NOVA IMS
Information Management School

MGI

Master Degree Program in

Information Management

The application of Anna Karenina principle on restaurants online reviews: a Portuguese case study

Mariya Igorivna Mukhacheva

Master Thesis

presented as partial requirement for obtaining the Master Degree in Information Management

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 restaurants online reviews: a Portuguese case study

by

Mariya Igorivna Mukhacheva

Master Thesis presented as partial requirement for obtaining the Master's degree in Information Management, with a specialization in Knowledge Management and Business Intelligence

Supervised by

Professor Flávio Pinheiro, Ph.D., NOVA IMS - Information Management School

November, 2023

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, 30/11/2023]

ABSTRACT

After an experience with a service or product, consumers tend to share their impressions online. For companies is crucial to keep track of this information by collecting and analyzing it. Due to the large volume, unstructured nature, and emotional polarity, online reviews require automatic analytical treatment. Text mining, topic modelling, and sentiment analysis are challenging approaches to handle the need to obtain knowledge from online reviews. However, in this context, questions may arise regarding the efficiency of the automatic interpretation of online reviews. The main objective of this study is to understand if the effect of TAK (The Anna Karenina) principle is observed in online reviews of Portuguese restaurants collected from TripAdvisor. We used methods such as Latent Dirichlet Allocation and Sentiment Analysis. TAK principle is applied in different areas, so on the one hand it can be considered universal for any context. Nevertheless, it is important to note that the principle was formed recently, and it is only at the beginning of its path of application. Further studies are needed to test whether TAK principle is verified in various contexts and if it is confirmed it would mean that the TAK limits the application of topical analysis when the nature of the data has great variability, that is, mostly customer dissatisfaction topics. This study contributes to important discoveries that have managerial implications in the sense of proposing specific strategies and approaches that can improve the customer experience, and theoretically by contributing to the universality of TAK principle.

KEYWORDS

Anna Karenina Principle; Topic Modelling; Sentiment Analysis; Portuguese restaurants; TripAdvisor;

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

BERT	**Bidirectional Encoder Representations from Transformers**
e-WOM	**Electronic Word-of-Mouth**
LDA	**Latent Dirichlet Allocation**
MAD	**Microbiome-Associated Diseases**
STEM	**Science, Technology, Engineering, and Math**
TAK	**The Anna Karenina principle**
VADER	**Valence Aware Dictionary and sEntiment Reasoner**
VOC	**Voice of the Customer**
WOM	**Word-of-Mouth**

1. INTRODUCTION

"Happy families are all alike; every unhappy family is unhappy in its own way."

Anna Karenina, Leo Tolstoy

Social media platforms increasingly influence the relationship between customers and companies (Sweeney et al., 2008; Golmohammadi et al., 2019). Customers share their experience of products and services through online reviews, creating a huge volume of opinionated data and contributing to e-WOM (electronic Word-of-mouth) (Golmohammadi et al., 2019). Such data are crucial for consumers' decisions because they serve as social proof that establishes an emotional connection and consequently help the consumer to make more informed choices (Golmohammadi et al., 2019; Ruiz-Mafe et al., 2018; Marcolin et al., 2020). Regarding restaurants, e-WOM is an important source for influencing consumers' decisions (Zhang et al., 2010). Positive reviews affect reputation and sales (Eastin et al., 2011; Golmohammadi et al., 2019). As such, customer satisfaction is an essential point for companies. Therefore, they must collect information from online reviews, analyze it, and extract important insights to understand the customers' needs and sentiments better.

However, due to their large volume and unstructured nature, online reviews require advanced analytical methods to extract information on the main topics reviewers discuss and their expressed sentiments. Kirilenko et al. (2021) suggest applying TAK principle to study topics from tourist reviews. TAK principle is based on the well-known first lines of Tolstoy's novel "Anna Karenina". In a nutshell, the principle of Anna Karenina states that a successful marriage requires success in many essential aspects —physical attraction, financial compatibility, parenting philosophies, religious beliefs, relationships with in-laws, etc. — while failed marriages can fail in many ways by failing in only a few of those. As a result, "happy families" are all the same, while "unhappy families" are each unhappy in their own way.

Kirilenko et al. (2021) used Tripadvisor online reviews to show that reviews with 4- to 5-stars are generally easily interpreted by the LDA (Latent Dirichlet Allocation) topic model: 67% of topics on average were "good" interpreted, and only 6% "poor." However, on average, 42% of negative reviews were poorly interpretable, and only 22% were highly interpretable. That means the variability of topics from satisfied tourists is smaller than those from dissatisfied tourists. The findings suggest that the larger number of issues in negative reviews could result in limitations for automated topic modeling. The study further confirms these findings using a dataset of reviews from the Terracotta Army Museum in China.

In fact, there is a lack of research about the effect of TAK principle on online reviews. There is also a gap in the literature regarding how this effect relates with review sentiment analysis. Sentiment analysis is widely used by researchers to detect text polarity, it can be an alternative or can be used as a complementation to star rating (M. Tardin et al., 2020; S. Al-Natour, Al-Natour & Turetken, 2020; B. Liu, 2012). Is the sentiment extracted from reviews a more important indicator of TAK principle effect than the score rating?

In the restaurant business, many factors contribute to successful outcomes. Previous studies on online restaurant reviews have mainly focused on four key aspects of service quality, namely food, service, atmosphere, and price. These factors have demonstrated stability in restaurant reviews over time (Rita et al., 2023). According to Rita et al., (2023) food and service were the most mentioned items in reviews of restaurants. In addition, consumers put food taste and quality first in their search for the perfect restaurant (Zhang et al., 2010; Bilgihan et al., 2018). However, menu, environment, operational efficiency, cleanliness, innovation, price, and location are important factors that may need to be mentioned when customer experience is good. These factors are often set aside when the customer is satisfied with the meal and everything else. Hence, is this behavior a pattern in positive reviews? And in a scenario of dissatisfaction, when something went wrong, or several things went wrong, does the nature of reviews follow a different pattern? The answers to these questions are an essential contribution to the universality of TAK and in the fields related to customer satisfaction. If TAK is confirmed, it would mean that the polarity of reviews is a factor to consider when is applied unsupervised topical analysis algorithms.

Here, and following the work of Kirilenko et al. (2021), we study if the TAK principle applies to reviews of Portuguese restaurants. The remainder of this thesis is organized in the following manner: Chapter 2 provides a literature review of the concept and application of TAK principle and the importance of online reviews for companies and customers; Chapter 3 discusses the data used for this study and methodologies employed, such as VADER (Hutto et al., 2014) for Sentiment Analysis and LDA for topic modeling; Chapter 4 provides results and their discussion; Chapter 5 provides the main conclusions of this work, contributions, limitations, and future research opportunities.

2. LITERATURE REVIEW

2.1. Concept of Anna Karenina Principle

TAK principle can be considered a phenomenon, which means it can be observed and studied, and in fact, it isn't just observed in marriages. It is also studied in other fields. Jared Diamond, a professor of geography at the University of California, popularized the TAK principle in his book "Guns, Germs, and Steel" (1997). Diamond explains that the TAK principle is applied to animal domestication and defines six reasons for failed domestication (diet, growth rate, problems of captive breeding, nasty disposition, tendency to panic, and social structure). For example, zebras have never been domesticated because they become impossibly dangerous as they age. Thus, many wild animal species reached the first stage of animal-human relations, leading to domestication. A century ago, the British scientist Francis Galton concisely explained this discrepancy: "It would appear that every wild animal has had its chance of being domesticated, that those few which fulfilled the above conditions were domesticated long ago, but that the large remainder, who sometimes fail in only one small particular, are destined to perpetual wildness so long as their race continues" (Galton, 1883).

The article "Females and STEM: The Anna Karenina Principle" proposes that the TAK principle can help explain the gender disparity in STEM (science, technology, engineering, and math) fields. The principle suggests that it is more effective in complex systems to examine the multiple necessary conditions that must be met rather than look for a single direct cause. The paper concludes that playing a gender role in society requires a complex list of necessary conditions for success in STEM, which may be longer for girls than for boys (Tanas et al., 2020).

Bornmann Lutz and Werner Marx (2011) use the TAK principle to explain success in science. Success in science requires multiple key aspects to be satisfied, and failure occurs if even one part is missing. The paper focuses on three critical areas where limited resources often hinder success: (1) the evaluation of research grant proposals and manuscripts through peer review, which requires scarce resources such as funding and journal space; (2) the citation of publications, which depends on their reception and is thus a scarce resource; and (3) the discovery of new scientific knowledge, which requires recognition as a scarce resource. Achieving success in these areas is contingent on meeting several key resource allocation prerequisites. Failure to meet any of these prerequisites results in the rejection of grant proposals, manuscript submissions, published papers, or scientific discoveries.

Dwayne R.J. Moore (2001) affirms that successful ecological risk assessments are alike; every unsuccessful ecological risk assessment fails in its way. The author argues that the TAK principle applies to ecological risk assessments involving multiple stressors. To be successful, the assessment must have societal and political support, consider a wide range of stressors and options, follow a rigorous process, have resources for model development and data collection, and have an adaptive management strategy. Failure to meet any of these criteria prevents the use of multiple stressor assessments in effective decision-making (Moore, 2001).

Another curious study is how the TAK principle helps understand which variables are important to explain variance. By the author's interpretation, the TAK principle implies that while no single factor ensures success, multiple factors can lead to failure. TAK bias is a logical consequence. When we observe the variables that best explain the variance in the universe, we mostly analyze successful observations, creating a model for success rather than focusing on failure. The important variables may become constant at the senary of success, but they start to fluctuate in the scenario of failure. Survivors, by definition, may have little to no variation in the most crucial factors for survival. This can lead to overlooking the most crucial variables for survival. The Anna Karenina bias focuses on attributes with the lowest variation, which could be the most crucial variables for survival, as low variation implies a requirement for survival. Additionally, this bias implies that studying many non-survivors is needed to identify all relevant variables related to survival (Shugan, 2007).

The TAK principle is also can be applied to financial markets. Baur (2022) defined happy companies as companies with positive returns and unhappy companies as companies with negative returns and then looked at whether the happy companies were more like each other than the unhappy ones. The research used various measures, such as return correlations, cross-sectional returns, and cross-sectional volatility dispersion, and found that the principle did not hold true in most cases. However, it was found that unsuccessful companies had higher average return volatility than successful companies, suggesting that they were under more stress. This conclusion may have implications for investors, and unhappy companies can impact portfolio risk (Baur, 2022).

Zhanshan Ma (2020) suggests that the TAK principle can be used in biology. In terms of the MAD (microbiome-associated diseases), the TAK principle may be translated into a hypothesis: all “healthy” microbiomes (of healthy individuals) are alike; each diseases-associated microbiome is “sick” (of a patient with MAD) in its own way. Studies demonstrate that the TAK principle exists in 50% or more of the MAD cases (Ma, 2020).

The TAK principle can be justified mathematically. When looking at the dynamics of correlation and variance in systems, “in crisis, typically, even before obvious symptoms of crisis appear, correlation increases, and, at the same time, variance (and volatility) increases too” (Gorban et al., 2010). This means that well-adapted systems resemble one another, but all non-adapted systems are non-adapted in its own way.

2.2. Anna Karenina Principle on Online Reviews

Kirilenko et al. (2021) argue that the growth of Web 2.0 has opened the door to new research and services based on reviews shared by tourists about their experience and identifies two important areas of review analysis: (1) Sentiment Analysis, which has been widely researched (2) Topic Analysis, that is not yet fully understood. Topical analysis aims to understand, categorize, and summarize the content of a set of documents (corpus). In this field, applications of Latent Dirichlet Allocation in tourism are gaining popularity, leaving behind methods such as Singular Value Decomposition. However, the limitations of the Latent Dirichlet Allocation method are still largely unknown.

According to Kirilenko et al. (2021), the principle of Anna Karenina is still very little applied in tourism. The principle suggests that reviews from satisfied customers are "more similar" than those from dissatisfied customers. This does not mean that there are fewer topics in positive reviews than in negative ones. Instead, positive topics are more likely to be shared across multiple reviews. In practice, this leads to less separability of topics in negative reviews and, as a result, to the appearance of topics that are poorly interpreted.

Based on 14,273 reviews about tourist attraction points, Kirilenko et al. (2021) confirmed the hypothesis that Latent Dirichlet Allocation analysis results on the datasets of low customer satisfaction are significantly less interpretable than the analysis of positive reviews. Thus, the effect of the Anna Karenina Principle was confirmed, which consists of limiting the use of topic modeling to identify the main topics of customer dissatisfaction. However, more research is needed to confirm this result.

2.3. IMPORTANCE OF ONLINE REVIEWS

WOM (Word-of-Mouth) refers to the exchange of experiences and opinions between consumers. It is, thus, a process of social influence relying on interpersonal communication of behaviors or attitudes (Sweeney et al., 2008). Online and through digital platforms, WOM is often called e-WOM (electronic Word-of-Mouth), defined as any positive or negative statements made by customers about a product, service, or company accessible to many people and institutions, and that can result in social influence (Ruiz-Mafe et al., 2018).

There are two main types of e-WOM: information-oriented contexts — such as consumer reviews on e-commerce platforms — and review websites (Eastin et al., 2011). Moreover, e-WOM can also be individual- or emotional-oriented when consumer opinions are shared on social media and online communities (Eastin et al., 2011). The e-WOM attracts particular attention due to its influence on consumers' buying behaviors and company performance. In particular, when the services are intangible, are produced and consumed concurrently, cannot be evaluated before consumption, and have experience and credence attributes that are difficult to evaluate (Golmohammadi et al., 2019). Thus, there appears to be an urgent need for service consumers to reduce pre-purchase uncertainty through access to enhanced information retrieval about service (Golmohammadi et al., 2019).

An online review is a spontaneous and usually free-of-charge opinion regarding an experience (Marcolin et al., 2020). It is common for customers to express their satisfaction or disappointment regarding a particular product by writing a review. Nowadays, online reviews can be found in many places and grouped as business and location-based, product-based, niche-specific reviews (like movies, restaurants, jobs, and companies), blogs, and social media platforms. In the late 1990s, Amazon introduced a 5-star rating system that has since become a commonly utilized tool (Bilgihan et al., 2018). Generally, ratings consist of a numerical evaluation (with 5-stars indicating the highest level of satisfaction and 1-star indicating the lowest level) and a written review, in which they can give more detailed feedback on the product, including pros and cons. Which, in turn, helps the user to choose the best service or product according to their needs.

Results show that interaction with other online community members is the most significant factor for determining positive e-WOM. Online communities such as TripAdvisor and Booking have empowered consumers to engage in product-related e-WOM and have become a valuable means of promoting marketing and e-commerce efforts (Ruiz-Mafe et al., 2018). According to TripAdvisor, 96% of users consider reading reviews important when booking a hotel, and 83% reference reviews before making their buying decision (Feinstein, E., 2018). Before deciding where to eat, 90% of users conduct online research on restaurants, which is more than any other type of business (ReviewTrackers, 2022). Online reviews can also affect business efficiency. The study by Kim et al. (2016) about the impact of social media reviews on restaurant business has shown that overall customer ratings and online reviews significantly impact restaurant performance. Online reviews are an effective source of information about customer feedback that enriches e-WOM. They are also important for customer research methods such as the VOC (Voice of the Customer).

The VOC is a customer research method that organizations employ to learn more about their customers' needs. It helps companies understand service quality from the customer perspective: what customers want and why they make specific decisions (Marcolin et al., 2020). As the name suggests, VOC focuses on customer feedback and the shared impression. There are many ways to get customer feedback, for example, through surveys or interviews, but text data like customer online reviews on their initiative can be very useful. It is probably the best way to identify their sentiments and thoughts about the products or services. The collection of reviews is a great source for VOC insights analytics. With the growing adoption of tools and platforms that allow customers to share opinions about previous experiences, this source is increasingly used as a decision tool for service or product choice (Marcolin et al., 2020).

A good reputation plays an important role in organizations. The optimal approach to building trust involves a comprehensive understanding of customers, which can be achieved through feedback collection. By analyzing the feedback provided by customers, potential customers can form a reliable perception of the reputation of a company or product. Online review users with more experience can influence others to take the final purchase step or promote consumers in the preliminary stages of the purchase journey. Showing innovative ideas to new customers can form initial preferences and stimulate positive emotions and purchase intentions (Ruiz-Mafe et al., 2018). In addition, the review lets the company know their clients' opinions regarding products and services. Businesses need to collect the most relevant comments from their customers to define the future strategy of bringing buying decisions forward, increasing conversion, reinforcing brand loyalty, and decreasing product returns. In the context of the reputation of companies, online reviews play an important role, notably negative reviews, but this conclusion is not so evident. Previous research indicates that, in general, negative e-WOM could have damaging consequences for services. At the consumer level, negative e-WOM is seen to be more credible than positive e-WOM. However, the negative effects can be moderated by contextual factors such as psychological distance and prior investments in information search, which could change the impact of negative e-WOM. When consumers feel psychologically closer to the service, i.e., they perceive the service or product as more personally relevant or important to them, the temporal investments in information search reduce the negative e-WOM impact on their behavior. Hence, the impact is amplified when the consumer is distal (Golmohammadi et al., 2019). Nonetheless, low-involvement consumers are more motivated by review quantity, positively affecting online shoppers' acquiring purpose (Sheng-Hsien Lee,2009).

Other studies showed that customers do not trust companies with lower than 4-star ratings (ReviewTrackers, 2022). However, as described in Figure 1, statistics show that reviews about food registered positive sentiments in 70% of online reviews, which is an important indicator in the food industry. Although negative feedback can have a significant impact, it does not represent the majority of reviews in the food industry. Any kind of feedback is helpful. It is possible to transform a bad customer experience into a good one with appropriate online review management through consumer interaction. However, it is a management that needs to be improved because only 23% of restaurateurs reach out to the reviewer when dealing with in-person feedback or negative reviews (ReviewTrackers, 2022).

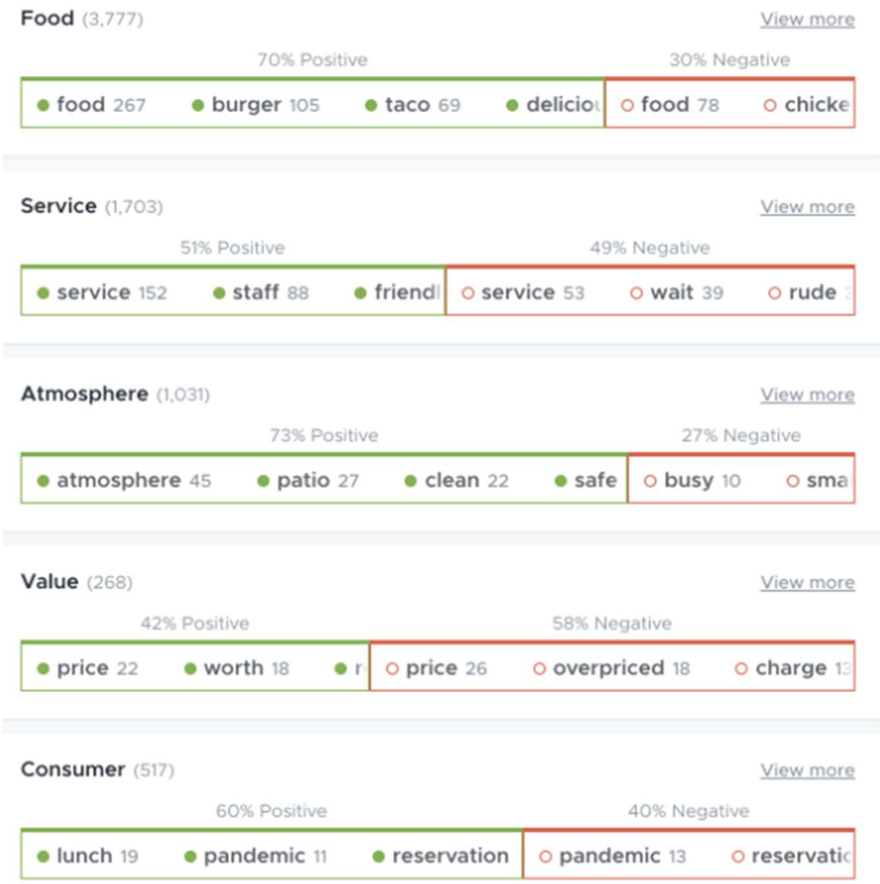

Figure 1– Analysis by ReviewTrackers of reviews sentiments across main experiences during the last 6 month (ReviewTrackers, January 9, 2022).

An important aspect in examining reviews is their "tone" or "sentiment", which reflects the polarity of opinion. Star ratings are widely used to quickly indicate the sentiment of a review. However, they have limitations, often being biased and not matching the sentiment expressed in the review text. Therefore, it is important not to consider star ratings as the only indicator of sentiment, but to take a deeper look at the reviews. This highlights the importance of automatic sentiment analysis as a technique for identifying the true sentiment of text, serving as an alternative or complement to star ratings (Al-Natour & Turetken, 2020).

Academics in marketing, information systems, and tourism often use online reviews from location-based review sites in their research. According to a recent market study, TripAdvisor is among the top 5 websites consumers use to find local business information (Brightlocal, 2020). TripAdvisor is a comprehensive travel guidance platform that helps users plan their trips and provides information on everything from routes and hotels to local gastronomy and attractions. With over 988 million reviews and opinions in 43 markets and 22 languages, TripAdvisor is a valuable resource for consumers and businesses, providing insights into flights, facilities, experiences, and more (Tripadvisor, 2021). Restaurants and cafes are the businesses consumers most frequently seek online advice (Brightlocal, 2020).

3. DATA AND METHODS

Figure 2 presents the workflow of the next steps, defining the main guidelines for the implementation and analysis process to produce results that answer the research question.

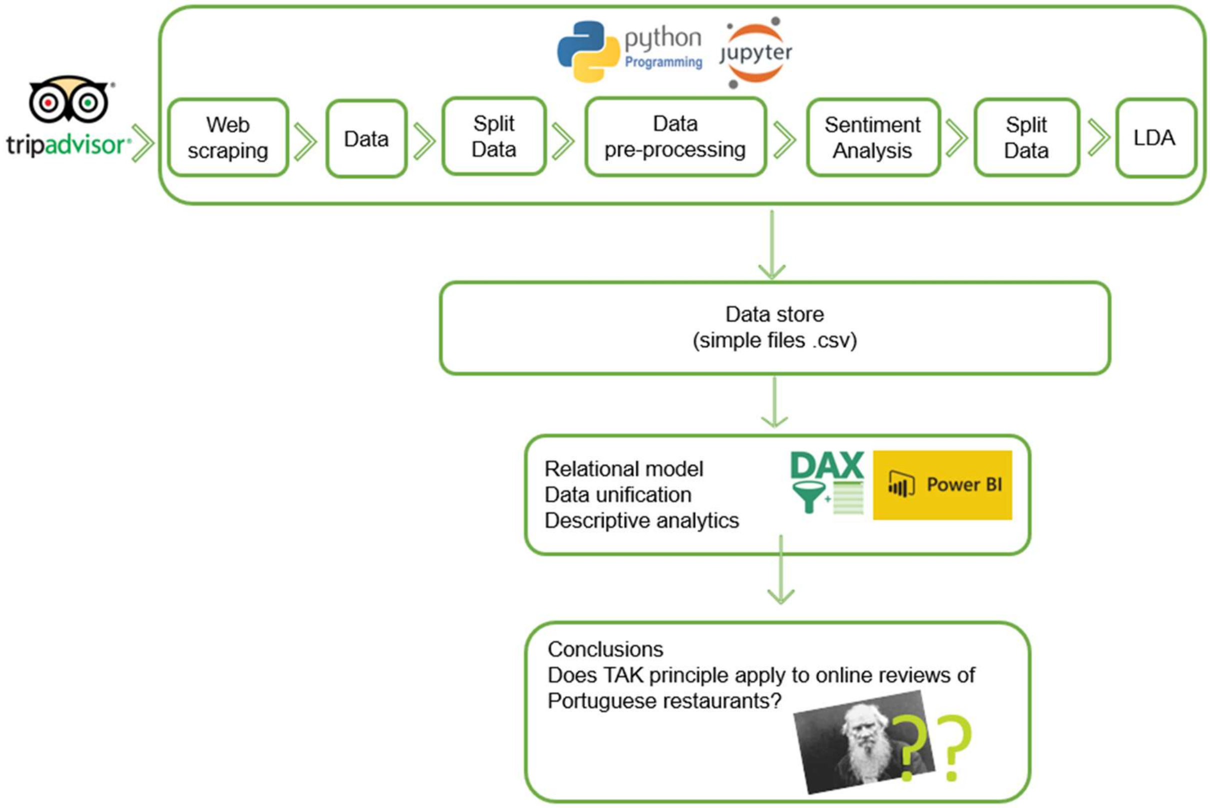

Figure 2 – The framework of Methodology.

The data obtained through web scraping was split into subsets by star ranking and then subjected to pre-processing. Then, we use VADER (Hutto et al., 2014) to identify the reviews' polarity as a complement to the users' star classifications. Hence, user classification and sentiment analysis score divide the data into positive and negative review subsets. Subsequently, we use LDA to identify the topics in each subset. The following subsections describes in more detail such steps as data description, pre-processing process, application of VADER (Hutto et al., 2014), and application of LDA.

3.1. Data Description

The basis of this work is to analyze online reviews from TripAdvisor about Portuguese restaurants, see Table 1. For this purpose, data was sourced through web scraping. We used a Python script to access TripAdvisor and extract relevant data, see Table 1. Since, in most cases, English reviews are displayed as the default, see Figure 3 – Panel A, we decided to extract only the first and second pages of reviews written in English for each restaurant sorted by date in descending order, which corresponds to a maximum of 30 most recent reviews per restaurant. This approach was also taken to avoid distorting the review data due to a high volume of reviews from a small number of restaurant outliers, as shown in Figure 3 – Panel B, 29% of restaurants have 2 or fewer reviews, and 76% have 30 or fewer reviews.

Table 1 – The data sourced from TripAdvisor.

Bubbles	**Bubble Score of review (1- to 5-star rating)**
RatingDt	Rating date
Quotes	Quotes of review
Review	Full review text
Region_URL	Region URL
Restaurant_URL	Restaurant URL
Adress	Adress of restaurant
Gastronomy	Gastronomy of restaurant

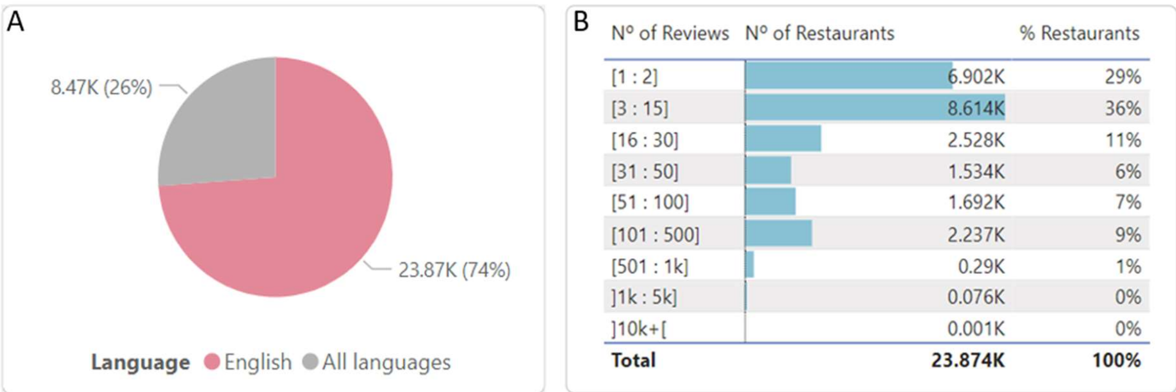

Figure 3 – Panel A shows the percentage of restaurants with reviews written in English by default. Panel B shows the number of reviews per restaurant written in English by default.

The dataset includes approximately 307k online reviews in English for approximately 24k Portuguese restaurants with dates between 2007 and 2022, however, most of the reviews are concentrated between 2017 and 2022 (Figure 4 – Panel A). Of the collected reviews, 60% have a rate of 5-stars, 21% have 4-stars, and the remaining 19% consist of 7% 3-star reviews, 4% 2-star reviews, and 7% 1-star reviews (Figure 4 – Panel B).

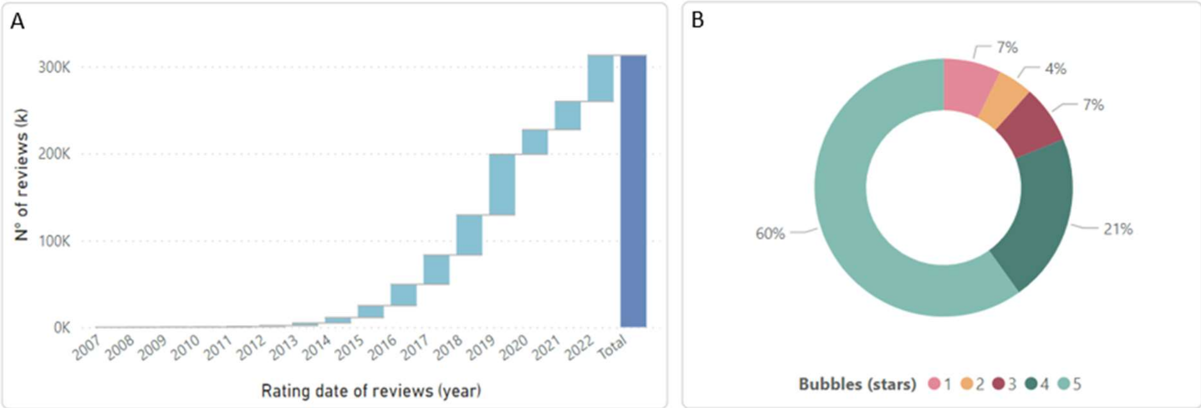

Figure 4 – Panel A shows the number of collected reviews per year. Panel B shows the percentage of reviews divided by bubbles (stars: 1-star to 5-star rating).

It is important to note that there are few restaurants with an average score below 3.5-stars, accounting for only 8%. On the other hand, there are few restaurants with a maximum average score of 5-stars, accounting for only 9%. In general, the statistics shows that, on average, 83% of the restaurants have rating 4-star and above (Figure 5).

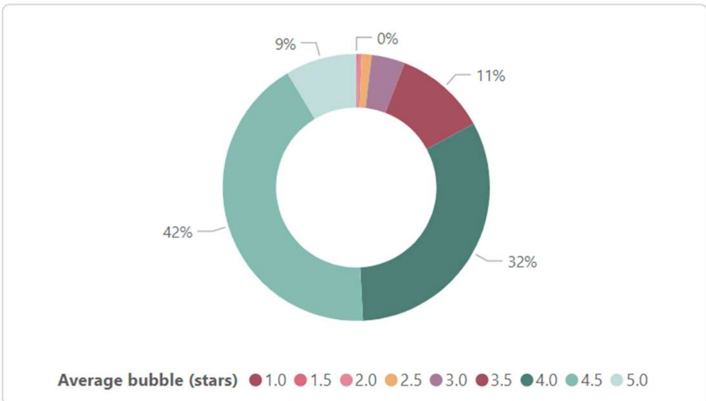

Figure 5 – The rounded average score per restaurant.

Regarding the geographic distribution of restaurants, Lisbon is the most represented region, accounting for 27% of restaurants, followed by the North region with 26%, see Figure 6 – Panel B. Concerning the number of reviews, Algarve and Lisbon are the regions with the highest number of reviews, accounting, respectively, for 29% and 28% of the total reviews, see Figure 6 – Panels A and C. Such density of reviews in the Algarve and Lisbon can be explained by the fact that it is a tourist attraction area.

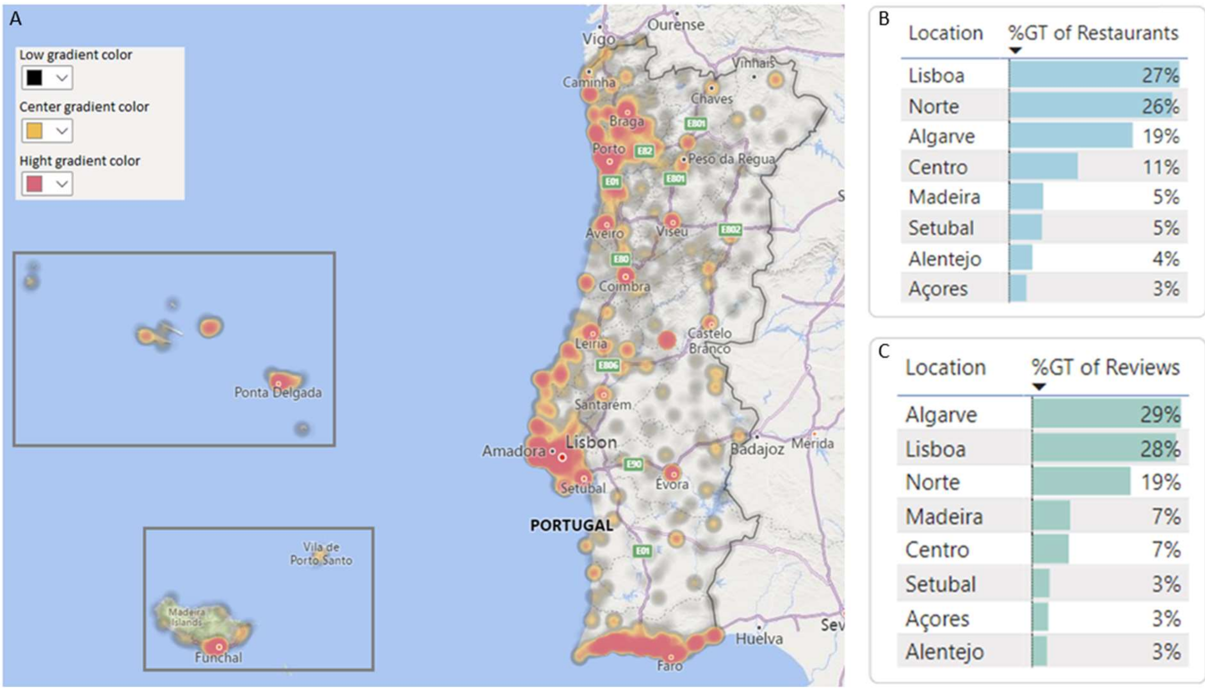

Figure 6 – Panel A shows the Heat map of a number of reviews per localization; gradient: black - low; yellow - middle; rose – high. Panel B shows the percentage of restaurants per location. Panel C shows the percentage of reviews per location.

To obtain a working dataset, we followed several steps: transformation of words to lower case and removal of punctuation. Tokenization with the removal of short and long tokens, as well as English stop words, numbers, and characters; stemming; cleaning of tokens by parts-of-speech (POS), retaining just nouns (indicative of product features) and adjectives (represent sentiments); filtering of the most frequent (in more than 70% of documents) and rare words (less than 1.5% of documents. The above steps are exemplified in Figure 7 (Atkinson-Abutridy, 2022; Kirilenko et al., 2020).

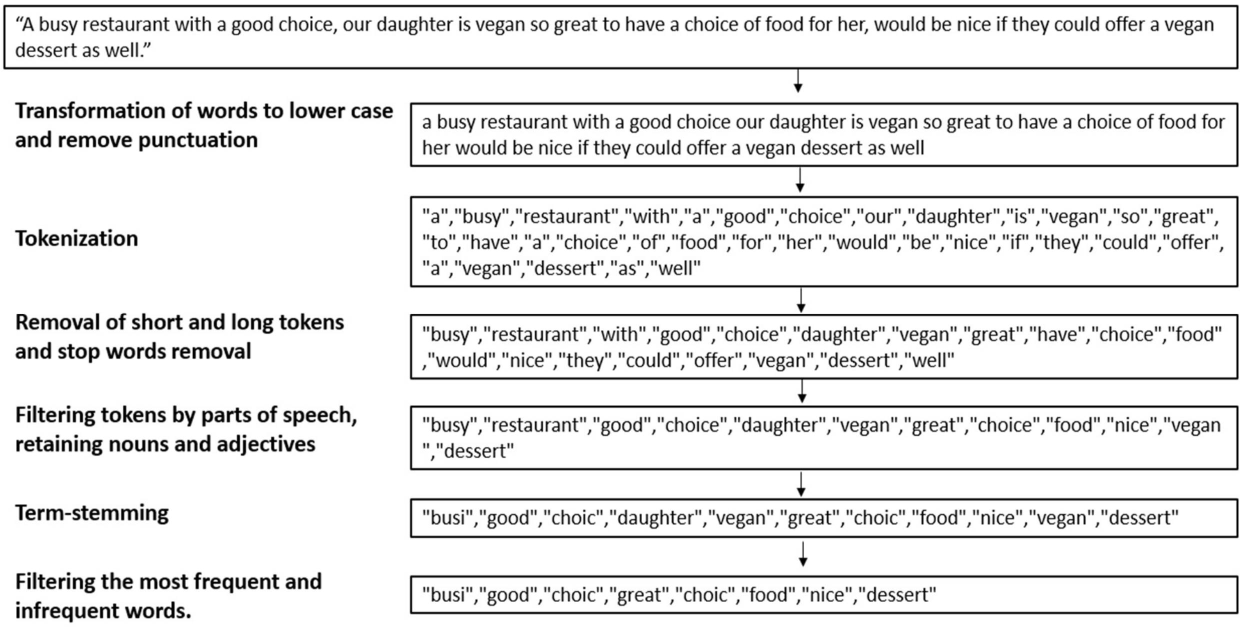

Figure 7 – Standard pre-processing steps.

3.2. SENTIMENT ANALYSIS AND TOPIC MODELLING

Hutto & Gilbert (2014) found that VADER (Valence Aware Dictionary and Sentiment Reasoner) performed better than other sentiment analysis tools when analyzing tweets for sentiment. VADER (Hutto et al., 2014) is a lexicon rule-based sentiment analysis tool specifically designed to analyze the sentiment of text written on social media. The VADER (Hutto et al., 2014) analyzer is trained on a human-curated sentiment lexicon that includes words and phrases commonly used in social media. It also considers the intensity of the sentiment conveyed by these words and phrases, as well as the presence of negations and other linguistic features that can influence the sentiment of a sentence.

The pre-trained VADER (Hutto et al., 2014) model was used in each review to calculate the associated sentiment. The output of VADER (Hutto et al., 2014) is a sentiment score named "compound" that represents a given text's positivity, negativity, and neutrality. This score is a continuous value between -1 (most negative) and +1 (most positive). The following thresholds were defined to classify the reviews: positive: "compound" ≥ 0.05; neutral: -0.05 < "compound" < 0.05; negative: "compound" ≤ -0.05.

We use LDA to identify topics in each subset of reviews (Atkinson-Abutridy, 2022). The key idea behind LDA is that documents are generated through a probabilistic process involving two assumptions: that similar topics tend to use similar words and that each document can be represented as a mixture of multiple topics. LDA requires the number of topics K as an input parameter. In that sense, we use several approaches to identify the optimal number of K* topics, namely the "elbow method” in combination with the Perplexity and Coherence scores (Kirilenko et al., 2021). Perplexity measures the model's ability to predict the words in a document based on identified topics. For optimal k, the perplexity should be as small as possible. Coherence measures the semantic quality of topics, *i.e.*, evaluates the frequencies with which the high-ranking and lower-ranking words composing the same topic tend to appear together in documents. For optimal K, coherence should be as high as possible. Moreover, according to the literature, extracting too many topics should be avoided (Cao et al., 2009). As such, we limit the maximum number of topics to 15.

4. RESULTS AND DISCUSSION

The amount of information in text format on the Internet is growing exponentially. Online reviews are increasingly influencing the decisions of both companies and consumers. It is extremely important to have automatic mechanisms for identifying key topics, but it is also important to have results close to reality.

In addition to the primary goal, the description of the data, namely the number of restaurants, their geographic distribution, and the number of reviews per restaurant, gives a global view of the food service industry in Portugal. Around 24k restaurants with reviews written in English were identified, of which approximately 81% were positive, with 4- and 5-star ratings. A large concentration of restaurants can be found in areas of large cities: Lisbon; Porto; Braga; and in several cities in the Algarve region. In gastronomic terms, more than 55% of restaurants offer Portuguese or European food. Looking at each restaurant separately, we can see that most restaurants (approximately 90%) have 100 or fewer online reviews written in English. This means that only approximately 10% of restaurants may experience a lack of automatic methods of extracting knowledge from reviews.

The Anna Karenina principle suggests that it is important to separate success from failure, as these two groups behave differently. Successful cases tend to be similar, while each unsuccessful case has unique reasons for failure. The research question of this thesis is whether the Anna Karenina principle applies to the approximately 307k reviews of Portuguese restaurants extracted from TripAdvisor. In the case of online restaurant reviews, success is a 5- and 4-star review, that is, total or mostly positive customer satisfaction. 1- and 2-star reviews are considered cases of failure, total or mostly negative customer dissatisfaction. Of course, in the real world, middle-of-the-road scenarios and intermediate 3-star ratings reflect partial customer satisfaction. But even within a 1-star review, a small “success” can be found, and inversely. On the other hand, customer feedback is not always sufficiently explanatory; in other words, the textual content may not correspond to the classification by stars and may only serve as a supplement. Therefore, it is important to identify the sentiment based on the content of the review.

Before deep analysis began, the data was divided into two groups: positive reviews (4- and 5-stars) and negative reviews (3-, 2-, and 1-star) based on star rating, with significantly different sample sizes: one with approximately 248k positive reviews and another with about 59k negative reviews. Then, for each group, data pre-processing was applied, consisting of three main blocks: text unification and stop word filter; filtering by word class, leaving only nouns and adjectives; filtering rare and common words.

This work aims to understand whether the nature of positive reviews differs from negative ones. That's why we started by looking at the word count of the reviews, which is a metric to gauge whether customers tend to write more negative reviews. In detail, distinct variables were established to analyze the raw and pre-processed data to enable quantification and comparison of outcomes across datasets, see Table 2.

Table 2 – Constructed metrics: number of words per review

Name of metric	Tested Data
N_Words_0	Number of words in the original reviews (without pre-processing).
N_Words_1	Number of words in the pre-processed reviews without filtering of the most frequent and rare words.
N_Words_2	Number of words in the pre-processed reviews (complete method).

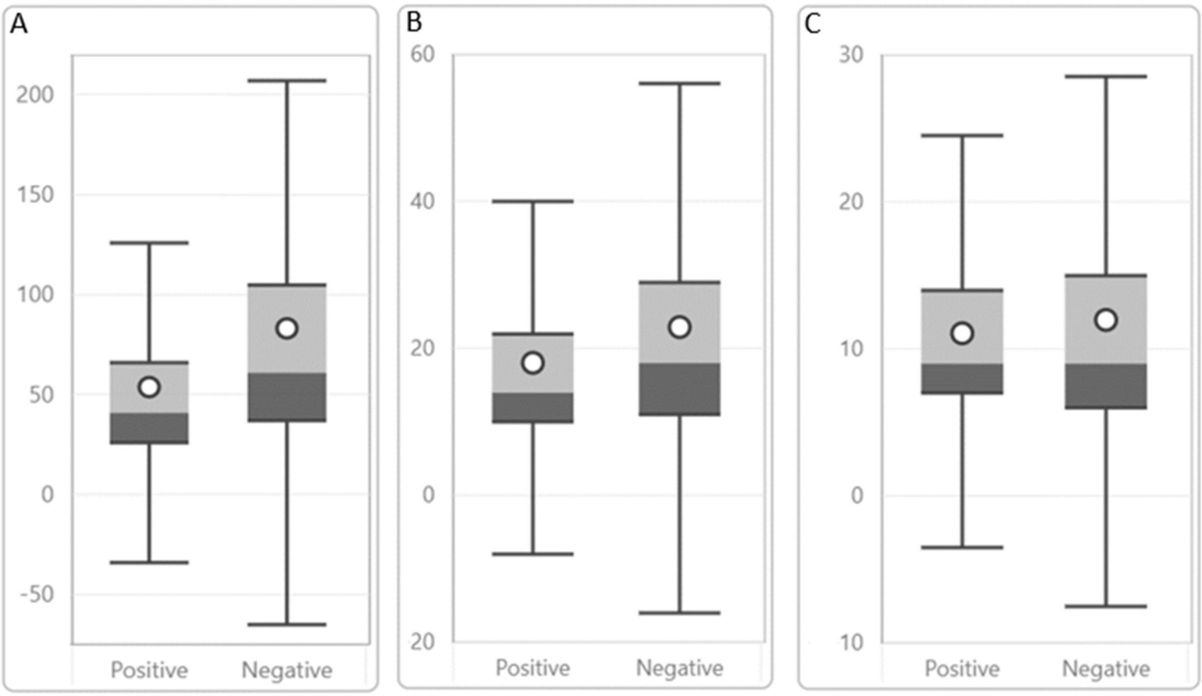

Figure 8 – Boxplot of number of words per reviews per subset of positive and negative reviews. Panel A shows the distribution of the number of words in the original reviews (without pre-processing). Panel B shows the distribution of the number of words in the pre-processed reviews without filtering of the most frequent and rare words. Panel C shows the distribution of the number of words in the pre-processed reviews (complete method).

Table 3 – Statistics from Figure 8

Panel of Figure 8	Data	Q_0	Median	Average	Q_3	Minimum	Maximum	Outliers
Panel A	Positive reviews	28	44	57	70	-35	133	6%
	Negative reviews	37	61	83	-64	-64	205	6%
	Total	29	45	61	-42	-42	147	6%
Panel B	Positive reviews	10	15	18	22	-8	40	5%
	Negative reviews	11	18	23	28	-15	54	5%
	Total	10	15	19	23	-10	43	5%
Panel C	Positive reviews	7	10	12	15	-5	27	4%
	Negative reviews	6	10	13	16	-9	31	5%
	Total	7	10	12	15	-5	27	4%

Figure 8 shows the Boxplot of the length of reviews per subset, which displays the distribution of the dataset based on the summary of statistics such as the minimum, the maximum, the median, the average, and the first and third quartiles, see Table 3.

It is observed that the mean is higher than the median in both subsets, and the percentage of outliers per star is the same, therefore, it has the same effect on all data. Negative reviews' interquartile range (IQR) is larger than positive reviews. The mean and median of negative reviews are higher than those of positive reviews. The text volume of negative reviews (an average of 83 words) is significantly larger than that of positive reviews (an average of 57 words). Negative reviews tend to have more varied words than positive reviews. It can mean a higher volume of terms, indicating that the number of topics mentioned in negative reviews is greater than in positive reviews. Therefore, indicates that positive reviews disperse into a smaller population of topics. That is, reviews are more similar. Although negative reviews diverge on a wider range of topics, negative reviews are less similar. However, a large amount of text is not always more informative. Therefore, the number of words in the review was also analyzed in the pre-processed data.

When we compare the differences between the two subsets (average number of words in the subset of negative reviews – the average number of words in the positive subsets), it becomes clear that there is a much more significant difference between reviews without pre-processing; when we pre-process reviews, this difference is minimized (Figure 9). As a result, we further confirm that in the pre-processed data the same trend continues to exist, however with a reduced difference. The amount of content removed in negative reviews is greater than in positive ones. Considering that the last step is to exclude all rare words, that is, those words with a low occurrence, we can conclude that there are more rare words in negative reviews than in positive ones. This suggests that negative reviews are less similar to each other than positive reviews.

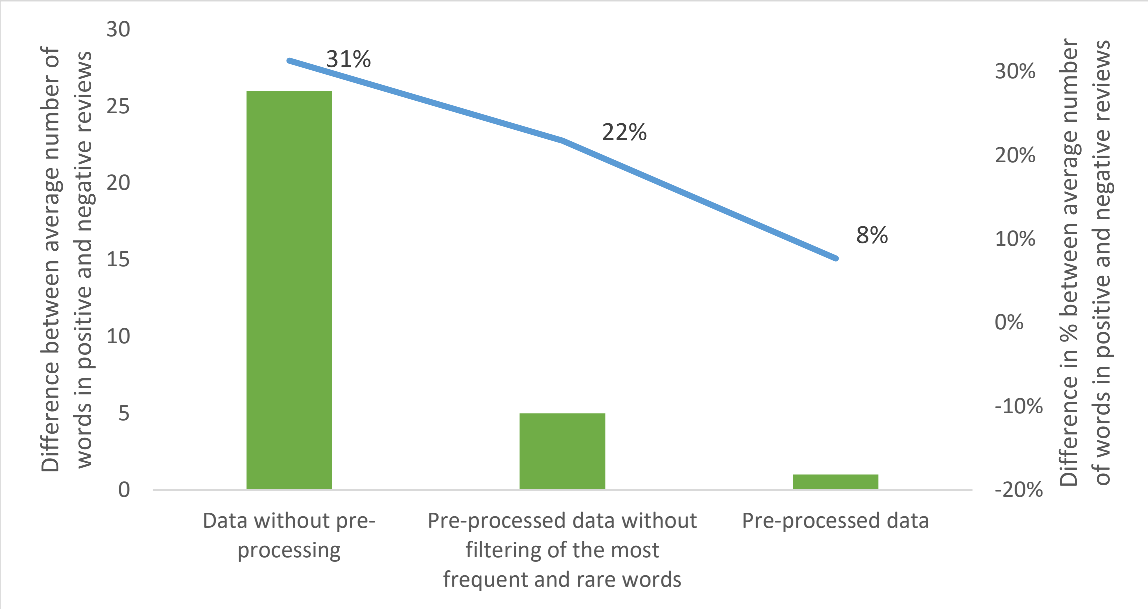

Figure 9 – Comparison of averages between subsets of positive and negative reviews by pre-processing phase

Hypothesis tests were conducted to understand if a difference between subsets is significant statistically. Non-parametric tests can be used with ordinary or categorical variables, where the sample does not need to be normally distributed and are consequently more flexible. The non-parametric Mann-Whitney U test was applied to data without outliers to test the null hypothesis (H_0) of that the two populations of negative and positive reviews are equal. Based on the hypothesis test results, we can affirm that the number of words in positive reviews is smaller than those in negative reviews with statistical significance, p=0.00 (Table 4).

Table 4 – Results of Mann-Whitney U test

Hypotheses	**Tested Data**	**Mann-Whitney U Statistic**	**p-Value**	**Result**
H_1: The Number of words in positive reviews < The Number of words in negative reviews	Number of words in the original reviews (without pre-processing).	9072557.5	0.0	Reject the null hypothesis.
	Number of words in the pre-processed reviews without filtering of the most frequent and rare words.	10707056.5	0.0	Reject the null hypothesis.
	Number of words in the pre-processed reviews (complete method).	11764342.5	0.0	Reject the null hypothesis.

To better understand the main terms of the final pre-processed data, a dictionary was created of different words for each subset of positive and negative reviews. The dictionary contains 324 words, of which 199 belong to both, positive and negative reviews, 31 belong exclusively to positive reviews, and 94 belong exclusively to negative reviews. In Figure 10, the number of words associated exclusively with negative reviews is higher than those associated exclusively with positive reviews. People may be more descriptive when they are dissatisfied. Furthermore, it may indicate that negative experiences vary more than positive ones. Positive reviews can be more repetitive, with many people using similar words to express their satisfaction, while negative experiences can vary widely, requiring different words to capture the nuances.

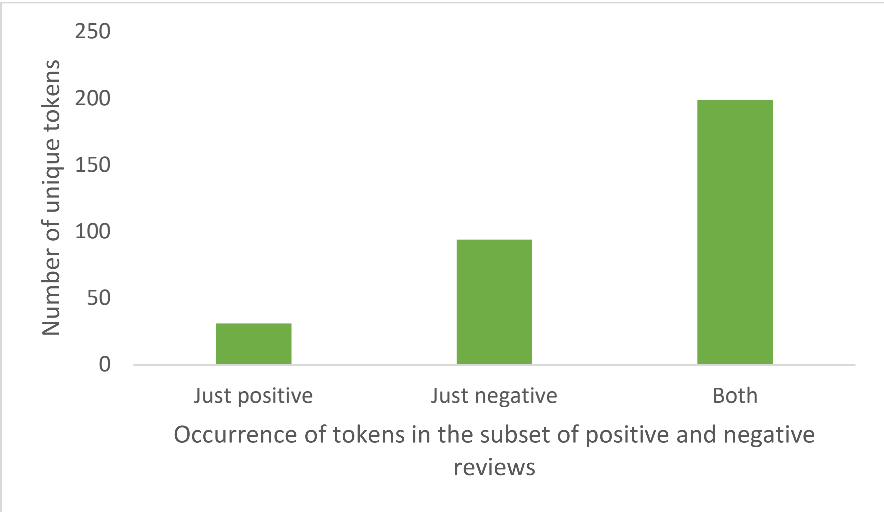

Figure 10 – Occurrence of tokens

We can approach this analysis from an alternative perspective. The dictionary is formed based on common words between reviews. The positive subset of the reviews dictionary contains 230 unique terms, and the negative subset contains 293 unique terms. While the overall sizes of the two dictionaries show no significant disparity, there is a noteworthy contrast in the sizes of the subsets of positive and negative reviews. This suggests that negative reviews are less similar to each other than positive reviews.

Table 5 displays the Top 10 most frequent words and their relative frequency of appearance within the reviews. As expected, some of the most repeated and exclusive words in positive reviews are "Fantastic", "Perfect, and “Wonder”, which are associated with positive effects, indicating that, in general, the customer's experience was good. On the other hand, negative reviews present most repeated exclusive words such as “Disappoint”, “Poor”, “Rude”, and “Terrible” which reflect dissatisfaction and negative sentiment. The common words used in positive and negative reviews such as “Good”, “Food”, “Place”, “Service”, “Great”, “Friend” are very frequent in both subsets. Still, they do not transmit negative sentiment. To understand the true context, topic analysis must be carried out.

Table 5 – Top 10 most frequent exclusive words from positive reviews and negative reviews.

Position	Positive reviews		Negative reviews		Total	
	Word	Frequency	Word	Frequency	Word	Frequency
1	perfect	8%	disappoint	10%	food	65%
2	fantastic	7%	poor	9%	good	56%
3	wonder	7%	rude	6%	place	37%
4	home	4%	terrible	5%	great	36%
5	tapa	2%	manager	4%	service	36%
6	authentic	2%	wrong	4%	friend	29%
7	chocolate	2%	piece	4%	staff	26%
8	superb	2%	kitchen	4%	nice	24%
9	relax	2%	half	3%	wine	20%
10	fabulous	2%	empty	3%	time	20%

The criteria used to classify reviews into positive and negative were based on customer classification. This has limitations regarding the formation of homogeneous groups in terms of the sentiments they convey. Therefore, we used sentiment analysis to identify the polarity of reviews. We tested VADER (Hutto et al., 2014) on both non-pre-processed and pre-processed datasets.

The result based on non-pre-processed data was observed to be closest to the user rating classification for both positive and negative reviews. However, the results are unfavorable when examining the negative review, as illustrated in Tables 6 and 7. When the data is pre-processed, it leads to the loss of some valuable information. For example, the string “not that great” becomes “great” because “not” and “that” are stop words. The results show that almost 98% of the positive reviews were accurately classified according to the user rating. Then, as ratings decline, the proportion of negative sentiment in the reviews increases. However, there are still reviews with positive sentiment in a subset of negative reviews; 1-star reviews have 32% positive reviews, and 2-star reviews have 53% positive reviews, indicating that VADER (Hutto et al., 2014) is not capturing negative sentiment. It can mean that the content of reviews does not transmit negative sentiment on average.

Based on the results of the sentiment analysis, the subsets were divided as follows: 98% of positive reviews (with 4- and 5-stars) have a strong positive tone, so this subset was not changed. The subset of negative reviews was divided into two sets: a set whose sentiment, as determined by VADER (Hutto et al., 2014), is positive or neutral, and another set whose sentiment is negative.

Table 6 – Application of VADER to the original reviews (without pre-processing).

Star Ranking of reviews	Average Sentiment	Percentage of reviews by sentiment			Average sentiment of reviews		
		Negative	Neutral	Positive	Negative	Neutral	Positive
Negative	0.12	41%	3%	55%	-0.62	0.00	0.67
1	-0.24	65%	3%	32%	-0.67	0.00	0.60
2	0.08	44%	4%	53%	-0.58	0.00	0.64
3	0.49	18%	3%	79%	-0.49	0.00	0.72
Positive	0.86	1%	1%	98%	-0.36	0.00	0.88
4	0.81	2%	1%	97%	-0.37	0.00	0.84
5	0.88	1%	1%	99%	-0.36	0.00	0.90
Total	0.73	9%	1%	90%	-0.59	0.00	0.86

Table 7 – Application of VADER to the pre-processed reviews (without filtering of most frequent and rare words).

Star Ranking of reviews	Average Sentiment	Percentage of reviews by sentiment			Average sentiment of reviews		
		Negative	Neutral	Positive	Negative	Neutral	Positive
Negative	0.25	30%	6%	64%	-0.54	0.00	0.65
1	-0.10	54%	8%	38%	-0.58	0.00	0.55
2	0.28	28%	6%	66%	-0.48	0.00	0.62
3	0.59	9%	5%	87%	-0.38	0.00	0.71
Positive	0.81	1%	2%	97%	-0.30	0.00	0.83
4	0.77	1%	3%	96%	-0.30	0.00	0.80
5	0.82	1%	2%	98%	-0.30	0.00	0.85
Total	0.71	6%	3%	91%	-0.51	0.00	0.81

Table 8 – Statistics of final subsets.

Subsets	Number of reviews	Average sentiment	Average star	Median
Subset of positive reviews (4- and 5-star)	248k	0.86	4.7	5
Subset of negative reviews (1-, 2- and 3-star) with negative sentiment	24k	-0.62	1.6	1
Subset of negative reviews (1-, 2- and 3-star) with positive sentiment	34k	0.64	2.3	3

Table 8 presents the results after one more dividing of negative reviews. Notably, by splitting subset of negative reviews, two sets of reviews with more homogeneous sentiments were obtained: one with extremely negative sentiment and another with intermediate positive sentiment. It is interesting to note that the average score follows the sentiment rating in the same direction. Regarding the size of the subsets, the two subsets of negative reviews have similar sizes, while the subset of positive reviews is much larger.

LDA topic modeling was used to extract the main topics that customers talk about. It is a valuable tool that allows you to segment reviews into specific categories. This allows you to explore which specific aspects of restaurants are most frequently discussed. In the context of the Anna Karenina principle, it helps to highlight behavior in cases of success and failure.

The combination of two methods — LDA topic modelling and sentiment analysis—provides a richer view by allowing contextualization of sentiment in online reviews. Allows for a deeper understanding of customer experience. However, the goal is to identify relatively common topics, given that the data spans different restaurants.

The next step is to apply LDA. In this step, it was necessary to compile a corpus and a dictionary of reviews and discover the optimal topics for K that must be between 10 and 15 (Kirilenko et al., 2021). To create a corpora and dictionary for each subset of reviews. A dictionary is created by processing text data (reviews) and assigning a unique ID to each word in a corpus. This allows to match words to their respective identifiers and vice versa efficiently. Table 9 shows that the number of unique tokens is the same in all subsets.

Table 9 – Number of unique tokens per subset

Subsets	Number of unique tokens
Subset of positive reviews (4- and 5-star)	230
Subset of negative reviews (1-, 2- and 3-star) with negative sentiment	293
Subset of negative reviews (1-, 2- and 3-star) with positive sentiment	293

Table 10 displays the Top 10 most frequent words and their relative frequency of appearance within the reviews. The corpus is a bag of words, where each document is represented as a list of tuples of unique ID words and frequency. The word "food", as expected, is the most frequently used in the reviews. In the case of positive sentiment reviews, the word “good” is also commonly used. We can notice that half of the most common words in 4- and 5-star reviews are adjectives, which is not the case for negative reviews. In negative reviews, most of the terms are nouns. This means that negative reviews are more diverse in terms of themes. In terms of positive reviews, in addition to positive adjectives, positively criticized terms refer to "Place", "Service", "Staff", "Wine" and "Price". As for negative criticism, moreover, to adjectives, relevant terms are “Service”, “Place”, “Staff”. There are more specific terms such as "Table", "Waiter", "Time", "Food" and "Dish" as well.

Table 10 – Top 10 most frequent words in positive reviews

Position	Positive reviews		Negative reviews with positive sentiment		Negative reviews with negative sentiment	
	Word	Frequency	Word	Frequency	Word	Frequency
1	food	62%	food	72%	food	72%
2	good	55%	good	62%	servic	39%
3	great	40%	place	39%	place	35%
4	place	35%	servic	38%	tabl	28%
5	servic	34%	nice	29%	good	27%
6	friend	31%	tabl	26%	waiter	26%
7	staff	26%	time	24%	time	25%
8	nice	23%	staff	24%	staff	22%
9	wine	21%	friend	21%	meal	18%
10	excel	20%	price	21%	dish	16%

By analyzing the 10 most frequently occurring words from the three subsets, we can conclude that in the subset of positive sentiment reviews, half of the terms are positive adjectives and the other half are main topics such as "food", "place", "service", " staff”, "wine", in the subset of negative reviews, the majority are nouns such as “food”, “site”, “service”, “staff”, “table”, “waiter”, “time”, “food", "dish". This suggests that reviews with a negative sentiment contain more terms, while positive reviews often use positive adjectives to describe the experience.

Next, we estimate the optimal number of topics using the perplexity and coherence scores combined with the elbow method, see Figure 11 and Figure 12. Perplexity and coherence scores were plotted against the number of topics between 2 and 19, however, a number of topics between 10 and 15 must be selected (Kirilenko et al., 2021). In terms of perplexity, perplexity improves for all subsets of reviews as the number of topics increases. The most significant improvement occurs from 14 topics to 15 for all subsets. Regarding the coherence score, the most significant improvement occurs when passing from 14 topics to 15 for the positive review subset and from 11 to 12, and 9 to 10 topics for the remaining negative review subsets.

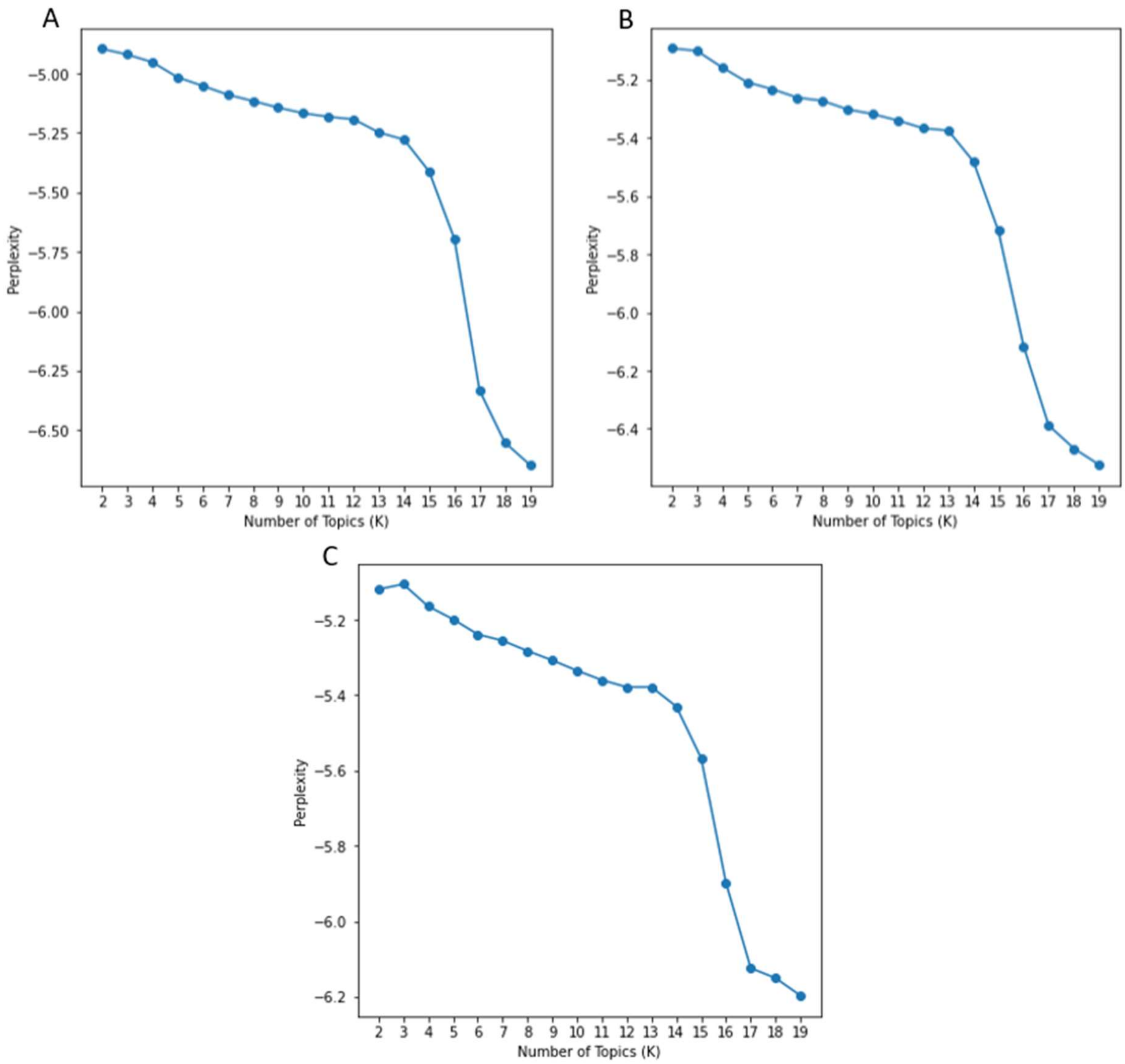

Figure 11 – Evaluating a topic model according to perplexity. Panel A shows the perplexity for subset of positive reviews. Panel B shows the perplexity for subset of negative reviews negative sentiment. Panel C shows the perplexity for subset of negative reviews positive sentiment.

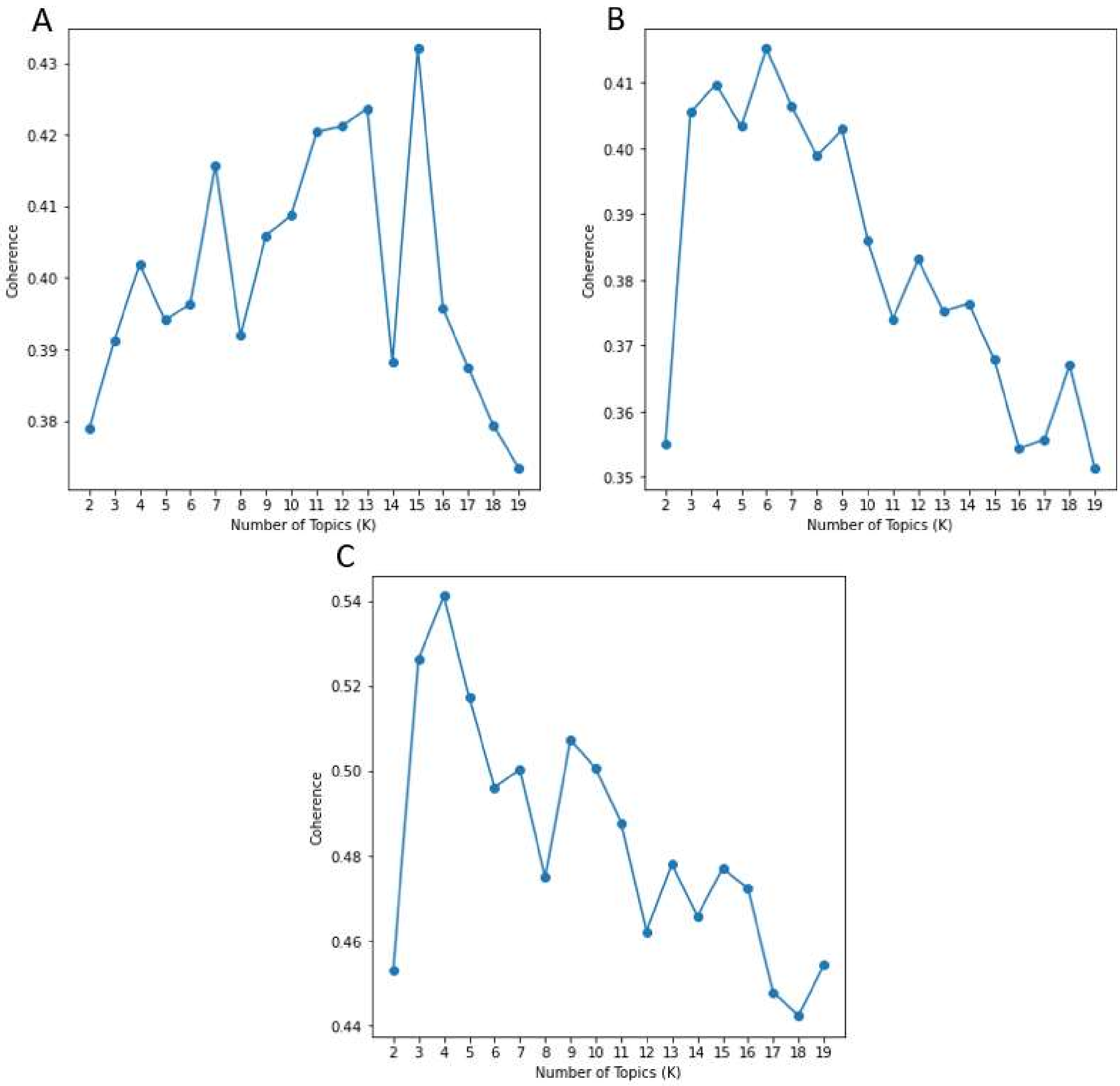

Figure 12 – Evaluating a topic model according to coherence. Panel A shows the coherence for subset of positive reviews. Panel B shows the coherence for subset of negative reviews negative sentiment. Panel C shows the coherence for subset of negative reviews positive sentiment.

It was decided to choose K* equal to 15 topics for the Positive Review Subset (4- and 5-star) as the optimal one; 12 topics for a subset of negative reviews (1-, 2-, and 3-star) with negative sentiment; 10 topics for the subset of negative reviews (1, 2 and 3 stars) with positive sentiments, see Table 11.

Table 11 – Results of Elbow method

Subsets	**Number of topics (K)**	**Perplexity**	**Coherence**
Subset of positive reviews (4- and 5-star)	15	-5.69	0.43
Subset of negative reviews (1-, 2- and 3-star) with negative sentiment	12	-5.37	0.38
Subset of negative reviews (1-, 2- and 3-star) with positive sentiment	10	-5.36	0.48

The LDA Model returned as results the topics generated for each subset of reviews, the relevant terms, and loadings of each review on each topic (from 0 to 1). In the Figure 13, each curve represents a single topic. The horizontal axis represents the documents sorted in order of decreasing probability of representing a particular topic, and the vertical axis represents the probability. In positive reviews, topics that can handle loadings greater than 0% for more reviews in approximately 80% of reviews were formed. Regarding negative reviews, topics "drop" from 60%. This means that approximately 41% of negative reviews have topics that are not represented (0%) in those reviews. This suggests that negative reviews tend to be more distant or less similar in terms of the topics they discuss when compared to positive reviews.

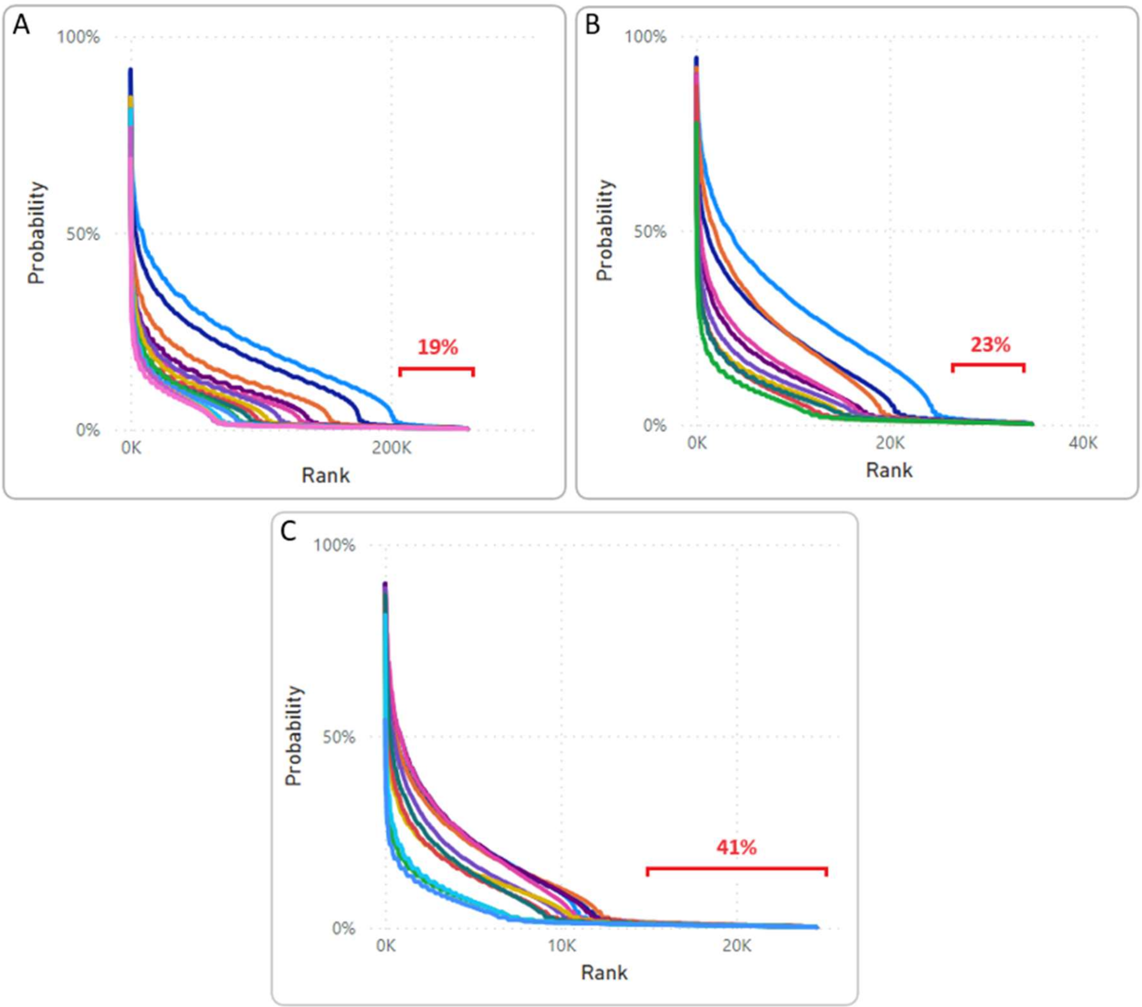

Figure 13 – Panel A shows the distribution of topic-loading reviews for a subset of positive reviews. Panel B shows the distribution of topic-loading reviews for a subset of negative reviews with positive sentiment. Panel C shows the distribution of topic-loading reviews for a subset of negative reviews with negative sentiment.

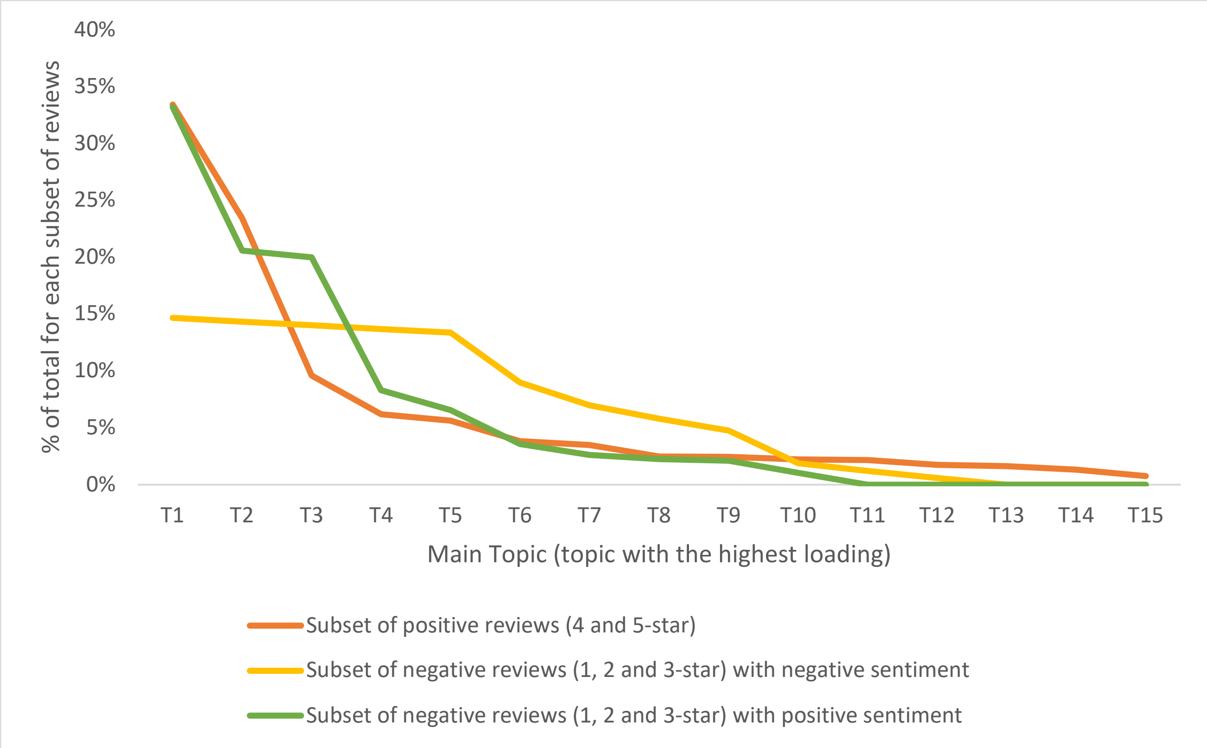

Figure 14 – Comparison of subsets in terms of main topics (the maximum loading)

When the maximum loading from each subset (Figure 14), in the case of reviews with positive sentiment, the first two topics cover a significant part, approximately 57% and 54% of these reviews. On the other hand, negative reviews require four different topics to cover approximately 57% of reviews. This suggests that the sentiment of reviews plays a substantial role in the context of TAK principle impact. Namely, negative reviews with a positive sentiment behave approximately as positive reviews. It seems that the sentiment of the reviews is a more relevant discriminator of the TAK effect than the star rating. The positive reviews tend to focus more on a smaller set of topics, while negative reviews are spread across a wider range of topics.

Tables 12, 13 and 14 show the resulting topics and the corresponding most relevant terms for each subset of reviews. Moreover, the relative percentage of each topic in terms of terms used is presented. The higher the relative percentage of a topic, the more significant it is regarding review coverage.

Table 12 – The positive reviews subset topics and the main terms for each topic.

ID Topic and description	Top 5 most relevant terms	Percentage of total of tokens
T1: Great, Food, and Staff	food; friend; great; staff; delici;	17%
T2: Good, Service, Wine, and Price	good; servic; wine; price; owner;	15%
T3: Excellent, Menu, Beer, and Value	excel; menu; perfect; beer; valu; good;	10%
T4: Amazing, Local, and Flavors	amaz; local; fresh; tasti; fantast;	7%
T5: Nice, Lunch, Family, and Full	nice; lunch; famili; home; full;	7%
T6: Place, Time, Little, and Variety	place; time; littl; good; varieti;	6%
T7: Coffee, Fish, Atmosphere and Cake	coffe; fish; atmospher; cake; welcom;	6%
T8: Meals, Night, Busy, and Weiter	meal; night; first; busi; waiter;	5%
T9: Dish, Portuguese, Small, and People	dish; portugues; small; peopl; year	5%
T10: Lovely, Table and Service	love; tabl; servic; mani; trip;	4%
T11: Wonderful, Desserts and Sweet	wonder; dessert; delici; main; sweet;	4%
T12: Well, Café, Portion, and View	well; cafe; reason; portion; view;	4%
T13: Drinks, Special, Time, and Happy	drink; area; special; time; happi;	3%
T14: Worth, Visit, Simple, and Real	worth; visit; simpl; kind; real;	3%
T15: Traditional, Different, Chicken and Location	tradit; differ; chicken; coupl; locat;	3%

Table 13 – The negative reviews with negative sentiment subset topics and the main terms for each topic.

ID Topic and description	Top 5 most relevant terms	Percentage of total of tokens
T1: Meal, Little, and Bill	meal; food; dish; littl; bill;	13%
T2: Food Selection and Dishes	fish; steak; meat; dish; salad;	13%
T3: Table, Time, Order and Drink	tabl; time; minut; order; drink;	12%
T4: Food, Place, Good, Quality and Service	food; place; good; quality; service;	10%
T5: Service, Food, Poor, and Experience	service; food; poor; slow; experi;	10%
T6: Nice, Food, and Coffee	nice; food; good; coffee; great;	9%
T7: Disappointment, Chicken, Good, and Reviews	disappoint; chicken; good; review; cours;	8%
T8: Price, Wine, Menu, and Small	price; wine; menu; small; euro;	8%
T9: Customer, Owner, Staff, and Rude	people; custom; owner; staff; rude;	7%
T10: Main, Starter, Full, Money, and Room	main; starter; full; money; room;	4%
T11: Pizza, Cold, and Hotel	pizza; cold; hotel; choic; open;	4%
T12: Portion, Fresh, Piece, and Server	portion; fresh; piec; person; server;	3%

Table 14 – The negative reviews with positive sentiment subset topics and the main terms for each topic.

ID Topic and description	Top 5 most relevant terms	Percentage of total of tokens
T1: Food, Good, and Service	food; good; service; staff; friend;	20%
T2: Dishes and Wine	dish; wine; main; meal; steak;	17%
T3: Place, Food, Good, and Service	place; food; good; nice; service;	14%
T4: Tables, Little, Time, and Waiter	tabl; little; time; minut; waiter;	10%
T5: Small, Local, Fish, Good, and Lunch	small; local; fish; good; lunch;	10%
T6: Drinks, View, and Dinner	drink; area; view; dinner; choic;	7%
T7: Time, Tasty, Busy, and Hotel	time; tasti; busi; excel; hotel;	6%
T8: Bread, Chicken, Cheese, Sauces, and Cheap	bread; chicken; chees; sauc; cheap;	6%
T9: Average, Coffee, Beer, and Fine	average; coffee; beer; reason; fine;	5%
T10: Portuguese, Love, Cold, Value, and Cream	portugues; love; cold; valu; cream;	4%

Among the reviews with a positive sentiment, the foremost two topics of note, accounting for more than 54% of the reviews, topics related to “Good”, “Food”, “Service”, “Dishes, Wine”, and “Price”. Although reviews with negative sentiment require more topics, the first two represent less than 30% of the reviews, to represent more than half of the reviews it is necessary four topics, that represent 57% of the reviews. Related topics are "Meal and Bill", "Food Selection", "Timing and Order", and "Food, Place, Quality, and Service".

We conducted a manual topic identification test based on reviews randomly collected from each subgroup. From each subgroup, 20 reviews were collected, read, and analyzed, and based on human interpretation, one to three topics were selected that best represented the reviews. For each review, the main topics were identified based on a predetermined loading threshold of 20%. Subsequently, an assessment was conducted to confirm the correspondence between the key terms associated with these topics and the actual content of the reviews. The results are presented in Table 15.

Table 15 – Result of manual interpretability between topic and reviews

Subsets	Loadings of main topics	Human interpretability (percentage of well-identified topics)
Subset of positive reviews (4- and 5-star)	38%	55%
Subset of negative reviews (1-, 2- and 3-star) with negative sentiment	57%	65%
Subset of negative reviews (1-, 2- and 3-star) with positive sentiment	62%	95%

The results show that the average load and interpretability of 4- and 5-star reviews are lower compared to negative reviews, 38% and 55% (success rate), respectively. However, the sizes of the subsets vary significantly. Nevertheless, it is valid to compare the remaining two groups of reviews as they have similar dimensions but opposite sentiments. The results show that positive sentiment review subsets achieve better results in terms of average topic loading by review and review interpretability: 62% versus 57% and 95% versus 65% (success percentage), respectively.

5. CONCLUSIONS

The main purpose of this work is to understand whether the Anna Karenina principle influences online reviews of Portuguese restaurants. To achieve this goal, we use sentiment analysis and LDA topic modelling techniques to determine the sentiment of reviews and identify topics that represent the content of reviews. As a working basis, we obtained a large pool of reviews from different restaurants, and by extracting themes from these reviews, we ultimately extracted general information. We extracted themes about overall successes or failures among the reviews. In the case of negative reviews, information about common problems in restaurants was extracted, and in the case of positive reviews, information about common aspects that contribute to customer satisfaction.

The results show that compared to positive online reviews, negative reviews have a larger volume of text and topics related to dissatisfaction. This causes greater difficulty in understanding the essence of specific dissatisfaction factors, as it requires a much more detailed analysis. It is important to mention that the sentiment identified in the reviews emerged as a more relevant indicator than the star ranking. Negative reviews (1- to 3-stars) with positive sentiment showed results close to positive reviews (4- and 5-stars). Most reviews with positive sentiment fall into two topics related to food, service, menu, price, and wine, but to represent the majority of negative reviews (with negative sentiment), at least four topics are required related to food and portion; details about the type of food (bread, chicken, cheese, sauces); details about the service (table, time, order); location and quality.

Under the research conducted in the restaurant sector, on the one hand, negative reviews are more difficult to analyze. Still, on the other hand, they contain more valuable information and are more detailed than positive reviews. From a practical perspective, these findings can be used by the restaurant industry to improve restaurant management and influence future strategy. Still, they also contribute to the universality of Anna Karenina's principle.

Some limitations were encountered in this study concerning the split of the review dataset. Initially, the pool was divided into two groups: a group of positive reviews (4- and 5-stars); a group of negative reviews (3-, 2-, and 1-star), sentiment analysis was applied, which helped to identify the need to separate the group of negative reviews by sentiment. Based on the topic modeling results, negative reviews with positive sentiment appear to be closer to positive reviews. This raises a question that has not yet been deeply analysed. Do positive reviews with a negative sentiment have a different nature than positive reviews with a positive sentiment?

Furthermore, uncertainties persist regarding the analysis of the TAK effect within subsets divided by sentiment segments (Very Positive, Positive, Neutral, Negative, Very Negative). Could the prominence of the TAK effect be increased under reviews segmented in this way? This consideration holds significant implications for how review platforms must prioritize and assign scores to restaurants. Rather than relying only on star ratings, an approach based on the sentiment expressed in reviews could offer a more nuanced and accurate assessment of restaurant performance. Additionally, since most reviews are positive, comparing results on of topic modelling between positive and negative reviews is difficult since the groups have completely different dimensions. More research is needed, and this principle needs to be tested in universes where failures are not a minority.

Another limitation is that the analysis was conducted to collect general information due to the large number of reviews. It made sense to study reviews by a restaurant or for a specific restaurant and thus obtain results focused on specific themes specific to that restaurant. Another limitation is that only one part of the available TripAdvisor reviews was retrieved, meaning there is potential for online review bias, and additional validation is required.

From a technical perspective, LDA is a topic modeling method with limitations such as a request to specify the number of topics in advance, using the bag-of-words model, which disregards word order and document structure and, in this way, does not capture the semantic meaning of words and phrases. These limitations can be overcome by applying BERT (Bidirectional Encoder Representations from Transformers). According to Jeet Rawat et al. (2022), Google introduced BERT, a sophisticated sentence embedding method that encodes the semantic meaning of documents.

Customer dissatisfaction is an area of great importance as it directly impacts the customer's decision. In the business world, this is a priority area that requires a previously implemented methodology to solve the often unpredictable problems or combinations of problems associated with customer dissatisfaction. This analysis provides a comprehensive understanding of the application of Anna Kerenina’s principle in online reviews of Portuguese restaurants. The main techniques used were sentiment analysis and LDA topic modeling, which, when combined, give better results. The results show the complexity inherent in this analysis and the wealth of information that can be extracted from customer opinions.

In a broader context, this study enriches academic knowledge and offers general practical and actionable insights for restaurant owners, managers, and other industry professionals. As we move forward, it is important to recognize that the Anna Karenina principle, sentiment analysis, and topic modeling techniques continue to evolve. This study may serve as a starting point for future research to further explore the nuances of online reviews in various fields. Ultimately, our research supports the idea that, as with Tolstoy's unhappy families, the experience of each dissatisfied customer is more complex in terms of the factors that led to dissatisfaction than the experience of a satisfied customer.

REFERENCES

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

Atkinson-Abutridy, J. (2022). Text Analytics: An Introduction to the Science and Applications of Unstructured Information Analysis (1.ª ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003280996

Baur, D. G. (2022). The Anna Karenina principle and stock prices. Journal of Behavioral and Experimental Finance, 33, 100602. https://doi.org/10.1016/j.jbef.2021.100602

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

Bornmann, L., & Marx, W. (2011). The Anna Karenina principle: A concept for the explanation of success in science. https://doi.org/10.48550/ARXIV.1104.0807

BrightLocal. (2020). Local Consumer Review Survey. BrightLocal. Retrieved [2023], from https://www.brightlocal.com/research/local-consumer-review-survey/

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. Neurocomputing, 72(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Eastin, M. S., Daugherty, T., & Burns, N. M. (Eds.). (2011). Handbook of Research on Digital Media and Advertising: User Generated Content Consumption. IGI Global. https://doi.org/10.4018/978-1-60566-792-8

Feinstein, E. (2018). OTA's vs. direct hotel bookings: Which is the leading trend for 2018? TravelDailyNews International. https://www.traveldailynews.com/post/otas-vs-direct-hotel-bookings-which-is-the-leading-trend-for-2018

Galton, F. (1883; 1907). Human faculty. Retrieved [2023], from https://galton.org/books/human-faculty/SecondEdition/text/web/human-faculty4.htm

Golmohammadi, A., Mattila, A. S., & Gauri, D. K. (2020). Negative online reviews and consumers' service consumption. Journal of Business Research, 116, 27–36. https://doi.org/10.1016/j.jbusres.2020.05.004

Gorban, A. N., Smirnova, E. V., & Tyukina, T. A. (2010). Correlations, risk and crisis: From physiology to finance. Physica A: Statistical Mechanics and Its Applications, 389(16), 3193–3217. https://doi.org/10.1016/j.physa.2010.03.035

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

Jeet Rawat, A., Ghildiyal, S., & Dixit, A. K. (2022). Topic modelling of legal documents using NLP and bidirectional encoder representations from transformers. Indonesian Journal of Electrical Engineering and Computer Science, 28(3), 1749. https://doi.org/10.11591/ijeecs.v28.i3.pp1749-1755

Kim, W. G., Li, J. (Justin), & Brymer, R. A. (2016). The impact of social media reviews on restaurant performance: The moderating role of excellence certificate. International Journal of Hospitality Management, 55, 41–51. https://doi.org/10.1016/j.ijhm.2016.03.001

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

Kherwa, P., & Bansal, P. (2018). Topic Modeling: A Comprehensive Review. ICST Transactions on Scalable Information Systems, 0(0), 159623. https://doi.org/10.4108/eai.13-7-2018.159623

Liu, B. (2012). Sentiment Analysis and Opinion Mining. Springer International Publishing. https://doi.org/10.1007/978-3-031-02145-9

Ma, Z. (Sam). (2020). Testing the Anna Karenina Principle in Human Microbiome-Associated Diseases. iScience, 23(4), 101007. https://doi.org/10.1016/j.isci.2020.101007

Marcolin, C. B., Becker, J. L., Wild, F., Behr, A., & Schiavi, G. (2021). Listening to the voice of the guest: A framework to improve decision-making processes with text data. International Journal of Hospitality Management, 94, 102853. https://doi.org/10.1016/j.ijhm.2020.102853

Moore, D. R. J. (2001). The Anna Karenina Principle Applied to Ecological Risk Assessments of Multiple Stressors. Human and Ecological Risk Assessment: An International Journal, 7(2), 231–237. https://doi.org/10.1080/20018091094349

ReviewTrackers. (2022). Online reviews survey. ReviewTrackers. Retrieved [2023], from https://www.reviewtrackers.com/reports/online-reviews-survey/

Rita, P., Vong, C., Pinheiro, F., & Mimoso, J. (2023). A sentiment analysis of Michelin-starred restaurants. European Journal of Management and Business Economics, 32(3), 276–295. https://doi.org/10.1108/EJMBE-11-2021-0295

Ruiz-Mafe, C., Bigne-Alcañiz, E., Sanz-Blas, S., & Tronch, J. (2018). Does social climate influence positive eWOM? A study of heavy-users of online communities. BRQ Business Research Quarterly, 21(1), 26–38. https://doi.org/10.1016/j.brq.2017.12.001

Shugan, S. M. (2007). Editorial—The Anna Karenina Bias: Which Variables to Observe? Marketing Science, 26(2), 145–148. https://doi.org/10.1287/mksc.1070.0274

Sweeney, J. C., Soutar, G. N., & Mazzarol, T. (2008). Factors influencing word of mouth effectiveness: Receiver perspectives. European Journal of Marketing, 42(3/4), 344–364. https://doi.org/10.1108/03090560810852977

Tanas, L., Winkowska-Nowak, K., Roszak, J., & Potega Vel Zabik, K. (2020). Females and STEM: The Anna Karenina Principle [Commentary]. IEEE Technology and Society Magazine, 39(1), 12–15. https://doi.org/10.1109/MTS.2020.2967485

Tardin, Matheus & Perin, Marcelo. (2020). The impact of COVID-19 on the Brazilian food service industry: topic modelling of online reviews.

Tripadvisor. (2021). About Us. Retrieved [2021], from https://tripadvisor.mediaroom.com/us-about-us

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 IMS

Information Management School

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

Universidade Nova de Lisboa