, as they reveal a greater diversity and distinctiveness among failure factors than success factors. However, we have some evidence suggesting that the findings need to be more robust to make a definitive assertion regarding the applicability of the AKP in this context.

The research results align closely with the previous studies, addressing key limitations and clarifying prior assumptions. However, further enhancing the credibility of this research could be achieved by exploring alternative embeddings, dimensional reduction, or clustering models beyond those currently employed. Such comparisons would help assess the robustness of our findings, allowing us to determine whether the weaker statistical results arise from the dataset itself or methodological limitations. Trying another type of metric to measure the variability within the clusters, such as silhouette score or entropy, would have also made this analysis more robust.

Additionally, a more homogeneous dataset with an even distribution of positive and negative reviews would be ideal, although challenging to obtain. Reducing the number of positive reviews to achieve balance might limit the study's real-world applicability, while gathering more negative reviews is difficult, as poorly performing restaurants may close, resulting in

fewer negative reviews to extract, making it difficult to obtain the real opinions of customers fully.

Another limitation encountered was the measurement of intra-cluster cosine distances. The initial planned approach aimed to calculate the intra-cluster cosine distance by considering all pairwise distances between points within each cluster and averaging those distances, giving a more robust representation. However, a simplified alternative was adopted due to the computational complexity associated with the high number of reviews. This method involves calculating the distance between each point and its assigned cluster centroid and then computing the mean of these distances. This adjustment was necessary to ensure computational feasibility while still providing a measure of cluster cohesion.

For future investigations into the applicability of the Anna Karenina Principle (AKP) in this context, it is advisable to explore another clustering methods, in conjunction with various dimensionality reduction and embedding techniques. Furthermore, augmenting the analysis of results with additional metrics, such as entropy and silhouette scores, would provide a more comprehensive understanding of the data. To improve the identification of success and failure topics, applying generative AI algorithms like GPT for reviewing and describe customer feedback automatically would be an alternative to the qualitative analysis performed in this study.

Given the ongoing advancements in state-of-the-art sentiment analysis and text mining techniques, integrating these methods into this context would be prudent. Such an approach has the potential to yield deeper insights into customer behavior and perceptions within the Portuguese restaurant sector.

BIBLIOGRAPHICAL REFERENCES

- Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic Topic Modeling: An Experimental Study on BERTopic Technique. *Procedia Computer Science*, 189, 191–194. <https://doi.org/10.1016/j.procs.2021.05.096>
- Allaoui, M., Kherfi, M. L., & Cheriet, A. (2020). Considerably Improving Clustering Algorithms Using UMAP Dimensionality Reduction Technique: A Comparative Study. In A. El Moataz, D. Mammass, A. Mansouri, & F. Nouboud (Eds.), *Image and Signal Processing* (Vol. 12119, pp. 317–325). Springer International Publishing. https://doi.org/10.1007/978-3-030-51935-3_34
- Al-Natour, S., & Turetken, O. (2020). A comparative assessment of sentiment analysis and star ratings for consumer reviews. *International Journal of Information Management*, 54, 102132. <https://doi.org/10.1016/j.ijinfomgt.2020.102132>
- Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics* (arXiv:2008.09470). arXiv. <https://doi.org/10.48550/arXiv.2008.09470>
- Bilgihan, A., Seo, S., & Choi, J. (2018). Identifying restaurant satisfiers and dissatisfiers: Suggestions from online reviews. *Journal of Hospitality Marketing & Management*, 27(5), 601–625. <https://doi.org/10.1080/19368623.2018.1396275>
- Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4), 77–84. <https://doi.org/10.1145/2133806.2133826>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null), 993–1022.
- Bornmann, L., & Marx, W. (2011). The Anna Karenina principle: A mechanism for the explanation of success in science. *CoRR*, *abs/1104.0807*.
- Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Advances in Knowledge Discovery and Data Mining* (pp. 160–172). Springer. https://doi.org/10.1007/978-3-642-37456-2_14

- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805). arXiv. <https://doi.org/10.48550/arXiv.1810.04805>
- Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7. <https://doi.org/10.3389/fsoc.2022.886498>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. <https://doi.org/10.48550/ARXIV.2203.05794>
- Hutto, C., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216–225. <https://doi.org/10.1609/icwsm.v8i1.14550>
- Hwang, J., & Zhao, J. (2010). Factors Influencing Customer Satisfaction or Dissatisfaction in the Restaurant Business Using AnswerTree Methodology. *Journal of Quality Assurance in Hospitality & Tourism*, 11(2), 93–110. <https://doi.org/10.1080/15280081003800355>
- Kirilenko, A. P., Stepchenkova, S. O., & Dai, X. (2021). Automated topic modeling of tourist reviews: Does the Anna Karenina principle apply? *Tourism Management*, 83, 104241. <https://doi.org/10.1016/j.tourman.2020.104241>
- Le, Q. V., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents* (arXiv:1405.4053). arXiv. <https://doi.org/10.48550/arXiv.1405.4053>
- Liu, K., Bellet, A., & Sha, F. (2019). *Similarity Learning for High-Dimensional Sparse Data* (arXiv:1411.2374). arXiv. <https://doi.org/10.48550/arXiv.1411.2374>
- McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. <https://doi.org/10.48550/arXiv.1802.03426>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (arXiv:1301.3781). arXiv. <https://doi.org/10.48550/arXiv.1301.3781>

- Mukhacheva, M. I. (n.d.). *The application of Anna Karenina principle on restaurants online reviews: A Portuguese case study.*
- Pacheco, L. (2018). An analysis of online reviews of upscale Iberian restaurants. *Dos Algarves: A Multidisciplinary e-Journal*, 32. <https://doi.org/10.18089/DAMEJ.2018.32.3>
- Sangaraju, V. R., Bolla, B. K., Nayak, D. K., & Kh, J. (2022). *Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings.* <https://doi.org/10.48550/ARXIV.2205.07259>
- Schuckert, M., Liu, X., & Law, R. (2015). Hospitality and Tourism Online Reviews: Recent Trends and Future Directions. *Journal of Travel & Tourism Marketing*, 32(5), 608–621. <https://doi.org/10.1080/10548408.2014.933154>
- Tabinda Kokab, S., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. *Array*, 14, 100157. <https://doi.org/10.1016/j.array.2022.100157>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (arXiv:1706.03762). arXiv. <https://doi.org/10.48550/arXiv.1706.03762>
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers* (arXiv:2002.10957). arXiv. <https://doi.org/10.48550/arXiv.2002.10957>
- Wang, Y., Kim, J., & Kim, J. (2021). The financial impact of online customer reviews in the restaurant industry: A moderating effect of brand equity. *International Journal of Hospitality Management*, 95, 102895. <https://doi.org/10.1016/j.ijhm.2021.102895>
- Wu, C., Che, H., Chan, T. Y., & Lu, X. (2015). The Economic Value of Online Reviews. *Marketing Science*, 34(5), 739–754. <https://doi.org/10.1287/mksc.2015.0926>

Zanetta, L. D., Xavier, M. C., Hakim, M. P., Stedefeldt, E., Zanin, L. M., Medeiros, C. O., & da Cunha, D. T. (2024). How does the consumer choose a restaurant? An overview of the determinants of consumer satisfaction. *Food Research International*, 186, 114369. <https://doi.org/10.1016/j.foodres.2024.114369>

Zhang, Z., Ye, Q., Law, R., & Li, Y. (2010). The impact of e-word-of-mouth on the online popularity of restaurants: A comparison of consumer reviews and editor reviews. *International Journal of Hospitality Management*, 29(4), 694–700. <https://doi.org/10.1016/j.ijhm.2010.02.002>



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa